AI: an international dialogue

Note of a Royal Society-National Academy of Sciences workshop

23 and 24 May 2019

THE ROYAL SOCIETY

Summary

As AI technologies advance at pace, governments across the world are developing national strategies to harness their economic benefits for national advantage and societal wellbeing.

The potential of AI to improve the lives and livelihoods of people across the world is significant, from helping address climate change and food security, to supporting elderly care and healthcare delivery. However, these technologies also have the potential to reinforce social divisions and biases, with particular implications for those already at the margins of society. There is now widespread consensus that action is needed to create an environment of careful stewardship of AI technologies, to ensure that their benefits are brought into being safely and rapidly, and that these benefits are shared across society.

AI's broad potential economic impact has implications for innovation, trade, competition, and ways of working and earning. While sensational early estimates of the impact of AI on employment dominated early policy debates on this subject, recent research has painted a more nuanced picture of the inter-related political, social, and institutional forces that shape how technologies are adopted and their impact on work. Competition policy, for example, will influence how markets help share the benefits of AI technologies, with current systems favouring a 'winner takes most' dynamic.

Patterns of adoption within organisations will also be complex, shaped by interactions between people and AI systems that might have unanticipated consequences, and collaborations between those with technical and domain knowledge will be important in ensuring that such systems are deployed effectively.

Better understanding the impact of AI on the global economy will require new ways of measuring economic activity and international data flows, and further analysis of how existing frameworks for international trade might apply to, or have to be adopted for, AI-enabled goods and services.

Experience of previous waves of technology change suggests that dramatic changes to living and working can arrive before significant changes to productivity or economic growth. In the near-term, AI's most significant effects could therefore be on societal wellbeing. AI systems are being developed and deployed in policy areas where complex social and political forces are at work, and there are already examples – from recruitment to policing – of how such systems could reinforce existing social biases. AI technologies are also augmenting the online information environment, in ways that interact with existing social relationships and human cognitive biases, potentially helping to propagate misinformation. While technical approaches to addressing concerns about bias and online fakes are advancing, additional measures to ensure AI does not have a detrimental impact, especially on vulnerable groups, will be necessary.

Changing technological capabilities are also challenging current notions of privacy. Many people's daily activities are now recorded in some way – through smart devices, for example – and there is a tension between managing the privacy concerns that this can cause, while also allowing individuals and organisations to reap the value from data. Different nations have adopted different responses to these challenges, and both regulation and technology can help play a role in managing the risks associated with data use.

If treated as a form of civic architecture, AI could be advanced in ways that bolster social cohesion and democracy. Such AI would be open and operate in the service of all in society; it would work to explicit, transparent rules and roles; systems that use it would enable institutional memory and learning, and support processes of oversight and accountability.

Levers for action to promote the safe and rapid use of AI technologies can operate over multiple levels – from the research culture that shapes the priorities of AI technologists and how they work, to national policies that advance the deployment of AI, and international agreements that set standards for its use. These levers interact with a range of fundamental rights – privacy, free speech, equality, for example – and require expertise from across research disciplines to develop, in addition to effective engagement with diverse publics.

Advances in science have long relied on international flows of people and ideas. International collaboration connects communities, helping ensure that science and technology advance in ways that benefit all in society. At a time when questions about trust in 'experts' and debates about how the benefits of technological progress are shared across society are again at the fore of public and policy debate, careful stewardship of the development of AI is especially important.

The international nature of AI research means it is well-placed to support international collaboration. This could be stimulated by action to define grand challenges in areas where AI could be applied for social good. It would also be catalysed by the creation of international datasets for research to support a new wave of AI applications, which benefit publics across the world. Such collaborations could be further supported through the definition of foundational values to provide a common ground for the development of ethical frameworks, and to translate these into tangible actions to shape the development of AI.

Given the scope of this challenge, the coming years will bring a need for researchers and policymakers to:

- **Prioritise:** identify areas of pressing need or significant benefit, and work to create solutions to these needs.
- **Specialise:** advance policy debates by focussing on specific use-cases, in order to move beyond high-level statements and develop application-specific responses.
- **Mobilise:** support mechanisms or infrastructures that advance international cooperation and progress in key areas of interest.

Progress in key areas could help create an environment of careful stewardship of AI technologies, in which the benefits of these powerful tools are shared across society. Such progress will require engagement across companies, investors, governments, researchers, and publics to create a vision for the development of AI that benefits society, and to advance collective action across public and private sectors – and international boundaries – in order to bring this into being. By defining global challenges against which AI could be deployed, designing incentives and opportunities for progress in those areas, and developing structures for dialogue across countries and research communities, business, governments and researchers can shape the development of AI for societal benefit.

Context: the changing policy and research landscapes

AI research has undergone rapid expansion in recent years, with a new wave of excitement about the potential of these technologies, created by advances in the power and sophistication of AI techniques based on machine learning. As the field advances, AI methods are being applied across sectors, with significant economic benefits at stake.

Governments around the world are developing national strategies that seek to bring the benefits of AI into being safely and rapidly, harnessing the economic benefits of these technologies for national advantage and societal wellbeing. These national strategies share many commonalities: supporting research, advancing applications in priority areas, building skills, and addressing core ethical concerns.

As they move from being a research domain to one applied at scale, AI technologies are interacting with some of society's most complex issues and institutions.

The potential of AI to improve the lives and livelihoods of people across the world is significant, from helping address climate change and food security, to supporting elderly care and healthcare delivery. However, these technologies also have the potential to reinforce social divisions and inequalities, with particular implications for those already at the margins of society.

Action is therefore needed to create an environment of careful stewardship, in which the benefits of AI are brought into being safely and rapidly, and shared across society.

Purpose of the Royal Society-National Academy of Sciences workshop AI: an international dialogue

Advances in science have long relied on international flows of people and ideas. Scientific collaborations can help to build trust between nations, providing an environment for the free exchange of ideas between people, regardless of cultural, national or religious backgrounds.

At a time when questions about trust in 'experts' and debates about how the benefits of technological progress are shared across society are at the fore of public and policy debate, maintaining strong international collaborations can play an important role in connecting communities and ensuring that science and technology advances in ways that benefit all in society.

In this context, the Royal Society and US National Academy of Sciences co-convened a workshop – *AI: an international dialogue* – on 23 and 24 May 2019. This meeting set out to identify areas in which further research is needed to advance understandings of AI and its societal implications, to consider what further policy activities may be necessary in these areas, and to explore the role international collaboration can play in both of these[1].

1. While the US National Academy of Sciences (NAS) assisted in supporting and organising this meeting, it did not author this summary, and the views expressed here do not necessarily represent those of the NAS. 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.

Understanding AI: current and near-term capabilities

The field of AI is developing at pace. Rapid growth in the number of research submissions to conferences and in the numbers of people enrolling in related computer science courses points to rising interest across countries. Major technology companies, meanwhile, are investing heavily in recruiting staff with AI skills, and in their own research programmes.

AI is the science of making computer systems that can perform tasks that are typically thought to require some level of human intelligence. Instead of being programmed step-by-step, these systems are able to learn how to achieve an objective, given data about the task at hand.

Today's AI systems build on a long-history of human fascination with machine intelligence, and its potential benefits and risks. In 340BCE, speculating about a future in which automata would perform many of the tasks carried out by humans, Aristotle predicted that intelligent machines would ultimately render human workers redundant. A similar sentiment was echoed by Turing in 1951, just one year after his seminal paper suggested that intelligent machines could be made real by creating computers that learn.

Alongside these long-term narratives about the impact of machine intelligence on people, there have been waves of hype surrounding the field. The 1956 Dartmouth Conference, which saw the coining of the term 'artificial intelligence', coincided with a period of optimism about the potential of symbolic systems[2] and neural networks[3] as methods to achieve intelligence.

While the next two decades saw many of these techniques falter, the 1980s saw the rise of expert systems[4] as a method of achieving intelligence. As enthusiasm for expert systems waned, on account of weaknesses that meant these systems often did not function well in the 'real world', the 1990s and 2000s saw a new wave of interest in probabilistic systems and neural networks. This influenced the development of deep learning[5] systems in the last decade and supported advances in machine learning that have catalysed recent attention on the field.

While the technique of deep learning has received significant attention in recent years, today's AI comprises several major technical areas, which – in addition to deep learning – include:

- Probabilistic reasoning, which combines deductive logic with probability theory to manage uncertainty.
- Supervised learning, an approach to machine learning which relies on training data that has labelled pairs of inputs and outputs.
- Reinforcement learning, an approach to machine learning in which an agent learns to interact with its environment, receiving inputs, and making sequential decisions so as to maximise future rewards.

2. A research domain that considers how symbols are used to communicate ideas and information.
3. A computer model with a form that was originally inspired by early work on understanding the nervous system.
4. Software that uses encoded knowledge from human experts to make a decision.
5. A machine learning method which composes details together to obtain more abstract, higher level, features of the data through composition of mathematical functions. Powerful modern deep learning algorithms often involve a large number of these levels.

Coupled with increasing data availability and compute power, current AI techniques provide powerful tools that have supported progress in a range of areas, including:

- Diagnosis, monitoring, and prediction of complex systems – from jet engines to intensive care, and global seismic monitoring for the Nuclear Test Ban Treaty;
- Credit scoring and financial management, such as fraud detection, or financial malfeasance, such as blackmail;
- Speech recognition and machine translation;
- Robotics and automated driving; and
- Science and research, for example the application of machine learning to analyse data from large-scale particle physics experiments, or to the challenge of understanding the patterns by which proteins fold.

While these systems have achieved many successes, important challenges remain. Many AI tools are, for example, vulnerable to adversarial attacks, which use malicious inputs to prompt a malfunction in the system.

As the field progresses, promising research directions are developing AI that can be trained using fewer examples, requiring less data to learn how to carry out a function. Some researchers are also finding value in revisiting and reassessing key concepts in older techniques, such as symbolic reasoning, to make more powerful systems.

Over the next decade, there will likely be further advances that enable the application of AI in areas such as:

- The use of robotics in unstructured environments – at home, in fields, in mines, and on roads;
- Web-scale automated extraction and question-answering;
- Global vision systems, via satellite imagery; and
- Intelligent personal assistants in daily life, for example helping arrange travel and manage diaries; in education, where they could help plan lesson structure and content; and in healthcare.

Notwithstanding this progress, several major conceptual breakthroughs would be necessary in order to develop AI with human-level capabilities. These include language understanding, the integration of learning and knowledge, long-range theory, and cumulative collection of concepts and knowledge.

Even without achieving human-level intelligence, the consequences of AI for individuals, communities, and societies are potentially profound. Systems today are creating convincing fake images, text and video online, which are influencing how people relate to the information they see; they are automating tasks in ways that shape the nature of work; they are enabling the development of automated weapons systems; and they are fostering new forms of dependency on technology.

At this stage, research and policy both play important roles in shaping the direction of technology development, and in helping to share the benefits of AI technologies across society.

AI and the global economy

Reshaping economies

By providing new means of innovating or producing goods and services, AI technologies could become an entrant in the pantheon of general purpose technologies – such as electricity and IT – whose widespread adoption brings significant reorganisations of economic activity within nations and across borders. As a tool in existing businesses, AI could optimise systems and automate processes. As an agent of innovation, it could reshape traditional sectors, while also opening new areas of economic activity, with potentially significant implications for work and the economy.

Sensational early estimates of the impact of AI on employment put the substitution of human labour for AI centre-stage in many public and policy debates about its economic effects. In the Royal Society's (2016) public dialogues on machine learning, for example, replacement of human labour by machines emerged as one of the top areas of concern that participants expressed about the impact of AI technologies on their lives. Recent studies continue to vary in their estimates, but generally predict that between 10-30% of jobs are likely to be subject to some level of automation, with the extent of these effects varying across countries.

Subsequent research has painted a more nuanced picture of the inter-related economic, political, social, and institutional forces that shape how technologies are adopted across economies, and how this shapes the nature of their impact on employment and working life. The nature and level of automation – and the extent to which this automation has a substitution effect on human labour – is contingent on factors including:

- Flexibility of labour market institutions, and the relationships between workers and employers;
- Demand for AI-produced or AI-enabled goods and services from businesses and consumers;
- Regulatory frameworks;
- Business incentives to adopt AI;
- The skills mix in the economy, and the extent to which people have support to move to new roles or work with new technologies;
- Whether complementary investments – for example in infrastructure – are necessary to implement AI-enabled solutions;
- Barriers to market entry for new firms.

In this context, technology is not a unique or overwhelming force, but one influenced by social, political and economic factors, which will vary across nations. Predicting the impact of AI on work – and who may be at risk of being economically disadvantaged by the widespread adoption of AI – is therefore challenging.

Examining market dynamics today offers insights into the patterns of economic growth arising from AI adoption, and how the benefits of these technologies are shared across society. Such studies yield two key insights:

- There are signs that business dynamism is slowing. US Bureau of Labour Statistics figures show that patterns of business entry and exit into markets are changing, with declining numbers of businesses being created or folded.
- Markets seem to be becoming more concentrated and less competitive.

In the UK, a Government review of the state of competition in digital markets[6] concluded that updated competition policies would be necessary to address the new challenges posed by the digital economy. This review noted a range of difficulties in enforcing current competition policy in digital sectors, including:

- The importance of matching to the services provided by platforms – bringing service users and providers together – which means that these often work most effectively in dense population markets. There are also indirect network effects, with service users and providers both benefiting the more participants there are from both groups in a platform.
- Price structures in the market, which make it difficult to judge the effectiveness of competition policy in traditional ways, and the importance of scaling for business viability, with businesses often loss-making up to a critical mass.
- Data as a barrier to entry, and the need for businesses to access data from a large number of customers in order to improve their service offering using current AI technologies.

Together, these factors create a 'winner takes most' dynamic that favours market concentration and the rise of a small number of large companies.

Frameworks for trade

The economic dominance of large technology companies can already be seen in rankings of the world's largest companies; lists that typically include Apple, Amazon, Alphabet, Microsoft, Facebook, Tencent, and Alibaba[7]. In the context of international trade, this is perhaps not an unusual pattern: in most sectors, a small number of firms tend to be responsible for the majority of trade, and it seems reasonable to expect that a similar pattern will emerge as international trade in AI-enabled goods and services develops.

Patterns of trade – who trades with whom, and for what – are affected by a range of technical, business, political, and cultural factors. Questions about the quality and value of goods are intertwined with questions about trust and politics. While in some cases issues of trust might require careful relationship-building to overcome, in others the answer might be more procedural: if a company wishes to be a trusted trading partner, it may need to meet specifications that prove it is worthy of such trust. When trading food goods, for example, the provable absence of infectious disease is an important criterion in allowing access to a market.

AI-specific international trade rules do not yet exist. There are, however, frameworks for trading in Intellectual Property or services that could form the basis of such a regime. In understanding how current trade rules apply to AI-enabled goods and services, a number of issues currently being contested within the World Trade Organisation (WTO) offer insights:

- States have deployed the WTO's national security exception in a wide range of circumstances. This provision allows a State to take actions that "it considers necessary" to protect essential security interests. The breadth of actions allowed within this wording often means that, once invoked, the national security exception is difficult for others to contest. Given the strategic importance ascribed to AI technologies by many nations, this exception may come into play when considering future trading relationships. Several current cases in front of the WTO panel are testing the boundaries of the exception, and the results of these may illustrate the limits on WTO members' abilities to bypass standard WTO rules in the name of national security[8].
- The WTO's Technical Barriers to Trade (TBT) Agreement aims "to ensure that technical regulations, standards, and conformity assessment procedures are non-discriminatory and do not create unnecessary obstacles to trade"[9]. However, within this agreement there are provisions that allow states to pursue "legitimate policy objectives". There may be scope to challenge restrictive trade practices on the basis of what constitutes an unnecessary obstacle or a legitimate objective.

6. HM Treasury (2019) Unlocking digital competition: report of the digital competition expert panel. Available at: https://www.gov.uk/government/publications/unlocking-digital-competition-report-of-the-digital-competition-expert-panel
7. See, for example, this ranking based on market value: Statista (2019) Top companies in the world by market value 2018. Available at: https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-value/
8. See, for example: DS512 (Russia – Measures concerning traffic in transit); DS526 (UAE – Measures relating to trade in goods and services, and trade-related aspects of intellectual property rights)
9. Details available at: https://www.wto.org/english/tratop_e/tbt_e/tbt_e.htm

As States navigate this evolving policy area, there may be opportunities for further international collaboration through the WTO's plurilateral process and Information Technology Agreements. Advancing such collaborations will require both further policy development to understand the technical barriers to trade and a political narrative that helps build support for further agreements with key decision-makers.

Understanding the new economy

The interactions between AI and the global economy are complex. AI adoption is occurring against a backdrop of highly variable productivity growth, in which a small number of companies are growing rapidly, and the productivity gap between these – usually technology-adopting companies – and 'the rest' is growing, with implications for jobs and earnings. At the same time, concern is growing about the population of people who have been left behind by recent economic growth, and the impact of rising income inequality on social cohesion.

Understanding the dynamics of this new economy is made more complex by the difficulties of measuring activity in the digital economy. Without understanding of how data flows across borders, how value is created by platforms, how work is changing, how markets are structured and new business models develop, or how to value data, it is difficult to develop effective policy responses. At the same time, if AI is an agent for a new wave of innovation, then there may be further, longer-term consequences with impacts that are hard to predict.

While the economic impact of General Purpose Technologies is significant, it is not necessarily rapid: it was almost 50 years before electricity had a measurable impact on the economy, and the relationship between IT and economic growth has been complicated to measure[10]. Transformational change requires not only technological progress and adoption, but also complementary investments – for example in infrastructure and ways of organising companies– as well as changes to policies and behaviours. Dramatic changes to ways of living and working can therefore come before significant changes to GDP. In this context, AI's most significant effects in the near-term could be on how society is ordered.

AI and the global economy: questions for further research and policy development

Trade

What measures or statistics are necessary to understand cross-border data flows and trade? How can these be reliably collected?

In what ways do existing WTO frameworks apply to AI? What tariff codes might be relevant? What lessons come from previous new technologies?

How do cultures and values shape policy approaches to trade?

Competition

How can existing policy structures support healthy competition in digital markets?

How should regulators respond to the growth of a small number of digital businesses? (How big is too big?)

Innovation

In what ways do AI technologies challenge current structures for managing IP? Are existing provisions under WIPO sufficient?

Jobs and employment

What data sources describe the impact of AI on employment? How can these be used to better inform policymaking?

10. Studies of the British Industrial Revolution suggest that productivity growth was quite modest in the decades following major inventions such as the steam engine and spinning mule; growth acquired momentum in the latter half of the 19th century – decades later. While technology ultimately contributes to economic growth, there is frequently a time lag between technology change and productivity increases. This is discussed further in the context of the impact of AI on work in the Royal Society and British Academy's (2018) evidence synthesis: AI and work: implications for individuals, communities, and society, available at https://royalsociety.org/topics-policy/projects/ai-and-work/

AI and social cohesion

Making AI that works for society

The application of AI to a range of public policy challenges could bring great benefits for all in society (Table 1). Many of these grand challenges would benefit from international collaboration, bringing together diverse expertise and different datasets to generate new insights and create tools that work for a range of users. In natural language processing, for example, multilingual datasets are vital in creating systems that can work for users across geographies.

In order to advance applications in these areas, action is needed to create secure shared data resources and frameworks for research collaborations.

TABLE 1

Examples of AI applications for social good.

AI application	
Public health	Healthcare applications of AI are already emerging, with AI-enabled tools enhancing diagnosis or improving monitoring for conditions including dementia, cancer and eye disease. Systems to support care and monitoring for the elderly are also in development, with high demand from many countries with aging populations.
Education	AI has already helped create a range of online educational tools, particularly in higher education. Many schools are now looking to AI to help improve teaching and learning, by providing tailored learning plans and feedback, by providing new types of course, or helping bridge language barriers.
Sustainability and the environment	Climate change, loss of biodiversity, water and air quality, and sustainable agriculture are complex research and policy challenges. AI can help researchers better understand these challenges – providing new insights from analysis of climate data to understand its local impacts, for example. It can also help develop new tools to tackle sustainability issues, including monitoring tools to help track endangered species in the wild and analysis of satellite images to monitor temporal changes in the environment.
Tackling online crime	Criminal groups have been early adopters of digital technologies, using online fora for activities such as terrorism, sex trafficking, and abuse, and exploiting digital systems to conduct online cyber-attacks, leading to data ransom. AI can help analyse the content of websites in order to identify and track those involved in these criminal activities.

AI and existing social biases

Hype surrounding the potential of AI to boost economic growth and promote societal wellbeing is increasingly tempered by growing understanding of the risks associated with AI technologies, often as a result of its disparate societal impact. Prominent examples of this impact have included:

- Women being less likely than men to be shown adverts for high-paid jobs through search engines[11].
- Racial disparities in how algorithmic risk assessment tools in the justice system predict the likelihood of recidivism[12].
- Hiring tools that penalise CVs from women candidates[13].
- Immigrant risk assessment tools that only recommend 'detention'[14].
- Mis-use of personal data for microtargeting of political advertising[15].
- The use of social media systems to spread hate speech and incite violence[16].

Many of these examples illustrate the potential of AI to reinforce existing social divisions or biases, with consequences for equality and social cohesion.

The datasets on which AI technologies are trained reflect society, and contain the biases that were embedded in processes, relationships, or social structures at the point of data collection. When this data is used to develop AI systems, the resulting AI systems reflect back the social and cultural structures or practices of the past; this means the biases that have shaped society in the past (or shape it today) can form the basis of predictions or recommendations about future action.

These issues with bias come to the fore as AI systems are used in policy areas where complex social and political forces are at work. The benefits and risks associated with these systems are unevenly distributed across society, with vulnerable communities potentially being further marginalised as a result of their deployment.

11. Datta ,A., Tschantz, M., and Datta, A. (2015) Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies, 2015 (1), 92 – 112. Available at: https://content.sciendo.com/view/journals/popets/2015/1/article-p92.xml
12. See, for example: MIT Tech Review (2019) AI is sending people to jail – and getting it wrong. Available at: https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/
13. See, for example: BBC (2018) Amazon scrapped 'sexist AI' tool. Available at: https://www.bbc.co.uk/news/technology-45809919
14. See, for example: Quartz (2018) US border agents hacked their risk assessment system to recommend detention 100% of the time. Available at: https://qz.com/1314749/us-border-agents-hacked-their-risk-assessment-system-to-recommend-immigrant-detention-every-time/
15. See, for example: The Guardian (2018) Cambridge Analytica Scandal 'highlights need for AI regulation'. Available at: https://www.theguardian.com/technology/2018/apr/16/cambridge-analytica-scandal-highlights-need-for-ai-regulation
16. See, for example, New York Times (2018) Facebook admits it was used to incite violence in Myanmar. Available at: https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

Sharing the benefits: policy and research responses

In seeking to resolve these issues, both technology-enabled and human-led solutions can play a role. A range of policy responses are already emerging, including, for example:

- Algorithmic impact assessments that help decision-makers, developers, and citizens to better understand the societal impact of an AI system, and how they can challenge decisions from such systems.
- Privacy laws adopted in California (the California Consumer Privacy Act, 2018), which gives consumers rights over how their data is used by businesses.
- New York City's Automated Decision Systems Task Force, which is scrutinising the use of algorithmic systems in local government decision-making in order to ensure they meet expected equality standards.
- The European Union's General Data Protection Regulation, which governs the use of personal data in the EU.
- The G7's proposed International Panel on AI, which is expected to facilitate international collaboration in areas including 'trust in AI'.

New research is also developing ways of managing bias in data, for example by removing sensitive information before that data is used to develop AI systems, or by taking into account standard operating characteristics and recommended usage for datasets that are made available for open use[17].

However, many of these technical attempts to remove bias from AI are very narrow 'fixes', and remain difficult to apply in some of the areas where fairness matters most, which are typically some of the most complex policy areas. 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 issues of fairness are evolving, but fairness remains a challenging issue. In progressing these ideas, further work will be required to improve access to the data and AI systems that are deployed in systems where decisions have a significant impact on people or society.

This action might include archiving and retention policies for organisations using these systems, in order to better understand the cumulative impact of their use, and processes for verification and validation of AI systems, to ensure they work well for all users. The research culture within developer communities also influences how individuals and organisations respond to fairness issues; better understanding – and changing – the incentives that underpin AI development could support more effective responses and help build better AI systems.

Advancing these ideas requires a sophisticated public and policy dialogue. Such a dialogue would bring to the table a wider range of expertise to fully examine the impacts of widespread AI adoption. Examples might include those studying on the environmental impact of technology; researchers from the social sciences and humanities; human rights advocates; workers' coalitions; and those representing marginalised communities.

17. See, for example the Datasheets for Datasets project. Information available at: https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf

Changes to the information environment and their implications

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 in ways that were not previously possible, exchanging views and gathering information from sources to which they would not previously have been exposed. There are new opportunities for communication, 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.

An area of particular current concern is the extent to which AI-enabled social media contributes to the spread of misinformation, and the consequences for democratic debate and electoral outcomes that follow. Patterns in the prevalence and influence of such 'computational propaganda' vary across nations and localities. Studies indicate, for example:

- That images, memes, and videos spread through WhatsApp and Facebook were particularly prevalent in the 2019 Indian elections, with widespread circulation of junk news and misinformation[18].
- That, even though the overall number of junk news stories in circulation during the 2019 EU Parliamentary elections was low, individual junk news stories can attract significantly more attention than other items[19].
- That automated 'bots' played a role in shaping online debates in the 2016 US Presidential election, with some indications that junk news was more concentrated in swing states[20].

Advances in AI technologies continue to contribute to these concerns, with new AI tools that generate realistic images or videos of people, or reports of events. At present, there is little evidence to suggest that AI-generated disinformation is widespread. However, as Deepfake content – AI-generated text, video, and audio – becomes more sophisticated, a race is emerging between those developing the tools to create online fakes and those pursuing the tools to identify them.

The public and policy debates about AI and democracy that follow suggest that these changes to the information environment influence the practice of democracy. These debates propose that rational communication of information is necessary for voters to make rational choices: changing patterns in information exchange can therefore change political opinions, influencing voting patterns, electoral outcomes, and the institutions that sustain them. Under this vision, engineers play an important role in creating spaces where information can move unimpeded, allowing citizens to seek information to shape their choices through rational public debate.

However, insights from behavioural economics and associated disciplines show that people are not rational. Emotional, cultural, and social factors play a role in shaping democratic debates and voting behaviour. AI interacts with each of these in ways that are not yet well-understood, reinforcing or exploiting existing social interactions and cognitive biases.

18. Further detail at: https://comprop.oii.ox.ac.uk/research/india-election-memo/
19. Further detail at: https://comprop.oii.ox.ac.uk/research/eu-elections-memo/
20. Further detail at: https://www.oii.ox.ac.uk/blog/algorithms-bots-and-political-communication-in-the-us-2016-election/

AI as civic architecture

Institutions can be rational in circumstances where individuals are not. In the context of shifting technologies, social norms, and behaviours, the rules and principles on which institutions are founded can be a force for reason in the democratic system.

An alternative democratic response could therefore pursue AI as a form of civic architecture or institution, which: is open and operates in the service of all in society; works to explicit, transparent rules and roles; enables institutional memory and learning; and works to established processes of oversight and accountability.

There are already indications of how such civic architecture could be progressed:

- International standards-setting organisations provide a space where groups can collaborate and set in place systems that bring about technical, social, and institutional change.
- The open data movement in the UK has worked to support the creation of institutional data repositories with high standards for input and maintenance.
- Ways of bringing accountability into the processes that rely on AI-enabled systems – such as algorithmic impact assessments – are emerging.

There is also a need for spaces where people can develop civic networks or new civic institutions that allow individuals from different backgrounds to engage as citizens on common endeavours.

In this context, questions about information exchange become less about the means of circulation, but instead about the ways in which institutions can ensure that citizens have access to trustworthy information.

AI and social cohesion: questions for further research and policy development

What data is needed to advance AI in application areas that could bring widespread societal benefits? How can these datasets be created and managed? What further action is necessary to support research in these areas?

What does the history of General Purpose Technologies show about how the benefits of AI can be shared across society?

In what application areas might new standards be needed to support the safe and rapid deployment of AI technologies?

Trustworthy AI

Privacy, data use, and existing governance structures

The ways in which individuals, communities, and societies think about privacy and its value are not fixed. While policy debates about privacy might previously have been framed in terms of the line between the public and private sphere in the context of celebrity, the volume of data that people generate every day and the range of potential users of that data has expanded the boundaries of these debates. Data might now have value to a range of users – individuals, governments, businesses – while also holding potentially sensitive information, and both people and organisations are increasingly aware of the harms caused by data misuse.

Changing technological capabilities and patterns of technology use are also challenging current notions of privacy. Many people's daily activities are now recorded in some way by smart devices; personal data is collected in new and potentially unexpected ways, for example by self-driving cars, facial recognition in smart cities or retail, and wearable devices; machine learning and advanced analytics can re-identify individuals in datasets previously considered to be anonymised; and sophisticated algorithmic tools can use data from different sources to target advertising or services in ways that might create concerns about privacy and personal profiling.

There is a tension between managing these privacy concerns, while also allowing individuals and organisations to exploit the value in data. One study, for example, found that privacy regulations had a negative impact on the adoption of digital healthcare tools: US-based regulation reduced hospital electronic medical records adoption by 24%, with implications for the effectiveness of healthcare delivery[21].

Different nations have adopted different approaches to managing concerns about data privacy:

- China's rules governing personal data use operate across multiple levels. The National People's Congress has passed a series of laws on cybersecurity, network information protection, and consumer rights, in addition to introducing criminal laws that prohibit unlawful collecting or selling of personal information. These are further supported by administrative provisions (such as the Provisions on Protection of Personal Information of Telecommunications and Internet Users) and national and industry standards.
- In the EU, the General Data Protection Regulation (GDPR) regulates the processing of personal data relating to individuals in the EU by an individual, company, or organisation. This includes provisions relating to automated decision-making and consent for data use. Implementation of the Regulation is supported by advice and guidance from national authorities, such as the UK's Information Commissioner's Office.
- While the US does not have similar federal data-specific regulations, it does have frameworks governing 'unfair or deceptive acts' in trade[22] and laws for protecting children's privacy online[23]. States across the US have also put in place state-level governance frameworks, such as the California Consumer Privacy Act (2018).

Regulatory consistency can be a barrier to international collaboration. To help address such concerns, some bilateral agreements are already in place. For example, data transfers between the US and the EU are governed by the EU-US Privacy Shield.

Technology can also play a role in helping to manage certain privacy risks. A suite of methods often referred to as 'Privacy Enhancing Technologies' are attracting increasing attention, owing to their ability to enable greater sharing and use of data in a privacy-preserving, trustworthy manner. These methods include design principles that embed privacy management in product development, anonymisation technologies such as differential privacy, and emerging approaches to AI such as federated machine learning.

21. Miller, A. and Tucker, C. (2011) Can health care information technology save babies? Journal of Political Economy 119 (2), 289 - 324
22. Including the Federal Trade Commission Act Section 5 (Unfair or Deceptive Acts or Practices). Details at: https://www.federalreserve.gov/boarddocs/supmanual/cch/ftca.pdf
23. The Children's Online Privacy Protection Rule. Details available at: https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule

Transparency and explainability

Once trained, some of the most sophisticated AI systems – notably those based on deep learning – are 'black boxes' whose methods are accurate, but difficult to interpret. Although these can produce statistically reliable results, the end-user will not necessarily be able to explain how these results have been generated or what particular features of a case have been important in reaching a final decision.

Where the decisions informed by these technologies have a significant impact – personally or socially – accuracy is unlikely to be sufficient to secure public confidence in their use. In such contexts, understanding how a decision was reached can play an important role. This requires some level of explainability or interpretability of the system.

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 and deploying AI.

Explainability features in many of the ethics frameworks that have been proposed to guide the development of AI, often with the aim of empowering those subject to an AI-enabled decision-making system to contest the outcome of a decision or process. In the EU, the GDPR implies some form of 'right to an explanation' for those subject to automated decision-making. However, it is not yet clear what this means in practice, and a body of case law may be required before the boundaries of this 'right' are clearer.

Different AI methods can support different types of explainability. Some methods are interpretable by design: these tend to be shallow decision trees, the workings of which can be well-characterised, but which do not get the leverage from vast amounts of data that techniques such as deep learning allow. Another approach is to create tools that interrogate complex AI, or to create decomposable systems that allow the output of a system to be examined at different stages of its production.

These different methods and approaches have different benefits and limitations. To create effective explainable AI, it is therefore important to consider the context in which it will be deployed, and to ensure that AI systems are designed to meet the needs of different users in that context.

AI adoption and the organisational context

Such design concerns will become increasingly important as AI is deployed in organisations or contexts that are far-removed from the technical expertise that created the system. This context influences not only the ways in which AI development is shaped by existing social biases, but also the ways in which humans and AI systems interact.

Patterns of trust in AI are complex. One study, for example, suggested that AI-generated online profiles were deemed less trustworthy than their human counterparts[24]. Others, meanwhile, have found that people tend to defer to the decisions made by AI systems[25], trusting their outputs even in situations where the AI appears not to be well-functioning[26].

These interactions can lead to unexpected patterns of use. In 2017, for example, New York Police Department Officers – having been unable to find a photo match in their facial recognition system for a suspect using surveillance camera footage – used a celebrity photo to have the system generate a list of possible candidates. While the use of such 'probe images' is not unlawful, it raised questions about the reliability of the results from AI systems that were not designed to be used in this way[27].

Automation of routine processes can also have unanticipated organisational effects. Current AI methods are best suited to automate routine tasks in stable environments. Automation of such tasks can help improve productivity, allowing workers to take on different roles. However, such tasks are often the basis for training junior workers. Without these opportunities, it can be more difficult for workers to develop advanced skills[28]. These technological 'fixes' can therefore come at a cost to the effectiveness of current workplaces and routines.

In order to address these concerns, AI development will need to make use of a wider range of expertise, including those with domain and organisational knowledge. Those implementing AI, meanwhile, will need to consider how to strengthen organisational capacity to make use of these technologies, and to support their responsible use.

Trustworthy AI: questions for further research and policy development

How do people respond to AI systems in practice? What are the implications for ways of working and domestic life?

Who has agency in the development and adoption of AI in organisations? What structures can help ensure accountability?

How can AI systems be interrogated? How can researchers create more explainable AI systems?

How do preferences for privacy vary across countries? What are the costs and benefits of different privacy regimes?

How can advances in AI research and policy help address issues of bias? What technical approaches are necessary? What structures for dialogue and engagement can help?

24. Jakesch, M., French, M., Ma, X., Hancock, J., and Naaman, M. (2019) AI-mediated communication: how the perception that profile text was written by AI affects trustworthiness. CHI 2019. Available at: http://www.mauricejakesch.com/pub/chi2019__ai_mc_camera_ready.pdf
25. Cabitza, F., Rasoini, M. and Gensini, G. (2017) Unintended consequences of machine learning in medicine. JAMA 318 (6) 517-518 https://jamanetwork.com/journals/jama/article-abstract/2645762
26. For example: Toon, J. (2016) Trust a robot in a fire? Available at http://www.rh.gatech.edu/front-office/trust-robot-fire
27. Garvie, C. (2019) Garbage in, garbage out. Available at: https://www.flawedfacedata.com/
28. For example, one study of robotic surgery found that the introduction of robots resulted in changes to the effectiveness of human trainee learning practices. Beane, M. (2018) Shadow learning: building robotic surgical skill when approved means fail. Administrative Science Quarterly. Available at: https://journals.sagepub.com/doi/abs/10.1177/0001839217751692?journalCode=asqa

The national and international dynamics of AI

Research culture and international mobility

Science is a global endeavour, and has always been international and collaborative. High-quality AI research can be found across the globe, in both academia and in business, and there is a strong culture of international mobility amongst AI researchers. Many international organisations or conferences work through geographically-dispersed memberships, and many research projects draw from institutions across multiple continents. This international collaboration on research and development has been a driving force in the development of international relationships across the sector.

Mobility is supported by the field's approach to open science. Disciplines such as machine learning place high priority on the rapid publication of results and methods, open sourcing tools and data in order to support rapid proliferation of technical capabilities.

These ways of working could lay the foundations for strong international collaborations on projects that advance the use of AI for social good. Access to international datasets in these areas of interest and the development of internationally-agreed standards in relevant application areas could further advance this cause.

As AI research grapples with how to respond to the widespread societal implications of the technology it helps develop – and seeks to create systems that benefit all in society – there is renewed focus on the composition of the research community. While the impacts of AI are broad, the base for its production is narrow, which has implications for the dynamics of its development. Further action is needed to diversify the AI development community, ensuring people from a wide range of backgrounds have the opportunity to contribute, and to hold those developing AI-enabled products and services to account in ways that ensure AI works well for a wide range of users.

The national context

In recent years, many countries have published national strategies for the development of AI. These often proclaim high levels of investment in these technologies, with the aim of harnessing their economic benefits, building a strong skills base to support their development and deployment, and advancing applications that could bring wider social benefits. For example:

- In the US, the American AI Initiative includes a suite of policy programmes to support technology development while protecting national security interest.
- The UK Government has established the Office for AI and the AI Sector Council to oversee policy development in relation to the UK's AI sector, while a new Centre for Data Ethics and Innovation will advise government on the responsible use of data and data-enabled technologies.
- China's New Generation of AI Development Plan sets out a range of government initiatives to increase investment in AI research and development, while a National Team of industry partners will work to advance AI in priority areas.

Experience of other emerging science and technology issues shows that early adoption does not guarantee continued support by all, or most, of the public. While it is not clear whether public awareness of AI is widespread – a 2016 Royal Society survey found that only 9% of people had heard the term 'machine learning' – there are signs that interest in these technologies is growing[29].

The last five years have seen a number of early efforts to create spaces for informed public dialogue about AI and its implications. In the UK, for example, in 2016 and 2017, the Royal Society carried out the first UK public dialogues on machine learning. These brought together AI researchers with demographically representative groups from across the UK. Their results showed that context was key in how members of the public evaluated the risks and benefits of AI, participants asking questions about who was developing the technology, with what purpose, and with what distribution of risks. Research programmes with a US focus have illustrated that members of the public considered a wide range of governance challenges to be important, with university researchers and the US military considered to be most trusted to develop AI in the public interest[30].

29. For example, over 4000 people attended the Royal Society's You and AI lecture series in person, with over 100,000 further online interactions.
30. https://www.fhi.ox.ac.uk/aipublic2019/

Continued public confidence in the systems that deploy AI technologies will be central to their continued success, and therefore to realising the benefits that these technologies promise across sectors and applications. In conditions where public knowledge about the specifics of the science and technology is limited, perceptions are likely to be informed by personal experiences and by popular narratives about the future.

However, many of these popular narratives focus on a limited number of concerns, which often do not reflect the complexity of the technology or its societal implications. Hype surrounding the potential of AI can create expectations that the technology is not able to fulfil, leading to disillusionment. By contrast, stories about the negative consequences of AI in the long term could overshadow issues that are already creating challenges today.

With major social and economic consequences at stake, it is important that public debate be well-founded. Building such a dialogue will require spaces for conversations between publics, researchers, and policymakers, and new narratives about the development and application of AI that reflect the experiences of different groups as well as technological realities.

Ethics principles and foundational values

As policy approaches for developing safe and beneficial AI emerge, organisations and governments have published a raft of ethical AI principles. These seek to put human wellbeing at the centre of technological progress, shaping the nature of AI technologies and directing their deployment in pursuit of this aim.

Many of these ethical codes contain elements that have long been a feature of public life: transparency and accountability; fairness and equality; safety and security. These usually draw from notions of human rights, democracy, and the rule of law, alongside a desire to enhance competitiveness and economic growth.

The ways in which these ideas are framed often reflects the local social context of their creator, with implications for the breadth of support that can rally behind them. In order to build a consensus that shapes the development of AI across geographical boundaries, further work is needed to identify ethical framings that speak to people from a wide range of backgrounds. Such foundational values would be open, inclusive, and adaptive, and would provide a common ground from which diverse cultural and political viewpoints could develop frameworks for the beneficial development of AI.

Two such foundational values could be harmony and compassion.

Harmony requires individuals or organisations to consider themselves in relation to others, taking into account social relations as well as individual needs. It suggests a creative fusion, with actors collaborating, understanding, and sharing in way that builds strength through diversity.

Compassion requires that actors work to address power disparities, taking a responsibility of care for disadvantaged groups, and for the environment. It requires systems to be open and inclusive, and to work for the benefit of all.

With these common framings established, further detailed ethics principles can be translated into actions to reform education systems, develop participatory product design processes, develop multi-stakeholder networks to projects for social good, and help build a well-founded public dialogue.

National and international dynamics of AI: questions for further research and policy development

How can researchers and policymakers measure progress in AI in order to better understand areas of potential social benefit and risk, and the actions needed in these areas?

What foundational values should underpin the development of AI, and how can international consensus be rallied behind these?

What type of science diplomacy is necessary to ensure the safe and rapid development of AI technologies?

International collaboration in AI

There is now widespread consensus that action is needed to create an environment of careful stewardship of AI technologies, to ensure that their benefits are brought into being safely and rapidly, and that these benefits are shared across society. Researchers, industry and policymakers will play a role in creating such conditions.

History provides examples of how scientists can collaborate across borders in order to influence the path along which technologies progress. The Pugwash conferences on the development of nuclear technologies, for example, provided opportunities for scientists from both sides of the Iron Curtain to communicate with each other, sharing technical knowledge and insights into emerging applications. By meeting as individuals, rather than representatives of states, scientists could collaborate in ways that helped develop cooperative approaches to tackling concerns about the development of nuclear technologies, even in the midst of intense political conflict between states. Insights generated through the meetings have informed a variety of international agreements, such as the 1962 Limited Test Ban Treaty, 1968 Nuclear Non-proliferation Treaty, and 1972 Antiballistic Missile Treaty.

Such scientific diplomacy can play an important role in advancing international policy dialogues. Science diplomacy can take different forms:

- Science in diplomacy: science advice to inform and support foreign policy objectives;
- Diplomacy for science: diplomacy to facilitate international scientific cooperation; and
- Science for diplomacy: scientific cooperation to facilitate international relations[31].

AI researchers can play a role across these, working with research partners and governments to develop relationships and insights that advance research and inform policy.

AI research and policy today covers a huge range of issues, from the politics of AI technology (addressing concerns about bias and fairness; building safe and secure systems; ensuring explainability and transparency) to domestic policy (building skills; data governance; support for research) and international political economy (competition policy; global trade and data flows; international security), as well as a variety of application-specific challenges (data access; standardisation; regulation). These issues interact with a range of fundamental rights – privacy, free speech, equality, for example – and require expertise from across research disciplines, in addition to effective engagement with diverse publics.

Levers for action to promote the safe and rapid use of AI technologies can also operate over multiple levels – from the research culture that shapes the priorities of AI technologists and how they work, to national policies that advance the deployment of AI, and international agreements that set standards for its use.

31. See, for example: Turekian, V., Gluckman, P., Kishi, T. and Grimes, R. (2018) Science Diplomacy: a pragmatic perspective from the inside. Available at: http://www.sciencediplomacy.org/article/2018/pragmatic-perspective

Given the scope of this challenge, in the coming years researchers and policymakers will need to:

- **Prioritise:** identify areas of pressing need or significant benefit, and work to create solutions to these needs.
- **Specialise:** advance policy debates by focussing on specific use-cases, in order to move beyond high-level statements and develop application-specific responses.
- **Mobilise:** support mechanisms or intermediary infrastructures that advance international cooperation and progress in key areas of interest.

Progress in key areas could help create an environment of careful stewardship of AI technologies, in which the benefits of these powerful tools are shared across society. Such progress will require engagement across companies, investors, governments, researchers, and publics to create a vision for the development of AI that benefits society, and to advance collective action across public and private sectors – and international boundaries – in order to bring this into being. By defining global challenges against which AI could be deployed, designing incentives and opportunities for progress in those areas, and developing structures for dialogue across countries and research communities, business, governments and researchers can shape the development of AI for societal benefit.

Annex

Steering group

Royal Society Fellows who contributed to a Steering Group that developed content for the *AI: an international dialogue* workshop are listed below. Members acted in an individual and not a representative capacity, contributing to the project on the basis of their own expertise.

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

Robin Grimes FRS FREng
Professor of Materials Physics, Imperial College London

Peter Dayan FRS
Director, Max Planck Institute for Biological Cybernetics

Gil McVean FRS FMedSci
Professor of Statistical Genetics and Director of the Big Data Institute, University of Oxford

Fellows of the US National Academy of Sciences also contributed to this Steering Group, and the Society would like to expires its thanks to them.

Royal Society staff

Dr Natasha McCarthy
Head of Policy, Data

Jessica Montgomery
Senior Policy Adviser

Workshop participants

The *AI: an international dialogue* workshop was conducted under the Chatham House rule. Names of participants are therefore not listed in this document. The Society would like to express its thanks to all those who presented and participated at this workshop.

The Royal Society
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
W royalsociety.org

Registered Charity No 207043
Issued: September 2019 DES6339

Cover Image