THE
ROYAL
SOCIETY

中國科学院
CHINESE ACADEMY OF SCIENCES

The Royal Society and Chinese Academy of Sciences policy dialogue on AI

Summary note of an online event on 28 – 29 September 2020

Background

This note provides a summary of discussions at an online event on artificial intelligence (AI), held on 28 and 29 September 2020, as part of a series of policy dialogues between the Royal Society and the Chinese Academy of Sciences. Talks and discussions outlined the state of the art and future research directions in AI; AI ethics; and applications of AI to grand challenges such as health and climate change.

This note highlights key points made in discussions, demonstrating areas of common interest, diverse areas of research strength and opportunities for collaboration that emerged from these discussions. It is not intended as a verbatim of discussions and it does not represent the views or positions of any participants or organisations who took part.

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.

Chinese Academy of Sciences

The Chinese Academy of Sciences (CAS) is the national academy of sciences in China. Comprising of a comprehensive research and development network, a merit-based learned society and a system of higher education, CAS brings together scientists and engineers from China and around the world to address both theoretical and applied problems using world-class scientific and management approaches.

CAS comprises 104 research institutes, 12 branch academies, three universities and 11 supporting organizations in 23 provincial-level areas throughout the country. These institutions are home to more than 100 national key labs and engineering centres as well as nearly 200 CAS key labs and engineering centres. Altogether, CAS comprises 1,000 sites and stations across the country.

State of the art, challenges and future directions

The UK and China both have great strength in AI research. Research in the foundations of AI has resulted in many advances in areas such as computer vision, speech and language. AI researchers, both in the UK and China, have also described a set of scientific and technical challenges that have yet to be solved to deliver the full potential of the technology.

Generations of AI

Progress in AI has come in 'generations', each with their own paradigms. This is illustrated by the fact that AI today often refers to deep learning methods, a specific branch of machine learning using artificial neural networks; and 'tensors', a type of data structure used in linear algebra. Such paradigms enable progress within boundaries but can also create inertia, with innovation both in software and hardware necessary in order to break the mould. A new generation of AI (Generation 3) is focusing on creating frameworks to work with natural objects, such as molecules represented by a graph structure.

In 2015, the Institute of Automation in Beijing opened a new centre for brain-inspired intelligence. In 2017, China started a development plan for new generation AI. Among its goals, the plan seeks to: close the technology gap with leading countries by the end of 2020; fund research, advance technological development and applications at leading level by 2025; and by 2030, be a global leader in research and ethical principles for AI development. This will require comprehensive planning of basic research, technology development, infrastructure and education.

Measures of progress and remaining challenges

Much of the recent progress in AI is attributable to an increase in the amount of data to train algorithms and the 300,000x increase in compute power seen between AlexNet in 2013 and AlphaGoZero in 2017. In computer vision, for example, there is a correlation between the number of input parameters and accuracy.

Accuracy is only one aspect of progress in AI however, and assessing advances in AI can helpfully be done in relation to real world applications, in a human context, as well as measurements against a benchmark. Aspects such as energy use, cost, fairness and privacy protection also matter when it comes to real world uses of AI. They constitute a set of research challenges, together with interpretability, causality, verification, security and human-machine interaction[1, 2]. For example, deep learning runs the risk of being a 'black box', and there is ongoing research using mathematics and physics to understand how neural networks work.

Fairness appears to be a particularly pressing challenge. For instance, systems for cropping photos have been shown to perform better for white skin. Training face recognition models with a set including numerous images with non-white skin is not enough to avoid errors. The issue is not just with the empirical data set, it is also about the underlying theory. It requires finding theories to address worst-case issues, such as poor performance in recognising non-white faces.

[1] The Royal Society, *Machine learning: the power and promise of computers that learn by example*, 2017. See: https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf (accessed 4 February 2021).

[2] The Royal Society, *Explainable AI: the basics*, 2019. See: https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf (accessed 4 February 2021).

Challenges in real-world applications of AI

Real-world applications of AI require overcoming a number of challenges. For example, contrary to the bounded development of AI in a lab, real-world applications may work with an unlimited number of hypotheses, while relying on limited computing resources in the form of mobile or edge computing. Real-world data can be in limited supply, and training AI on such 'small data' may require labelled data which is expensive. There is a trend towards building smaller AI models using so-called 'small data' rather than big data, as many real-world problems are 'small data' problems. This requires metalearning and multitasking – centred on the idea of using knowledge from outside the immediate problem. Generative models generalise well on small data, as well as providing better explainability and reliability than other forms of deep learning.

The dynamic nature of data in the wild means that AI training will need to be sequential and incremental. Real-world data is not perfect and AI systems need to deal with data disturbance. AI systems today often lack in semantic understanding, which means they struggle to extract contextual information and, for example, do not readily utilise human knowledge and common sense. There are ongoing projects based on adaptive perception and learning in open, dynamic environments, based on integrating knowledge in neural networks.

Taken together, this means that real-world applications of AI must handle complex systems, with many dynamically interrelated elements and incomplete information. In addition, applications need to demonstrate reliability and support human acceptance and interaction. For applications with a high safety requirement, 'safety critical systems', being able to demonstrate the robustness of an AI model is very important.

While for some challenges it may be appropriate for AI systems to be put to the test directly in real-world applications to identify what works and what does not, for many applications it is necessary to first check for and address harms, bias and potential unintended consequences before release.

Energy cost

Current AI methods relying on big data can be particularly energy intensive. Deep learning is built on basic functions, including a 'multiply' function which contributes to its high energy consumption. Future developments should seek to go beyond such primitive elements. For example, brain-inspired AI models and computing promise a lower energy cost, while offering greater robustness and explainability[3].

Human-machine interaction

AI systems have the potential to change the way people think, behave and feel. The specific changes may be based on previous experience – for example people tend to either accept all or reject all machine-generated recommendations, based on their prior experience of such systems. But there may be longer term trends as we advance towards AGI – for example, some say that machines might be going through a humanisation process while humans might be going through mechanisation due to the way they interact with AI systems. To understand how human-machine interaction is changing human behaviour and experience, there is a need for a well-designed longitudinal study about the long-term effects of AI on people and society.

Artificial General Intelligence

While the current generation of AI research is overall focused on AI applied to specific tasks, or narrow AI, there is also some ongoing research towards Artificial General Intelligence (AGI). For instance, bringing cognitive empathy to AI models is the subject of active research in China. Researchers expect a number of benefits, for example, brain-inspired cognitive empathy models could improve AI safety. Other researchers caution that discussions relating to AGI can fuel fears and jeopardize the benefits to be gained from existing narrow AI. There are clearly complex questions about the potential impact of AGI, and researchers from both countries are collaborating with the Leverhulme Centre for the Future of Intelligence, regarding long term strategic research on the risks of AGI.

[3] The Royal Society, *Digital technology and the planet: Harnessing computing to achieve net zero*, 2020. See: https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet (accessed 4 February 2021).

Trust in AI

Many of the AI research challenges discussed above are critically linked to trust in AI. Researchers from both China and the UK recognise clear societal issues relating to the use of AI and data, including issues of privacy, security, accountability and trust, alongside ethical and legal implications, including trustworthy access to data. Both the UK and China have also established institutions, and developed sets of principles, to ensure that these issues are central to AI development.

While there are many commonalities in the response to these issues, trust in AI, and the perception of ethical and societal issues associated with AI, can vary depending on a number of cultural, social and environmental factors. Therefore, perception of these issues, and what might be acceptable ways to resolve them, does differ between in China and in the UK. For example, in China each individual who owns a mobile phone has a health code that they have needed throughout the COVID-19 pandemic to prove they were healthy, and the government is responsible to protect the corresponding data; meanwhile the UK deployed a a contact tracing app with a privacy-preserving design. Awareness of such differences is important because an AI project that is not 'culturally intelligent' might not be relevant to a particular culture.

Skills and diversity

When training machine learning and AI programmers, and encouraging greater adoption in industry, consideration of impacts on society should always be thoroughly integrated. Incorporating diversity at every stage in the AI process, from its creation to its deployment, is essential to appropriately manage its benefits and potential consequences[4]. For example, AI currently tends to average a lot of data points, and it could be made more relevant by better reflecting diversity in society.

Diversity should be included as part of any ethical framework for AI. It is imperative to have teams building AI that reflect the diversity of society, in terms of gender, age, ethnicity, disability, and skills. People who are neither mathematicians, nor scientists, nor engineers in the domain of AI, also need to be trained to be able to have a meaningful understanding of AI and interaction with AI. Scholarships could be devised to this end, and a large proportion targeted at underrepresented groups. There is a need to develop simple frameworks to enable diverse stakeholders to audit AI, although how it would be achieved and resourced is unclear.

Regulation and standards

The right approach for the governance of AI, and the right level for regulation, are still being debated. Evidently, government regulation cannot solve the issue of diversity in research. It is an internal matter of self-regulation.

In the world of work, AI will often be deployed as an automated decision-support tool. This will require providing professionals with training regarding the regulatory background for using AI in the sector – in a similar way that doctors are trained to understand the regulatory aspects of using MRI. Looking at the medical sector, privacy is a key issue. There are existing regulations, for example GDPR in Europe, but currently anecdotal evidence suggests a lot of people do not understand the spirit of the law.

There is a role for international governance, and in particular shared standards, although lessons must be learned from other areas to make these effective. For example, there may be lessons to learn from how standards evolved in the manufacturing and financial sector; and useful comparisons to be drawn with unregulated spaces such as computer science. This could be the subject of a a potential research project.

[4] World Economic Forum, *Here's why AI needs a more diverse workforce*, Ronit Avni and Rana el Kaliouby, 2020. See: https://www.weforum.org/agenda/2020/09/ai-needs-diverse-workforce (accessed 4 February 2021).

Principles in practice

There are more than 70 different AI principles worldwide and there are many areas where they overlap, however when focusing on the detail, it is apparent that there are a number of important differences. For example, harmony does not appear as a theme in Western principles, but harmony is a focus in the Beijing AI principles[5]. This reflects the importance of harmony in Chinese culture. Harmony is also a common concept or principle in Japanese culture. For example, the Japanese concept of 'Wa' designates harmony between social groups and means that a focus be placed on harmonious communities over personal interests – this includes harmony between people and AI. Japan is taking the approach of building AI as a partner to society. The current view in China, in most scenarios, is reported to be closer to using AI as a tool for society, while there are also perspectives and efforts on building Human-AI symbiotic society. Western countries' perceptions of AI has the potential to oscillate between a tool for, and a competitor to humans.

There are many AI ethics research groups and centres, including for example: a £150m donation to the centre for humanities at the University of Oxford to include an institute for ethics in AI[6], a China-UK Research Centre for AI Ethics and Governance, the UK's Centre for Data Ethics and Innovation (CDEI) and the Ada Lovelace Institute. However, principles must be turned into practice and, for instance, in organisations, the C-Suite must meet to discuss how they incorporate ethical principles into daily operations, as well as consider how to bring in customers and business partners.

[5] See Beijing AI Principles. See: https://www.baai.ac.cn/news/beijing-ai-principles-en.html (accessed 4 February 2021).

[6] Institute for AI in Ethics. See: https://www.schwarzmancentre.ox.ac.uk/ethicsinai (accessed 4 February 2021).

AI for grand challenges

For the field of AI to be trusted, it must also demonstrate its ability to harness the technology to tackle societal challenges. There are, for example, international initiatives to use AI for sustainable development goals[7]. While it will be important to develop mechanisms to identify major challenges of relevance to society, two areas currently stand out: AI for health and care, and AI for the planet.

AI for health and care

AI has applications in brain imaging and in neuropsychiatry, for example to provide objective diagnosis, individualised treatment and targeted therapies for conditions such as schizophrenia or Parkinson's disease. In disorders of consciousness, eg where severe brain injury is a cause, AI systems have been developed that can predict with 88% accuracy which patients would regain consciousness and recover.

Progress in this area builds on an explosion of neuroimaging data – including clinical data, the fine-grained mapping of the brain or 'brain atlas'[8], and advances in AI. The hope is also that advances in this field will lead to new models for brain inspired computing.

Ultrasound imaging is another area of medical image analysis which has seen significant progress over the years. AI and image analysis can help democratise the use of ultrasound devices, which have become increasingly small but are not yet widely used. Nimble training is necessary to rapidly bring healthcare staff up to speed.

Progress in AI for ultrasound imaging has relied on supervised deep learning, meaning neural networks trained on labelled examples. In a latest generation of medical image analysis, there is an emerging interdisciplinary approach, using less data and bringing in more contextual knowledge.

Sonography data science has developed recently. This enables learning from best practice by experienced sonographers – eg understanding what tasks take time and what tasks people find hard by using eye gaze tracking.[9] Studying the gaze of sonographers yields insight that goes beyond the information sonographers would have provided when asked how they analyse a sonogram. It is then possible to use such 'video gaze models' to proactively guide sonographers to find information in scans and to label them.

A number of challenges remain, such as in annotating data and evaluating algorithms. Researchers and practitioners will need to factor in potential imbalances and bias in the data, for example due to different sites, devices or populations. Establishing a way to evaluate clinical AI will be essential, in particular to assess its interpretability and trustworthiness.

AI for the planet

AI also has great potential to help tackle the pressing environmental challenges relating to climate change. Already used in weather forecasting, AI has been used to improve the predictive ability of seasonal forecasting and modelling of long-range spatial connections across multiple timescales.

Applications to atmospheric science include more accurate forecasting of extreme weather events, and AI to forecast the air quality of cities. 'Nowcasting' is another area of development, whereby very short-range forecasting is becoming ever more precise, for example to forecast precipitation.

[7] Think tank, *AI for Sustainable Development Goals*. See: http://www.ai-for-sdgs.academy (accessed 4 February 2021).

[8] China's Brainnetome Atlas project. See: https://atlas.brainnetome.org (accessed 4 February 2021).

[9] Chatelain, P., Noble, A.J. *et al*, *Evaluation of Gaze Tracking Calibration for Longitudinal Biomedical Imaging Studies*, 2018. See: https://www.researchgate.net/publication/327451311_Evaluation_of_Gaze_Tracking_Calibration_for_Longitudinal_Biomedical_Imaging_Studies (accessed 4 February 2021).

Challenges remain, in particular finding ways to best build on existing physics models and embedding prior knowledge. There are highly complex models of cloud physics and integrating these into AI models requires a research effort in itself. Applying AI to climate science calls for collaboration between computer scientists and physicists.

Another area of application of data is towards monitoring the environment from the perspective of the carbon cycle. This requires a well-established data infrastructure, as well as the development of technology that can be trusted to measure carbon emissions[10].

For example, researchers have used circumstantial evidence on emissions drop during COVID lockdowns[11]. This was not straightforward as current emissions monitoring, on a yearly basis, hardly supports such a granular analysis. Using a range of data, together with data science, the research showed different sectors of the economy changed their activity to different extents during the first lockdown in Spring 2020. There was a small decrease in power production, a bigger drop of about a third of emissions in industry, a halving of surface transport emissions, and a three fourth drop for aviation. However, this drop in overall emissions was sustained over a relatively short period of time only, and overall emissions for 2020 were estimated to be only 4 – 7% lower than for 2019.

There are obstacles to making data available for addressing challenges, such as the commercial value of data, cost in processing the data, and of course privacy. Researchers should be working with private companies to expand the use of data and therefore research opportunities. There is also value in research and development on alternatives to real-world data such as synthetic data, which could be part of the solution.

[10] The Royal Society, *Digital technology and the planet: Harnessing computing to achieve net zero*, 2020. See: https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet (accessed 4 February 2021).

[11] Le Quéré, C., Jackson, R.B., Jones, M.W. *et al*, *Temporary reduction in daily global CO2 emissions during the COVID-19 forced confinement*, 2020. See: https://www.nature.com/articles/s41558-020-0797-x (accessed 4 February 2021).

Opportunities for collaboration

The discussion noted the importance of collaborations between the UK and China on AI, alongside a number of potential areas for collaboration including health and climate.

Interdisciplinary AI research and partnerships

Interdisciplinary research is valuable to ensure that other areas of science benefit from using AI methods. This can promote advances across science, with health and environmental research being examples. To ensure AI embeds ethics, and to ensure it is fit for purpose to tackle real-world challenges, there is a clear need to identify mechanisms to promote interdisciplinary research. This could be through multi-disciplinary programmes of research, that would include ethics and social sciences from the start. This will require crossing disciplinary and cultural barriers.

Pushing the frontiers of AI

Collaboration can help to make advances in the foundations of AI – such as developing methods that are less energy intensive, that learn from human intelligence, or that work with small data.

Learning from diversity

Collaboration should be sensitive to, and learn from, the differences in approach between the UK and China, learning from this diversity to help develop AI that better reflects the complexity of society.

Sharing best practice, tools, techniques, and ethical concerns can enable learning between researchers in China and the UK. This might start with relationship building, especially between early career scientists from each country. Promoting or creating research Fellowships – for example focused on research at the interface between science, ethics and philosophy – could foster collaborations. Funding schemes can also enable a focused approach to progressing the frontiers of AI. Further discussions, for example through online events, will enable exploration of these areas for collaboration, and of the kinds of Fellowships and funding schemes that can support sustained research relationships.

THANKS
The Society would like to express its thanks to all those who participated in discussions.