ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

POST-ELECTION BRIEFING: BALANCING AI INNOVATION AND REGULATION

RSS manifesto ask

An investment of £250 million over the next parliament for AI (artificial intelligence) open-source development including support for an open-source unit in government, an open-source fund for business and developing open-source software – to democratise AI and allow the UK to compete internationally.

Summary

The AI ecosystem is dominated by big tech companies who pursue closed-source approaches to AI (in that their code and data sources are not widely accessible to third parties). This risks reducing trust in the technology to the point where it becomes challenging for UK businesses to effectively innovate – and it is already challenging for these companies to compete with large multinational companies. Investment in open-source resources is important both to build trust and to enable UK businesses to compete.

What's the problem?

The market for AI is dominated by big tech companies – based outside the UK and with minimal democratic oversight. These companies have developed closed AI systems (using the public's data and copyrighted material). We believe that a reliance on closed systems and a lack of democratic oversight risks corroding public trust in the technology.

At the moment the AI ecosystem is dominated by a few big tech companies. They have developed closed-source AI applications – the most well-known example being Chat-GPT. Chat-GPT is a closed AI system in the sense that the system's software is securely held by OpenAI and is only shared with a limited set of partners for the purpose of vetting. The company provides access to the AI interface as a service, but the process by which Chat-GPT produces its responses to queries is closed to wider scrutiny. This means, for example, that the public do not have access to information about the data that has been used to train the AI model (and owners of data cannot easily check whether the system is using their proprietary information).

This is problematic for two main reasons. First, the size and power of the big tech companies makes it a very hard environment for new, UK-based AI companies to compete. Advancing openness in the approach to AI is important for enabling competition – open-source approaches breed innovation, transparency and trust and UK businesses that promote these values need to be empowered to compete against multinational behemoths.

Second, the closed approach makes it harder to identify negative impacts of AI systems. There is limited diversity in the people developing AI – the tech industry remains predominantly white and male – and this can feed through both in terms of datasets used to train AI models (if you have biased inputs, then you are liable to have biased outputs) and in terms of a low awareness of alternative perspectives, which means that AI products may not be especially nuanced in communicating with people from a range of cultural backgrounds. With a closed approach it is impossible for the public or external experts to assess the inputs of an AI system with an eye towards potential adverse impacts. We have seen this repeatedly with how statistics and data are used that this lack of transparency breeds low trust – if the UK is to benefit from AI innovation it is vital to consciously cultivate trust in the technology.

How to fix it

Investing in open source is important for two reasons: first, it would allow the UK to scale-up AI systems and compete internationally with big tech companies and China; second, the open approach is important for transparency and building trust in the technology.

We call on the government to invest £250 million over the parliament in open-source development to:

- Support a government-funded unit of open-source developers who would:
    - Contribute to open-source software projects of strategic importance to the UK.
    - Provide internships to help build and transfer skills.
    - Support the use of open-source in the UK through knowledge sharing, training and building the community.
    - Collaborate internationally with similar projects such as OLMo (open Language Model) at the Allen Institute for AI.
    - Disseminate knowledge about large language models within the UK tech community.
- Introduce a new Fund – modelled on Germany's Sovereign Tech Fund and Prototype Fund – to support open source projects outside government.
- Support the development of open-source software.

To have maximum impact this investment would need to be carefully targeted on the needs of the UK community – it should not be used on things that already exist or that are easy for companies to build themselves.

One criticism of open-source approaches to AI is that they come with greater risk of abuse – that there are greater security risks than with closed systems. It is true that there are risks associated with open-source AI. However, that should not be an argument for not backing it in the ways we have set out – it is an argument for being cautious and regulating in a manner that works for open-source approaches. We would note that there are also risks associated with the closed approach: closed models can be abused by bad actors (and it is harder to see when this might happen) and deployed without full awareness of the system's biases and potential adverse impacts. A black box approach, with a few big tech companies acting as gatekeepers, has its own risks and is not a sensible path forward.

It is important, though, that while investing in open source the government emphasise backing systems that are easy to use and easy to trust. Combined with a competitive market environment, an open approach to AI can help ensure that advancements are accessible, inclusive and transparent.

RSS manifesto ask

The AI Safety Institute to fund a unit dedicated to developing new evaluation methodologies to keep pace with technological advances.

Summary

Robust and trustworthy evaluation of AI (artificial intelligence) is important both to manage the risks associated with the technology and build public confidence. The UK, through the AI Safety Institute, is already establishing itself as an international leader on evaluation of AI – but there is more to be done. The technology is very fast moving, and it is very challenging to evaluate it in scenarios that accurately represent the real world. These are the types of challenges that statisticians and data scientists have experience in solving, and the UK has a depth of expertise in evaluation methodologies. We call for the AI Safety Institute to invest in a unit dedicated to developing new evaluation methodologies to help keep pace with technological advances.

What's the problem?

It is widely recognised that the increasing use of AI brings with it a range of risks – from the potential to entrench societal biases through to apocalyptic scenarios. Evaluation has a crucial role to play in assessing the performance and capabilities of AI systems and ensuring that we protect against these risks.

An example helps to demonstrate the importance of evaluation. Consider the application of AI in healthcare for medical imaging. AI models, particularly deep learning-based algorithms, have shown significant promise in analysing medical images and assisting healthcare professionals to make accurate diagnoses. Evaluation is important here for many reasons – here we highlight three illustrative examples. First, the accuracy of the AI is crucial – because errors can have serious consequences for patients – and proper evaluation is needed to ensure that the performance is at least at the level of human doctors. Second, the AI must be able to generalise – medical images can vary greatly and evaluation is necessary to check the ability of the AI to generalise well. Third, AI tools need to be evaluated to ensure that they are properly balancing sensitivity (the ability to detect true positive cases) with specificity (the ability to avoid false positives) – if the balance between sensitivity and specificity isn’t right you risk either missing cases or unnecessarily worrying a large number of patients.

The AI Safety Institute has already made some good progress around evaluations – providing, eg, a new platform to evaluate, strengthen and accelerate global safety evaluations. It has developed a software, Inspect, which enables testers to assess specific capabilities of individual models. This is a welcome move and it is a strong first step towards securing a leadership role for the UK in AI evaluation.

However, evaluation of generative AI tools (those capable of producing images, text etc) is difficult because the technology is changing at a rapid pace and key questions – like whether an AI tool is fair – involve concepts that are difficult to tie down. There are also limitations to current evaluation methodologies in the context of generative AI tools using Large Language Models (LLMs), which prevent us understanding their capabilities and limitations. There are a few particularly pressing problems:

- Existing evaluation metrics can overly focus on narrow and task-specific benchmarks, overlooking the models' broader performance across various domains and tasks.
- The lack of standardised evaluation criteria makes it challenging to compare results across different studies – risking inconsistent and potentially misleading conclusions.

- Evaluation datasets frequently fail to represent the complexity and diversity of real-world language usage, resulting in models that perform well on artificial data but struggle to generalise to practical scenarios.
- Evaluation needs to be continuous – model performance changes over time, but a lot of evaluation work is designed for one-off checks pre-deployment. That model is fine for standard algorithms, but is not sufficient when dealing with AI tools that change post-deployment.
- Communication of evaluations in a way that is helpful to very different audiences – eg, regulators, teams using LLMs to design products, people using or affected by those products – is challenging but vital if evaluations are to have impact.
- The assessment of biases and fairness in LLMs is an ongoing challenge, with current methods often insufficient to capture subtle biases or to understand the potential legal and ethical implications of the model's outputs.

Addressing these limitations is crucial to ensure that evaluation of AI tools aligns closely with real-world requirements and to give the public confidence in how the technology is used.

How to fix it

Statistics has an important role to play in improving the evaluation of AI systems. New statistical approaches can provide more informative, nuanced, and reliable assessments of models' performance and behaviour. Statisticians are also expert at uncertainty estimation – offering users more reliable confidence intervals and uncertainty measures for AI predictions. Statistical methods can also systematically quantify biases across different demographic groups, facilitating more robust and fair AI models. Statisticians are also expert at communicating the outcome of evaluations appropriately for their audience.

The UK's statistical community has experience of establishing international evaluation leadership in the context of health, in the 1990s. Then, there were a great many trials that were being funded but there was no clear centre of expertise on methodology for evaluating the evidence in terms of cost/benefit. As a result the NHS R&D Health Technology Assessment Programme was set up in 1993 to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies was produced in the most efficient way for those who use, manage and provide care. This was a massive success, establishing the UK as a world leader in evaluating healthcare. Given the UK's track record in healthcare evaluation, there is an opportunity to establish ourselves as genuinely world leading in AI evaluation methodology as well.

We call on the AI Safety Institute to invest in a unit of statisticians and data scientists dedicated to developing new evaluation methodologies to keep pace with technological advances.

RSS manifesto ask

A public register of cases where a complex algorithm or artificial intelligence (AI) tool is used in the public sector – with risk assessments carried out in cases where the tool directly impacts on citizens (eg, its use in facial recognition or in informing decisions around welfare payments).

Summary

Complex algorithms and AI tools have the potential to make public service delivery more productive. However, if the use of these tools becomes widespread without public scrutiny (as seems to be happening in, eg, the justice sector) there is a risk that the public will lose trust in the technology. Public trust is required if the tools are to be used in decisions that impact on individuals. Transparency is vital in building public trust. The starting point should be a register of all cases where complex algorithms or AI tools are used in the public sector. Risk assessments should be conducted and made public in cases where the tool directly impacts citizens, so the public can see the decisions that have been made and what steps have been taken to de-risk the use of such tools.

What's the problem?

Complex algorithms and AI tools have a wide range of possible applications in the public sector – and there is huge potential for the technology to improve productivity and help make better decisions. Used well this technology can have positive impacts – potentially helping in a wide range of areas from crime detection, to healthcare diagnoses to assisting welfare entitlement decisions.

The risk is that the technology gets deployed in these areas and used to make decisions that impact people's lives without proper scrutiny – that the practice of AI-assisted decision-making proliferates across the public sector without proper oversight. Indeed a recent House of Lords Select Committee report, *Technology Rules?*, suggests that this has already happened in the justice system where they highlight "a new Wild West, in which new technologies are developing at a pace that public awareness, government and legislation have not kept up with" (p3).

There is currently a lack of transparency in how complex algorithms and AI tools are being used in the public sector. Finding out about how the technology is used by government currently relies on piecemeal individual investigation in areas of interest – a recent example is Big Brother Watch's work to uncover the Department of Work and Pensions's use of algorithms in assessing potentially fraudulent cases. This approach is not satisfactory, and risks undermining public trust.

We saw the importance of transparency in algorithms during the pandemic, when Ofqual attempted to use an algorithm to assign grades to students in the absence of exams. There was a lack of transparency around the algorithm in advance of people receiving their grades, which meant a missed opportunity to engage with experts. To their credit, however, they did release the details of the algorithm on results day – this enabled experts (including the RSS) to assess the algorithm and determine that the algorithm was not robust enough to bear the weight being put on it. It was subsequently abandoned.

The root issue is that using complex algorithms and AIs for the public good is only partly about the quality of the models themselves. You could have an algorithm or AI tool that works exactly as intended and that is based on full and representative data – but that is only part of the story. Decisions will be made in the design of the systems that embed political judgements. In the case of Ofqual's algorithm, these judgements were around how much grade inflation to allow and how to be fair to individual students. The public need to be able to know when complex algorithms and AI tools are being used in a way that affects them and what decisions have been made in how they are developed and used.

How to fix it

Transparency is key. After the Ofqual algorithm fiasco, the RSS asked the Office for Statistics Regulation to conduct a review of the case and to draw lessons for future use of statistical modelling in policy making. Their report, *Ensuring statistical models command public confidence*, drew three key lessons for organisations developing algorithms (p.62):

- Be open and trustworthy: this means being transparent about aims of the model and the model itself, including limitations, and acting on feedback.
- Be rigorous and ensure quality: this means ensuring that there are clear governance processes and accountability, involving subject matter and technical experts and ensuring that both data used as inputs and any outputs are quality assured.
- Meet the need and provide public value: in this context it is particularly important to engage with affected groups to test and ensure the acceptability of any new approach.

These are all important, but we wish to emphasise the first point as fundamental. Without transparency about what models are used and what they are aiming to do, we risk a situation in which the public lose faith in the technology.

A key starting point is the creation of a publicly available register that sets out where complex algorithms and AI tools are being used. In cases where citizens are directly affected by the technology risk assessments should be conducted – and made available – setting out decisions that have been made, limitations of the model, the possible impact this will have and what mitigations have been put in place.