ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

POST-ELECTION BRIEFING: BALANCING AI INNOVATION AND REGULATION

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.