After the Summit: progress in public policy on AI

Position Paper 2023-03

December 2023

Ben Hawes, Wendy Hall

The UK's international Artificial Intelligence Safety Summit has answered some questions and sparked new ones. This is a good moment to reflect on what it delivered, what it didn't cover, and how to influence development of AI in the future, in the interests of societies globally.

First, it's great to be able to report that the Summit was in many ways a success, indeed more of a success than many people thought it could be. It was arranged and delivered fast. It had to manage difficult questions about the scope and the invitee list. There were good reasons to fear that it might not be more than a superficial, passing event. It is greatly to the credit of the organisers that it became more than that.

The Summit could also easily have been submerged among other recent developments, because there have been enough of those. The last month has been busy for AI and AI policy, even in the context of a packed year so far.

Immediately before the summit, the United Nations announced a new high-level advisory council on AI, and I'm proud to say that they invited me to be a member.

And then two days before the Summit, the White House issued President Biden's Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. The order "establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more."

The Executive order sets out expansive, complex and diverse ambitions for AI in the USA, including on equity, civil rights and impacts on workers. It is a major step forward. The EU AI Act has been the subject of very heated debate within and between EU institutions. It has now passed, though the nature of recent debates shows how difficult it is for legislation to keep up with technology developments. The US had previously made much less ground in comparison on proposals for government action and legislation on AI. That has now changed, and in the UK we will need to track how those ambitions are taken forward in practice, and how potential conflicts between economic and social aspirations are managed.

In that context, the UK AI Safety Summit did well not to slip quickly from sight in the technology news cycle. In fact it did much more. It brought together 28 countries and the EU to agree the Bletchley Declaration. The Declaration includes a shared resolution to “support an internationally inclusive network of scientific research on frontier AI safety that encompasses” and to “sustain an inclusive global dialogue that engages existing international fora and other relevant initiatives and contributes in an open manner to broader international discussions”.

Those countries included China, India, Brazil, Indonesia and Nigeria, and so the Declaration managed to reach well beyond western technologically-advanced democracies, to include a much more substantial proportion of the world’s population. The commitments are broad and high level. Many collaborations in channels will be needed to make them really meaningful, but that breadth of inclusion represents a welcome step forward.

The UK also gained a new AI Safety Institute, to be developed from the existing Frontier AI Taskforce, to “act as a global hub on AI safety, leading on vital research”. The Institute gained an immediate boost with a commitment that US institutions, notably the National Institute of Standards and Technology (NIST), which will also host an AI Safety Institute, would collaborate with it. The governments of Germany, Singapore and Canada expressed support for the new Institute, as did some major international AI companies. As yet, we don’t know exactly where the UK AI Safety Institute will sit in government, or what form those international collaborations will take. Those collaborations could become an important source of common approaches and shared expertise.

It was announced at the Bletchley Summit that the next Summit will be in Korea in March, and the one after that will be in France in November.

What didn’t the Summit focus on?

As many commentators have pointed out, the focus of the Summit was on research on longer term risks, not on known existing harms and how those could play out in economies and societies. As the Ada Lovelace Institute put it, “many within industry, academia and civil society have rejected the Summit’s focus as overly narrow and insufficiently attentive to the wide range of AI harms people are currently experiencing – without adequate protection.”

The Declaration overtly recognises the concerns felt in common across countries, that are not within the scope of its resolutions.

AI also poses significant risks, including in those domains of daily life. To that end, we welcome relevant international efforts to examine and address the potential impact of AI systems in existing fora and other relevant initiatives, and the recognition that the protection of human rights, transparency and explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, bias mitigation, privacy and data protection needs to be addressed. We also note the potential for unforeseen risks stemming from the capability to manipulate content or generate deceptive content. All of these issues are critically important and we affirm the necessity and urgency of addressing them.

For many people, these issues are what's the matter with AI. It is reasonable to argue that these should have been within the core focus of the Summit. It is also reasonable to suspect that the focus of the Summit on longterm risks was strongly influenced by major international technology companies, who have been known to downplay the known risks of AI applications and to head off regulation that addresses them.

On the other hand, you might argue that international collaboration to improve shared knowledge of longterm risk is a valuable gain in itself, and that it was not apparently imminent in any other forum. Work on longterm risks will very likely improve understanding of the full spectrum of risks. UK AI development should also certainly gain from having a global centre for AI safety located here, in terms of access to expertise and investment.

It is also not obvious that achieving that collaboration has materially held back international collaboration on the near-term and comparatively better understood risks. Collaboration on those risks has begun in other convening organisations, and (as I'll say more about) it is complex and tied up with many other political issues and challenges.

So I suggest we welcome the progress the Summit made, while not allowing its limited focus to constrain international collaboration to make AI responsible to societies. Focus on longterm risk should not take resources or political attention away from the other risks.

Future AI: collaboration and inclusion

Along similar lines, the Summit was criticised for representing some interests and not others. It included governments, major AI companies and some research institutions. It did not include many non-governmental organisations which champion human rights, safety and security.

It is more difficult to disagree with this. Collective action to deliver responsible AI across sectors and places will need the participation of civil society bodies and other non-governmental organisations. That includes organisations that work in the interests of poorer and marginalised groups across countries. Even within research, responsible AI will need the involvement of researchers from social sciences and a wide range of other disciplines, not only researchers in digital technology.

It is right to talk about who was not included the Summit, but let's do that as a spur to bringing those organisations and groups into future processes where they can have an impact on decisions and directions.

There are already good reasons to feel that the Summit's impact has been positive beyond its direct focus and outputs. The Summit has already galvanised many people and organisations who were not invited to it. A really impressive range of events happened outside the Summit, including the AI Fringe. Some of those events brought a wider set of perspectives to bear on the same themes as the Summit. Others brought together expertise on the issues outside the Summit's scope. I took part in an excellent working group on India-UK collaboration on responsible and trustworthy AI at the Royal Society. I hope the views and ideas that were expressed in all of those events are being captured and will be brought into future collaborations.

The Summit and the activity around it formed a peak in AI policy debate, but from now on we should probably assume that the debate won't really stop. At the end of November, a Private Member's Bill on AI regulation was introduced in the House of Lords. We can also expect the Government's response on issues raised in the AI Regulation White Paper consultation, which will may be the next public step in development of the AI regulatory framework for AI. The Department for Science, Innovation and Technology has issued new business guidance to boost skills and unlock benefits of AI, a welcome effort to encourage wider engagement and capability building.

If we collectively want socially beneficial uses of AI to become the norm, then we should help grow mechanisms that can influence governments, particularly governments with more limited internal resources to devote to that. In time, AI will have impacts across the principal areas of public policy and decision-making. There is no one model for managing this, because AI will have such a wide set of impacts, but there is a need to decide how to bring AI into existing channels and organisations for collaboration, and for bringing more and broader social issues into AI policy.

In recent years we have already seen governments trying to work out whether they need wholly new units to address AI, or to bring AI expertise into existing ones. The more successful approaches generally seem to involve a pragmatic mix: some combination of central AI strategic leadership, understanding within each department of the implications and opportunities for their responsibilities, and expertise to employ AI to fulfil those responsibilities better and in new ways.

A good example of pragmatism is the UK's Digital Regulation Cooperation Forum. A group of sector regulators have recognised that digital technologies (increasingly including AI) will have different impacts across (for example telecoms and the financial sector, but that much can be learned and adapted from one sector to another, and there will be new questions around the overlaps between sectors.

Large companies are going through a similar evolution, recognising the need for expertise in overall AI strategy and foresight, and in practical application.

Inter-governmental organisations, non-governmental organisations and convening bodies will need to address similar questions. If there are important voices and interests missing in public policy on AI, then there is a need for practical proposals to include and empower those voices and interests. The recent United Nations advisory council is one welcome addition.

None of this is simple. Inclusivity in policy-making is also political. Positions taken by governments about it will reflect their political assumptions, and will shift over time.

The balance between competition and collaboration between nations may also fluctuate over time. There may be a case for collaborative international activity to reach a common understanding of what - in public policy on AI - is currently addressed by international collaboration, what is not, and which interests do not have influence, and arguably should. A better shared view of what greater inclusivity could look like, and how it could work in practice, would be more powerful than appeals to the principle.

A shared view of the opportunities would be supported by understanding of what examples of comparatively inclusive policy-making, particularly in relation to technology, have been successful in the past. There are some models. Human rights have gained broad international support, and have been developed over time, including in relation to emerging technologies, including the internet.

Another model is the Sustainable Development Goals, which alongside human rights represent internationally supported objectives to aim AI towards, and by which to measure its impacts. So, it is encouraging that the new United Nations advisory body will “offer diverse perspectives and options on how AI can be governed for the common good, aligning internationally interoperable governance with human rights and the Sustainable Development Goals”. The UN’s declared interest in AI is not restricted to limiting harms: there is also clear enthusiasm for using AI in pursuit of the SDGs.

Crucially, the SDGs offer a set of positive directions to guide policy. That seems missing in a lot of recent debate about AI. Listening to estimates of the value that AI could add to economies, you could be forgiven for thinking that it offers only economic growth and at the expense of additional risks and inequality. We will collectively need to use more imagination than this. We should demand more exploration and more international collaboration on how AI can positively improve lives, societies and the environment. If we only think about risks, we may not even understand the full spectrum of risks.

Working together on risks is vital, but so is working together to imagine new ways that AI can help people, and build healthier, stronger and fairer societies.

About the Authors

- Ben Hawes, technology policy consultant and Associate Director at the Connected Places Catapult
- Professor Dame Wendy Hall, Regius Professor of Computer Science at the University of Southampton

About WSI

The Web Science Institute (WSI) draws together the University's world-class, interdisciplinary, sociotechnical expertise in Web Science, Data Science and Artificial Intelligence. We act as a focus for international esteem as we create new opportunities to bring faculties, schools, and disciplines together to leverage the unique role of online technologies in tackling global challenges, including the challenges posed by society's use of those technologies themselves.

The WSI was established to study the evolution of the Web and society but has evolved into an institute that specialises in the sociotechnical study of the evolution of digital technologies and society in general, focussing currently on, but not restricted to, the new discipline of Human-Centred Artificial Intelligence (HCAI) as well as Web Science.

The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the University of Southampton, the Web Science Institute or its Board of Directors.

Web Science Institute, Building 32, Highfield Campus, University of Southampton, SO17 1BJ
wsi@soton.ac.uk (DOI: 10.5258/SOTON/WSI-WP007)