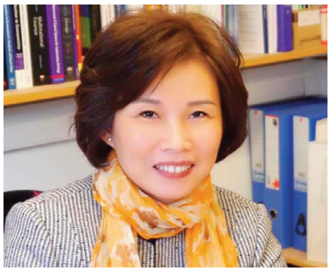
THE UK AI SAFETY SUMMIT: A 'HISTORIC MOMENT' OR LARGELY SYMBOLIC?

John Tasioulas

Professor Carissa Véliz

Felipe Thomaz

Alex Connock


Xiaolan-Fu

Robert Trager

Lulu Shi

Brent Mittelstadt

Ciaran Martin

Helen Margetts

On 1 and 2 November 2023, the UK Government hosted the first global AI Safety Summit, bringing together leading nations, technology companies, AI researchers, and multilateral organisations. The aim was to consider the risks of AI, especially at the frontier of development, and to discuss how these could be mitigated through international collaboration. But did it succeed? Leading AI researchers at Oxford University give their views.

EVERYONE AROUND THE TABLE

Some would argue that regardless of what was decided at the summit, simply convening such a broad range of stakeholders could be regarded as a success. "Perhaps the biggest achievement of the summit was that China was brought into this discussion" says **Professor John Tasioulas, Director of the Institute for Ethics in AI, Oxford University**. "This is absolutely vital, since there cannot be the meaningful global regulation of AI that is needed without China's participation."

This broad international representation led to all 28 attending countries and the European Union signing 'The Bletchley Declaration': a joint agreement on the need for international collaboration to ensure advanced AI technologies are developed safely. Whilst Prime Minister Rishi Sunak described this as "a landmark achievement" that would help "ensure the long-term future of our children and grandchildren", Professor Tasioulas believes that the Declaration will have little real impact unless it is translated into meaningful actions. "The concept of 'safety' is stretched in the Declaration to include not only avoiding catastrophe, but also securing human rights and the UN Sustainable Development Goals etc. This is a highly unstructured list of concerns. As such, the value of the Declaration may be largely symbolic. The heavy lifting still needs to be done to translate the Declaration's values into effective regulation."

COOPERATIVE REGULATION?

Nevertheless, some progress may have been made towards a framework for regulation. At the summit, a group of 11 Government signatories and eight leading AI companies - including Meta, Google DeepMind, and OpenAI – agreed to collaborate on testing the latest AI products before their public release. As a voluntary initiative, however, this may ultimately lack any real teeth to hold big tech companies to account.

According to **Associate Professor Carissa Véliz, from Oxford's Institute for Ethics in AI**, if the UK wants to be taken seriously as a leader in AI regulation, "it would do well to properly regulate AI within its own borders. It should also give less prominence to tech executives who, by definition, cannot regulate themselves—their financial conflict of interest disqualifies them."

In this respect, many felt that the Prime Minister's decision to host Elon Musk for a near hour-long interview only highlighted the technology sector's excessive influence. "As a political scientist, I was uncomfortable that a head of state should be asking questions of a tech mogul and not the other way around" says **Helen Margetts, Professor of Society and the Internet at Oxford University**. "Brilliant though he may be at technological development, Musk has shown himself to lack skills in understanding the social world – ill befitting him to tackle the central question of the summit: how will frontier AI affect all of us?"

Felipe Thomaz, Associate Professor of Marketing at Oxford University, adds that the new agreement may even play directly into the hands of the largest tech companies. "This announcement represents an incredibly successful year-long lobbying effort by the largest AI players and providers globally, who were deeply concerned about the ease of entry by homebrewed competitors into their arena" he says. "By requiring government approval prior to public testing and product releases, these governments have raised very tall barriers to enter the AI economy. This will favour the largest companies who already have inroads into government and who already spent the previous year accelerating their own R&D with the knowledge of this development."

A NEW UK AI SAFETY INSTITUTE

During the summit, the UK Government announced the creation of a new UK AI Safety Institute, with the mission "to minimise surprise to the UK and humanity from rapid and unexpected advances in AI." Although a welcome development, there are concerns that this will have too limited a scope to address AI threats on a global scale. "The use of AI by bad actors is one risk that can indeed be partially mitigated by AI safety institutes of this kind, by spy agencies checking new models and so forth" says **Dr Alex Connock from Oxford's Saïd Business School** and author of 'The Media Business and Artificial Intelligence.' "Although even that will be hard as Large Language Models increasingly become something you can tune and run on a laptop, and there are actually countries out there who didn't make the summit that nonetheless might want to use unregulated models of their own."

Xiaolan Fu, Professor of Technology and International Development at Oxford University, adds: "I hope that the Institute also considers AI safety in different contexts, taking into consideration low-income countries to make sure AI is safe and working for good in all countries at different levels of development."

One of the Institute's first activities will be to host an expert writing group chaired by Yoshua Bengio (one of three so-called 'godfathers of AI') to produce a 'State of the Science' Report on the capabilities and risks of advanced AI. According to **Professor Robert Trager, Director of the Oxford Martin AI Governance Initiative**, both the new Institute and the commissioning of the report demonstrate a "real show of leadership" by the UK Government. "A state of the science consensus report could potentially play a role similar to the IPCC in the climate area. A key to success there will be producing findings more quickly than the IPCC does, and happily Bengio appreciates this urgency" he says.

TOO MUCH FOCUS ON FRONTIER TECHNOLOGIES?

Despite the extreme risks posed by highly advanced AI, the summit has been criticised for focusing too much on frontier AI rather than the ways AI is already disrupting society. **Dr Lulu Shi, from Oxford's Department of Education**, says: "Concentrating on the long-term risks, such as risks of human extinctions, is dangerous as it leads the debate away from the very real and already existing risks that AI is causing, such as those caused by surveillance technologies which have been punishing people from already marginalised groups. At no point was social justice put at the center of the discussion during the summit."

According to **Professor Brent Mittelstadt from the Oxford Internet Institute**, frontier AI should not be used as an excuse to avoid regulating the well-established harms of today's AI systems. "The decision to focus the Safety Summit on frontier AI and long-term existential risks, cybersecurity, and terrorism, meant that an exceptional portfolio of research on AI ethics, regulation, and safety, was effectively being ignored. We know the risks that AI poses now, and we've developed ways to address them, so why is there such a reluctance to take any steps towards hard regulation?"

NOT PERFECT – BUT A START

Even with these criticisms, the Bletchley Summit was an important start in an ongoing process, argues **Ciaran Martin, Professor of Practice in the Management of Public Organisations, Oxford University**. "It is easy to criticise, but don't let the perfect be the enemy of the good. This was an important initiative and the British Government deserves credit for its global leadership" he says. "The alternative was not a better event – the alternative was nothing at all. Going forward, we will need to broaden the conversation and make sure it's not captured by the existing tech giants. But Bletchley was a good start."

As Professor Véliz observes, ultimately time will decide whether the first Safety Summit will be judged a success or a failure. "Thus far, the event has had a symbolic function. However, sometimes symbolism weighs enough to make a difference. If the summit eventually leads to an adequate and binding international agreement on AI ethics, then it will have been a success. But if all we are left with are a few nice photos, a toothless and vague declaration, and well wishes, then it will have been a failure indeed." ■

Author Dr Caroline Wood is a Communications Manager for Research and Innovation at Oxford University.