"Ethics and Philosophy and Technology Are Colliding Right Now": What To Do about Artificial Intelligence

Shashi Jayakumar

Artificial Intelligence has in recent months arrived at a point where the probabilistic prediction of a next token gives a seemingly fantastic illusion – one almost of sentience. This and other recent developments with generative AI and what are known as foundation models have given rise to concern and hand-wringing in equal measure, with some experts prepared to countenance the possibility that uncontrolled efforts in AI development might even pose an existential risk – fears that would have been impossible to take seriously a decade ago.

There is concern, but eye-rolling too. In the course of discussing the limitations of generative AI and ChatGPT, Meta's chief AI scientist Yann LeCun has observed that such models are not very intelligent – and not even as smart as a dog – because they are solely trained on language.

Other respected voices who play a leading role in AI development acknowledge that we are quickly coming to a pass that many thought would be decades away (or, might not eventuate at all). Demis Hassabis, founder of Google DeepMind, recently signed an open letter alongside Sam Altman, who leads OpenAI, and Dario Amodei, former vice-president of research at OpenAI and now CEO of Anthropic, and other major names in the AI industry. The brief but urgent letter simply states that mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks, such as pandemics and nuclear war.

Dangers

Progress on foundation models and large language models (LLMs) has clearly happened much faster than anyone has expected, and it is possible – even likely – that there may be large jumps in capability which take even the creators (not to say governments) by surprise. Those who favour regulation (over self-regulation) highlight several factors: first, that frontier AI may have unexpected, difficult-to-detect dangerous capabilities; second, that models deployed for broad use can be difficult to reliably control and prevent from being used to cause harm; and third, that there is the possibility of models proliferating rapidly, enabling circumvention of safeguards.

Given that generative AI can be used by those with ill intent, the risks to public safety need to be studied. Academics and researchers have already begun to point out that generative AI and foundation models could potentially be used to do many things – for example, to synthesise dangerous substances, to design novel chemical/biological weapons, or to turbocharge disinformation campaigns at scale.

International Norms

As competing tech companies roll out ever more powerful foundation models, the potential ramifications of this trend, by their sheer magnitude, had compelled key figures – even those at the forefront of AI technology in the private sector – to concede that some sort of control mechanism is needed. Jack Clark, the co-founder of Anthropic, speaking to the UN Security Council (when the UNSC met in July 2023 for its first discussion of AI), said that private companies should not be allowed to dominate the development of AI. UN Secretary-General António Guterres stated at the same meeting that the United Nations should create a new international body to help govern the use of AI. Guterres

also pledged to convene an advisory council that will develop proposals for regulating AI more broadly by the end of the year.

But here, battle lines are clearly being drawn. China's UN ambassador, Zhang Jun, while stating that Beijing supports a central coordinating role for the United Nations in establishing guiding principles for AI, also observed that international norms should still allow countries to establish national-level regulations; he also blasted unnamed "developed countries" for trying to achieve dominance in AI.

There is every likelihood that the coming AI norms/regulation debate will be similar to what has been going on for some time now for cyber norms. Lip service will be paid to the notion of having rules of the road, but major nations will have their own internal calculations on what these norms should look like, all the while taking steps to ensure that any agreed norms are either non-binding or do not impinge on national self-interest.

We are thus very far away from a scenario where countries will be able to verify each other's compliance with potential future international agreements on advanced machine learning development. This will take decades, if it happens at all. The same could be said of other ideas that have been floated:

(i) There have been calls, for example, for a pause in the development of foundation/frontier models. This is unrealistic. Senior US defence officials, for example, have poured cold water on the prospect of such a pause, pointing out that adversaries are likely to ignore any such understanding and use the time to pull ahead in the AI arms race. As the US Defense Department's chief digital and AI officer, Craig Martell, observed, "This is a very bizarre world. Ethics and philosophy and technology are colliding right now in a way that we feared that they would but they haven't before … We have the absolute pressure to maintain leadership in this domain. … Think about six months ago — if we stopped for six months, we're going to lose that leadership."

(ii) Some experts, concerned that we are not approaching the issue of the danger of frontier models with the level of seriousness needed, have suggested that we need an international coalition banning large AI training runs, with extreme measures to have this ban actually put into effect. Some like the AI researcher Eliezer Yudkowsky have gone as far as to suggest tracking the sales of graphics processing units (GPU) and monitoring all data centres. But there is no evidence at the level of international bodies (such as the United Nations) that there is the will or broad consensus to undertake these measures.

The fact that an overarching, binding, international agreement on AI development is probably decades away is not in itself cause for complete passivity, and, to their credit, certain nations and bodies have started to enact codes of conduct and guidelines, especially where they can find common ground. The United States and European Union are attempting, it appears, to push a joint AI voluntary code of conduct to provide safeguards, perhaps as a stopgap while new laws (such as the EU AI Act, which will have binding measures on high-risk AI systems) are developed. But as informed observers have pointed out, by the time the EU AI Act is rolled out – most likely in 2026 – the foundation models are likely to have advanced beyond recognition.

ASEAN

ASEAN leaders have realised that the regional grouping needs to find some common ground on these issues, agreeing in February on the need to develop a "Guide on AI Governance and Ethics" (which,

according to reports, may be ready by end 2023). Details are scant at present, but the guide may address balancing the economic benefits of the technology, with some suggestion as well that AI's potential to further the weaponisation of misinformation will be tackled.

If ASEAN discussions on cyberspace and the slow journey to regional cyber maturity are anything to go by, coming to a common understanding on the risks and opportunities that AI brings will take a considerable amount of time; the forthcoming ASEAN guide will by the same token almost certainly be a set of denominators that all ASEAN member states can for the time being agree on, in short, a low-hanging fruit.

In developing its guide, ASEAN will need to learn from the West, but also from China, which has made noteworthy and methodical attempts to regulate AI. Recent rules unveiled by the Cyberspace Administration of China (CAC) on generative AI (which were scheduled to have taken effect in August) were not as stringent as initially feared and seem overall to be reasonably well-thought-through attempts to balance the public interest, the needs of technology companies and the overarching security of the state.

Singapore is likely to be a key player (as it was for moving the discussion of cyber norms forward within ASEAN), but it should have its eye on the wider picture too: regional hesitancy should not hold it back. If, at the technical level, it becomes feasible to establish an international laboratory like the European Organization for Nuclear Research (CERN) that would have controlling oversight over research on artificial general intelligence or AGI (as some, like DeepMind's Demis Hassabis have suggested) and can act as a centre for the testing of AI in controlled sandboxes, Singapore agencies would presumably want to ensure in advance (and, necessarily, behind the scenes) that there is sufficient political capital to become part of such a consortium.

But there are other aspects to the geopolitics of AI that ASEAN nations will not want to be caught in the middle of. According to one report, there has been discussion among Internet entrepreneurs in China concerning how AI models could be exported to countries like Singapore for training on more-advanced hardware before being re-exported back to China. This might seem puzzling: China should on the face of it have little difficulty developing cutting-edge models domestically. But if (as seems possible) this is part of a longer-term plan to circumvent future moves by the United States to constrain China's progress in AI (through further reinforcing existing export curbs on AI chips to China), then a plausible scenario might see Singapore coming under pressure from the United States on this front.

Singapore's Approach

Some initiatives, including the Public Consultation for the Proposed Advisory Guidelines on Use of Personal Data in AI Recommendation and Decision Systems of Singapore's Personal Data Protection Commission (PDPC), might on the surface seem to suggest that the more developed strand of thinking among agencies has to do more with the business use cases in the private sector, rather than with the risks of AI.

That said, Singapore agencies are clearly making strides when it comes to understanding what has been happening in the wider global conversation and implementing best practices in the country. In May 2022, the Information Media Development Authority (IMDA) and the PDPC launched an AI governance testing framework and software toolkit, AI Verify, to enable businesses to check the implementation of AI models against a set of 11 internationally recognised principles, including transparency, safety and fairness – all principles that other international AI frameworks have

coalesced around. The AI Verify Foundation, launched in June 2023, builds on this, with a key aim being to develop an open-source community to contribute to AI testing frameworks.

A discussion paper on generative AI issued by IMDA and the Singapore-based AI company Aicadium in June 2023 (focused overall on guidance for policymakers and leaders in private sector organisations) tackles key issues of explainability, ethics and governance in the course of rolling out AI solutions. In what is a useful sign of growing maturity when it comes to addressing the downsides of AI, the discussion paper carries pertinent comments on safety and alignment research:

> … as AI models become more powerful, we need to ensure that human capacity to control AI systems keeps pace. Development in safety and alignment lags that of generative AI development. Policymakers need to invest strategically to accelerate safety and alignment research especially in more advanced techniques, to enable interpretability, controllability and robustness. This effort should also nurture centres of knowledge in Asia and other parts of the world, to complement the ongoing efforts in the US and EU.

Besides mentioning the need to monitor the development of very powerful AI, the paper gives a brief mention to the issue of AI powering disinformation. It is telling that a grant call issued in 2023 by AI.sg, the national programme overseeing the growth of Singapore's AI capabilities, specifically invites proposals on generative AI and disinformation (with a second theme being fairness).

Overall, these are positive developments – not long ago, those working in fields such as AI safety, fairness or the use of AI for nefarious purposes were considered players in a niche field.

The next few years will be crucial, as AI.sg and other bodies such as IMDA will most likely need to find ways to strike a balance between, on the one hand, facilitating the roll-out of AI solutions (including generative AI) that bring benefits to society and, on the other hand, addressing developments that could lead to negative societal disamenity. Should the proliferation of frontier AI models through open-sourcing be checked, as some have suggested? Or, should government agencies be able to enforce rules either on development or usage of such models, either by tracking or licensing? Should watermarking of products that are seemingly original but in fact made through generative AI be mandated?

Even while policymakers grapple with these questions, there are other moves that could be considered, especially if Singapore aims to be a leader in safe and responsible AI. If, as seems possible (or likely) we really are on the cusp of era-defining change brought on by AI, it is not just government that should gear itself, but the whole of society too. Introductory principles of ethics in technology should be introduced in schools, perhaps as an adjunct to ongoing efforts to inculcate cyber hygiene and social media literacy. But alongside this effort, agencies should also consider finding ways (perhaps as part of future national conversations or consultations) to introduce the future concept of the AI-inflected SMART Nation to the broader public. And members of the public – perhaps in increasing numbers – may in time want more transparency and responsible disclosure when it comes to wanting to know how, and when, they are being impacted by AI roll-outs.

Facing up to AI – both its promise and peril – will not simply be a whole of government effort; it will be a whole of society one, too.