HOW DO WE MANAGE THE RISKS FROM ADVANCED AI SYSTEMS WHILE STILL ALLOWING PROGRESS?

Dr Jan Brauner, from the University of Oxford's Department of Computer Science, outlines the steps policy makers should be taking now to ensure we can safely reap the benefits of AI in the future. This article is based on a paper Dr Brauner recently co-authored as part of an international consensus of global experts for the journal *Science*[1], and quotes from this.

Despite the constant stream of media headlines related to artificial intelligence (AI), it is difficult to grasp the sudden, explosive growth in the capabilities of these technologies. In 2019, GPT-2 could not even reliably count to ten. Now, only four years later, deep learning systems are able to create hyper-realistic scenes on demand, write software and code, generate advice, and combine language and image processing to steer robots.

And there is no sign of this progress slowing. Tech companies, backed by enormous cash reserves, are racing to create ever more powerful AI systems. Their stated aim is to create systems that exceed human abilities in most cognitive work and can automate most labour. Many leading experts find it possible that, within this decade or the next, generalist AI systems will broadly outperform humans in most important domains.

Advanced AI systems have the potential to help us tackle major challenges, such as in health or climate change. Advanced AI offers vast opportunities. But strong AI capabilities also imply large-scale societal risks, including rapid job displacement, amplified social injustice, automated misinformation, and large-scale cyber and biological threats. These risks demand urgent recognition and action, so that we are adequately prepared for the largest risks *before* we have to face them. **Climate change took decades to be acknowledged and confronted, but for AI, decades could be too long.**

AUTONOMOUS AI SYSTEMS AMPLIFY SOCIETY-SCALE RISKS

Harms such as misinformation and discrimination from algorithms are already evident today; other harms show signs of emerging. A particularly urgent issue is the need to proactively address the rapidly-evolving threats from autonomous AI: systems that can plan, act in the world, and pursue goals. While current AI systems have limited autonomy, companies are working to change this. For example, the non-autonomous GPT-4 model was quickly adapted to browse the web, design and execute chemistry experiments, and utilize software tools, including other AI models. **Once realised, autonomous AI will radically amplify the current risks with AI, besides creating new potential harms.**

With highly advanced autonomous AI, we risk creating systems that pursue undesirable goals. Worryingly, no one currently knows how to reliably align AI behaviour with complex values. Once autonomous AI systems pursue undesirable goals, embedded by malicious actors or by accident, we may be unable to rein them in. Even now, we struggle to detect and control relatively simple computer worms and viruses: advanced autonomous AI systems will have strong skills in critical domains such as hacking, social manipulation, deception, and strategic planning, and thus be much harder to control.

To advance undesirable goals, future autonomous AI systems could use undesirable strategies—whether learned from humans or developed independently—as a means to an end. AI systems could gain trust and resources to achieve their goals. They could manipulate or otherwise influence important decision-makers, and could find allies in humans or other AI systems. Future AI systems could insert and then exploit security vulnerabilities to control the computer systems behind our communication, media, banking, supply-chains, militaries, and governments. In a worst-case scenario of open conflict between AI systems and humanity, AI systems could threaten with or use autonomous or biological weapons. There is also the risk that humans voluntarily hand over control: companies and militaries may outsource more and more key functions to AI systems, for the sake of efficiency.

We could lose control over advanced AI systems, leading to rapid escalation of harms like widescale cybercrime and social manipulation. Continued unchecked AI advancement could ultimately result in catastrophic loss of human life, devastation of Earth's ecosystems, and the marginalisation or even extinction of humanity.

A PATH FORWARD TO SAFE AND ETHICAL AI

But all is not lost, if we act now. Alongside 22 world-leading AI scientists and governance experts from the US, China, EU, UK, and other countries, I was recently part of a global effort to develop a comprehensive response to manage the risks presented by advanced AI systems [1]. Together, our recommendations present a viable way forward to ensure progress in AI development is safe and ethical, and establish effective government oversight. The key recommendations of this framework are:

- **Industry labs should invest in safe, ethical AI and develop if-then plans for further scaling.** Leading labs should allocate at least one third of their AI research and development resources to ensure the safety and ethical use of AI systems. This level of investment in AI safety would be on par with the resources devoted to increasing AI capabilities. As a stopgap measure until binding regulations are complete, AI labs should also commit to rigorous and independently scrutinised scaling policies that set out the safety measures they will take if specific dangerous capabilities are found in their AI systems.
- **Governments should allocate at least one third of their AI research and development resources to ensure the safety and ethical use of frontier AI systems**. This includes oversight and honesty, robustness, interpretability and transparency, inclusive AI development, addressing emerging challenges, evaluations for dangerous capabilities, evaluations of alignment, risk assessment, and resilience.
- **Governments must establish oversight of the AI industry by governments and civil society**. This includes mandating that AI labs report frontier AI training runs, providing legal protections for whistleblowers at major AI labs, creating a registry of frontier AI systems that are in training or deployment, and requiring labs to report incidents where AIs displayed harmful behaviour or novel dangerous capabilities.
- **Require auditing of frontier AI systems during training and before deployment**. Labs should give regulators and independent auditing bodies the access needed to evaluate these systems in development for dangerous capabilities.
- **AI system developers and owners must be held legally liable for harms from their frontier AI systems that can be reasonably foreseen and prevented**. This includes harms resulting from deploying highly capable AI systems whose behaviour cannot be reliably predicted.

Despite our best efforts to test and evaluate advanced AI systems, we cannot simply assume they are safe until proven otherwise. Current testing methods are far from foolproof and can easily overlook issues. Moreover, it is unclear if governments can rapidly develop the extensive expertise required to thoroughly assess the full scope of an AI system's capabilities and potential societal risks. Therefore, the burden of proof should fall on the developers of frontier AI systems to convincingly demonstrate, through structured arguments grounded in evidence, that their systems will remain within acceptable risk boundaries. By making such **"safety cases"**, AI companies would be following best practices for safety-critical industries such as aviation, medical devices, and military software.

Governments should build capacity, standards, and regulatory authorities to address the risks posed by future AI systems with exceptionally dangerous capabilities, such as the ability to circumvent human control. Amongst others, governments should be prepared to:

- **Establish a licensing system for training AI systems that are unusually resource-intensive and risky**.
- **Empower regulators to pause the further development of an AI system**, if it demonstrates sufficiently dangerous capabilities during training.

- **Mandate access controls for frontier AI systems and their training code**.
- **Require cyber security measures for actors that will hold access to dangerous frontier AI systems,** to prevent model proliferation. Given the utility of advanced AI for economic gain and for malicious use, AI labs will need security measures of the highest standard.

While there have been some promising initial efforts in these directions, society's current response falls far short of what is needed given the potential for transformative and rapid AI progress that many experts anticipate. As AI capabilities continue to grow, so too do the risks. Huge investments are flowing into making AI more powerful, but far less into making AI safe and mitigating its negative impacts. Realising the benefits of AI for humanity will require reorienting our priorities. This will only be realised if there is a concerted effort by both tech companies and government to ensure that these technologies are developed ethically and safely. **The time to act is now.**

Reference

[1] Bengio, Yoshua, et al. "Managing extreme AI risks amid rapid progress." Science (2024): eadn0117. https://www.science.org/doi/10.1126/science.adn0117

Jan Brauner is a final year PhD student in Computer Science at the University of Oxford. He has worked on a range of topics, from AI safety to applications of AI systems in public health. He has published over 25 peer-reviewed publications, including in high-profile outlets such as Science, Nature Communications, PNAS, ICML, ICLR, and NeurIPS. His work has been cited in federal bills, presented at institutions like the Africa CDC, the OECD Global Science Forum, and the UK Cabinet Office, and featured in media outlets such as Forbes, Guardian, Vox, and TIME. ■