nature

AI safety regulation need not stifle innovation

Early, robust and proportionate regulation is crucial for the technology's success. Done right, it could even inspire progress.

How to maximize the benefits of artificial intelligence (AI) while minimizing its harms is a worldwide preoccupation. This week, the spotlight is on the United Kingdom, where Prime Minister Rishi Sunak is hosting an international AI Safety Summit, welcoming influential figures from industry, policy and research. The meeting is being held north of London at Bletchley Park, the home of one of the world's first programmable computers.

In a speech at the Royal Society in London last week, Sunak said: "Right now, the only people testing the safety of AI are the very organisations developing it." He might have added that researchers outside those corporations who want to study AI safety struggle to do so because of a lack of access to the companies' data.

Nonetheless, the idea that companies should not be, as Sunak put it, "marking their own homework", is right for a technology that poses a known risk to jobs and uses algorithms that often reinforce bias and discrimination. Governments want to attract flagship companies — Elon Musk, the owner of X (formerly Twitter), is reported to be attending the summit. But when it comes to AI safety, independent scrutiny is required, and that means regulation — a word that governments are reluctant to use.

US President Joe Biden's executive order on AI safety, published this week, doesn't mention it (see go.nature.com/46sbeno). Importantly, it does state that developers of "the most powerful AI systems" must notify the government and share safety data. Moreover, safety standards will be set by the US National Institute of Standards and Technology.

Sunak has said that the United Kingdom is not in a rush to regulate. "This is a point of principle — we believe in innovation," he said. But innovation and regulation need not be in opposition. The world needs both.

Fortunately, there is a wealth of literature on regulation — from banking to medicines to food to road safety — on which governments can draw. For instance, researchers at Google DeepMind in London have laid out some lessons that civil nuclear technology could have for AI, using the example of the International Atomic Energy Agency (H. Law & L. Ho *Nature Rev. Phys.* https://doi.org/k3h2; 2023).

Every technology is different, but some fundamental principles have emerged over decades of regulatory experience. One is the necessity for transparency, and for regulators to have access to complete data to allow them to make the right decisions. Another is the need for legally binding standards for monitoring, compliance and liability.

> "Everyone seems to be drawing on their favourite historical analogies."

The 2008 global financial crisis shows what can happen when regulators don't look closely at relevant data. Regulators didn't know, or were unable to detect, that banks and insurance companies had invested hundreds of billions of dollars in loans in opaque 'black box' financial products that were, ultimately, dependent on risky credit — until it was too late. Relatively few people understood how these products had been created or what their systemic risks were, as Andrew Haldane and Robert May described in the aftermath (A. G. Haldane & R. M. May *Nature* **469**, 351–355; 2011).

Other mainstays of regulation include registration, regular monitoring, reporting of incidents that could cause harm, and continuing education, for both users and regulators. Road safety offers lessons here. The car has transformed the lives of billions, but also causes harm. To mitigate risks, vehicle manufacturers need to comply with product safety standards; vehicles must be tested regularly; and there is compulsory driver training and licensing, along with an insurance-based legal framework to assess and apportion liability in the case of accidents. Regulation can even spur innovation. The introduction of emissions standards inspired the development of cleaner vehicles.

Sunak last week announced £100 million (US$121.5 million) in funding for AI safety research, as well as plans to set up an AI-safety research institute. He is making another £100 million available to develop AI for use in health care. The details released so far are sketchy, but these announcements are welcome. There are many strands of opinion on the balance of AI's benefits and risks, and all parties must work together towards a consensus. This needs to be led by researchers and mediated by international organizations, allowing evidence to lead decision-making and giving all countries an equal voice and an equal stake in the outcome.

Crucially, the safety of AI cannot be a matter for those working in computational disciplines to shoulder alone. Researchers who study ethics, equality and diversity in science, public engagement and technology policy all need to have a seat at the table. Social scientists from these areas should have been front and centre at the summit. Many are instead gathering — along with computer scientists — for a separate series of events called AI Fringe in London.

At this early stage of AI's development, everyone seems to be drawing on their favourite historical analogies. Sunak talked of the need for a body in the style of the Intergovernmental Panel on Climate Change, which was set up in 1988. Others have mentioned the Asilomar conference, at which researchers gathered in California in 1975 to warn of the necessity of biosafety controls for DNA modification. Although useful, what these examples can offer the AI challenge is limited. For one thing, corporations and their research staff were not the dominant actors in climate science and DNA modification in the way they are in AI.

Governments and corporations should not fear regulation. It enables technologies to develop and protects people from harm. And it need not come at the cost of innovation. In fact, setting firm boundaries could spur safer innovation within them.