When to Regulate AI*

Simon Chesterman

The better part of a century ago, science fiction author Isaac Asimov imagined a future in which robots have become an integral part of daily life. In this speculative world, a safety device was built into these morally neutral robots in the form of three laws of robotics. The first is that a robot may not injure a human, or through inaction allow a human to come to harm. Secondly, orders given by humans must be obeyed, unless doing so would conflict with the first law. Thirdly, robots must protect their own existence, unless doing so conflicts with the first or second laws.

The three laws are a staple of the literature on regulating new technology although, like the Turing Test, they are more of a cultural touchstone than serious scientific proposal. Among other things, the laws presume the need only to address physically embodied robots with human-level intelligence — an example of the android fallacy. They have also been criticised for putting obligations on the technology itself, rather than on the people creating it.

As the European Union prepares to enact the first major legislation governing AI — even as the rush to deploy generative AI appears to have lowered the threshold for tech companies to release potentially dangerous products to market — it is timely to address the question of when regulation of AI is appropriate.

To Regulate, or Not to Regulate?

Governments generally regulate activities either to address market failures or support social or other policies.

In the context of AI systems, market justifications for regulation include addressing information inadequacies between producers and consumers of technology, as well as protecting third parties from externalities — harms that may arise from deploying AI. In the case of autonomous vehicles, for example, we are already seeing a shift of liability from driver to manufacturer, with a likely obligation to maintain adequate levels of insurance.

Regulation is not simply intended to facilitate markets, however. It can also defend rights or promote social policies, in some cases imposing additional costs. Such justifications reflect the moral arguments for limiting AI. In the case of bias, for example, discrimination on the basis of race or gender is prohibited even if it is "efficient" on some other measure.

Similarly, the prohibition on AI systems making kill decisions in armed conflict is not easily defended on the utilitarian basis that this will lead to better outcomes; these systems may eventually be more compliant with the law of armed conflict than humans. The prohibition stems, instead, from a determination that morality requires that a human being take responsibility for such choices.

Different considerations may restrict the outsourcing of certain functions to AI — notably certain public decisions, the legitimacy of which depends on the process by which they are made as much as

the efficiency of the outcome. Even if an AI system were believed to make superior determinations than politicians and judges, inherently governmental functions that affect the rights and obligations of individuals should nonetheless be undertaken by office-holders who can be held accountable through political or constitutional mechanisms.

A further reason for regulating AI is more procedural in nature. Transparency, for example, is a necessary precursor to effective regulation. Although not a panacea and bringing additional costs, requirements for minimum levels of transparency and the ability to explain decisions can make oversight and accountability possible.

Against all this, governments may also have good reasons *not* to regulate a particular sector if it would constrain innovation, impose unnecessary burdens or otherwise distort the market.

Different political communities will weigh these considerations differently, although it is interesting that regulation of AI appears to track the adoption of data protection laws in many jurisdictions.

The United States, for example, has largely followed a market-based approach, with relatively light-touch sectoral regulation and experimentation across its 50 states. That is true also of data protection, where a general federal law is lacking but particular interests and sectors, such as children's privacy or financial institutions, are governed by statute. In the case of AI, towards the end of the Obama administration in 2016, the US National Science and Technology Council argued against broad regulation of AI research or practice. Where regulatory responses threatened to increase the cost of compliance or slow innovation, the council called for softening them, if that could be done without adversely impacting safety or market fairness.

That document was finalised six months after the European Union enacted the General Data Protection Regulation (GDPR), with sweeping new powers covering both data protection and automated processing of that data. The EU approach has long been characterised by a privileging of human rights, with privacy enshrined as a right after the Second World War, laying the foundation for the 1995 Data Protection Directive and later the GDPR. Human rights is also a dominant theme in EU considerations of AI, although there are occasional murmurings that this makes the continent less competitive.

China offers a different model again, embracing a strong role for the state and less concern about the market or human rights. As with data protection, a driving motivation has been sovereignty. In the context of data protection, this is expressed through calls for data localisation — ensuring that personal data is accessible by Chinese state authorities. As for AI, Beijing identified it as an important developmental goal in 2006 and a national priority in 2016. The State Council's New Generation AI Development Plan, released the following year, nodded at the role of markets but set a target of 2025 for China to achieve major breakthroughs in AI research with "world-leading" applications — the same year forecast for "the *initial* establishment of AI laws and regulations".

Many were cynical about China's lack of regulation — its relaxed approach to personal data has often been credited as giving the AI sector a tremendous advantage. Yet laws adopted in 2021 and 2022 incorporated norms closely tracking principles also embraced in the European Union and

international organisations. More generally, such projections about future regulation show that, for emerging technologies, the true underlying question is not *whether* to regulate, but *when*.

The Collingridge Dilemma

Writing in 1980 at Aston University in Birmingham, England, David Collingridge observed that any effort to control new technology faces a double bind. During the early stages, when control would be possible, not enough is known about the technology's harmful social consequences to warrant slowing its development. By the time those consequences are apparent, however, control has become costly and slow.

The climate emergency offers an example of what is now termed the Collingridge Dilemma. Before automobiles entered into widespread usage, a 1906 Royal Commission studied the potential risks of the new machines plying Britain's roads; chief among these was thought to be the dust that the vehicles threw up behind them. Today, transportation produces about a quarter of all energy-related CO_2 emissions and its continued growth could outweigh all other mitigation measures. Although the COVID-19 pandemic had a discernible effect on emissions in 2020 and 2021, regulatory efforts to reduce those emissions face economic and political hurdles.

Many efforts to address technological innovation focus on the first horn of the dilemma — predicting and averting harms. Research institutes have been established to evaluate the risks of AI, with some warning apocalyptically about the threat of general AI.

If general AI truly poses an existential threat to humanity, that threat could justify a ban on research, comparable to restrictions on biological and chemical weapons. No major jurisdiction has yet supported a ban, however, either because the threat does not seem immediate or due to concerns that it would merely drive that research elsewhere. When the United States imposed limits on stem cell research in 2001, for example, one of the main consequences was that US researchers in the field fell behind their international counterparts. A different challenge is that if regulation targets near-term threats, the pace of technological innovation can result in regulators playing an endless game of catch-up. Technology can change exponentially, while social, economic and legal systems tend to change incrementally.

Collingridge himself argued that instead of trying to anticipate the risks, more promise lies in laying the groundwork to address the second aspect of the dilemma: ensuring that decisions about technology are flexible or reversible. This is also not easy, presenting what some wags describe as the "barn door" problem of attempting to shut it after the horse has bolted.

New Rules?

If Asimov's three laws had avoided or resolved all the ethical dilemmas of machine intelligence, his literary career would have been brief. In fact, the very story in which they were introduced focuses

on a robot that is paralysed by a contradiction between the second and third laws, resolved only by a human putting himself in harm's way to invoke the first.

A blanket rule not to harm humans is obviously inadequate when forced to choose between the lesser of two evils. Asimov himself later added a "zeroth" law, which provided that a robot's highest duty was to humanity as a whole. In one of his last novels, a robot is asked how it could ever determine what was injurious to humanity as a whole. "Precisely, sir", the robot replies. "In theory, the Zeroth Law was the answer to our problems. In practice, we could never decide."

The demand for new rules to deal with AI is often overstated. Yet some new rules will be required, at least in the areas of human control and transparency. Human control requires limits on the kinds of AI systems that can be developed, guarding against the deployment of AI systems that are uncontainable or uncontrollable.

On the question of transparency, accountability of government officials requires a limit on the use of opaque processes. Above and beyond that, measures such as impact assessments, audits and an AI ombudsperson could mitigate some harms and assist in ensuring that other harmful conduct can be attributed back to legal persons capable of being held to account.

As AI becomes more sophisticated and pervasive — and as harms associated with AI systems become more common — demands for more than ethical restrictions on AI will increase. The precise nature of those laws will vary from jurisdiction to jurisdiction. The only safe bet is that there are likely to be more than three.

* *This article draws on material considered at greater length in* We, the Robots? Regulating Artificial Intelligence and the Limits of the Law *(Cambridge University Press, 2021) and in a chapter for* The Handbook of the Ethics of AI, *edited by David J. Gunkel (Edward Elgar, forthcoming).*