NUS Law Working Paper No 2023/014

From Ethics to Law: Why, When, and How to Regulate AI

Simon Chesterman

chesterman@nus.edu.sg

[May 2023]

Note: The views expressed in each paper are those of the author or authors of the paper. They do not necessarily represent or reflect the views of the National University of Singapore.

From Ethics to Law: Why, When, and How to Regulate AI

Simon Chesterman[1] (0000-0002-3599-4573)

[1] National University of Singapore. This chapter draws on material considered at greater length in *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law* (Cambridge University Press, 2021).

Abstract (150 words)
The past decade has seen a proliferation of guides, frameworks, and principles put forward by states, industry, inter- and non-governmental organizations to address matters of AI ethics. These diverse efforts have led to a broad consensus on what norms might govern AI. Far less energy has gone into determining how these might be implemented — or if they are even necessary. This chapter focuses on the intersection of ethics and law, in particular discussing why regulation is necessary, when regulatory changes should be made, and how it might work in practice. Two specific areas for law reform address the weaponization and victimization of AI. Regulations aimed at general AI are particularly difficult in that they confront many ‘unknown unknowns’, but the threat of uncontrollable or uncontainable AI became more widely discussed with the spread of large language models such as ChatGPT in 2023. Additionally, however, there will be a need to prohibit some conduct in which increasingly lifelike machines are the victims — comparable, perhaps, to animal cruelty laws.

Keywords (6 keywords): artificial intelligence; ethics; law; regulation; markets; compliance

The better part of a century ago, science fiction author Isaac Asimov (1942) imagined a future in which robots have become an integral part of daily life. At the time, he later recalled (1982, pp. 9-10), most robot stories fell into one of two genres. The first was robots-as-menace: technological innovations that rise up against their creators in the tradition of *Frankenstein*, but with echoes at least as far back as the Greek myth of Prometheus, the subtitle of Mary Shelley’s 1818 novel. Less commonly, a second group of tales considered robots-as-pathos — lovable creations that are treated as slaves by their cruel human masters; morality tales about the danger posed not by humanity's creations, but by humanity itself.

Asimov's contribution was to create a third category: robots as industrial products built by engineers. In this speculative world, a safety device is 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 that would conflict with the first law. Thirdly, robots must protect their own existence, unless that conflicts with the first or second laws.

The three laws are a staple of the literature on regulating new technology though, like the Turing Test, they are more of a cultural touchstone than serious scientific proposal (Anderson, 2008).[1] Among other things, the laws presume the need only to address physically embodied robots with human-level intelligence — an example of the android fallacy.[2] They have also been criticized for putting obligations on the technology itself, rather than the people creating it (Balkin, 2017). Here it is worth noting that Asimov's laws were not 'law' in the sense of a command to be enforced by the state. They were, rather, encoded into the positronic brains of his fictional creations: constraining what robots *could* do, rather than specifying what they *should*.

More importantly, for present purposes, the idea that relevant ethical principles can be reduced to a few dozen words, or that those words might be encoded in a manner interpretable by an AI system, misconceives the nature of ethics and of law. Nonetheless it was reported in 2007 that Korea had considered using them as the basis for a proposed Robot Ethics Charter. This was one of many attempts to codify norms governing robots or AI since the turn of the century, accelerating in the wake of the First International Symposium on Roboethics in Sanremo, Italy, in 2004. The European Robotics Research Network produced its 'Roboethics Roadmap' in 2006, while the first multidisciplinary set of principles for robotics was adopted at a 'Robotics Retreat' held by two British Research Councils in 2010.

The years since 2016 in particular saw a proliferation of guides, frameworks, and principles focused on AI. Some were the product of conferences or industry associations, notably the Partnership on AI's Tenets (2016), the Future of Life Institute's Asilomar AI Principles (2017), the Beijing Academy of Artificial Intelligence's Beijing AI Principles (2019), and the Institute of Electrical and Electronics Engineers (IEEE)'s Ethically Aligned Design (2019). Others were drafted by individual companies, including Microsoft's Responsible AI Principles, IBM's Principles for Trust and Transparency, and Google's AI Principles — all published in the first half of 2018.

Governments have been slow to pass laws governing AI. Several have developed softer norms, however, including Singapore's Model AI Governance Framework (2019), Australia's AI Ethics Principles (2019), China's AI Governance Principles (2019), and New Zealand's Algorithm Charter (2020). At the intergovernmental level, the G7 adopted the Charlevoix Common Vision for the Future of Artificial Intelligence (2018), the OECD issued its Recommendation of the Council on Artificial Intelligence (2019), and the European Union published Ethics Guidelines for Trustworthy AI (2019), precursor to the draft AI Act circulated in 2021. Various parts of the UN system have adopted documents, most prominently UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021). Even the Pope endorsed a set of principles in the Rome Call for AI Ethics (2020).

What is striking about these documents is the overlapping consensus that has emerged as to the norms that should govern AI (Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019). Though the language and the emphasis may differ, virtually all those written since 2018 include variations on the following six themes:

1. *Human control* — AI should augment rather than reduce human potential, and remain under human control.

2. *Transparency* — AI systems should be capable of being understood and their decisions capable of being explained.

3. *Safety* — AI systems should perform as intended and be resistant to hacking.

4. *Accountability* — Though often left undefined, calls for accountable or responsible AI assume or imply that remedies should be available when harm results.

5. *Non-discrimination* — AI systems should be inclusive and 'fair', avoiding impermissible bias.

6. *Privacy* – Given the extent to which AI relies on access to data, including personal data, privacy or personal data protection is often highlighted as a specific right to be safeguarded.

Additional concepts include the need for professional responsibility on the part of those developing and deploying AI systems, and for AI to promote human values or to be 'beneficent' (Luciano et al., 2018, pp. 696-697). At this level of generality, these amount to calls for upholding ethics generally or the human control principle in particular. Some documents call for AI to be developed sustainably and for its benefits to be distributed equitably, though these more properly address how AI is deployed rather than what it should or should not be able to do.

None of the six principles listed above seems controversial. Yet, for all the time and effort that has gone into convening workshops and retreats to draft the various documents, comparatively little has been applied to what they mean in practice or how they might be implemented. This is sometimes explicitly acknowledged and addressed, with the justification that a document is intended to be applicable to technologies as yet unknown and to address problems not yet foreseen.

A different question yields a more revealing answer, which is whether any of these principles are, in fact, necessary. Calls for accountability, non-discrimination, and privacy essentially amount to demands that those making or using AI systems comply with laws already in place in most jurisdictions. Safety requirements recall issues of product liability, with the additional aspect of taking reasonable cybersecurity precautions. Transparency is not an ethical principle as such but a condition precedent to understanding and evaluating conduct (Turilli & Floridi, 2009). Together with human control, however, it could be a potential restriction on the development of AI systems above and beyond existing laws.

Rather than add to the proliferation of principles, this chapter shifts focus away from the question of *what* new rules are required for regulating AI. Instead, the three questions that it will attempt to answer are *why* regulation is necessary, *when* changes to regulatory structures (including rules) should be adopted, and *how* they might be implemented.

To Regulate, or Not to Regulate?

In theory, governments regulate activities to address market failures, or in support of social or other policies. In practice, relationships with industry and political interests may cause politicians to act — or refrain from acting — in less principled ways (Baldwin et al., 2011, pp. 15-24). Though the troubled relationship between Big Tech and government is well documented (Alfonsi, 2019), this section will assume good faith on the part of regulators and outline considerations relevant to the choices to be made.

In the context of AI systems, market justifications for regulation include addressing information inadequacies as 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. This provides a model for civil liability for harm caused by some other AI systems — notably transportation more generally (including drones) and medical devices — under product liability laws (Mondello, 2022).

Regulation is not simply intended to facilitate markets, however. It can also defend rights or promote social policies, in some cases imposing additional costs (Prosser, 2006). 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 on some other measure 'efficient'. 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 (Chesterman, 2020).

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 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. Though 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 (Auld et al., 2022; Ugur, 2013). Different political communities will weigh these considerations differently, though 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, toward 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 (*Preparing for the Future of AI*, 2016, p. 17).

That document was finalized 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 characterized 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 (*EU White Paper on AI*, 2020, p.

10), though there are occasional murmurings that this makes the continent less competitive (Justo-Hanani, 2022; Pehrsson, 2016).

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 localization — ensuring that personal data is accessible by Chinese state authorities (Chander & Lê, 2015; Liu, 2020; Selby, 2017). 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' (*国务院关于印发新一代人工智能发展规划的通知 [State Council Issued Notice of the New Generation Artificial Intelligence Development Plan]*, 2017).

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 (Roberts et al., 2021). Yet laws adopted in 2021 and 2022 incorporated norms closely tracking principles also embraced in the European Union and international organizations (Hine & Floridi, 2022; Yang & Yao, 2022). 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 (1980, p. 19) 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 (*Royal Commission on Motor Cars*, 1906). Today, transportation produces about a quarter of all energy-related CO_2 emissions and its continued growth could outweigh all other mitigation measures. Though 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 (Liu et al., 2019).

Many efforts to address technological innovation focus on the first horn of the dilemma — predicting and averting harms. That has been the approach of most of the principles discussed at the start of this chapter. In addition to conferences and workshops, 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, it could justify a ban on research, comparable to restrictions on biological and chemical weapons. No major jurisdiction has imposed a ban, however, either because the threat does not seem immediate or due to concerns that it would merely drive that research elsewhere. (The 2023 open letter calling for a 'pause' in the development of large language models will be considered in the section 'Drawing Red Lines', below.) 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 (Murugan, 2009). 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 (Downes, 2009, p. 2). For these reasons, the principles discussed at the start of this chapter aim to be future-proof and technology-neutral. This has the advantage of being broad enough to adapt to changing circumstances, albeit at the risk of being so vague as to not offer meaningful guidance in specific cases.

Collingridge himself argued (pp. 23-43) 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.

This section considers two approaches to the timing of regulation that may offer some promise in addressing or mitigating the Collingridge Dilemma: the precautionary principle and masterly inactivity.

An Ounce of Prevention

A natural response to uncertainty is caution. The precautionary principle holds that if the consequences of an activity could be serious but are subject to scientific uncertainties, then precautionary measures should be taken or the activity should not be carried out at all (Aven, 2011). The principle features in many domestic laws concerning the environment and has played a key role in most international instruments on the topic. The 1992 Rio Declaration, for example, states that '[w]here there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation' (Rio Declaration, 1992). In some implementations, the principle amounts to a reversal of the burden of proof: those who claim an activity is safe must prove it to be so (Le Moli et al., 2017).

Critics argue that the principle is vague, incoherent, or both. A weak interpretation amounts to a truism, as few would argue that scientific certainty is required for precautions to be taken; a strong interpretation is self-defeating, since precautionary measures can themselves have harmful effects (Boyer-Kassem, 2017). In a book length treatment denouncing it as ‘European’, Cass Sunstein (2005, pp. 109-115) outlines the predictably irrational ways in which fears play out in deliberative democracies, notably the over-valuation of loss and the reactive nature of public opinion with regard to risk. That said, the notion that there are at least *some* risks against which precautionary steps should be taken before they materialize or can be quantified is widely accepted.

In the context of AI, the precautionary principle is routinely invoked with regard to autonomous vehicles (Smith, 2016, p. 572), lethal autonomous weapons (Bhuta & Pantazopoulos, 2016, pp. 290-294), the use of algorithms processing personal data in judicial systems (*European Ethical Charter on the Use of AI*, 2018, p. 56), and the possibility of general AI turning on its human creators (Maas, 2018). Only the last is a proper application of the principle, however, in that there is genuine uncertainty about the nature and the probability of the risk. The precise failure rate of autonomous vehicles may be unknown, for example, but the harm itself is well understood and capable of being balanced as against the existing threat posed by human drivers. As for lethal autonomous weapons, opponents explicitly reject a cost-benefit analysis in favour of a bright moral line with regard to decisions concerning human life; though there are ongoing debates about the appropriate degree of human control, the ‘risk’ itself is not in question. Similarly, wariness of outsourcing public sector decisions to machines is not founded — or, at least, not *only* founded — on uncertainty as to the consequences that might follow. Rather, it is tied to the view that such decisions should be made by humans within a system of political accountability.

Nevertheless, as indicated earlier, it is telling that, despite the risks of general AI, there has thus far been no concerted effort to restrict pure or applied research in the area. More promising are calls that implicitly focus on the second horn of Collingridge’s dilemma: requirements to incorporate measures such as a kill switch, or attempts to align the values of any future superintelligence with our own. These can be seen as applications of the principle that human control should be prioritized. If a path to general AI becomes clearer, they should become mandatory.

Masterly Inactivity

Another response to uncertainty is to do nothing. Refraining from action may be appropriate to avoid distorting the market through pre-emptive rulemaking or delaying its evolution through lengthy adjudication. The term sometimes used to describe this is ‘masterly inactivity’. With origins in nineteenth century British policy on Afghanistan, it suggests a watchful restraint in the face of undesirable alternatives (Adye, 1878; Roy, 2015, p. 69). (Britain’s involvement in Afghanistan, it should be noted, ended in humiliating defeat.)

In the context of AI, for many governments this amounts to a ‘wait and see’ approach. Yet there is a difference between passively allowing events to play out and actively monitoring and engaging with an emerging market and its actors. Government engagement in the processes that led to the principles described at the start of this chapter is an example, as is the encouragement of industry associations to develop standards and research into governance possibilities (Auld et al., 2022).

Inactivity may also amount to a buck-passing exercise. Even if governments choose not to regulate, decisions with legal consequences will be made — most prominently by judges within the common law tradition, who exercise a law-making function. Such decisions are already influencing norms in areas from contracts between computer programs and the use of algorithms in sentencing to the ownership of intellectual property created by AI. This can be problematic if the law is nudged in an unhelpful direction because of the vagaries of how specific cases make it to court. It is also limited to applying legal principles after the event — ‘when something untoward has already happened’, as the British House of Commons Science and Technology Committee warned (*Robotics and Artificial Intelligence, Fifth Report of Session 2016–17*, 2016).

Masterly inactivity, then, is not a strategy. Properly used, however, it may buy time to develop one.

Regulatory Approaches

Regulation is a contested concept and embraces more than mere ‘rules’. A leading text (Baldwin et al., 2011, p. 3) distinguishes three distinct modalities of regulation that are useful in considering the options available. First, regulation can mean a specific set of commands — binding obligations applied by a body devoted to this purpose. Secondly, it can refer to state influence more broadly, including financial and other incentives. Broader still, regulation is sometimes used to denote all forms of social or economic suasion, including market forces. The theory of ‘smart regulation’ has shown that regulatory functions can be carried out not only by institutions of the state but also professional associations, standard-setting bodies, and advocacy groups. In most circumstances, multiple instruments and a range of regulatory actors will produce better outcomes than a narrow focus on a single regulator (Guihot et al., 2017; Gunningham & Grabosky, 1998). These modalities of regulation can interact. An industry may invest in self-regulation, for example, due to concerns that failure to do so will lead to more coercive regulation at the hands of the state.

Regulation is not limited to restricting or prohibiting undesirable conduct; it may also enable or facilitate positive activities — ‘green light’ as opposed to ‘red light’ regulation (Harlow & Rawlings, 2009, pp. 1-48). ‘Responsive regulation’ argues in favour of a more cooperative relationship, encouraging regulated parties to comply with the goals of the law rather than merely strict rule compliance (Ayres & Braithwaite, 1992). Other approaches emphasize efficiency: risk-based and problem-centred regulatory techniques seek to prioritize the most important issues — though identification, selection, and prioritization of future risks and current problems involve uncertainty as well as normative and political choices (Baldwin & Black, 2016).

The tools available to regulatory bodies may be thought of in three categories also: traditional rulemaking, adjudication by courts or tribunals, and informal guidance — the latter comprising standards, interpretive guides, and public and private communications concerning the regulated activity. Tim Wu (2011) once provocatively suggested that regulators of industries undergoing rapid change consider linking the third with the first two by issuing 'threats' — informally requesting compliance, but under the shadow of possible formalization and enforcement.

Many discussions of AI regulation recount the options available — a sliding scale, a pyramid, a toolbox, and so on — but the application is either too general or too specific. It is, self-evidently, inappropriate to apply one regulatory approach to all of the activities impacted by AI. Yet, it is also impractical to adopt specific laws for every one of those activities. A degree of clarity may, however, be achieved by distinguishing between three classes of problems associated with AI: managing some risks, proscribing others, while in a third set of cases ensuring that proper processes are followed.

Managing Risks

Civil liability provides a basis for allocating responsibility for risk — particularly in areas that can be examined on a cost-benefit basis. This will cover the majority, perhaps the vast majority, of AI activities in the private sector: from transportation to medical devices, from smart home application to cognitive enhancements and implants. The issue here is not new rules but how to apply or adapt existing rules to technology that operates at speed, autonomously, and with varying degrees of opacity. Minimum transparency requirements may be needed to ensure that AI systems are identified as such and that harmful conduct can be attributed to the appropriate owner, operator, or manufacturer. Mandatory insurance will spread those risks more efficiently. But the fundamental principles remain sound.[3]

For situations in which cost-benefit analysis is appropriate but the potential risks are difficult to determine, regulatory 'sandboxes' allow new technologies to be tested in controlled environments. Though some jurisdictions have applied this to embodied technology, such as designated areas for autonomous vehicles, the approach is particularly suited to AI systems that operate online. Originating in computer science, a virtual sandbox lets software run in a manner that limits the potential damage if there are errors or vulnerabilities. Though not amounting to the immunity that Ryan Calo once argued (2011) was essential to research into robotics, sandboxes offer 'safe spaces' to trial innovative products without immediately incurring all the normal regulatory consequences. The technique has been most commonly used with respect to finance technology (or 'fintech'), enabling entrepreneurs to test their products with real customers, fewer regulatory constraints, reduced risk of enforcement action, and ongoing guidance from regulators (Fenwick et al., 2017, pp. 591-593; Zetzsche et al., 2017, p. 45). Pioneered by Britain in 2016, it is credited with giving London a first-mover advantage in fintech and has since been copied in other jurisdictions around the world (Allen, 2019, p. 580).

Drawing Red Lines

In some cases, however, lines will need to be drawn as to what is permissible and what is not. These red lines will, in some cases, go beyond merely applying existing rules to AI. Linked with the ethical principle of maintaining human control, an obvious candidate is prohibiting AI from making decisions to use lethal force.

Yet even that apparently clear prohibition becomes blurred under closer analysis. If machines are able to make every choice up to that point — scanning and navigating an environment, identifying and selecting a target, proposing an angle and mode of attack — the final decision may be an artificial one. Automation bias makes the default choice significantly more likely to be accepted in such circumstances. That is not an argument against the prohibition, but in favour of ensuring not only that a human is at least 'in' or 'over' the loop but also that he or she knows that accountability for decisions taken will follow him or her. This is the link between the principles of human control and accountability — not that humans will remain in control and machines will be kept accountable, but that humans (and other legal persons) will continue to be accountable for their conduct, even if perpetrated by or through a machine.

The draft AI Act of the European Union also seeks to prohibit certain applications of AI — notably real-time biometric surveillance, technologies that manipulate or exploit individuals, and social scoring (*AI Act (EU)*, 2021). The last item appeared to be at least partly a critique of China's social credit system, which has been criticized as an Orwellian scheme of surveillance and harbinger of a dystopian future (Mac Síthigh & Siems, 2019).

A discrete area in which new rules will be needed concerns human interaction with AI systems. The lacuna here, however, is not laws to protect us from them but to protect them from us. Anodyne examples include those adopted in Singapore in early 2017, making it an offence to interfere with autonomous vehicle trials. These are more properly considered as an extension of the management of risk associated with such technologies. More problematic will be laws preserving human morality from offences perpetrated against machines. At present, for example, it is a crime to torture a chimpanzee but not a computer.

As ‘social robots’ become more prevalent — in industries from eldercare to prostitution — it may be necessary to regulate what can be created and how those creations may or may not be used and abused.

In 2014, for example, Ronald Arkin ignited controversy by proposing that child sex robots be used to ‘treat’ paedophiles in the same way that methadone is used by heroin addicts (Hill, 2014). Though simulated pornography is treated differently across jurisdictions,[4] many have now prohibited the manufacture and use of these devices through creative interpretations of existing laws or passing new ones such as the CREEPER Act in the United States (Danaher, 2019).

As lifelike embodied robots become more common, and as they play more active roles in society, it will be necessary to protect them not merely to reduce the risk of malfunction but because the act of harming them will be regarded as a wrong in itself. The closest analogy will, initially, be animal cruelty laws. This is, arguably, another manifestation of the android fallacy — purchasing a lifelike robot and setting it on fire will cause more distress than deleting its operating system. Moving forward, however, the ability of AI systems to perceive pain and comprehend the prospect of non-existence may change that calculation (Anshar & Williams, 2021; Ashrafian, 2017).[5]

This raises the question of whether red lines should be established for AI research that might bring about self-awareness — or the kind of superintelligence sometimes posited as a potential existential threat to humanity (Bostrom, 2014). Though many experts have advocated caution about the prospect of general AI, few had called for a halt to research in the area until March 2023, when the Future of Life Institute issued an open letter — signed by Elon Musk among others — calling for a six month pause on the development of generative AI, in the form of large language models ‘more powerful than GPT-4’ (*Pause Giant AI Experiments: An Open Letter*, 2023), referring to the generative pre-trained transformer chatbot developed by OpenAI. The letter received much coverage but did not appear likely to result in an actual halt to research. Tellingly, no government has issued such a call — though Italy did ban ChatGPT due to concerns about its use of personal data (Satariano, 2023), and China announced restrictions on similar technology if it risked upsetting the social and political order (China Releases Draft Measures for the Management of Generative Artificial Intelligence Services, 2023).

As Bostrom and others have warned, there is a non-trivial risk that attempts to contain or hobble general AI may in fact bring about the threat they are intended to avert. A ‘precautionary principle’ approach might be, therefore, to stop well short of such capabilities. Yet general AI seems far enough beyond our present capacities that this would be an excessive response if implemented today.

In any case, a ban in one jurisdiction may not bind another. Short of an international treaty, with a body competent to administer it, unilateral prohibition would be ineffective (Chesterman, 2021).

Limits on Outsourcing

Limiting the decisions that can be outsourced to AI is an area in which new rules are both necessary and possible.

One approach is to restrict the use of AI for inherently governmental functions. There have been occasional calls for a ban on government use of algorithms, typically in response to actual or perceived failures in public sector decision-making. These include scandals over automated programs that purported to identify benefit fraud in Australia (Doran, 2020) and the Netherlands (Government's Fraud Algorithm SyRI Breaks Human Rights, Privacy Law, 2020), and the Covid-19 university admissions debacle in Britain (Satariano, 2020).

Other jurisdictions have prohibited public agencies from using specific applications, such as facial recognition. San Francisco made headlines by prohibiting its use by police and other agencies in 2019, a move that was replicated in various US cities and the state of California but not at the Federal level. As in the case of data protection, Washington has thus far failed to enact broad legislation (despite several attempts) while Europe approached the same question initially as an application of the GDPR and then incorporated a ban on real-time remote biometric identification in publicly accessible spaces into the draft AI Act. China, for its part, has far fewer restrictions on facial recognition — though the government has acknowledged the need for greater guidance and there has been at least one (unsuccessful) lawsuit (Lee, 2020).

Banning algorithms completely is unnecessary, not least because any definition might include arithmetic and other basic functions that exercise no discretion. More importantly, it misidentifies the problem. The issue is not that machines are *making* decisions but that humans are *abdicating responsibility* for them. Public sector decisions exercising inherently governmental functions are legitimate not because they are correct but because they are capable of being held to account through a political or other process.

Such concerns activate the first two principles discussed at the start of this chapter: human control and transparency. A more realistic and generalizable approach to the regulation of AI in the public sector is escalating provisions for both in public sector decision-making. An early example of this was Canada's provisions on transparency of administrative decisions (Directive on Automated Decision-Making, 2019). A similar approach was taken in New Zealand's Algorithm Charter (*Algorithm Charter (NZ)*, 2020). Signed by two dozen government agencies, the Charter included a matrix that moves from optional to mandatory based on the probability and the severity of the impact on the 'wellbeing of people'. Among other provisions, mandatory application of the Charter requires ‘human oversight’, comprising a point of contact for public inquiries, an avenue for appeals against a decision, and ‘clearly explaining the role of humans in decisions informed by algorithms’. It also includes provisions on transparency that go beyond notions of explainability and include requirements for plain English documentation of algorithms and publishing information about how data are collected, secured, and stored.

These are important steps, but insufficient. For such public sector decisions, it is not simply a question of striking ‘the right balance’, as the Charter states, between accessing the power of algorithms and maintaining the trust and confidence of citizens. A more basic commitment would guarantee the means of challenging those decisions — not just legally, in the case of decisions that violate the law, but also politically, by identifying human decision-makers in positions of public trust who can be held to account through democratic processes for their actions or inaction.

One of the most ambitious attempts at regulation of this space — still being debated at the time of writing — is the EU draft AI Act. As written, it adopts an expansive definition of AI and applies to all sectors except for the military. Intended to be horizontal legislation, it would provide baseline rules applicable to all use-cases, with stricter obligations being possible in sensitive areas (such as the medical sector). It also classifies AI applications by risk: low-risk applications are not regulated at all, while escalating requirements for assessment prior to release on the market apply to medium- and high-risk applications. As indicated earlier, certain applications would be prohibited completely.

Optimists hope that the AI Act may enjoy the ‘Brussels effect’ and shape global AI policy, in the way that the EU GDPR shaped data protection laws in many jurisdictions (Siegmann & Anderljung, 2022). Critics have highlighted the extremely broad potential remit of the legislation to a wide range of technologies, as well as the vagueness of some of its key proscriptions — such as whether recommendation algorithms and social media feeds might be considered ‘manipulative’ (Veale & Zuiderveen Borgesius, 2021). Others have pointed to the risks of general purpose AI and the need to regulate it, linked to the concerns raised about large language models discussed earlier (Gebru et al., 2023).

Conclusion

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 (Asimov, 1942) 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.[6]

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 (1986), 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. Ryan Abbott, for example, has argued (2020, pp. 2-4) that the guiding principle for regulatory change should be AI legal neutrality, meaning that the law should not discriminate at all between human and AI behaviour. Though provocatively simple, the full import of such a ‘rule’ is quickly abandoned: personality is not sought for AI systems, nor are the standards of AI (the ‘reasonable robots’ of the title) to be applied to human conduct. Rather, Abbott’s thesis boils down to a case-by-case examination of different areas of AI activity to determine whether specific sectors warrant change or not.

This is a sensible enough approach, but some new rules of general application will be required, primarily to ensure the first two ‘principles’ quoted at the start of this chapter — human control and transparency — can be achieved. Human control requires limits on the kinds of AI systems that can be developed. The precautionary principle offers a means of thinking about such risks, though the clearest decisions can be made in bright line moral cases like lethal autonomous weapons. More nuanced limitations are required in the public sector, not constraining the behaviour of AI systems but limiting the ability of public officials to outsource decisions to them. On the question of transparency, accountability of government officials also requires a limit on the use of opaque processes. Above and beyond that, measures such as impact assessments, audits, an AI ombudsperson could mitigate some harms and assist in ensuring that others 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 — demand for more than ethical restrictions on AI will increase. This chapter has sought to move debate away from abstract consideration of what rules might constrain or contain AI behaviour, to the more practical challenges of why, when, and how regulators may choose to move from ethics to laws. 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.

References

Abbott, R. (2020). *The Reasonable Robot: Artificial Intelligence and the Law*. Cambridge University Press.

Adye, M.-G. J. (1878, 18 October). England, Russia, and Afghanistan. *The Times*.

Alfonsi, C. (2019). Taming Tech Giants Requires Fixing the Revolving Door. *Kennedy School Review*, *19*, 166.

Algorithm Charter for Aotearoa New Zealand. (2020). https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter

Allen, H. J. (2019). Regulatory Sandboxes. *George Washington Law Review*, *87*, 579.

Anderson, S. L. (2008). Asimov's ''Three Laws of Robotics'' and Machine Metaethics. *AI & Society*, *22*, 477.

Anshar, M., & Williams, M.-A. (2021). Simplified Pain Matrix Method for Artificial Pain Activation Embedded into Robot Framework. *International Journal of Social Robotics*, *13*(10), 187.

Ashrafian, H. (2017). Can Artificial Intelligences Suffer from Mental Illness? A Philosophical Matter to Consider. *Science and Engineering Ethics*, *23*, 403.

Asimov, I. (1942, March). Runaround. *Astounding Science Fiction*, 94.

Asimov, I. (1982). *The Complete Robot*. Doubleday.

Asimov, I. (1986). *Foundation and Earth*. Doubleday.

Auld, G., Casovan, A., Clarke, A., & Faveri, B. (2022). Governing AI Through Ethical Standards: Learning from the Experiences of Other Private Governance Initiatives. *Journal of European Public Policy*, *29*(11), 1822.

Aven, T. (2011). On Different Types of Uncertainties in the Context of the Precautionary Principle. *Risk Analysis*, *31*, 1515.

Ayres, I., & Braithwaite, J. (1992). *Responsive Regulation: Transcending the Deregulation Debate*. Oxford University Press.

Baldwin, R., & Black, J. (2016). Driving Priorities in Risk-Based Regulation: What's the Problem? *Journal of Law and Society*, *43*, 565.

Baldwin, R., Cave, M., & Lodge, M. (2011). *Understanding Regulation: Theory, Strategy, and Practice* (2nd ed.). Oxford University Press. (1999)

Balkin, J. M. (2017). The Three Laws of Robotics in the Age of Big Data. *Ohio State Law Journal*, *78*, 1217.

Bhuta, N., & Pantazopoulos, S.-E. (2016). Autonomy and Uncertainty: Increasingly Autonomous Weapons Systems and the International Legal Regulation of Risk. In N. Bhuta, S. Beck, R. Geiβ, H.-Y. Liu, & C. Kreβ (Eds.), *Autonomous Weapons Systems: Law, Ethics, Policy* (pp. 284). Cambridge University Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Boyer-Kassem, T. (2017). Is the Precautionary Principle Really Incoherent? *Risk Analysis*, *37*, 2026.

Calo, M. R. (2011). Open Robotics. *Maryland Law Review*, *70*, 571.

Chander, A., & Lê, U. P. (2015). Data Nationalism. *Emory Law Journal*, *64*, 677.

Chesterman, S. (2020). Artificial Intelligence and the Problem of Autonomy. *Notre Dame Journal on Emerging Technologies*, *1*, 210.

Chesterman, S. (2021). Weapons of Mass Disruption: Artificial Intelligence and International Law. *Cambridge International Law Journal*, *10*, 181.

China Releases Draft Measures for the Management of Generative Artificial Intelligence Services. (2023, 13 April). *National Law Review*.

Collingridge, D. (1980). *The Social Control of Technology*. Frances Pinter.

Danaher, J. (2019). Regulating Child Sex Robots: Restriction or Experimentation? *Medical Law Review*, *27*, 553.

Directive on Automated Decision-Making, (2019). www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Doran, M. (2020, 6 February). Robodebt Legal Warning Came on Same Day Scheme Was Suspended by Federal Government. *ABC News (Australia)*. https://www.abc.net.au/news/2020-02-06/robodebt-illegal-scheme-suspended/11939810

Downes, L. (2009). *The Laws of Disruption: Harnessing the New Forces that Govern Life and Business in the Digital Age*. Basic Books.

European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment. (2018). www.coe.int/cepej

Fenwick, M., Kaal, W. A., & Vermeulen, E. P. M. (2017). Regulation Tomorrow: What Happens When Technology Is Faster Than the Law? *American University Business Law Review*, *6*, 561.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. cyber.harvard.edu/publication/2020/principled-ai

Gebru, T., Hanna, A., Kak, A., Myers West, S., Gahntz, M., Khan, M., & Talat, Z. (2023). *General Purpose AI Poses Serious Risks, Should Not Be Excluded From the EU's AI Act*. ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act

Government's Fraud Algorithm SyRI Breaks Human Rights, Privacy Law. (2020, 5 February). *DutchNews.nl*.

Guihot, M., Matthew, A. F., & Suzor, N. P. (2017). Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. *Vanderbilt Journal of Entertainment & Technology Law*, *20*, 385.

Gunningham, N., & Grabosky, P. (1998). *Smart Regulation: Designing Environmental Policy*. Clarendon Press.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines*, *30*, 99.

Harlow, C., & Rawlings, R. (2009). *Law and Administration* (3rd ed.). Cambridge University Press.

Hill, K. (2014, 14 July). Are Child Sex-Robots Inevitable? *Forbes*. https://www.forbes.com/sites/kashmirhill/2014/07/14/are-child-sex-robots-inevitable

Hine, E., & Floridi, L. (2022). New Deepfake Regulations in China Are a Tool for Social Stability, but at What Cost? *Nature Machine Intelligence*, *4*, 608.

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, *1*, 389.

Justo-Hanani, R. (2022). The Politics of Artifcial Intelligence Regulation and Governance Reform in the European Union. *Policy Sciences*, *55*, 137.

Le Moli, G., Vishvanathan, P. S., & Aeri, A. (2017). Whither the Proof? The Progressive Reversal of the Burden of Proof in Environmental Cases Before International Courts and Tribunals. *Journal of International Dispute Settlement*, *8*, 644.

Lee, S. (2020). *Coming into Focus: China's Facial Recognition Regulations*. https://www.csis.org/blogs/trustee-china-hand/coming-focus-chinas-facial-recognition-regulations

Liu, J. (2020). China's Data Localization. *Chinese Journal of Communication*, *13*, 84.

Liu, Y.-H., Liao, W.-Y., Li, L., Huang, Y.-T., Xua, W.-J., & Zeng, X.-L. (2019). Reduction Measures for Air Pollutants and Greenhouse Gas in the Transportation Sector: A Cost-Benefit Analysis. *Journal of Cleaner Production*, *207*, 1023.

Luciano, F., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*, 689.

Maas, M. M. (2018). Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment. *Proceedings of 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*, 223.

Mac Síthigh, D., & Siems, M. (2019). The Chinese Social Credit System: A Model for Other Countries? *Modern Law Review*, *82*, 1034.

Mondello, G. (2022). Strict Liability, Scarce Generic Input and Duopoly Competition. *European Journal of Law and Economics*, *54*, 369.

Murugan, V. (2009). Embryonic Stem Cell Research: A Decade of Debate from Bush to Obama. *Yale Journal of Biology and Medicine*, *82*, 101.

Pause Giant AI Experiments: An Open Letter. (2023). futureoflife.org/open-letter/pause-giant-ai-experiments/

Pehrsson, U. (2016, 17 June). Europe's Obsession with Privacy Rights Hinders Growth. *Politico*.

Preparing for the Future of Artificial Intelligence. (2016). https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. (2021). (COM/2021/206 final). eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Prosser, T. (2006). Regulation and Social Solidarity. *Journal of Law and Society*, *33*, 364.

Rio Declaration on Environment and Development. (1992). In (Vol. UN Doc A/CONF.151/26 (Vol. I), Annex I).

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation. *AI & Society*, *36*, 59.

Robotics and Artificial Intelligence, Fifth Report of Session 2016–17. (2016). (HC 145). https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

Roy, K. (2015). *War and Society in Afghanistan: From the Mughals to the Americans, 1500–2013*. Oxford University Press.

Royal Commission on Motor Cars. (1906). (Cd 3080-1).

Satariano, A. (2020, 20 August). British Grading Debacle Shows Pitfalls of Automating Government. *New York Times*.

Satariano, A. (2023, 31 March). ChatGPT Is Banned in Italy Over Privacy Concerns. *New York Times*.

Selby, J. (2017). Data Localization Laws: Trade Barriers or Legitimate Responses to Cybersecurity Risks, or Both? *International Journal of Law and Information Technology*, *25*, 213.

Siegmann, C., & Anderljung, M. (2022). The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market. *arXiv.org*. arxiv.org/abs/2208.12645

Smith, B. W. (2016). Regulation and the Risk of Inaction. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving: Technical, Legal and Social Aspects* (pp. 571). Springer.

Sunstein, C. (2005). *Laws of Fear: Beyond the Precautionary Principle*. Cambridge University Press.

Turilli, M., & Floridi, L. (2009). The Ethics of Information Transparency. *Ethics and Information Technology*, *11*, 105.

Ugur, M. (Ed.). (2013). *Governance, Regulation, and Innovation: Theory and Evidence from Firms and Nations*. Edward Elgar.

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, *4*, 97.

White Paper on Artificial Intelligence. (2020). (COM(2020) 65 final). ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Wu, T. (2011). Agency Threats. *Duke Law Journal*, *60*, 1841.

Yang, F., & Yao, Y. (2022). A New Regulatory Framework for Algorithm-Powered Recommendation Services in China. *Nature Machine Intelligence*, *4*, 802.

Zetzsche, D. A., Buckley, R. P., Barberis, J. N., & Arner, D. W. (2017). Regulating a Revolution: From Regulatory Sandboxes to Smart Regulation. *Fordham Journal of Corporate & Financial Law*, *23*, 31.

国务院关于印发新一代人工智能发展规划的通知 [State Council Issued Notice of the New Generation Artificial Intelligence Development Plan]. (2017). (Guofa [2017] No 35). www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

[1] See also chapter 2 by Eileen Hunt Botting.

[2] See also chapter 10 by Eleanor Sandry.

[3] See also chapter 5 by Cindy Friedman.

[4] Images of a wrong (abuse of children or acts of violence) are generally prohibited. The question is whether a simulation itself is a wrong. The US Supreme Court, for example, has struck down provisions of the Child Pornography Prevention Act of 1996 that would have criminalized such 'speech' that 'records no crime and creates no victims by its production'.

[5] See also chapter 4 by Josh Smith.

[6] The robot initially tries to comply with a weakly-phrased order that would entail its own certain destruction and ends up stuck in an 'equilibrium' — quoting Gilbert and Sullivan, for reasons that are never explained — until the need to save a human life breaks it free.