# Introduction

Artificial intelligence (AI) and concerns about its potential impact on humanity have been with us for more than half a century. The term entered the discourse in 1956 at a Dartmouth College symposium; early research explored topics like proving logic theorems, deducing the molecular structure of chemical samples, and playing games such as draughts. A dozen years later, Stanley Kubrick's film *2001: A Space Odyssey* offered an iconic vision of a machine empowered to override the decisions of its human counterparts, the HAL 9000's eerily calm voice explaining why a spacecraft's mission to Jupiter was more important than the lives of its crew.

Both AI and the fears associated with it advanced swiftly in subsequent decades. Though worries about the impact of new technology have accompanied many inventions, AI is unusual in that some of the starkest recent warnings have come from those most knowledgeable about the field – Elon Musk, Bill Gates, and Stephen Hawking, among others. Many of these concerns are linked to 'general' or 'strong' AI, meaning the creation of a system that is capable of performing any intellectual task that a human could – and raising complex questions about the nature of consciousness and self-awareness in a non-biological entity.

The possibility that such an entity might put its own priorities above those of humans is non-trivial, but this book focuses on the more immediate challenges raised by 'narrow' AI – meaning systems that can apply cognitive functions to specific tasks typically undertaken by a human.[1]

[1] For a discussion of attempts to define AI, see Stuart J Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (3rd edn, Prentice Hall 2010) 1–5. Four broad approaches can be identified: acting humanly (the famous Turing Test), thinking humanly (modelling cognitive behaviour), thinking rationally (building on the logicist tradition), and acting rationally (the rational-agent approach favoured by Russell and Norvig, as it is not dependent on a specific understanding of human cognition or an exhaustive model of what constitutes rational thought). On the Turing Test itself, see chapter five, introduction.

A related term is 'machine learning', a subset of AI that denotes the ability of a computer to improve on its performance without being specifically programmed to do so.[2] The program AlphaGo Zero, for example, was merely taught the rules of the notoriously complex board game Go; using that basic information, it developed novel strategies that have established its superiority over any human player.[3]

The field of AI and law is fertile, producing scores of books, thousands of articles, and at least two dedicated journals.[4] In addition to the more speculative literature on what might be termed robot consciousness,[5] much of this work describes recent developments in AI systems,[6] their actual or potential impact on the legal profession,[7] and normative ques-

[2] This process may be supervised or unsupervised, or through a process of reinforcement: Kevin P Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press 2012) 2. See the discussion of human-in-the-loop and other models in chapter two, section 2.3, and the discussion of bias in machine learning in chapter three, section 3.2.1.

[3] David Silver et al, 'Mastering the Game of Go without Human Knowledge' (2017) 550 Nature 354. A subsequent iteration of the program, MuZero, was not even taught the rules of Go and other games. Julian Schrittwieser et al, 'Mastering Atari, Go, Chess, and Shogi by Planning with a Learned Model' (2020) 588 Nature 604.

[4] *Artificial Intelligence and Law* (Springer, 1992–); *RAIL: The Journal of Robotics, Artificial Intelligence & Law* (Fastcase, 2018–).

[5] See generally Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014); Mark O'Connell, *To Be a Machine: Adventures among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death* (Granta 2017); David J Gunkel, *Robot Rights* (MIT Press 2018). On legal personality of AI systems, see also Samir Chopra and Laurence F White, *A Legal Theory for Autonomous Artificial Agents* (University of Michigan Press 2011); Gabriel Hallevy, *When Robots Kill: Artificial Intelligence under Criminal Law* (Northeastern University Press 2013); John Frank Weaver, *Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws* (Praeger 2014); Gabriel Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems* (Springer 2015); Visa AJ Kurki and Tomasz Pietrzykowski (eds), *Legal Personhood: Animals, Artificial Intelligence and the Unborn* (Springer 2017). See further chapter five, section 5.3.

[6] Recent edited collections in this vein include Ryan Calo, A Michael Froomkin, and Ian Kerr (eds), *Robot Law* (Edward Elgar 2016); Patrick Lin, Keith Abney, and Ryan Jenkins (eds), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (Oxford University Press 2017); Woodrow Barfield and Ugo Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar 2018); Marcelo Corrales, Mark Fenwick, and Nikolaus Forgó (eds), *Robotics, AI and the Future of Law* (Springer 2018); Markus D Dubber, Frank Pasquale, and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (Oxford University Press 2020); Martin Ebers and Susana Navas (eds), *Algorithms and Law* (Cambridge University Press 2020); Thomas Wischmeyer and Timo Rademacher (eds), *Regulating Artificial Intelligence* (Springer 2020).

[7] See, eg, Richard Susskind, *The Future of Law: Facing the Challenges of Information Technology* (Oxford University Press 1996); Richard Susskind, *The End of Lawyers?*

tions raised by particular technologies – driverless cars,[8] autonomous weapons,[9] governance by algorithm,[10] and so on. A still larger body of writing overlaps with the broader fields of data protection and privacy, or law and technology more generally.

The bulk of that literature tends to concentrate on the activities of legal practitioners, their potential clients, or the machines themselves.[11] The objective here, by contrast, is to focus on those who seek to *regulate* those activities and the difficulties that AI systems pose for government and governance. Rather than taking specific actors or activities as the starting point, this book emphasizes structural problems that AI poses for meaningful regulation as such.

The term 'regulation' is chosen cautiously. Depending on context, its meaning can range from any form of behavioural control, whatever the origin, to the specific rules adopted by government that are subsidiary to

*Rethinking the Nature of Legal Services* (Oxford University Press 2008); Dory Reiling, *Technology for Justice: How Information Technology Can Support Judicial Reform* (Leiden University Press 2010); Richard Susskind, *Tomorrow's Lawyers: An Introduction to Your Future* (Oxford University Press UP 2013); Kevin D Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age* (Cambridge University Press 2017); Richard Susskind, *Online Courts and the Future of Justice* (Oxford University Press 2019); Simon Deakin and Christopher Markou (eds), *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence* (Hart 2020).

[8] See, eg, James M Anderson et al, *Autonomous Vehicle Technology: A Guide for Policymakers* (RAND 2014); Markus Maurer et al (eds), *Autonomous Driving: Technical, Legal and Social Aspects* (Springer 2016); Hannah YeeFen Lim, *Autonomous Vehicles and the Law: Technology, Algorithms, and Ethics* (Edward Elgar 2018).

[9] See, eg, Nehal Bhuta et al (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press 2016); Alex Leveringhaus, *Ethics and Autonomous Weapons* (Palgrave Macmillan 2016); Stuart Casey-Maslen et al, *Drones and Other Unmanned Weapons Systems under International Law* (Brill 2018); Wolff Heintschel von Heinegg, Robert Frau, and Tassilo Singer (eds), *Dehumanization of Warfare: Legal Implications of New Weapon Technologies* (Springer 2018).

[10] Christopher Steiner, *Automate This: How Algorithms Came to Rule Our World* (Penguin 2012); Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015); Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books 2016).

[11] There are some exceptions, notably focusing on the private law challenges posed by AI and robotics. See, especially, Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (Springer 2013); Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019); Mark Chinen, *Law and Autonomous Machines: The Co-evolution of Legal Responsibility and Technology* (Edward Elgar 2019); Ryan Abbott, *The Reasonable Robot: Artificial Intelligence and the Law* (Cambridge University Press 2020); Matthew Lavy and Matt Hervey, *The Law of Artificial Intelligence* (Sweet & Maxwell 2020); Dominika Ewa Harasimiuk and Tomasz Braun, *Regulating Artificial Intelligence: Binary Ethics and the Law* (Routledge 2021).

legislation.[12] In the United States, regulation is often asserted to mean a burden that is the opposite of free markets; in the academic literature, competing visions of regulation posit it as being either the infringement of private autonomy or a collaborative enterprise.[13] Across the various definitions, much of the literature discusses the different roles that specific regulators can and should play in economic and political activities.

For present purposes, the focus will be on public control of a set of activities.[14] This embraces two important aspects. The first is the exercise of control, which may be through rules, standards, or other means including supervised self-regulation. The second is that such control is exercised by one or more public bodies. These may be the executive, the legislature, the judiciary, or other governmental or intergovernmental entities, but the legitimacy of this form of regulation lies in its connection – however loose – to institutions of the state. The emphasis on public control highlights avoidance of its opposite: a set of activities that would normally be regulated falling outside the effective jurisdiction of any public entity because those activities are being undertaken by AI systems. Regulation need not, however, be undertaken purely through law in the narrow sense of the command of a sovereign backed up by sanctions.[15] It also includes economic incentives such as taxes or subsidies, recognition or accreditation of professional bodies, and other market-based mechanisms.[16]

One question that arises in this context is the extent to which AI systems themselves might have a role to play in regulation.[17] A central argument of the book, however, is that primary responsibility for regulation must fall to states. This embraces both a negative and a positive aspect. The negative aspect is that, in the near term, states should not

[12] Barry M Mitnick, *The Political Economy of Regulation: Creating, Designing, and Removing Regulatory Forms* (Columbia University Press 1980); Anthony Ogus, *Regulation: Legal Form and Economic Theory* (Hart 2004); Robert Baldwin, Martin Cave, and Martin Lodge (eds), *The Oxford Handbook of Regulation* (Oxford University Press 2010).

[13] Tony Prosser, *The Regulatory Enterprise: Government, Regulation, and Legitimacy* (Oxford University Press 2010) 1–6.

[14] Cf Philip Selznick, 'Focusing Organizational Research on Regulation' in Roger Noll (ed), *Regulatory Policy and the Social Sciences* (University of California Press 1985) 363.

[15] John Austin, *The Province of Jurisprudence Determined* (first published 1832, Cambridge University Press1995) 18–37.

[16] Robert Baldwin, Martin Cave, and Martin Lodge, *Understanding Regulation: Theory, Strategy, and Practice* (2nd edn, Oxford University Press 2011) 3.

[17] See Lawrence Lessig, *Code: Version 2.0* (Basic Books 2006).

outsource inherently governmental functions to entities (AI or otherwise) that are beyond their control.[18] The positive aspect is that, moving forward, effective management of the risks associated with AI will require international co-operation and co-ordination. Primary does not mean exclusive responsibility, however. Technology companies already play an outsized role in determining standards; this role will doubtless expand as AI systems become more complex. Yet the legitimacy of those standards and their incorporation into regulatory structures will be greatest, and they will be most effective, when endorsed by publicly accountable institutions.

The book is written for a global audience, but it is striking that the vast majority of the published material relies almost exclusively on the laws of Europe and the United States. That is understandable, given the economic importance of these jurisdictions and their sway in establishing global standards, directly or indirectly, in many fields related to technology. The two regimes also offer interesting points of comparison, with human rights concerns shaping the European response while market-based approaches hold sway in the United States. In the field of AI, however, China is – or soon will be – the dominant actor.[19] The book therefore examines the Chinese approach and the relationship between that dominance and the far more limited regulation within China. Another prominent Asian jurisdiction considered is Singapore, which has long sought to position itself as a rule of law hub to attract investment. As in the case of data protection law,[20] Singapore's government has explicitly set the goal of regulation as being to attract and encourage AI innovation.[21]

Such a public law perspective has been sorely lacking in debates over regulation of AI to date, while international law and institutions have been left out almost entirely.[22] The book builds on the author's past work

[18] See generally Simon Chesterman and Angelina Fisher (eds), *Private Security, Public Order: The Outsourcing of Public Services and Its Limits* (Oxford University Press 2009).

[19] See 腾讯研究院 [Tencent Research Institute] and 中国信通院互联网法律研究中心 [China ICT Internet Law Research Center], 人工智能：国家人工智能战略行动抓手 *[Artificial Intelligence: National Artificial Intelligence Strategy]* (Renmin University Press 2017); Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Houghton Mifflin Harcourt 2018).

[20] Simon Chesterman (ed), *Data Protection Law in Singapore: Privacy and Sovereignty in an Interconnected World* (2nd edn, Academy 2018).

[21] Model Artificial Intelligence Governance Framework (2nd edn, Personal Data Protection Commission, 2020).

[22] For a discussion of the various non-binding frameworks that have been proposed, see chapter seven, introduction.

looking at public authority in times of crisis – ranging from humanitarian intervention and transitional administration, when a state turns on its population or collapses entirely,[23] to the outsourcing of security to private actors and the expansive powers asserted by intelligence agencies in response to terrorism.[24] AI may not yet pose a threat on such a scale, but lessons on how to manage risk, draw red lines, and preserve the legitimacy of public authority are useful now – and will be essential if it ever does.

## Outline of the Book

The book is organized around the following sets of problems: How should we understand the challenges to regulation posed by the technologies loosely described here as 'AI systems'? What regulatory tools exist to deal with those challenges and what are their limitations? And what more is needed – rules, institutions, actors – to reap the benefits offered by AI while minimizing avoidable harm?

Part I groups the challenges to regulation into three broad categories.

The first, considered in chapter one, is speed. Since computers entered into the mainstream in the 1960s, the efficiency with which data can be processed has raised regulatory questions. This is well understood with respect to privacy. Data that was notionally public – divorce proceedings, say – had long been protected through the 'practical obscurity' of paper records.[25] When such material was available in a single hard copy in a government office, the chances of one's acquaintances or employer finding it were remote. Yet when it was computerized and made searchable through what ultimately became the Internet, practical obscurity disappeared. Today, high-speed computing poses comparable threats to existing regulatory models in areas from securities regulation to competition law, merely by enabling lawful activities – trading in stocks, or comparing and adjusting prices, say – to be undertaken more quickly than previously conceived possible. Many of these questions are practical

[23] Simon Chesterman, *Just War or Just Peace? Humanitarian Intervention and International Law* (Oxford University Press 2001); Simon Chesterman, *You, the People: The United Nations, Transitional Administration, and State-Building* (Oxford University Press 2004).

[24] Simon Chesterman and Chia Lehnardt (eds), *From Mercenaries to Market: The Rise and Regulation of Private Military Companies* (Oxford University Press 2007); Simon Chesterman, *One Nation under Surveillance: A New Social Contract to Defend Freedom without Sacrificing Liberty* (Oxford University Press 2011).

[25] *United States Department of Justice v Reporters Committee for Freedom of the Press*, 489 US 749, 762 (1989).

rather than conceptual and apply to technologies other than AI. Nevertheless, current approaches to slowing down decision-making – through circuit-breakers to stop trading, for example – will not address all of the problems raised by the speed of AI systems.

A second set of challenges is the increasing autonomy of those systems, exposing gaps in regulatory regimes that assume the centrality of human actors. Yet surprisingly little attention is given to what is meant by 'autonomy' and its relationship to those gaps. Driverless vehicles and autonomous weapon systems are the most widely studied examples, but related issues arise in algorithms that allocate resources or determine eligibility for programmes in the private or public sector. Chapter two develops a novel typology that distinguishes three lenses through which to view the regulatory issues raised by autonomy: the practical difficulties of managing risk associated with new technologies, the morality of certain functions being undertaken by machines at all, and the legitimacy gap when public authorities delegate their powers to algorithms.

Chapter three turns to the increasing opacity of AI. As computer programs become ever more complex, the ability of non-specialists to understand them diminishes. Opacity may also be built into programs by companies seeking to protect proprietary interests. Both such systems are capable of being explained, albeit with recourse to experts or an order to reveal their internal workings. Yet a third kind of system may be naturally opaque: some machine learning techniques are difficult or impossible to explain in a manner that humans can comprehend. This raises concerns when the process by which a decision is made is as important as the decision itself. For example, a sentencing algorithm might produce a 'just' outcome for a class of convicted persons. Unless the justness of that outcome for an individual defendant can be explained in court, however, it is, quite rightly, subject to legal challenge. Separate concerns are raised by the prospect that AI systems may mask or reify discriminatory practices or outcomes.

This is, of course, a non-exhaustive list of the challenges posed by AI. Among others on the horizon are the likely displacement of large segments of the workforce and the possibility of artificial general intelligence raising meaningful questions about the rights of 'smart robots'.[26] Nor does this study seek to examine the broader ethical implications of AI taking on greater roles in society, or the regulation of cyberspace, virtual

[26] See above n 5.

worlds, and so on.[27] Similarly, it will not attempt to cover fully the potential impact of blockchain or distributed ledger technology.[28] The more modest aim is to use the problems identified in this part to highlight gaps in existing regulatory models with a view to seeing whether the tools at our disposal can fill them.

Part II, then, turns to those tools. Chapter four examines how existing laws can and should apply to emerging technology through attribution of responsibility. Legal systems typically seek to deter identifiable persons – natural or juridical – from certain forms of conduct, or to allocate losses to those persons. Responsibility may be direct or indirect: key questions are how the acts and omissions of AI systems can and should be understood. Given the complexity of those systems, novel approaches to responsibility have been proposed, including special applications of product liability, agency, and causation. More important and less studied is the role that insurance can play in compensating harm but also structuring incentives for action. Another approach is to limit the ability to *avoid* responsibility, drawing on the literature on outsourcing and the prohibition on transferring certain forms of responsibility – most notably the exercise of discretion in the public sector.

As AI systems operate with greater autonomy, however, the idea that they might themselves be held responsible has gained credence. On its face, the idea of giving those systems a form of independent legal personality may seem attractive. Yet chapter five argues that this is both too simple and too complex. It is simplistic in that it lumps a wide range of technologies together in a single legal category ill-suited to the task; it is overly complex in that it implicitly or explicitly embraces the anthropomorphic fallacy that AI systems will eventually assume full legal personality in the manner of the 'robot consciousness' arguments mentioned earlier. Though the emergence of general AI is a conceivable future

[27] See, eg, F Gregory Lastowka, *Virtual Justice: The New Laws of Online Worlds* (Yale University Press 2010); Andrew Sparrow, *The Law of Virtual Worlds and Internet Social Networks* (Gower 2010); Jacqueline Lipton, *Rethinking Cyberlaw: A New Vision for Internet Law* (Edward Elgar 2015); Andrew Murray, *Information Technology Law: The Law and Society* (3rd edn, Oxford University Press 2016); Paul Lambert, *Gringras: The Laws of the Internet* (5th edn, Bloomsbury 2018); Lilian Edwards (ed), *Law, Policy, and the Internet* (Hart 2019); Roxana Radu, *Negotiating Internet Governance* (Oxford University Press 2019); Frank Pasquale, *New Laws of Robotics: Defending Human Expertise in the Age of AI* (Belknap Press 2020).

[28] See, eg, William J Magnuson, *Blockchain Democracy: Technology, Law, and the Rule of the Crowd* (Cambridge University Press 2020); Fabian Schär and Aleksander Berentsen, *Bitcoin, Blockchain, and Cryptoassets* (MIT Press 2020).

scenario – and one worth taking precautions against – it is not a sound basis for regulation today.

Notions of foreseeability underpin another tool that has been embraced as a means of limiting the risks associated with AI: transparency. Chapter six considers the manner in which transparency and the related concept of 'explainability' are being elaborated, notably the 'right to explanation' in the European Union (EU) and a move towards explainable AI (XAI) among developers. These are more promising than the arguments for legal personality, but the limits of transparency are already beginning to show as AI systems demonstrate abilities that even their programmers struggle to understand. That is leading regulators to cede ground and settle for explanations of adverse decisions rather than transparency of decision-making processes themselves. Such a backward-looking approach relies on individuals knowing that they have been harmed – which will not always be the case – and should be supplemented with forward-looking mechanisms like impact assessments, audits, and an ombudsperson.

The final part of the book considers the rules and institutions required to address the inadequacies of existing tools and regulatory bodies.

As the preceding chapters demonstrate, existing norms, suitably interpreted, are able to deal with many of the challenges presented by AI. But not all. Chapter seven begins with a survey of guides, frameworks, and principles put forward by states, industry, and intergovernmental organizations. These diverse efforts have led to a broad consensus on half a dozen norms that might govern AI. Far less energy has gone into determining how these might be implemented – or if they are even necessary. Rather than contribute to norm proliferation, the chapter focuses on why regulation is necessary, when regulatory changes should be made, and how it would 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 uncontrollable or uncontainable AI could pose a threat far more serious than lethal autonomous weapon systems. 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.

The answers that each political community finds to the law reform questions posed may differ, but a larger threat in the very near future is that AI systems capable of causing harm will not be confined to one jurisdiction – indeed, it may be impossible to link them to a specific

jurisdiction at all. This is not a new problem in cybersecurity, but different national approaches to regulation will pose barriers to effective regulation exacerbated by the speed, autonomy, and opacity of AI systems. For that reason, some measure of collective action, or at least co-ordination, is needed. Lessons may be learned from efforts to regulate the global commons, as well as moves to outlaw at the international level certain products (weapons and drugs, for example) and activities (such as slavery and child sex tourism). The argument advanced here is that regulation, in the sense of public control, requires active involvement of states. To co-ordinate those activities and enforce global 'red lines', chapter eight posits a hypothetical International Artificial Intelligence Agency (IAIA), modelled on the agency created after the Second World War to promote peaceful uses of nuclear energy, while deterring or containing its weaponization and other harmful effects.

Chapter nine turns to the possibility that the AI systems challenging the legal order may also offer at least part of the solution. Here, China, which has among the least developed rules to regulate conduct by AI systems, is at the forefront of using that same technology in the courtroom. This is a double-edged sword, however, as its use implies a view of law that is instrumental, with parties to proceedings treated as means rather than ends. That, in turn, raises fundamental questions about the nature of law and authority: at base, whether law is reducible to code that can optimize the human condition or if it must remain a site of contestation, of politics, and inextricably linked to institutions that are themselves accountable to a public. For many of the questions raised, the rational answer will be sufficient; but for others, *what* the answer is may be less important than *how* and *why* it was reached, and *whom* an affected population can hold to account for its consequences.

## Precaution vs Innovation

Underlying the question of regulation is the need to balance precautionary steps against unnecessarily constraining innovation. A government report in Singapore, for example, highlighted the risks posed by AI, but concluded that 'it is telling that no country has introduced specific rules on criminal liability for artificial intelligence systems. Being the global first-mover on such rules may impair Singapore's ability to attract top industry players in the field of AI.'[29]

[29] Penal Code Review Committee (Ministry of Home Affairs and Ministry of Law, August 2018) 29. China, for its part, included in the State Council's AI development