**Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence**

Dr Michael Guihot, Anne Matthew and Dr Nicolas Suzor*



* Dr Michael Guihot, Senior Lecturer, Member, Commercial and Property Law Research Centre, Faculty of Law, Queensland University Technology ('QUT'); Anne Matthew, Lecturer, Member, Commercial and Property Law Research Centre, Faculty of Law, QUT; Associate Professor Nicolas Suzor, School of Law, QUT. Associate Professor Suzor is the recipient of an Australian Research Council DECRA Fellowship (Project Number DE160101542). The authors can be reached at a.matthew@qut.edu.au.

We Robot Conference at Yale University March 2017. Please do not cite without the authors' consent.

## 1. Introduction

When Google purchased DeepMind in 2014, its owners made it a condition of the sale that Google establish an ethics board to govern the future use of the artificial intelligence technology.[1] Google apparently agreed, but nothing is known about who the members of the board are or of the content of any discussions that the board might have had. On 20 July 2016, Google reported that it had deployed DeepMind's machine learning in a series of tests on one of its live data centres. The tests resulted in a reported 40% decrease in energy consumption for the centre while the AI was applied. Google's press release on the tests noted that 'because the algorithm is a general-purpose framework to understand complex dynamics, we plan to apply this to other challenges in the data centre environment and beyond in the coming months'.[2] Google also reported that 'working at Google scale gives us the opportunity to learn how to apply our research to truly global and complex problems, to validate the impact we can have on systems that have already been highly optimised by brilliant computer scientists, and - as our data centre work shows - to achieve amazing real-world impact too'.[3] Working at 'Google scale' presumably means using Google's worldwide infrastructure to test AI systems. If these results can be replicated more broadly so as to reduce the world's energy consumption, then humanity will reap the benefits. However, while the results of this application of AI appear laudable, what checks and balances were in place to govern its application? Were any risks of its application considered and ameliorated in the tests? Should society be concerned about the rampant research and development into AI by some of the world's biggest companies without ostensible regulation governing it?[4] One question that might be asked is, if regulation was put in place prematurely or without proper thought and consultation, would the obvious benefits in reduced energy consumption that might result from the general application of this program in other areas be retarded or lost? This paper asks, with the increase in societal concerns about the risk inherent in developing AI, is regulation of AI inevitable and if so what form will it take?

The pace of innovation in AI has far outstripped the pace of innovation in regulatory tools that might be used to govern it. This is often referred to as the pacing problem of regulation.[5] In these situations, regulation lags behind or in some circumstances 'decouples' from the technology it seeks to address.[6] The core challenge regulatory agencies face lies in the difficulty in understanding the social impacts of AI on a systems level, and engaging with

---

1 Alex Hern, *Whatever happened to the DeepMind AI ethics board Google promised?*, THE GUARDIAN, January 27, 2017, https://www.theguardian.com/technology/2017/jan/26/google-deepmind-ai-ethics-board (last visited Mar 13, 2017).

2 Google, DEEPMIND AI REDUCES GOOGLE DATA CENTRE COOLING BILL BY 40% DEEPMIND (2016), https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/ (last visited Mar 13, 2017).

3 Google, DEEPMIND COLLABORATIONS WITH GOOGLE DEEPMIND (2016), https://deepmind.com/applied/deepmind-for-google/ (last visited Mar 13, 2017).

4 This issue was raised in an article in Nature by Kate Crawford and Ryan Calo and was referred to as the "blind spot in thinking about AI". See Kate Crawford & Ryan Calo, *There is a blind spot in AI research*, 538 NATURE 311–313, 311 (2016).

5 See THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT: THE PACING PROBLEM, (Gary E. Marchant, Braden R. Allenby, & Joseph R. Herkert eds., 2011); Braden R. Allenby, *Governance and Technology Systems: The Challenge of Emerging Technologies*, *in* THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT 3–18 (Gary E. Marchant, Braden R. Allenby, & Joseph R. Herkert eds., 2011); Kenneth W Abbott, *Introduction: The Challenges of Oversight for Emerging Technologies*, *in* INNOVATIVE GOVERNANCE MODELS FOR EMERGING TECHNOLOGIES 1–16 (Kenneth W Abbott, Gary E. Marchant, & Braden R. Allenby eds., 2014).

6 Braden R. Allenby, *The Dynamics of Emerging Technology Systems*, *in* INNOVATIVE GOVERNANCE MODELS FOR EMERGING TECHNOLOGIES , 43 (Kenneth W Abbott, Gary E. Marchant, & Braden R. Allenby eds., 2013).

these impacts at every (or any) stage of development.[7] As the DeepMind example illustrates, the reasons why particular decisions involving the ways in which AI is developed and applied are made can be opaque, largely incomprehensible,[8] and sometimes even unknowable.[9] Research and development of AI can be, and apparently is already being, carried out in many different locations, at different times, in ways that are not highly visible — and at a scale that only large multinational companies such as Google can attain.

Current regulatory mechanisms, including laws governing tort, copyright, privacy, and patent among others, and regulations that govern other emerging technologies are either unsuitable or for other reasons cannot be applied to novel technological developments in areas such as the regulation of AI.[10] The challenge in regulating this field is magnified by the fundamental uncertainty about how AI will develop, and what challenges we will need to face in the future.[11] While the regulation of other emerging technologies is not directly applicable to AI, there is much that can be learnt from innovations in regulation in other fields. Brownsword considered the regulation of emerging technology more generally. He asked whether 'generic lessons' could be learned about the regulation of technologies with similar risk profiles such as the regulation of biotechnology and nanotechnology. He concluded that, while there are some regulatory tools that can be applied generally, each particular technology has its own profile of participants, risks, and norms that require regulation to be more nuanced when applied to novel technology.[12]

Public regulators such as governments face an unprecedented challenge in managing complex governance systems that include not only public regulatory agencies but also individuals, firms, market competitors, and civil society organisations that all might play some role in influencing the development of AI in different contexts. In order to meet the challenge of regulating AI, we suggest the need for finesse and new ways of thinking. For a government to influence the development of AI systems and successfully further the public interest, it must be able to understand and influence this complex and intricate web of actors that often have diverse goals, intentions, purposes, norms and powers.[13]

This paper seeks to identify opportunities for innovation in the work of public regulators. Too often law and legal analysis focus on setting the outer limits for technological innovation – from outright prohibitions on certain types of research to the allocation of liability for harm that results from the deployment of new technologies. Much less is known about how public regulators can successfully influence the iterative processes of innovation to more steadily guide the development of new technologies. The available tools for governing innovation in this way are still relatively crude. Tax policy and direct public funding are familiar methods of setting agendas for investment in research and development, but more targeted interventions are still at very early stages of development. Scholars have suggested that

---

7 Crawford and Calo, *supra* note 4.

8 Perri 6, *Ethics, regulation and the new artificial intelligence, part II: autonomy and liability*, 4 INF COMMUN SOC 406–434, 410 (2001).

9 FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).

10 See Allenby, *supra* note 6 at 20–21.

11 Gonenc Gurkaynak, Ilay Yilmaz & Gunes Haksever, *Stifling artificial intelligence: Human perils*, COMPUT. LAW SECUR. REV., 754–5 (2016), http://www.sciencedirect.com/science/article/pii/S0267364916300814 (last visited Nov 2, 2016).

12 Roger Brownsword, *So What Does the World Need Now? Reflections on Regulating Technologies*, *in* REGULATING TECHNOLOGIES 23–48, 30 (Roger Brownsword & Karen Yeung eds., 2008).

13 Julia Black, *Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a "Post-Regulatory" World*, 54 CURR. LEG. PROBL. 103–146, 105 (2001).

public regulators may be able to experiment with more rapid, temporary laws,[14] although the potential lack of legal certainty that results still causes unsolved problems for investors and other participants in the field.

Other regulation theorists are experimenting with different interventions in choice architecture to set the context and environment in which choices are made so as to promote regulatory goals.[15] Still others have proposed greater roles for regulatory agencies with specific expertise,[16] and understanding how these agencies may be able to better perform these roles is the focus of this paper. Whitt proposes applying his version of 'adaptive policymaking', where regulators 'tinker' with 'inputs, connectivity, incentives, and feedback'[17] to encourage firms to act in ways that further the public good.[18]

We provide examples of how public regulators in other fields are beginning to use an expanded regulatory toolset to address some of the same challenges that arise in regulating AI. In short, a regulatory intervention in the development of AI technology may include elements of risk governance; it needs to be flexible, adaptable, responsive, nimble, and must include some element of regulation through the use of technology.

In Part 2 of this paper, we outline the challenges that current and future developments in AI are likely to pose for regulators. We consider the urgency of developing effective mechanisms of regulation, and explain how these challenges are different in kind to challenges of regulating in other domains. In order to meet these challenges, we suggest that regulators will need to develop innovative strategies. We show that recent developments in how regulation is conceived go some way to identifying potential future strategies for regulators, but that more work is needed.

In Part 3, we turn to consider the specific requirements for effective regulation of AI. We highlight the different capabilities that public regulators must develop in order to successfully influence the design and deployment of AI technologies. We provide examples from different attempts to regulate new technologies to ground each of these requirements. In Part 4 we highlight the need for innovative approaches to regulate AI and set out some of the ways that a decentred approach might work. Part 5 describes innovative measures designed to address the resource and informational constraints faced by regulators. In Part 6 we conclude with a suggestion for cooperation between regulators and regulatees.

---

[14] Wulf A. Kaal, *Dynamic Regulation for Innovation*, *in* PERSPECTIVES IN LAW, BUSINESS AND INNOVATION (M Fenwick et al. eds., 2016), https://papers.ssrn.com/abstract=2831040 (last visited Nov 2, 2016); S. RANCHORDÁS, CONSTITUTIONAL SUNSETS AND EXPERIMENTAL LEGISLATION: A COMPARATIVE PERSPECTIVE (2015).

[15] Frederik J. Zuiderveen Borgesius, *Behavioural Sciences and the Regulation of Privacy on the Internet*, *in* NUDGING AND THE LAW-WHAT CAN EU LAW LEARN FROM BEHAVIOURAL SCIENCES (Alberto Alemanno & Lise Sibony eds., 2014), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2513771 (last visited Nov 4, 2016).

[16] Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J. LAW TECHNOL. 354–400 (2016).

[17] Richard S. Whitt, *Adaptive Policymaking: Evolving and Applying Emergent Solutions for U.S. Communications Policy*, 61 FED. COMMUN. LAW J. 483, 487 (2008).

[18] Other versions of adaptive policymaking to address deep uncertainty have been proposed using various models or approaches to policymaking. See for example the various adaptive approaches set out in Warren E Walker, Vincent AWJ Marchau & Darren Swanson, *Addressing Deep Uncertainty Using Adaptive Policies*, 77 TECHNOL. FORECAST. SOC. CHANGE 917–923 (2010); Warren E Walker, Adnan S Rahman & Jonathan Cave, *Adaptive Policies, Policy Analysis, and Policy-Making*, 128 EUR. J. OPER. RES. 282–289 (2001).

## 2. Artificial Intelligence: What does the future hold?

The Stanford Report into AI titled *Artificial Intelligence and Life in 2030* made the following finding:

> Contrary to the more fantastic predictions for AI in the popular press, the Study Panel found no cause for concern that AI is an imminent threat to humankind. No machines with self-sustaining long-term goals and intent have been developed, nor are they likely to be developed in the near future. Instead, increasingly useful applications of AI, with potentially profound positive impacts on our society and economy are likely to emerge between now and 2030, the period this report considers.[19]

However, in recent years, prominent figures have begun to warn about the need to ensure that the development and deployment of AI technology is effectively regulated. An open letter hosted by the Future of Life organisation recommends 'expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial' and argues that 'our AI systems must do what we want them to do'.[20] There are signs too from within the AI industry that precaution needs to be taken.[21] Bostrom recently noted that 'it may seem obvious now that major existential risks would be associated with … an intelligence explosion, and that the prospect should therefore be examined with the utmost seriousness even if it were known (which it is not) to have but a moderately small probability of coming to pass.'[22] A research priorities document attached to the Future of Life open letter and authored by Russell, Dewey and Tegmark argued that 'very general and capable AI systems operating autonomously to accomplish some task will often be subject to effects that increase the difficulty of maintaining meaningful human control'.[23] Elon Musk, the founder of Tesla, puts the case for caution succinctly:

> I think we should be very careful about artificial intelligence. If I had to guess at what our biggest existential threat is, it's probably that. So we need to be very careful. I'm increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don't do something very foolish.[24]

Russell, Dewey and Tegmark also set out two policy questions that they argue need to be addressed: '(1) what is the space of policies worth studying, and how might they be enacted? (2) Which criteria should be used to determine the merits of a policy?'[25] They

---

[19] PETER STONE ET AL., ARTIFICIAL INTELLIGENCE AND LIFE IN 2030 4 (2016), https://ai100.stanford.edu/2016-report (last visited Mar 14, 2017).

[20] AI Open Letter, FUTURE OF LIFE INSTITUTE, https://futureoflife.org/ai-open-letter/ (last visited Mar 13, 2017). The letter has been signed by over 8 000 people including Stuart Russell, Peter Norvig, Elon Musk, Stephen Hawking, Steve Wozniak, Laurent Orseau and many other prominent members of the AI community from Microsoft, DeepMind, Google, and Facebook.

[21] See Steve Omohundro, *Rational Artificial Intelligence for the Greater Good*, *in* SINGULARITY HYPOTHESES: A SCIENTIFIC AND PHILOSOPHICAL ASSESSMENT 161–179 (Amnon H Eden et al. eds., 2012); Stuart Russell, Daniel Dewey & Max Tegmark, *Research Priorities for Robust and Beneficial Artificial Intelligence*, AI MAGAZINE, 2015, at 105–114.

[22] NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 5 (First edition ed. 2014).

[23] Russell, Dewey, and Tegmark, *supra* note 21 at 111.

[24] Samuel Gibbs, *Elon Musk: artificial intelligence is our biggest existential threat*, THE GUARDIAN, October 27, 2014, https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat (last visited Mar 13, 2017).

[25] Russell, Dewey, and Tegmark, *supra* note 21 at 107.

propose that the qualities that these policies might have include 'verifiability of compliance, enforceability, ability to reduce risk, ability to avoid stifling desirable technology development, likelihood of being adopted, and ability to adapt over time to changing circumstances'.[26] These issues are only some of the issues that face regulators trying to address the regulation of AI and are some of the things considered in this paper.

The Future of Life Institute's conference in January 2017 drafted a list of 23 principles grouped under three headings: research issues, ethics and values, and longer-term issues. The principles that fall within the longer term issues include principle 22 titled 'Importance' that states 'Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources'. Principle 23 titled 'Risks' notes that 'Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact'. Further, principle 24 titled 'Recursive Self-Improvement' notes that 'AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures'.[27] These principles reflect the concern that even those within the industry hold about the development of AI and must be considered in any regulatory responses.

Another industry body, the Institute of Electrical and Electronics Engineers (IEEE), recently produced a discussion paper titled 'Ethically Aligned Design: A Vision for Prioritising Human Well-Being with Artificial Intelligence and Autonomous Systems'.[28] The Ethically Aligned Design project aimed to 'bring together multiple voices in the Artificial Intelligence and Autonomous Systems communities to identify and find consensus on timely issues'. Those issues address concerns about how to ensure that AI does not infringe human rights, that the decisions of autonomous systems are accountable and transparent, and that there are checks in place to minimise risks through enhanced education.[29] The ongoing pace of change, and the notoriously slow response of lawyers and regulators, creates real challenges for this type of multidisciplinary collaboration. So much so that, in a cri de coeur, the Ethically Aligned Design report noted that 'there is much to do for lawyers in this field that thus far has attracted very few practitioners and academics despite being an area of pressing need'.[30] The report calls on lawyers to be 'part of the discussions on regulation, governance, and domestic and international legislation in these areas.'[31]

The task of regulating the development and deployment of AI appears to be increasingly pressing. The AI Now Report prepared after the AI Now public symposium hosted by the White House and New York University's Information Law Institute in July 2016 set out several key recommendations for future work in AI development. One of those recommendations was to:

> Increase efforts to improve diversity among AI developers and researchers, and broaden and incorporate the full range of perspectives, contexts, and disciplinary backgrounds into the development of AI systems. The field of AI should also support

---

26 *Id.* at 107.

27 Future of Life Institute, ASILOMAR AI PRINCIPLES FUTURE OF LIFE INSTITUTE, https://futureoflife.org/ai-principles/ (last visited Mar 13, 2017).

28 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, STANDARDS ASSOCIATION, ETHICALLY ALIGNED DESIGN: A VISION FOR PRIORITIZING HUMAN WELLBEING WITH ARTIFICIAL INTELLIGENCE AND AUTONOMOUS SYSTEMS (2016), http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html (last visited Mar 14, 2017).

29 *Id.* at 5.

30 *Id.* at 89.

31 *Id.* at 89.

> and promote interdisciplinary AI research initiatives that look at AI systems' impact from multiple perspectives, combining the computational, social scientific, and humanistic.[32]

We take up the challenge of contributing to AI research from a legal and regulatory perspective in this paper. We set out below a number of concrete examples of potential and existing problems associated with the deployment of unregulated AI systems.

### (a) The singularity and the end of humanity

Perhaps the most visible fear about the development of AI is the existential threat to humanity as a result of the rise of super-intelligent machines.[33] In 1965 Good argued that society would be transformed by the invention of a machine with ultra-intelligence, the last machine that humans would ever need to make for themselves.[34] This *ex machina* in human image, would have a 'vast artificial neural circuitry' designed from an understanding of human intellect.[35] It would surpass human intelligence and be able to design even more intelligent machines.[36] Good boldly claimed that humanity's survival would depend upon the building of an ultra-intelligent machine.[37] His optimism was not shared by subsequent scholars such as Vinge, who saw super-intelligent machines not as saviours but as the advent of doomsday. Vinge's concern was that once the machine attained human level intelligence, it would not remain at that level for long and would reach superintelligence and beyond very quickly. Vinge argued that such a machine could become aware of its own superior intelligence. This event, which he described as the singularity, would spell the end of humanity.[38]

The biggest fears around the development of AI are not new and are not confined to fears of AI. Age-old concerns in human mythology about humans playing god-the-creator and the ramifications of that form the basis of stories such as the Golem stories. These stories have parallels with, and lessons for, the development of general AI. In the myth, a golem is created, often from clay, and imbued with life through 'a detailed statement of specific letter combinations that are required to bring about the "birth" of a golem'[39] — the algorithm. In some golem stories, the golem obtains superhuman strength and, uncontrolled, causes destruction and mayhem. The parallels to the creation of AI with super human intelligence are apt. A further parallel might be drawn with the desire to regulate or control these fears. For example, the golem stories also note that 'golems, while not human, were still bound by

---

[32] K CRAWFORD ET AL., THE AI NOW REPORT: THE SOCIAL AND ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN THE NEAR-TERM 5 (2016), https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3_RpmwKHu.pdf (last visited Nov 18, 2016).

[33] RAY KURZWEIL, THE SINGULARITY IS NEAR (2010); BOSTROM, *supra* note 22; Nick Bostrom, *When machines outsmart humans*, 35 FUTURES 759–764 (2003); JOHN VON NEUMANN & RAY KURZWEIL, THE COMPUTER & THE BRAIN (2012).

[34] Irving John Good, *Speculations Concerning the First Ultraintelligent Machine*, 6 *in* ADVANCES IN COMPUTERS 31–88, 31–32 (Franz L. Alt and Morris Rubinoff ed., 1966), http://www.sciencedirect.com/science/article/pii/S0065245808604180 (last visited Mar 14, 2017).

[35] *Id.* at 78.

[36] *Id.* at 33.

[37] *Id.* at 31.

[38] Vernor Vinge, *The coming technological singularity: how to survive in the post-human era*, *in* VISION 21: INTERDISCIPLINARY SCIENCE AND ENGINEERING IN THE ERA OF CYBERSPACE 11–22, 33 (1993), https://ntrs.nasa.gov/search.jsp?R=19940022855 (last visited Mar 14, 2017).

[39] STORYTELLING: AN ENCYCLOPEDIA OF MYTHOLOGY AND FOLKLORE, 204 (Josepha Sherman ed., 2008).

Jewish law. They would not take a person's life unless it was absolutely necessary, and they were not capable of falsehood.[40] Like Asimov's three laws of robotics,[41] we see here the birth of the idea of embedding legal codes within technical code.[42]

Existential concerns have piqued the minds of ethicists and philosophers since soon after work began on AI.[43] However, discussions about the legal ramifications of AI were typically slow to develop and early considerations of AI and the law only appear in the early 1980s.[44] Lehman-Wilzig was quick to draw parallels with golem stories and that of Frankenstein's monster in his comparisons to AI but he concentrated on the criminal sanctions that may be applicable for problems associated with AI. This is not the emphasis of this paper but, nevertheless, some of the problems he recognised, such as how to control AI, were obvious problems deserving of a regulatory response even at that stage.[45] Lehman-Wilzig noted that those developing AI programs 'will often be largely ignorant of quite what is going on inside, and thus will not know if and when the computer has learned too much, i.e. that the danger point has been passed.'[46] Lehman-Wilzig wrote:

> it is simply a matter of fact that almost all very large computer systems in use today have "many layers of poorly understood control structure and obscurely encoded knowledge." It is simply no longer possible for these systems' designers — or for anyone else — to understand what these systems have "evolved into," let alone to anticipate into what they will evolve.[47]

This problem has not diminished and has probably increased since 1981. Considering the problem more recently, Omohundro noted that 'rational systems are subject to a variety of ''drives'' including self-protection, resource acquisition, replication, goal preservation, efficiency, and self-improvement'.[48] Researchers in AI recognise that there is a potential risk that if autonomous AI is developed, it will be difficult for a human operator to maintain control.

Cultural shifts will be attendant upon AI taking on more corporeal forms, particularly those that are human-like with language capabilities.[49] Stanford's major study, *The One Hundred Year Study on Artificial Intelligence (AI100)* predicts that as driverless cars fall into common use, they will be formative of first public impressions of AI in a corporeal form.[50] This

---

40 *Id.* at 205.

41 The three laws of robotics appear in a short work of science fiction: ISAAC ASIMOV, RUNAROUND (1942); The short story was republished in a collection of stories on robots first published in 1950: ISAAC ASIMOV, I, ROBOT (2013). The three laws feature in stories throughout this collection.

42 LAWRENCE LESSIG, CODE: VERSION 2.0 (2nd edition ed. 2006).

43 See NORBERT WIENER, CYBERNETICS (2nd ed. 1961).

44 See Sam N Lehman-Wilzig, *Frankenstein Unbound: Towards a Legal Definition of Artificial Intelligence*, 13 FUTURES 442–457 (1981).

45 Note Good, *supra* note 34 at 33 where Good contemplated control when he suggested that the "intelligence explosion" following invention of intelligent machines would ensure that the first ultra-intelligent machine would be the last invention humans would need, "provided that the machine is docile enough to tell us how to keep it under control".

46 Lehman-Wilzig, *supra* note 44 at 446 citing; WIENER, *supra* note 43.

47 Lehman-Wilzig, *supra* note 44 at 446.

48 Omohundro, *supra* note 21.

49 See generally Crawford and Calo, *supra* note 4 at 313 where Crawford and Calo acknowledge the cultural shift associated with artificial intelligence. Will Knight, *AI's Unspoken Problem*, 119 MIT TECHNOLOGY REVIEW, 2016, at 28–37.

50 STONE ET AL., *supra* note 19 at 18–25.

experience will be an important one for AI, since we are on the cusp of surge of AI with a physical embodiment. The Stanford study predicts that the typical North American city will by 2030 feature personal robots, driverless trucks and flying cars.[51] As *ex machina* looks to come to pass, the attendant cultural shift will prompt not only debate around what it is to be human, and to think, but what rights should robots have? McNally and Inavatullah argue that the rights of robots deploying AI will become an increasingly pressing issue as we reconsider our relationship with the world around us and reconceptualise our own rights and the responsibilities expected of AI.[52] McNally and Inavatullah consider that the debate surrounding the rights of AI robots will be contextualised by an environment where they can be alive, or at least perceived to be alive by the humans with whom they interact.[53]

Perhaps to counter these concerns, Orseau and Armstrong, an engineer at DeepMind and a researcher into systemic risk respectively, recently published a paper detailing how DeepMind's engineers have developed a 'big red button', or an off switch for such an artificially intelligent reinforcement learning agent. As Orseau and Armstrong note, 'reinforcement learning agents interacting with a complex environment like the real world are unlikely to behave optimally all the time'.[54] They recognise 'concerns that a "superintelligent" agent may resist being shut down, because this would lead to a decrease of its expected reward'.[55] In those cases, it is important that the superintelligent agent acting on reinforced learning process will allow a human to interrupt its operation.

### (b) Current problems with AI

For the moment, the social ramifications of rampant, uncontrollable AI are still the imaginings of science fiction writers.[56] The current challenge for regulating AI is the proliferation in the capabilities of AI systems tasked with performing a specific function that could be performed by human intelligence.[57] As Domingos puts it, '[p]eople worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world'.[58] Developments in AI technology have been smouldering since research on it began shortly after World War II.[59] Today, AI is at the

---

[51] STONE ET AL., *supra* note 19, 18-23 (automated vehicles), 24-25 (home robots), 7, 18, 20 (flying vehicles).

[52] Phil McNally & Sohail Inayatullah, *The rights of robots*, 20 FUTURES 119–136, 120, 125–126, 134 (1988).

[53] See *Id.* at 125–126.

[54] Laurent Orseau & Stuart Armstrong, *Safely Interruptible Agents*, *in* UNCERTAINTY IN ARTIFICIAL INTELLIGENCE: 32ND CONFERENCE (2016).

[55] *Id.* at 2.

[56] See for example, Will Knight, *AI's Future Is Not So Scary*, 119 TECHNOLOGY REVIEW; CAMBRIDGE, 2016, at 17. As Knight puts it, at 17, "we can stop fretting that it's going to destroy the world like Skynet."

[57] KURZWEIL, *supra* note 33 at 459, where Kurzweil explains that there is an expectation that narrow AI will perform the task better or faster than human intelligence given the AI's capacity to manage and consider vast arrays of data and variables; Ben Goertzel, *Human-level artificial general intelligence and the possibility of a technological singularity*, 171 ARTIF. INTELL. 1161–1173, 1162 (2007). Goertzel notes that the distinguishing features of narrow AI are that it does not understand itself, the task, nor how to generalize or apply the knowledge it has learnt in performing the task beyond the specific problem. For example, a narrow AI program for diagnosing one type of cancer, would not itself be able to generalize its diagnostic insights to diagnose another type of cancer, though a human might be able to further develop the first AI for the subsequent purpose.

[58] PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD (2016), https://en.wikipedia.org/w/index.php?title=The_Master_Algorithm&oldid=742981158 (last visited Nov 18, 2016).

[59] John McCarthy, WHAT IS ARTIFICIAL INTELLIGENCE? (2007), http://www-formal.stanford.edu/jmc/whatisai/ (last visited Mar 13, 2017).

forefront of technology development. AI is used in driverless vehicles, speech and facial recognition, language translation, lip-reading, combatting spam and online payment fraud, detecting cancer, law enforcement, logistics planning, and language translation. Much of this AI is what can be described as narrow AI, that is, AI designed to solve a specific problem or familiar task, such as to play chess. These commercial applications of AI appear to be limitless and the world's largest technology companies are investing heavily in its potential. IBM's cognitive computing platform, Watson, has leapt from winning jeopardy to commercial and health applications.[60] DeepMind's AlphaGo recently defeated the human master of the complex Chinese board game Go; Microsoft has incorporated AI into its agents such as Cortana and Zo who can perform a dizzying array of tasks and answer seemingly unlimited questions using a mellifluous (female by design) computer generated voice;[61] Microsoft has also developed an algorithm named DeeperCoder that is capable of writing code to solve simple problems;[62] Facebook uses AI in its face recognition, language translation, camera effects and its research arm, Facebook Artificial Intelligence Research (FAIR), is said to be 'committed to advancing the field of machine intelligence'.[63] Joaquin Candela, Director of Engineering for Facebook's Applied Machine Learning (AML) group has stated that Facebook is working towards 'generalization of AI'.[64] Facebook's director of core machine learning Hussein Mehana has noted that the generalization of AI is capable of enhancing the speed at which applications can be built by 'a hundred-x magnitude', expanding possibilities for impact in fields ranging from medicine to transportation.[65]

---

[60] IBM describes Watson as "the world's first and most-advanced AI platform": Cognitive Computing - IBM Research, , http://research.ibm.com/cognitive-computing/ (last visited Mar 11, 2017); See also IBM WATSON, IBM WATSON: HOW IT WORKS (2014), https://www.youtube.com/watch?v=_Xcmh1LQB9I (last visited Mar 11, 2017); Video: IBM insiders break down Watson's Jeopardy! win, TED BLOG (2011), http://blog.ted.com/experts-and-ibm-insiders-break-down-watsons-jeopardy-win/ (last visited Mar 11, 2017); IBM, IBM WATSON: A SYSTEM DESIGNED FOR ANSWERS (2011), https://www.youtube.com/watch?v=cU-AhmQ363I (last visited Mar 11, 2017); STEPHEN BAKER, FINAL JEOPARDY: MAN VS. MACHINE AND THE QUEST TO KNOW EVERYTHING (2011); Jessica S Allain, *From Jeopardy! to Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems*, 73 LA. LAW REV. 1049–1070 (2012); Ryan Abbott, *I Think, Therefore I Invent: Creative Computers and the Future of Patent Law*, 57 BCL REV 1079, 1088–1091 (2016); Betsy Cooper, *Judges in Jeopardy: Could IBM's Watson Beat Courts at Their Own Game*, 121 YALE LJF 87 (2011); IBM is currently tasking Watson with learning how to help with the identification of melanoma, and is seeking peoples input to assist with timely, accurate detection. See IBM, IBM COGNITIVE - OUTTHINK MELANOMA - AUSTRALIA, https://www.ibm.com/cognitive/au-en/melanoma/ (last visited Mar 11, 2017); Commercial applications of Watson include, for example, ROSS Intelligence's software marketed to lawyers as "your own personal artificially intelligent researcher … that can effortlessly find the answer to any legal question"; ROSS can be asked questions in natural language, just as you would "any other lawyer". See ROSS INTELLIGENCE, MEET ROSS, YOUR BRAND NEW ARTIFICIALLY INTELLIGENT LAWYER 0.32-0.36 seconds (2016), https://www.youtube.com/watch?v=ZF0J_Q0AK0E (last visited Mar 10, 2017); Mark Gediman, *Artificial Intelligence: Not Just Sci-Fi Anymore*, 21 AALL SPECTR. 34–37, 35–36; Paul Lippe, *What We Know and Need to Know About Watson, Esq.*, 67 SCL REV 419 (2015).

[61] Microsoft, MICROSOFT'S AI VISION, ROOTED IN RESEARCH, CONVERSATIONS NEWS CENTRE, https://news.microsoft.com/features/microsofts-ai-vision-rooted-in-research-conversations/ (last visited Mar 13, 2017).

[62] Dave Gershgorn, MICROSOFT'S AI IS LEARNING TO WRITE CODE BY ITSELF, NOT STEAL IT QUARTZ, https://qz.com/920468/artificial-intelligence-created-by-microsoft-and-university-of-cambridge-is-learning-to-write-code-by-itself-not-steal-it/ (last visited Mar 20, 2017).

[63] Facebook, FACEBOOK AI RESEARCH (FAIR) FACEBOOK RESEARCH, https://research.fb.com/category/facebook-ai-research-fair (last visited Mar 14, 2017).

[64] Steven Levy, INSIDE FACEBOOK'S AI MACHINE BACKCHANNEL (2017), https://backchannel.com/inside-facebooks-ai-machine-7a869b922ea7 (last visited Mar 13, 2017).

[65] *Id.*

Advances in AI technology are vaulting toward the exponential as computer capacity and speed double every two years.[66]

Issues arising with current AI systems present a more immediate issue than the longer-term possibility of the development of superintelligent AI. These issues exist along a spectrum, such that some are likely to be dealt with by developers as they come to their attention. Others may be dealt with by end users of the system as they refine their use of the system and work with developers in overcoming issues as and when they arise. Some though, are capable of provoking a regulatory response. In this Part three examples of issues along the spectrum are considered: bias in law enforcement, reliance on AI in judicial decision making and privacy.

(i) Bias

The coalescing of AI and big data open significant possibilities for synthesis and analysis, but stand to compound existing problems with both. Serious concerns exist for bias as a ghost in the machine of AI and big data. These include unintended racism, sexism and discrimination.[67] Ajunwa, et al have proposed a model for the regulation of big data to address privacy concerns and to allow a pathway for the correction of erroneous assumptions made from an assemblage of big data.[68] Bias can be difficult to detect, but is important to question lest it 'become part of the logic of everyday algorithmic systems'.[69] These biases have arisen in a law enforcement context: Algorithms performing predictive risk assessments of defendants committing future crimes were making mistakes with risk scores for black defendants, giving them high risk scores at almost double the rate of white defendants.[70] Risk scores were mistakenly low for white defendants.[71] Bias also arises in the work of private platforms that filter, index, and sort online content and mediate communications.[72] Crawford sees at least some of this as a manifestation of a bias problem with data and calls for vigilance in AI system design and training to avoid built in bias.[73] Bias

[66] This is known as Moore's Law after the co-founder of Intel who predicted in 1965 that computing power would double every year (later revised to every two years). There is some speculation that this rate of change is no longer happening. See Tom Simonite, *Moore's Law is Dead. Now What?*, MIT TECHNOL. REV. (2016), https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/ (last visited Mar 14, 2017); See also PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD 287 (2015).

[67] Kate Crawford, *Artificial Intelligence's White Guy Problem*, THE NEW YORK TIMES, June 25, 2016, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html (last visited Mar 13, 2017); Kate Crawford, *Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics*, 41 SCI. TECHNOL. HUM. VALUES 77–92 (2016).

[68] Ifeoma Ajunwa, Kate Crawford & Joel S. Ford, *Health and Big Data: An Ethical Framework for Health Information Collection by Corporate Wellness Programs*, 44 J. LAW. MED. ETHICS 474 (2016).

[69] Crawford, *supra* note 67.

[70] Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.*, PROPUBLICA, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last visited Nov 18, 2016).

[71] *Id.*

[72] Tarleton Gillespie, *The Relevance of Algorithms*, *in* MEDIA TECHNOLOGIES: ESSAYS ON COMMUNICATION, MATERIALITY, AND SOCIETY 167–93 (Tarleton Gillespie, Pablo Boczkowski, & Kirsten Foot eds., 2013); Nicolas Suzor, *Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms*, (2017), https://osf.io/ymj3t/ (last visited Sep 30, 2016).

[73] Crawford, *supra* note 67; See also: Kate Crawford, DARK DAYS: AI AND THE RISE OF FASCISM (2017), http://schedule.sxsw.com/2017/events/PP93821 (last visited Mar 13, 2017).

issues such as these are unlikely to provoke a regulatory response if they are dealt with in AI system design.

(ii) Legal decision making

AI has potential for more extensive application to highly specific legal tasks such as sentencing and judicial interpretation in an effort to improve transparency and consistency in judicial decisions. Concerns exist for mechanistic reliance upon such AI systems and the capacity of such systems to influence and shape the behaviour of people involved in the decision making process. This is significant since it suggests that the use of AI systems in this context has a regulatory effect on the people involved in the decision making process informed by AI systems. It is for these reasons that use of AI systems in legal decision making attracts criticism.

AI systems relied upon in judicial decision making have been criticized as lacking capacity to exercise discretion and make situational value judgments. Decision-making in the application of legal principles necessarily involves discretion. Hall et al created an early model for decision making in sentencing acknowledging that sentencing relies on 'induction and intuition as well as the capacity to assess the social impact of the decision'.[74] These are not among AI's greatest strengths. Sunstein argues that AI is not capable of legal reasoning since the analogical reasoning involved requires evaluation of value judgments.[75] Bench-Capon and Prakken suggest that, at least in theory, AI ought to be capable of quite sophisticated legal reasoning given the structure and context of legal argument.[76] Two decades ago Leith argued that capacity to exercise discretion may limit the potential of AI to 'fully represent the richness of legal knowledge in any useful way'.[77] The development of expert legal systems seems to have entered into somewhat of an 'AI Winter' before emerging with a new-found maturity, still plagued by old problems. Leith, who remains sceptical as to the use of AI in legal expert systems, including their use in sentencing, acknowledges that expert legal systems are attractive given their promise to improve legal processes and suggests we ought to remain wary of their allure.[78] Schild proposes that AI should not make definitive legal decisions involving discretionary judgments, such as those that must resolve 'conflicting arguments' or 'ambiguous and contradictory evidence'; rather, Schild argues AI systems be utilised to better inform human decisions.[79] Zeleznikow agrees that where disputes involve interpretation of facts or data, the involvement of people in

---

[74] Maria Jean J. Hall et al., *Supporting discretionary decision making with information technology: a case study in the criminal sentencing jurisdiction*, , 9 (2005), http://www.uoltj.ca/articles/vol2.1/2005.2.1.uoltj.Hall.1-36.pdf (last visited Mar 12, 2017).

[75] CASS R. SUNSTEIN, OF ARTIFICIAL INTELLIGENCE AND LEGAL REASONING 5 (2001), https://papers.ssrn.com/abstract=289789 (last visited Mar 13, 2017).

[76] Trevor Bench-Capon & Henry Prakken, *Argumentation*, *in* INFORMATION TECHNOLOGY AND LAWYERS 61–80 (Arno R. Lodder & Anja Oskamp eds., 2006), http://link.springer.com.ezp01.library.qut.edu.au/chapter/10.1007/1-4020-4146-2_3 (last visited Mar 13, 2017).

[77] Philip Leith, *The Judge and the Computer: How Best "Decision Support"?*, 6 ARTIF. INTELL. LAW 289–309, 14 (1998).

[78] Philip Leith, *The rise and fall of the legal expert system*, 30 INT. REV. LAW COMPUT. TECHNOL. 94–106, 94, 101, 104 (2016); See also Philip Leith, *The Emperor's New Expert System*, 50 MOD. LAW REV. 128–132 (1987); Philip Leith, *Logic, Formal Models and Legal Reasoning*, 24 JURIMETRICS 334–356 (1984).

[79] URI J SCHILD, EXPERT SYSTEMS AND CASE LAW 19, 27, 193 (1992); John Zeleznikow, *Building decision support systems in discretionary legal domains*, 14 INT. REV. LAW COMPUT. TECHNOL. 341–356, 204 (2000).

decision making is vital.[80] As Simpson puts it, even if AI is able to approximate human discretion in sentencing decision making, the question that remains is the extent to which 'an algorithm can have a heart':[81] 'to what extent can such algorithms deal with the unexpected, quirky or unique individual that may require appeals to a sense of justice?'[82] These concerns animate Article 22 of the EU's General Data Protection Regulation, which creates a new right for individuals 'not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her'.[83] The implication, at least in Europe, is that from 2018 human must somehow be involved in decision making, although how effective this is likely to be remains to be seen.

Even where AI is used to support human decision making with the goal of increasing consistency, there are still substantial risks. Hall et al contend that consistency in legal decision making may not be desirable if it leads to standardization.[84] Oskamp and Tragter argue that standardization in automated legal decision making processes has a regulatory effect on people involved in the decision making process.[85] This regulatory impact may extend to an unintended chilling effect on individualization, even where the legislature intended there to be some flexibility.[86] People involved in the decision making process may have difficulty deviating from the standardization in order to have a heart, as Simpson puts it,[87] to 'introduce an element of humanity in special circumstances'[88] or to consider whether the decision is in the best interests of society.[89] Paliwala argues that transparency of these evaluations, or lack of thereof, is critical since, if the decision making process is opaque, the transformative impact will extend to the character and quality of the law itself, and its interrelationship with society.[90] Paliwala calls for legal decision making systems to be developed with more than just an understanding 'rule handling skills';[91] There also needs to be a deeper understanding of the social and ethical contextual framework and the needs of users of the decision making system.[92] Without such a deeper understanding, Paliwala warns that we risk 'forcing crude AI systems on society' leading to 'results which do not promote social justice and human rights.'[93] These concerns for the development of systems

---

80 Zeleznikow, *supra* note 79 at 350; See also Anja Oskamp & Maaike W. Tragter, *Automated Legal Decision Systems in Practice: The Mirror of Reality*, 5 ARTIF. INTELL. LAW 291–322, 312 (1997); Note Christopher Hart, Chairman, National Transportation Safety Board (USA), has stated his belief that autonomous vehicles are most likely to involve human co-pilots due to issues associated with trust, and the limitations of programing to deal with all potential eventualities including ethical challenges. See: Andrew Rosenblum, *Policing Driverless Cars*, 119 TECHNOLOGY REVIEW; CAMBRIDGE, 2016, at 15.

81 Brian Simpson, *Algorithms or advocacy: does the legal profession have a future in a digital world?*, 25 INF. COMMUN. TECHNOL. LAW 50–61, 56 (2016).

82 *Id.* at 56.

83 2016/679 (GDPR), to take effect from 25 May 2018.

84 Hall et al., *supra* note 74.

85 Oskamp and Tragter, *supra* note 80 at 293.

86 *Id.* at 293.

87 Simpson, *supra* note 81 at 56.

88 Hall et al., *supra* note 74 at 33.

89 Oskamp and Tragter, *supra* note 80 at 308.

90 Abdul Paliwala, *Rediscovering artificial intelligence and law: an inadequate jurisprudence?*, 30 INT. REV. LAW COMPUT. TECHNOL. 107–114, 112–113 (2016).

91 *Id.* at 112–113.

92 *Id.* at 112–113.

93 *Id.* at 112–113.

with deep contextual understanding will become more pressing where AI systems share information with each other in order to inform improvement of their own process.[94]

Reliance upon AI systems in judicial decision making enlivens long standing fears regarding reducing human processes to their most mechanistic. In 1973 Rose and Rose warned against reductionism of human intellect in generating better models for machine intelligence, given that the process would lend itself to identifying and seizing upon the most mechanistic traits of human intelligence.[95] Rose and Rose argued that the result would be neither beneficial for human intelligence nor human decision making since, once a process was reduced to its most mechanistic, it would make humans more compliant or programable to the process.[96] This is a warning of an unintended regulatory effect. Rose and Rose argued that predicating machine intelligence on human intelligence would blur the limits of one with the other in such a way that justifications for differentiation in the treatment of machines vis a vis humans would inevitably arise.[97]

Cooper has considered the manner in which the systematic reasoning of AI could be used by the judiciary in statutory interpretation so as to avoid ideological bias that judges may bring to their interpretation.[98] Cooper argues that while AI may be objective, since it lacks ideological bias, judicial decision making should involve some normative inputs of which AI is incapable, such as evaluating the absurdity of an interpretation.[99] AI is more likely to decide a statutory interpretation issue based on frequency and predictive systems.[100] Lippe et al argue that the information architecture of law is lacking, and this will significantly limit the extent to which narrow AI can assist with complex legal work.[101] AI is likely to struggle with the usefulness of predictive systems in fields of law where the law is unsettled, the subject of complex debate or frequent amendment.[102] Lippe et al suggest that neither human nor technology alone are sufficient to undertake complex legal work given the disconnect between the current manner in which lawyers create, access and develop legal argument and documents on the one hand, and the demands of integration and differentiation of processes and information in complex organizations on the other. Lippe et al propose that at this stage, 'the appropriate question is to determine what ensemble of humans and technology can most efficiently and accurately complete a given task'.[103]

The array of concerns surrounding the use of AI systems in judicial decision making are likely to be managed by the continual refinement of how AI systems are deployed by people

---

94 Consider for example, Amanda Schaffer, *Robots That Teach Each Other*, 119 TECHNOLOGY REVIEW; CAMBRIDGE, 2016, at 48–51 where Schaffer explains data sharing goals to improve robot ability; Will Knight, *Shared Robot Knowledge*, 119 TECHNOLOGY REVIEW; CAMBRIDGE, 2016, at 26.

95 Steven P. R. Rose & Hilary Rose, *"Do not adjust your mind, there is a fault in reality"— ideology in neurobiology*, 2 COGNITION 479–502, 498–499 (1973).

96 *Id.* at 498–499.

97 *Id.* at 499.

98 Cooper, *supra* note 60 at 97–99.

99 *Id.* at 99.

100 *Id.* at 99.

101 Paul Lippe, Daniel Martin Katz & Dan Jackson, *Legal by Design: A New Paradigm for Handling Complexity in Banking Regulation and Elsewhere in Law*, , 4, 13, 20 (2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2539315 (last visited Mar 10, 2017).

102 McGinnis and Pearce suggest that Dodd-Frank regulation may be a bridge too far for AI: John O. McGinnis & Russell G. Pearce, *The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services*, , 3042 (2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2436937 (last visited Mar 10, 2017).

103 Lippe, Katz, and Jackson, *supra* note 101 at 20.

in the decision making process: the judiciary and their administrators. Such decisions are capable of having a regulatory effect, shaping the development of AI systems for use in this context. Public regulatory response is unlikely to be required.

(iii) Privacy

Privacy concerns relating to existing AI systems may provoke a regulatory response. Privacy issues may surround disclosure of data sets to tech companies with AI capabilities, even where data is disclosed for a specific purpose and is technically compliant with current regulatory disclosure models. The potential regulatory response is a serious concern for AI as it may impede future possibilities for development of AI systems for the greater public benefit. The leaps in advancement that are the promise of AI may turn on the quality and quantity of information available to it to inform AI learning. In some areas, such as health, rich data sets are controlled in heavily regulated environments. Privacy concerns associated with disclosure are somewhat avoided where the information is volunteered to the tech company by the patients or general public themselves.[104] Volunteering of personal information for a particular purpose can be distinguished from situations where data sets held by public health services are shared with tech companies. Sensitivities surrounding well-intentioned disclosures may result in a regulatory response, even where the disclosure technically complies, or invites debate surrounding compliance, with existing regulatory processes.[105] Such a regulatory response may be because the existing regulatory compliance process did not contemplate the scale of the disclosure, use by AI systems or how the data might be used and stored by private entities not previously considered an interested stakeholder in that type of data at the time the regulatory process was settled. In

[104] For example, the public appeal for involvement in helping IBM's Watson gather information about personal experiences with melanoma. A media saturation campaign in Australia has promoted the sharing of personal stories on social media with Watson. Cancer screening to assist with Watson's data collection has been offered to the public on Bondi Beach, Sydney, Australia on 18-19 March 2017. Information being shared includes demographic information, family history with melanoma, photographs or text describing changes to the person's own skin. See: IBM, *supra* note 60.

[105] See for example the debate surrounding the disclosure of private health data of an estimated 1.3 million UK patients in a collaboration between DeepMind and the Royal Free London NHS Foundation Trust in the UK. For statements on the project from DeepMind and the Royal Free London NHS Foundation, see: DeepMind, WORKING WITH THE NHS TO BUILD LIFE SAVING TECHNOLOGY DEEPMIND, https://deepmind.com/ (last visited Mar 19, 2017); The Royal Free London NHS Foundation Trust, GOOGLE DEEPMIND: Q&A FOR PATIENTS ROYAL FREE LONDON NHS FOUNDATION TRUST, https://www.royalfree.nhs.uk/news-media/news/google-deepmind-qa/ (last visited Mar 19, 2017); DeepMind has provided information about its Independent Reviewers involved in the NHS project here: DeepMind, OUR INDEPENDENT REVIEWERS DEEPMIND, https://deepmind.com/ (last visited Mar 19, 2017); The relevant statute in the United Kingdom applicable to the disclosure of this type of data is the DATA PROTECTION ACT 1998 (UK), legislation primarily regulated by the Information Commissioner's Office (ICO). The ICO is currently reviewing these disclosures with the assistance of the National Data Guardian. DeepMind and the Royal Free London NHS Foundation Trust, have on their webpages referred to in this note stated their belief that they satisfied all appropriate regulatory processes for exchange of this data and this is the point that is the subject of review; The National Data Guardian has reported completed its report for the ICO. See: Jane Wakefield Cellan-Jones Dave Lee, Rory, *Google DeepMind's NHS deal under scrutiny*, BBC NEWS, March 17, 2017, http://www.bbc.com/news/technology-39301901 (last visited Mar 19, 2017); Note: Information Commissioner's Office, STATEMENT ON NHS DIGITAL (FORMERLY HSCIC) FOLLOW-UP REPORT INFORMATION COMMISSIONER'S OFFICE (ICO) (2017), https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/02/statement-on-nhs-digital-formerly-hscic-follow-up-report/ (last visited Mar 20, 2017); The debate surrounding this disclosure is explored in Julia Powles & Hal Hodson, *Google DeepMind and healthcare in an age of algorithms*, HEALTH TECHNOL. (2017), http://link.springer.com/10.1007/s12553-017-0179-1 (last visited Mar 20, 2017).

the worst case scenario, such a regulatory response may involve the imposition of a command and control model heavily restricting future access to such data sets.

In this Part we have outlined a few examples of problems arising with AI systems, some of which may provoke a regulatory response. As we set out in the next Part, regulating AI systems is an extremely difficult problem to solve. Formulating the regulatory response will be a challenging one for any regulator. As specific problems manifest, fear, anxiety or populist concerns, whether evidence based or not, may create an urge in the regulator to step in. We urge a considered, principled and consultative approach.

## 3. Regulating AI

Given the growing visibility of AI and the potential risks outlined above, there is increasing interest in how regulators might respond. Some of the risks seem remote, but stories that portray the catastrophic consequences of autonomous, self-aware AI such as those portrayed in science fiction as well as the prophecies of researchers such as Omohundro pervade the zeitgeist and have begun to induce a level of anxiety and fear that may well yet reach a tipping point in society's consciousness.[106] People can be particularly risk averse when they stand to lose something,[107] and the threats posed by AI may well soon generate sufficient alarm to prompt governments to regulate.

The threat that governments may respond to these fears and seek to interfere in AI development has given rise to important warnings against heavy-handed and clumsy regulation. Entrepreneurs and technological innovators maintain a healthy fear of regulation, which is often seen as red tape that hinders or stymies development. Thierer argues about what he terms 'permissionless innovation' that:

> … refers to the notion that experimentation with new technologies and business models should generally be permitted by default. Unless a compelling case can be made that a new invention will bring serious harm to society, innovation should be allowed to continue unabated and problems, if any develop, can be addressed later.[108]

Technology-rich industries have a long history of seeking to avoid the impulse to regulate that often accompanies widespread social fears about new technologies.[109] The key fear is that it may be too early to regulate AI, and that any regulation adopted today 'may hinder developments that could prove essential for human existence.'[110] Scholars and industry representatives have expressed important concerns about the limits of regulation in high technology industries, and AI poses its own specific challenges for regulators.

---

[106] Malcolm Gladwell identified the three characteristics that identify what he described as a tipping point, particularly in epidemics, as 'one, contagiousness; two, the fact that little causes can have big effects; and three, that change happens not gradually but at one dramatic moment', see MALCOLM GLADWELL, THE TIPPING POINT: HOW LITTLE THINGS CAN MAKE A BIG DIFFERENCE 9 (2000).

[107] Daniel Kahneman & Amos Tversky, *Prospect Theory: An Analysis of Decision Under Risk*, 47 ECONOMETRICA 263–292, 279 (1979).

[108] ADAM THIERER, PERMISSIONLESS INNOVATION: THE CONTINUING CASE FOR COMPREHENSIVE TECHNOLOGICAL FREEDOM 2 (2016), https://www.mercatus.org/system/files/Thierer-Permissionless-revised.pdf (last visited Mar 13, 2017).

[109] Adam D. Thierer, *Technopanics, threat inflation, and the danger of an information technology precautionary principle*, 14 MINN. J. LAW SCI. TECHNOL. 309 (2012).

[110] Gurkaynak, Yilmaz, and Haksever, *supra* note 11 at 753.

### 3.1. Regulating is hard to do

'Regulation is extraordinarily difficult'.[111] When we consider the regulation of new technologies, former justice of the High Court of Australia, Michael Kirby noted that 'the normal organs of legal regulation often appear powerless'.[112] Regulating the development of AI may yet be the hardest task for regulators to tackle. It begins with a definitional problem. Once intelligence is defined, estimations or approximations of those qualities form the benchmark of attempts to create or simulate it. Intelligence is often described in human terms as 'consciousness, self-awareness, language use, the ability to learn, the ability to abstract, the ability to adapt, and the ability to reason'.[113] The fuzzy nature of intelligence complicates settlement upon a generally accepted definition of 'artificial' intelligence. A working definition postulated by Scherer is that AI 'refers to machines that are capable of performing tasks, that if performed by a human, would be said to require intelligence'.[114] John McCarthy defined artificial intelligence as 'the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.'[115] Russell and Norvig define AI via characteristic traits differentiated by how they reflect expectations of human thinking and behaviour or (machine) rational thinking and behaviour.[116]

The problems that face regulators in this area include that the technology outpaces any attempt at controlling it. This pacing problem is a problem that plagues the regulation of technology generally and often leads to the technology disengaging or decoupling from the regulation that seeks to regulate it. The regulation becomes irrelevant. Scherer sets out four general problems with regulating research and development of AI as problems with (1) discreetness, that is 'AI projects could be developed without large scale integrated institutional frameworks', (2) diffuseness, that is AI projects could be carried out by diffuse actors in many locations around the world; (3) discreteness, that is, projects will make use of discrete components and technologies 'the full potential of which will not be apparent until the components come together'; and (4) opacity, that is, the 'technologies underlying AI will tend to be opaque to most potential regulators'.[117] These *ex ante* problems: discreetness, diffuseness, discreteness and opacity are but some of the many things that regulators of AI will need to consider. The development of AI poses other problems for those seeking to regulate its development. Scherer also proposed a system under which an agency would be set up to certify AI systems as safe,[118] and where such certified systems 'enjoy limited tort liability'[119] while uncertified systems would be subject to full liability. This approach concentrates on consequences of problems with AI and seeks to punish errant behaviour after it has occurred. This paper is more concerned with proposing solutions to regulating the development of AI *ex ante.* Even Scherer's approach may not be sufficiently comprehensive

[111] Bridget M Hutter, *A Risk Regulation Perspective on Regulatory Excellence*, *in* ACHIEVING REGULATORY EXCELLENCE 101–114, 101 (Cary Coglianese ed., 2017).

[112] Michael Kirby, *New Frontier: Regulating Technology by Law and "Code," in* REGULATING TECHNOLOGIES 367–388, 383 (Roger Brownsword & Karen Yeung eds., 2008).

[113] Scherer, *supra* note 16 at 360.

[114] *Id.* at 360.

[115] McCarthy, *supra* note 59 at 2.

[116] STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (3rd ed ed. 2016).

[117] Scherer, *supra* note 16 at 369.

[118] *Id.* at 394.

[119] *Id.* at 394.

to capture the multifarious regulatory problems associated with AI. Some of those problems are set out below.

### (a) The pacing problem

Regulators need to find the optimal middle ground between regulation that is ineffective because it cannot keep pace with the rate of innovation, and regulation that is too general to be meaningful in specific cases. Many have argued that in the face of the continuously increasing speed of innovation, legal and ethical oversight has lagged.[120] At the forefront of scientific discovery, artificial intelligence is affected by this issue more than other technologies. Attempts by regulators to address the pacing problem by future-proofing legislation often result in regulatory disconnect, where the laws are too general or vague to effectively serve their intended purpose or provide meaningful guidance regarding any specific technology.[121]

### (b) Information Asymmetry and the Collingridge Dilemma

Any innovating technology will pose a problem for regulators in the form of information asymmetries.[122] Even if lawmakers are able to obtain technical information from developers, most non-technical folk will still be at a loss to understand a product, let alone predict what impacts it may have on individuals, societies and economies.[123] This is the major cause of the pacing problem, but it is also an issue for courts trying to interpret and apply any legislation that has been implemented, as well as commentators and advocacy groups looking to hold companies accountable. The information problem forms the first half of the Collingridge Dilemma, which states that at the earliest stages of development of a new technology, regulation is difficult due to a lack of information, while in the later stages entrenchment of the technology results increased resistance to regulatory change, from users, developers and investors.[124]

---

[120] Marchant, G.E., Allenby, B.R., and Heckert, J.R. (eds) *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Springer, 2011); Fenwick, M., Kaal, W.A., and Vermeulen, E.P.M, 'Regulation Tomorrow: What Happens When Technology is Faster than the Law?' (2016) *Lex Research Topics in Corporate Law & Economics Working paper No. 2016-8*; Brownsword R, *Rights, Regulation, and Technological Revolution* (Oxford University Press, 2008); Laurie, G., Harmon, S.H.E., Arzuaga, F., 'Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty' (2012) 4(1) *Law, Innovation and Technology,* 1.

[121] Marchant, G.E., Allenby, B.R., and Heckert, J.R. (eds) *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Springer, 2011); Fenwick, M., Kaal, W.A., and Vermeulen, E.P.M, 'Regulation Tomorrow: What Happens When Technology is Faster than the Law?' (2016) *Lex Research Topics in Corporate Law & Economics Working paper No. 2016-8*; Brownsword R, *Rights, Regulation, and Technological Revolution* (Oxford University Press, 2008); Purdy R, 'Legal and Regulatory Anticipation and 'Beaming' Presence Technologies' (2014) 6(2) *Law, Innovation and Technology*, 147; Brownsword R, *Rights, Regulation, and Technological Revolution* (Oxford University Press, 2008).

[122] Brownsword R, *Rights, Regulation, and Technological Revolution* (Oxford University Press, 2008); Laurie G, Harmon SHE and Arzuaga F, 'Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty' (2012) 4(1) *Law, Innovation and Technology,* 1.

[123] Stephenson MC, 'Information Acquisition and Institutional Design' (2011) 124(6) *Harvard Law Review*, 1422; Bakhshi H, Freedman A and Heblich PJ, *State of Uncertainty: Innovation Policy through Experimentation* (NESTA, 2011); Mandel GN, 'Regulating Emerging Technologies' (2009) 1 *Law, Innovation and Technology*, 75.

[124] Collingridge D, *The Social Control of Technology* (Pinter, 1980); Laurie G, Harmon SHE and Arzuaga F, 'Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty' (2012) 4(1) *Law, Innovation and Technology,* 1.

The fact that the technology is opaque[125] also makes it easier for firms to hide wrongdoing and evade regulation. Volkswagen, for example, was able to create specific code to identify the tests use by regulators to measure emissions and make its car engines appear to run more cleanly than when under normal use. Similarly, recent reports suggest that Uber has created a version of their app that is specifically designed to identify regulators and prevent them from accessing the system to investigate concerns or collect evidence.[126]

### (c) Regulatory Delay and Legal Uncertainty

Many authors writing in the economics, innovation and management fields have described the impact of legal uncertainty around new innovations, and regulatory delay.[127] The latter occurs where legislators have hastily banned the commercialisation of new products in response to concerns in the public, acting even before enough research can be conducted to ascertain whether the concerns are well founded.[128] Investors and developers are left in the dark while legislators decide what to do, sometimes having to withdraw funding and resources from what might turn out to be a useful and lucrative innovation because they couldn't bear the risk.[129] Paradoxically, Sunstein argues that 'overregulation produces underregulation' — that is, imposing stringent regulations may lead to the administrators of the regulation not issuing regulations or refusing to enforce the regulations.[130]

Many commentators worry that social pressure to regulate the unfamiliar will unduly limit the social benefits that may come from innovation and development of AI. Sunstein in particular has warned against adopting what is known as the 'precautionary principle' to regulate risk.[131] People are nothing if not 'predictably irrational';[132] our assessment of risk is negatively impacted by traits such as loss aversion, the myth of a benevolent nature, the availability heuristic, probability neglect, and system neglect. Regulators should accordingly avoid instituting stifling innovation purely on the threat of unknown future risks, noting that 'unfamiliar risks produce far more concern than familiar ones, even if the latter are statistically larger'.[133] At the same time, Sunstein warns against regulatory inaction, noting that 'well-organised private groups sometimes demand conclusive proof of harm as a precondition for regulation; the demand should be firmly resisted because a probability of harm is, under many circumstances, a sufficient reason to act'.[134] Ultimately, Sunstein urges that 'a better approach would acknowledge that a wide variety of adverse effects may come

---

[125] PASQUALE, *supra* note 9.

[126] The Associated Press, *Uber Deploys Secret Weapon Against Undercover Regulators*, THE NEW YORK TIMES, March 3, 2017, https://www.nytimes.com/aponline/2017/03/03/us/ap-us-uber-dodging-authorities.html (last visited Mar 13, 2017).

[127] Bakhshi H, Freedman A and Heblich PJ, *State of Uncertainty: Innovation Policy through Experimentation* (NESTA, 2011); Mandel GN, 'Regulating Emerging Technologies' (2009) 1 *Law, Innovation and Technology*, 75; Braeutigam RR, 'The effect of uncertainty in the regulatory delay on the rate of innovation' (1979) 43(1) *Law & Contemporary Problems*, 98; Stephenson MC, 'Information Acquisition and Institutional Design' (2011) 124(6) *Harvard Law Review*, 1422.

[128] Braeutigam RR, 'The effect of uncertainty in the regulatory delay on the rate of innovation' (1979) 43(1) *Law & Contemporary Problems*, 98.

[129] Mandel GN, 'Regulating Emerging Technologies' (2009) 1 *Law, Innovation and Technology*, 75; Romano R, 'Regulating in the Dark' (2012) *Yale Law and Economics Research Paper No. 442.*

[130] Cass R Sunstein, *Regulatory Paradoxes*, 57 UNIV. CHIC. LAW REV. 407–441 (1990).

[131] Cass R Sunstein, *Beyond the Precautionary Principle*, 151 UNIV. PA. LAW REV. 1003–1058, 1003–1004 (2003).

[132] DAN ARIELY, PREDICTABLY IRRATIONAL: THE HIDDEN FORCES THAT SHAPE OURDECISIONS (Revised ed. 2009); see also DANIEL KAHNEMAN, THINKING, FAST AND SLOW (First paperback edition ed. 2013).

[133] Sunstein, *supra* note 131 at 1009.

[134] *Id.* at 1055.

from in action, regulation, and everything in between', noting the need to 'attempt to consider all of those adverse effects and not simply a subset'.[135]

### (d) Coordination Across Regulatory Bodies

The problem of coordinating the many regulatory bodies involved in a new technology plagues every innovating industry.[136] Given the increasingly interdisciplinary nature of AI research, it is no less a challenge for regulators in this field.[137] An AI regulatory regime would need to account for existing laws, governmental regulatory bodies, and self-regulatory industry bodies that enforce professional codes of ethics, and it needs to do this across the fields of neuroscience; neurobiology; mechanical, electrical and software engineering; psychology; innovation studies; and economics and finance, for a start.[138]

### (e) Agency Capture

Regulatory failure due to agency capture occurs where regulators become sympathic with the industry they are regulating. This can be the result of any number of factors, such as a high frequency of interaction between industry and regulators, industry reps 'buying off' regulators with gifts like free lunches or sponsorship to attend conferences, or a 'revolving door' for employees between regulatory agencies and industry.[139] While each of these problems are relatively common throughout innovating industries, the AI industry is particularly at risk of the revolving door issue.[140] The information asymmetry issue makes the knowledge and expertise acquired by employees of AI developers particularly valuable to regulators, who are likely to be interested in employing former AI developers when they can.

### (f) Limited enforcement mechanisms and Jurisdiction Shopping

Added to these complexities, the major players in the area such as Google, Facebook, Microsoft, and Apple are some of the biggest, most complex, and powerful corporations that the world has seen. They own and control what Marx might describe as the means of production in this field. That is, the vast array of super powerful computers and the phalanx of the world's best and brightest mathematicians and engineers required to churn the algorithms necessary to create AI. The power disparity between these players and government regulators, who often struggle to secure sufficient resources to operate,

---

[135] *Id.* at 1056.

[136] Davidson S and Potts J, 'Social costs and the institutions of innovation policy' (2015) SSRN; Bennett Moses L, 'Agents of Change: How the Law 'Copes' with Technological Change' (2011) 20(4) *Griffith Law Review,* 763.; Benjamin M and Rai AK, 'Fixing Innovation Policy: a Structural Perspective' (2008) 77 *George Washington Law Review*, 1; Mandel GN, 'Regulating Emerging Technologies' (2009) 1 *Law, Innovation and Technology*, 75.

[137] Bennett Moses L, 'Agents of Change: How the Law 'Copes' with Technological Change' (2011) 20(4) *Griffith Law Review,* 763.

[138] Milne CP and Tait J, 'Evolution along the Government—Governance Continuum: FDA's Orphans Products and Fast Track Programs as Exemplars of "What Works" for Innovation and Regulation' (2009) 64(4) *Food & Drug Law Journal,* 733; Phillips PWB, *Governing transformative technological innovation: who's in charge?* (Edward Elgar Publishing, 2007); Potts J, 'The national origins of global innovation policy and the case for a World Innovation Organization' (2015) SSRN.

[139] McGarity TO, 'MTBE: A Precautionary Tale' (2004) 28(2) *Harvard Environmental Law Review*, 281; Goldacre B, *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients* (Faber and Faber Inc., 2013).

[140] Goldacre B, *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients* (Faber and Faber Inc., 2013).

highlights the difficulties that might be faced by a regulator in trying to regulate these companies.[141]

(g) Little established ethical guidance, normative agreement, or regulatory precedent

The ethical and social implications of introducing robots into mainstream society is a very weighty question that remains largely unanswered, at least in a comprehensive way, even as the consequences are already unfolding.[142] AI has already been deployed in society in a wide variety of fields, from medical diagnostics to criminal sentencing to social media, rendering the question even more urgent.[143] Lin et al. broke down their analysis of robot ethics into the categories of safety and errors; law and ethics; and social impact.[144] In other work they consider the possibility and desirability of programming ethics into AI systems, the use of military robots, human-robot relationships, such as the use of robots as sex partners, caregivers, and servants.[145] A regulatory regime for the design and diffusion of robots in society requires first that we reach some agreement on what we want this new world to look like.

### 3.2. Evaluating the risk: the case for regulation

Regulation is often implemented as a means to avoid or limit risks to human health or safety, or to the environment or against some moral hazard such as gene manipulation.[146] The real risks of AI are yet to be known and perhaps unknowable. This necessarily makes them difficult to evaluate. Assessing risk involves subjective social attitudes and will often depend on the culture and values of society. Just as the nature of the technology developed is diffuse, so too are the culture, values and attitudes where it is developed and deployed. There are also risks associated with the regulatory response. For example, any regulation put in place may lose connection to the thing being regulated if there is not sufficient buy in from stakeholders; this is an inherent difficulty with the regulation of complex technologies poorly understood by potential state based regulators. Risk is associated with the

[141] This idea is a work in progress and may form the basis of a further paper on power-relations in regulating AI.

[142] Brundage M, 'Limitations and risks of machine ethics' (2014) 26(3) *Journal of Experimental & Theoretical Artificial Intelligence* 355; Crnkovic G and Çürüklü B, 'Robots: ethical by design' (2012) 14(1) *Ethics and Information Technology* 61; Perri 6, 'Ethics, Regulation and the New Artificial Intelligence, Part I: Accountability and Power' (2001) 4(2) *Information, Communication and Society* 199; Koops BJ, 'The Concepts, Approaches, and Applications of Responsible Innovation: An Introduction' *Tilburg Law School Research Paper No.*

[143] Margelisch A, *A State of the Art Report on Legal Knowledge-Based Systems* (Swiss Life, 1999); Borenstein J and Pearson Y, 'Companion Robots and the Emotional Development of Children' (2013) 5(2) *Law, Innovation and Technology* 172; Angwin J, Larson J, Mattu S and Kirchner L, 'Machine Bias: There's software used across the country to product future criminals. And it's biased against blacks.' (2016) ProPublica; Dickens BM and Cook RJ, 'Legal and ethical issues in telemedicine and robotics' (2006) 94(1) *International Journal of Gynaecology & Obstetrics* 73.

[144] Lin P, Abney K and Bekey G, 'Robot ethics: Mapping the issues for a mechanized world' (2011) 175(5) *Artificial Intelligence* 942.

[145] Lin P, Abney K and Bekey G, *Robot ethics: the ethical and social implications of robots* (MIT Press, Cambridge, Mass., 2012).

[146] See Beyleveld and Brownsword 'Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning' (2012) 4(1) Law, Innovation and Technology 35 at 35.

undercurrent of fear of stifling innovation by regulating too hard too early. Regulators of AI must tread carefully if regulation is to be successful since its regulation is inherently linked to these and other risks.[147] The discussion of the risks associated with the development of AI then requires us to consider regulating that risk; all regulatory action has potential consequences, and this includes where the regulatory response is to do nothing.[148]

AI has its own risk profile such that, if assessed at its height, AI presents inherent systemic risk.[149] What could potentially go wrong with AI? The potential risks have multiple unpredictable sources that need not necessarily include a 'singularity' scale event.[150] Immediate systemic risk issues present with existing AI systems. These risks are attenuated with uncertainty and not restrained by sector, domain or geography.[151] The integrated nature of AI's deployment into complex, interdependent systems or networks amplifies potential for risk impact, particularly where it is deployed in a pervasive way.[152] The more complex and non-linear these networks are, the eventualities may proliferate rapidly throughout the network affecting multiple stakeholders.[153] Systemic risks are problematic for regulation.

---

147 See Bridget M Hutter, *Risk, Regulation and Management*, *in* RISK IN SOCIAL SCIENCE 202–227, 205 (2006).

148 ORTWIN RENN, RISK GOVERNANCE: COPING WITH UNCERTAINTY IN A COMPLEX WORLD 1 (2008).

149 Baum provides an example of a systemic risk as the 14th century black plague in Venice, which was managed by the Venetians without knowledge or aforethought of germ theory or micro-biology: Seth D. Baum, *Risk and resilience for unknown, unquantifiable, systemic, and unlikely/catastrophic threats*, 35 ENVIRON. SYST. DECIS. N. Y. 229–236 (2015).

150 For a full discussion of systemic risk generally see: Marjolein BA van Asselt & Ortwin Renn, *Risk Governance*, 14 J. RISK RES. 431–449, 436 (2011); Systemic risk has been studied in a technology context. See: Jessica Carlo, Kalle Lyytinen & Richard Boland, *Systemic Risk, Information Technology Artifacts, and High Reliability Organizations: A Case of Constructing a Radical Architecture*, *in* ICIS 2004 CONFERENCE PROCEEDINGS , 686 (2004), http://aisel.aisnet.org/icis2004/56 (last visited Mar 16, 2017); Numerous studies have been conducted into how the law should deal with unknown risks. See for example, Jaap Spier, *Uncertainties and the state of the art: a legal nightmare*, 14 J. RISK RES. 501–510 (2011); Paradoxically, AI may be able to assist with the management of systemic risk. See: Jerzy Balicki et al., *Methods of Artificial Intelligence for Prediction and Prevention Crisis Situations in Banking Systems*, *in* PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FS '14) (2014).

151 These characteristics of systemic risk are identified and discussed in van Asselt and Renn, *supra* note 150 at 436.

152 *Id.* at 436.; For a full discussion of the systemic risks of networked technology, see Tomas Hellström, *Systemic innovation and risk: technology assessment and the challenge of responsible innovation*, 25 TECHNOL. SOC. 369–384 (2003). Note Karppi and Crawford have considered the connected nature of human communication and financial system algorithms. The eventual coalescence of big data and AI will compound this interconnectness of social systems. See: Tero Karppi & Kate Crawford, *Social Media, Financial Algorithms and the Hack Crash*, 33 THEORY CULT. SOC. 73–92, 87 (2016).

153 van Asselt and Renn, *supra* note 150 at 436; Carlo, Lyytinen, and Boland, *supra* note 150 at 686; Note Baum disagrees that AI (or aliens) could be considered an systemic risk since if either risk were to eventuate and achieve world domination, humanity would have lost control of its system and be rendered incapable of managing it. Thus any attempts to make systems more resilient to AI or alien invasion is misguided. Baum's view of the systemic risks of AI are predicated on a vision of the systemic risk being the singularity or harbinger of doom. We argue that this dismisses the systemic risk narrower AI systems might present. Note that Baum suggests that since in his view AI is not a systemic threat, appropriate risk management is "not to increase resilience of affected systems but to reduce the probability of the systems being affected in the first place": Baum, *supra* note 149 at 234.

Systemic risks are not unknown to public regulators,[154] but the potential size of the connectedness of the network in AI is unprecedented. Since the risk is nebulous, systemic risk does not lend itself to command and control models of regulation. It has been suggested that systemic risk be managed via 'a cautious and flexible strategy that enables learning from restricted errors, new knowledge, and visible effects, so that adaption, reversal, or adjustment of regulatory measures is possible', and an ongoing concerted effort by government, business and society to ensure that 'early warning systems' for risk eventualities are in place.[155] Such an approach would lend itself to the formulation of agreed principles synthesising 'what seems wise to do, or at least what needs to be seriously considered, in organising structures and processes to govern risks'.[156] The mix and interplay of stakeholders will be important in the formulation of any such principles to regulate the systemic risk, since it is non-state stakeholders that are at an information advantage in understanding the underlying matrix of science and technology. The necessarily diverse mix of stakeholders and heterogenous interests may make unified agreement on principles difficult.[157] Determining the regulatory approach to be taken with managing systemic risks of AI ought to include a consideration of social norms, and ethics since the risk is not isolated to technical or scientific matters.[158]

Settling on such a set of principles will involve an element of trust in the science and technology. This trust could be engendered by creating a culture of iterative and cooperative development. Progress could be smoothed by a culture of fidelity and transparency from those with technical knowledge and scientific expertise in AI. This may be made exceedingly difficult by barriers to transparency including the blackbox character of algorithms.[159]

---

[154] Carlo et al consider for example management of risk hazards associated with nuclear facilities. See: Carlo, Lyytinen, and Boland, *supra* note 150 at 686; Numerous studies have been conducted considering the resilience of infrastructure in the face of systemic risk from a number of eventualities. See for example, JONATHON CLARKE ET AL., RESILIENCE EVALUATION AND SOTA SUMMARY REPORT: REALISING EUROPEAN RESILIENCE FOR CRITICAL INFRASTRUCTURE (2015); Sabrina Larkin et al., *Benchmarking agency and organizational practices in resilience decision making*, 35 ENVIRON. SYST. DECIS. N. Y. 185–195 (2015); Julie Dean Rosati, Katherine Flynn Touzinsky & W. Jeff Lillycrop, *Quantifying coastal system resilience for the US Army Corps of Engineers*, 35 ENVIRON. SYST. DECIS. N. Y. 196–208 (2015); Nicole R. Sikula et al., *Risk management is not enough: a conceptual model for resilience and adaptation-based vulnerability assessments*, 35 ENVIRON. SYST. DECIS. N. Y. 219–228 (2015); Baum, *supra* note 149; Seth D. Baum et al., *Resilience to global food supply catastrophes*, 35 ENVIRON. SYST. DECIS. N. Y. 301–313 (2015); Daniel Dimase et al., *Systems engineering framework for cyber physical security and resilience*, 35 ENVIRON. SYST. DECIS. N. Y. 291–300 (2015).

[155] van Asselt and Renn, *supra* note 150 at 439.

[156] van Asselt and Renn, *supra* note 150; Principles suggested include communication and inclusion, integration, and reflection. See: RENN, *supra* note 148; Hutter, *supra* note 147 at 214–215; Carlo, Lyytinen, and Boland, *supra* note 150 at 686.

[157] Note Carlo et al, have made similar observations in their research on high reliability organizations: Carlo, Lyytinen, and Boland, *supra* note 150 at 694.

[158] Carlo, Lyytinen, and Boland, *supra* note 150.

[159] Crawford as observed that the "algorithmic black box" is compounded by the fact that "algorithms do not always behave in predictable ways". See: Crawford, *supra* note 67 at 77; In an analysis of societal impacts of algorithms, Karppi and Crawford suggest that instead of seeking to find transparency in algorithms, a better approach would be the development of "theories that address and analyze the broader sweep of their operations and impact as well as their social, policital and institutional contexts". See: Karppi and Crawford, *supra* note 152 at 74.

Earlier research has acknowledged that the reliability and fidelity of organisations involved ought to be evaluated including the intended use to which the technology might be put.[160] Armed with a deeper understanding, stakeholders involved in informing the regulatory approach will be better placed to ask the right questions to assuage, or at least contextualise, their understanding of the risk. Risk analysis generally involves striking a balance, and the promise of AI makes taking some risk worthwhile. Iterative and cooperative involvement of all stakeholders including public regulators is key to avoiding ill-considered command and control regulatory action and its unintended consequences.[161] Many of these tenets are included in models of self-regulation. However, there are limits to self-regulation as an effective regulatory tool.

### 3.3. The limits of self-regulation

The challenges of regulating fast moving technology are so great that industry self-regulatory approaches are often presented as the most effective mechanism to manage risk. Industry bodies are already forming to respond to fears about the ongoing deployment of AI systems in ways that could be interpreted as staving off clumsy and heavy-handed public regulation. One of the most prominent efforts is the Partnership on AI between Google, DeepMind, Facebook, Microsoft Apple, Amazon, and IBM, together with the American Civil Liberties Union and the Association for the Advancement of Artificial Intelligence (AAAI). The Partnership on AI's purpose statement is to 'benefit people and society',[162] and is said to have been '[e]stablished to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society'.[163] It has developed a series of tenets for the development of AI that commit its members to ongoing engagement with stakeholders to protect the privacy, security and other human rights of individuals. In doing so, the Partnership is taking on the role of a self-regulatory association, and potentially warding off more enforceable state-imposed regulatory obligations.[164]

---

[160] Phil Macnaghten & Jason Chilvers, *Governing Risky Technologies*, *in* CRICTICAL RISK RESEARCH: POLITICS AND ETHICS 99–124, 102 (Matthew Kearnes, Francisco Klauser, & Stuart Lane eds., 2012); See also Carlo, Lyytinen, and Boland, *supra* note 150.

[161] Note Sunstein outlined a number of paradoxes that can be brought about by inappropriate regulation: Imposing stringent regulations may lead to the regulator's own administrators failing to act or refusing to enforce the regulations. Further, "stringent regulation of new risks can increase aggregate risk levels" (at 418). By way of example, Sunstein noted that stringent regulation of nuclear facilities had "perpetuated the risks produced by coal, a significantly more dangerous power source" (at 418). Sunstein argued that these paradoxes (among others) must be borne in mind when introducing regulation. See: Sunstein, *supra* note 130 at 413, 418, 441.

[162] Partnership on AI, TENETS PARTNERSHIP ON ARTIFICIAL INTELLIGENCE TO BENEFIT PEOPLE AND SOCIETY, https://www.partnershiponai.org/tenets/ (last visited Mar 13, 2017).

[163] *Id.*

[164] Self regulation relies on the strong and "credible threat" of state-based regulation. See Christopher Kevin Walker, *Neoliberalism and the reform of regulation policy in the Australian trucking sector: policy innovation or a repeat of known pitfalls?*, 37 POLICY STUD. 72–92, 72 (2016); It has been argued that self regulation fails, or at least is unreliable without the ever-present threat of state-based sanctions: Ian Ayres & John Braithwaite, *Tripartism: Regulatory Capture and Empowerment*, 16 LAW SOC. INQ. 435–496 (1991); See also: DAVID P MCCAFFREY & DAVID W HART, WALL STREET POLICES ITSELF: HOW SECURITIES FIRMS MANAGE THE LEGAL HAZARDS OF COMPETITIVE PRESSURES (1998); Andrew A. King & Michael J. Lenox, *Industry Self-Regulation Without Sanctions: The Chemical Industry's Responsible Care Program*, 43 ACAD. MANAGE. J. 698–716 (2000); Jodi L. Short & Michael W. Toffel, *Coerced Confessions: Self-Policing in the Shadow of the Regulator*, 24 J.

Proponents of AI have sought to counter the fears long-expressed by science fiction authors by highlighting the positive and benign applications of AI already in place today.[165] Developers suggest that technical contingency plans, like DeepMind's 'big red button' are in place in case AI gets out of hand. The implication is that up to this limit – the 'nuclear option' of shutting down rogue AI completely – the developers of AI are already effectively regulating its development through initiatives like the Partnership on AI and the principles set out by IEEE. In this regard the Partnership on AI has endorsed the United States Government's Report, *Preparing for the Future of Artificial Intelligence.*[166] It is in the interests of the industry participants such as the Partnership on AI not to disavow the government's position. It shows that the industry is very capable of self-regulating and that it is in lock-step with the government and its public regulators.

It may well be that self-regulation will be effective in mitigating the most important risks of the development and deployment of artificial intelligence systems. But there is also a risk that self-regulation may not be sufficient. Certainly, it will be important to avoid regulation that is ineffective or unduly stymies research and development. But we suggest that governments need to consider these concerns now in order to protect public interests that industry-led regulation is not well-suited to addressing.[167]

---

LAW ECON. ORGAN. 45–71 (2008); CHRISTINE PARKER, THE OPEN CORPORATION: EFFECTIVE SELF-REGULATION AND DEMOCRACY (2010); JOSEPH V REES, REFORMING THE WORKPLACE: A STUDY OF SELF-REGULATION IN OCCUPATIONAL SAFETY. (1988); Neil A. Gunningham, Dorothy Thornton & Robert A. Kagan, *Motivating management: Corporate compliance in environmental protection*, 27 LAW POLICY 289–316 (2005); JAY A SIGLER & JOSEPH E MURPHY, INTERACTIVE CORPORATE COMPLIANCE: AN ALTERNATIVE TO REGULATORY COMPULSION (1988); PARKER, *supra* note; Jay P. Shimshack & Michael B. Ward, *Regulator reputation, enforcement, and environmental compliance*, 50 J. ENVIRON. ECON. MANAG. 519–540 (2005); Jackson et al reject that public regulation and self regulation are diametrically opposed choices, and argue that their relationship is symbiotic: GREGORY JACKSON ET AL., REGULATING SELF-REGULATION? THE POLITICS AND EFFECTS OF MANDATORY CSR DISCLOSURE IN COMPARISON (2017), https://papers.ssrn.com/abstract=2925055 (last visited Mar 15, 2017); Gunningham, Thornton, and Kagan, *supra* note; The literature acknowledges that self motivation and reputation figure in the motivational mix. See: Roland Bénabou & Jean Tirole, *Incentives and prosocial behavior*, 96 AM. ECON. REV. 1652–1678 (2006); The deterrent effect of public regulation has somewhat of a paradoxical effect on self regulation and can dampen other intrinsic and external motivating factors such as earnest goodwill and concern for reputation. See: IAN AYRES & JOHN BRAITHWAITE, RESPONSIVE REGULATION (1994); FIONA HAINES, CORPORATE REGULATION: BEYOND "PUNISH OR PERSUADE" (1997); Jodi L. Short & Michael H Toffel, *Making Self Regulation more than merely symbolic: the critical role of the legal environment*, 55 ADM. SCI. Q. 361–396 (2010); ROBERT BALDWIN, MARTIN CAVE & MARTIN LODGE, UNDERSTANDING REGULATION: THEORY, STRATEGY, AND PRACTICE 261–262 (2011), http://qut.eblib.com.au/patron/FullRecord.aspx?p=829488 (last visited Mar 16, 2017); EUGENE BARDACH & ROBERT A KAGAN, GOING BY THE BOOK: THE PROBLEM OF REGULATORY UNREASONABLENESS (2009).

[165] See for example the positive story about DeepMind creating a 40% energy saving in Google's data centres on page one of this paper.

[166] Partnership on AI Expresses Support for White House Report on Artificial Intelligence, PARTNERSHIP ON ARTIFICIAL INTELLIGENCE TO BENEFIT PEOPLE AND SOCIETY (2016), https://www.partnershiponai.org/2016/10/partnership-ai-expresses-support-white-house-report-artificial-intelligence/ (last visited Mar 20, 2017); See also: EXECUTIVE OFFICE OF THE PRESIDENT NATIONAL SCIENCE AND TECHNOLOGY COUNCIL COMMITTEE ON TECHNOLOGY, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (2016); NATIONAL SCIENCE AND TECHNOLOGY COUNCIL & NETWORKING AND INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT SUBCOMMITTEE, THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN (2016); The Administration's Report on the Future of Artificial Intelligence, WHITEHOUSE.GOV (2016), https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence (last visited Mar 20, 2017).

[167] Black, *supra* note 13 at 115; Susan S. Silbey & Jodi L. Short, *Self-Regulation in the Regulatory Void: "Blue Moon"or "Bad Moon"?*, 649 ANN. AM. ACAD. POL. SOC. SCI. 22–34 (2013); Rob Baggott, *Regulatory Reform in*

## 4. The need for regulatory innovation

Some academics have proposed that a great deal of the difficulty with regulating AI can be avoided simply by adapting existing liability regimes. The common law has long adjusted to changes in technology iteratively, and to a large extent, this incremental approach helps to minimise the risks of incorrect decisions in regulatory policy.[168] So, for example, concerns about potential harm caused by autonomous cars may be adequately dealt with simply by a judicial process that adapts tort law principles to place liability for harm on the entity that is most effectively able to mitigate the risk – the 'least cost avoider'. Proponents of an iterative, 'light touch' approach favour responding to concrete problems as they arise, either through incremental adjustments to the common law or careful, limited and predominantly sui generis legislation if and as required.[169] The attractiveness of this approach is that it avoids the necessity of evaluating prospective risks – ensuring that regulation is targeted and limited to clear harms that courts and legislatures are able to understand.

We suggest that there is a broader role for the state in influencing the development of AI systems, but that doing so well will require substantial innovation in regulatory policy. The concerns of many commentators about the development and deployment of AI suggest that a 'social systems analysis' approach is required that understands the operation of AI in a broad social context and is able to inform its ongoing development.[170] Also in this vein, the recent Stanford Report recommends a 'vigorous and informed debate' to 'steer AI in ways that enrich our lives and our society'.[171] How regulators may actually be able to steer AI development, however, is a crucial and, as yet, unanswered question. In this Part, we consider how regulatory agencies may be able to adopt strategies to 'nudge'[172] the development of AI. That is, before heavy 'command-and-control' regulation is introduced, regulators may be able to influence those responsible for designing and deploying AI systems to do so in a way that furthers the public interest.

In order to start to answer this question, we provide a review of both the regulatory theory literature and the legal literature on the regulation of technology. As will be shown, these theories have clear limitations when asked to respond to the development of new technologies. Regulatory theory that has developed over the last two decades such as responsive regulation and really responsive regulation are normative and propose what good regulation should look like. They are also responsive to the regulatory framework that is in place and as such are best used to guide interactions between regulators and the regulated when regulatory systems are already in place. Further, they are limited in their ability to regulate new technologies that exhibit the kinds of characteristics set out in Part 3 above.

---

*Britain: The Changing Face of Self-Regulation*, 67 PUBLIC ADM. 435–454 (1989); BALDWIN, CAVE, AND LODGE, *supra* note 164 at 259–280.

[168] See, e.g., Bowman DM, 'The hare and the tortoise: an Australian perspective on regulating new technologies and their products and processes', in G E Marchant et al. (eds), Innovative Governance Models for Emerging Technologies (Edward Elgar Publishing, 2013); Kaal WA, 'Dynamic Regulation for Innovation' (2016) U of St. Thomas (Minnesota) Legal Studies Research Paper No. 16-22.

[169] Ibid.

[170] Crawford and Calo, *supra* note 4.

[171] Stanford Report at 49.

[172] RICHARD H THALER & CASS R SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS (2009).

Further, while much can be learned from regulation of other emerging technologies, the regulation of AI is sui generis and is still in its nascent stages and will take a more nuanced set of regulatory approaches.

Regulators face an extremely difficult challenge in responding to AI. As discussed above, regulators find it difficult to keep up with the pace of change; do not have all the information they require; must avoid over-regulation and uncertainty; need to coordinate actions with other regulatory agencies; must guard against capture and cannot rely too heavily on specialist knowledge obtained from industry; have to make do with enforcement mechanisms that are only partially effective; and need to make clear and justifiable policy decisions in a field that is highly contested.

### 4.1. Regulating with limited resources in a decentralised environment

For a long time, regulation was thought of mainly in terms of legal commands and sanctions. The state, in the classical model of regulation, is a powerful entity that can command obedience through a monopoly on the legitimate use of force.[173] Over the last three decades, regulatory scholars and regulatory agencies have been grappling with the 'decentring' of regulation: a recognition that regulation is not the exclusive work of states, and state power to command obedience is in practice often severely limited.[174] As Black contends, 'complexity, fragmentation of knowledge and of the exercise of power and control, autonomy, interactions and interdependencies, and the collapse of the public/private distinction are the central elements of the composite 'decentred understanding' of regulation'.[175] The hallmarks of 'decentred' regulation as proposed by Black are that it is 'hybrid (combining governmental and non-governmental actors), multi-faceted (using a number of different strategies simultaneously or sequentially), and indirect'.[176] Black argued that decentred regulation:

> …should be indirect, focusing on interactions between the system and its environment. It should be a process of co-ordinating, steering, influencing, and balancing interactions between actors/systems, to organise themselves, using such techniques as proceduralization, collibration, feedback loops, redundancy, and above all, countering variety with a variety.[177]

It is now widely recognised that there are far more techniques in the regulation toolbox than 'command and control' style rules backed by sanctions.[178] Ayres and Braithwaite's concept of 'responsive regulation' sets out a graduated pyramid of interventions by the state in policing behaviour, in order to encourage and direct an efficient mix of regulatory work by private and public entities.[179] The responsive element of responsive regulation is that 'escalating forms of government intervention will reinforce and help constitute less intrusive

[173] THOMAS HOBBES, LEVIATHAN (2006); JOHN AUSTIN, THE PROVINCE OF JURISPRUDENCE DETERMINED (2nd ed. 1861).
[174] Black, *supra* note 13.
[175] Julia Black, *Critical reflections on regulation*, 27 AUSTL J LEG PHIL 1, 8 (2002).
[176] Black, *supra* note 13 at 111.
[177] *Id.* at 111.
[178] Black, *supra* note 175 at 4.
[179] AYRES AND BRAITHWAITE, *supra* note 164.

and delegated forms of market regulation'.[180] That is, responsive regulation still requires government to assert a 'willingness to regulate more intrusively' and by so doing 'can channel marketplace transactions to less intrusive and less centralised forms of government intervention'.[181] Ayres and Braithwaite proposed a pyramid of enforcement measures by government with the most intrusive command and control regulation at the apex and less intrusive measures such as self-regulation at the base. The threat of using escalating forms of responsive regulation, they suggest, could be used 'without abdicating government's responsibility to correct market failure'.[182] The threat relies on the government's ability to inflict varying degrees of discretionary punishment or other forms of persuasion within the pyramidal structure if there is compliance failure — this is referred to as tit-for-tat approach.[183] The critical effect of responsive regulation was to highlight developments in alternative means of regulating other than command and control – and therefore avoiding some of the more problematic effects of blunt regulatory tools.

In the context of regulating AI, the emphasis of responsive regulation on a strong regulatory state that is ultimately still able to direct behaviour with effective sanctions no longer fully reflects practical realities. It may still be an effective means of governing more stately industries such as the production of wool in Australia but, we argue, it is not sufficiently flexible and nuanced to apply to a dynamic environment such as the development of AI. Further, it relies on the power of the state to impose the ultimate sanction at the apex of the pyramid; that is, the command and control regulation of an industry if it is required to. The notion of government as at the apex of power structures is no longer applicable when considering the power and diffuseness of companies such as Google, Microsoft, Apple, and Facebook.

When considering the regulatory role of the contemporary state in 2007, Hood and Margetts listed four resources of regulation that governments have in differing degrees in differing contexts: nodality, authority, treasure, and organisation. Nodality, they argued, referred to the government's position 'in the middle of an informational social network'.'[184] This gives government a central presence as a receptor and distributor of information. It gives government access to 'the whole picture'. It 'equips government with a strategic position from which to dispense information, and likewise enables government to draw in information for no other reason than that it is a centre or clearing-house'.[185] Authority refers to the official power the government has to 'demand, forbid, guarantee, [and] adjudicate'.[186] It gives the government legal authority to determine the legality.[187] Treasure 'denotes the possession of a stock of monies or fungible chattels'.[188] Organisation 'denotes the possession of a stock of people with whatever skills they may have (soldiers, workers, bureaucrats), land, buildings,

[180] *Id.* at 4.
[181] *Id.* at 4.
[182] *Id.* at 5.
[183] See generally *Id.* at 38–41.
[184] CHRISTOPHER C HOOD & HELEN Z MARGETTS, THE TOOLS OF GOVERNMENT IN THE DIGITAL AGE 5 (2007).
[185] *Id.* at 6.
[186] HOOD AND MARGETTS, *supra* note 184.
[187] *Id.* at 6.
[188] HOOD AND MARGETTS, *supra* note 184.

materials, computers and equipment'.[189] Hood and Margetts argued that viewing government's role as through the interaction between these simplifies analysis of the role of government in regulation.[190] However, as noted above, a lot has changed in the last 10 years and the government's nodality, authority, treasure and organisation is depleted or limited and does not match that of the major technology companies when considering regulation of AI.

Part of the challenge of effectively regulating AI is to identify opportunities for regulatory agencies to influence other actors when these four resources are limited. Baldwin and Black point out that existing theories of regulation are limited in this regard, as they do not:

> say a great deal about how a regulator should deal with resource constraints, conflicting institutional pressures, unclear objectives, changes in the regulatory environment, or indeed how particular enforcement strategies might impact on other aspects of regulatory activity, including information gathering, and how regulators can or should assess the effectiveness of their particular strategies when any of these circumstances obtain.[191]

In these situations, there appears to be no choice other than decentred regulation and, where political influence and power exists in the industries that are subject of proposed regulation, self-regulation appears to be the default position.

Some academics have proposed a system of governance that may be described as peer governance. For example, Jessop defines governance as 'the reflexive self-organisation of independent actors involved in complex relations of reciprocal interdependence, with such self-organisation being based on continuing dialogue and resource-sharing to develop mutually beneficial joint projects and to manage the contradictions and dilemmas inevitably involved in such situations'.[192] Jessop emphasises the role of self-organisation of stakeholders to include:

> (1) the more or less spontaneous, bottom-up development by networks of rules, values, norms and principles that they then acknowledge and follow [and]; (2) increased deliberation and participation by civil society groups through stakeholder democracy, putting external pressure on the state managers and/or other elites involved in governance.[193]

While this is not the focus of our proposals in this paper, it may reflect what is occurring between the industry participants in the absence of clear regulatory approach from government.[194]

---

[189] *Id.*

[190] *Id.* at 12.

[191] Robert Baldwin & Julia Black, *Really Responsive Regulation*, 71 MOD. LAW REV. 59–94, 61 (2008).

[192] Bob Jessop, *Governance and meta-governance: On Reflexivity, requisite variety and requisite irony*, *in* GOVERNANCE AS SOCIAL AND POLITICAL COMMUNICATION 101 (Henrik P Bang ed., 2003).

[193] Bob Jessop, *State Theory*, *in* HANDBOOK ON THEORIES OF GOVERNANCE 71–85 (Christopher Ansell & Jacob Torfing eds., 2016). At 82

[194] See efforts at self-regulation set out in Part 3.3 above.

There is no shortage of advice for regulators about how they might address the challenges of regulating with limited resources – either in general, or specifically in the context of new technologies. We set out below the range of regulatory theories that have predominated over the last two decades. As we suggest, many of these theories are either inappropriate or would be ineffective when regulating AI.

**Table 1 – Theories of regulation**

| Theory | Guiding principles |
|---|---|
| **'Responsive Regulation':[195] a 'tit-for-tat' approach to enforce compliance by persuasion and education before escalating up a 'pyramid' of more punitive sanctions.** | Braithwaite, 2011:<br>1. Think in context<br>2. Listen actively (build commitment with stakeholders)<br>3. Engage with fairness<br>4. Praise those who show commitment<br>5. Signal a preference for support and education<br>6. Signal a range of escalating sanctions that may be used if necessary<br>7. Engage a wider network of partners as regulatory responses increase in severity<br>8. Elicit active responsibility from stakeholders where possible<br>9. Evaluate regulations and improve practices. [196] |
| **'Smart regulation'** | • Prefer a mix of regulatory instruments while avoiding 'smorgasboardism';<br>• Prefer less interventionist measures<br>• Escalate up a pyramid of sanctions when required (responsive regulation)<br>• Empower third parties to act as surrogate regulators<br>• Maximise opportunities for win-win outcomes by encouraging businesses to go 'beyond compliance'[197] |
| **'Risk-based regulation'** | Hampton Review:[198]<br>• Regulators, and the regulatory system as a whole, should use comprehensive risk assessment to concentrate resources on the areas that need them most;<br>• Regulators should be accountable for the efficiency and effectiveness of their activities, while remaining independent in the decisions they take;<br>• All regulations should be written so that they are easily understood, easily implemented, and easily enforced, and all interested parties should be consulted when they are being drafted; |

[195] AYRES AND BRAITHWAITE, *supra* note 164.

[196] John Braithwaite, *The Essence of Responsive Regulation*, 44 UBC LAW REV. 475–520 (2011).

[197] Neil Gunningham & Darren Sinclair, *Designing smart regulation*, *in* ECONOMIC ASPECTS OF ENVIRONMENTAL COMPLIANCE ASSURANCE. OECD GLOBAL FORUM ON SUSTAINABLE DEVELOPMENT (1999), http://www.oecd.org/env/outreach/33947759.pdf (last visited Mar 15, 2017).

[198] PHILIP HAMPTON, HAMPTON REVIEW ON REGULATORY INSPECTIONS AND ENFORCEMENT 7 (2005), http://webarchive.nationalarchives.gov.uk/content/20130129110402/http://www.hm-treasury.gov.uk/bud_bud05_hampton.htm (last visited Mar 15, 2017).

| | |
|---|---|
| | • No inspection should take place without a reason;<br>• Businesses should not have to give unnecessary information, nor give the same piece of information twice;<br>• The few businesses that persistently break regulations should be identified quickly, and face proportionate and meaningful sanctions;<br>• Regulators should provide authoritative, accessible advice easily and cheaply;<br>• When new policies are being developed, explicit consideration should be given to how they can be enforced using existing systems and data to minimise the administrative burden imposed;<br>• Regulators should be of the right size and scope, and no new regulator should be created where an existing one can do the work; and<br>• Regulators should recognise that a key element of their activity will be to allow, or even encourage, economic progress and only to intervene when there is a clear case for protection. |
| **'Regulatory craft' (focusing on problem solving)** | 1. Nominate potential problems for attention<br>2. Define the problem precisely<br>3. Determine how to measure impact<br>4. Develop solutions or interventions<br>5. Implement the plan with periodic monitoring, review, and adjustment<br>6. Close project, allowing for long-term monitoring and maintenance.[199] |
| **'Really Responsive regulation'** | Regulators should be responsive to:<br>• firms' compliance responses (Responsive regulation); but also<br>• the 'attitudinal settings' (operating and cognitive framework of the target of regulation)<br>• the institutional environment<br>• the 'logics of different regulatory strategies and tools'<br>• the regulatory regime's own performance and effects<br>• change in priorities, circumstances, and objectives.[200] |
| **'Really Responsive Risk-** | In applying risk-based regulation, regulators should: |

[199] MALCOLM K. SPARROW, THE REGULATORY CRAFT: CONTROLLING RISKS, SOLVING PROBLEMS, AND MANAGING COMPLIANCE 142 (2011).

[200] Baldwin and Black, *supra* note 191.

| | |
|---|---|
| **based regulation'** | • take attitudinal matters on board<br>• identify how attitudes vary across regulatory tasks;<br>• 'be clear about the degree to which any particular regulatory task can and should be guided by a risk-scoring system'[201]<br><br>Risk-based regulation must focus on:<br><br>• '*detecting* undesirable or non-compliant behaviour,<br>• *responding* to that behaviour by developing tools and strategies,<br>• *enforcing* those tools and strategies on the ground,<br>• *assessing* their success or failure, and *modifying* them accordingly'.[202] |

Meanwhile, others have offered different and sometimes more concrete suggestions for how regulatory agencies can deal with the particular difficulties of regulating fast moving technological change.

**Table 2 – Applications of strategies**

| Theory | Guiding principles |
|---|---|
| **'Adaptive policymaking'** | Regulation should be:<br><br>• Cautious<br>• Macroscopic<br>• Incremental<br>• Experimental<br>• Contextual<br>• Flexible<br>• Provisional<br>• Accountable<br>• Sustainable[203] |
| **One Hundred Year Study on AI** | Government should:<br><br>• Accrue greater technical expertise in AI<br>• Remove impediments to research on the social impacts of AI<br>• Increase public and private funding for research on the social impacts of AI<br>• Resist pressure for 'more' and 'tougher' regulation |

[201] Julia Black & Robert Baldwin, *Really Responsive Risk-Based Regulation*, 32 LAW POLICY 181–213, 193 (2010).
[202] *Id.* at 187. (emphasis in original).
[203] Whitt, *supra* note 17.

| | |
|---|---|
| | that stifles innovation or forces innovators to leave the jurisdiction<br>• Encourage a 'virtuous cycle' of accountability, transparency, and professionalization among AI developers<br>• Continually re-evaluate policies in the context of research on social impacts[204] |
| **Whitehouse report – *Preparing for the Future of Artificial Intelligence*** | Regulatory agencies should:<br>• Recruit and develop technical expertise in AI<br>• Develop a workforce with 'more diverse perspectives on the current state of technology'<br>• Use risk-assessment to identify regulatory needs<br>• Avoid increasing compliance costs or slowing development or adoption of beneficial innovations where possible<br>• Avoid premature regulation that could stifle innovation and growth[205] |
| **Experimental innovation policy (OECD report, 'Making Innovation Work')** | The quality and efficiency of public expenditure on regulation targeted at innovation can be improved by an experimental approach to policy-making. Regulators should accordingly:<br>• Embed diagnostic monitoring and evaluation into regulatory programmes at the outset<br>• Collaborate closely with private firms and non-governmental actors<br>• Share and compare results of policy experimentation with other jurisdictions[206] |

[204] STONE ET AL., *supra* note 19.

[205] NATIONAL SCIENCE AND TECHNOLOGY COUNCIL, COMMITTEE ON TECHNOLOGY, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE* (2016), https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/nstc/preparing_for_the_future_of_ai.pdf. *The authors note that this link to the report is no longer active. It is interesting to note that a search for the term "artificial intelligence" on the whitehouse.gov website returns the following response: "Sorry no results found for 'artificial intelligence'. Try entering fewer or broader query terms"; Note the report appears to have been archived with documents from the previous administration. See: EXECUTIVE OFFICE OF THE PRESIDENT NATIONAL SCIENCE AND TECHNOLOGY COUNCIL COMMITTEE ON TECHNOLOGY, *supra* note 166.

[206] MARK ANDREW DUTZ, MAKING INNOVATION POLICY WORK: LEARNING FROM EXPERIMENTATION (2014), http://gateway.library.qut.edu.au/login?url=http://dx.doi.org/10.1787/9789264185739-en (last visited Mar 16, 2017).

The theories outlined in the Table 1 provide a normative guide on what regulation should involve. They presuppose a regulatory environment already being in place. Table 2 lists strategies that might be put in place to deal with problems as they arise. They are more practically applicable rather than theoretical but there appears to be no end to suggestions about what might work as a regulatory approach to new technologies – including AI. Many scholars have suggested specific regulatory tools that may be useful in regulating AI. These include, for example:

- enhancing flexibility through 'temporary regulation' by using 'experimental legislation'[207] and sunset clauses to 'define adaptable goals and enable the adjustment of laws and regulations according to the evolution of the circumstances'.[208]
- Creating 'regulatory sandboxes' to allow firms to roll out and test new ideas 'without being forced to comply with the applicable set of rules and regulations'.[209]
- developing 'anticipatory rulemaking'[210] techniques that leverage feedback processes to enable 'rulemakers to adapt to regulatory contingencies if and when they arise because a feedback effect provides relevant, timely, decentralized, and institution-specific information ex-ante.'[211]
- Making increased use of data analysis to identify what, when, and how to regulate;[212]
- Utilising the iterative development of the common law to adapt rules to new technological contexts where possible, and developing new specialist regulatory agencies where they are particularly needed;[213]
- Using 'legal foresighting' to identify and explore possible future legal developments, in order to discover shared values, develop shared lexicons, forge a common vision of the future, and take steps to realise that vision;[214]
- Creating new multi-stakeholder fora to help overcome information and uncertainty issues that stifle innovation or inhibit effective regulation.[215]

While there is no shortage of advice for regulators, it is hard to actually distil a clear set of concrete recommendations from the wide and varied literature. Ultimately, one of the key

[207] (Recommending engaging in policy and regulatory experiments to compare different regulatory regimes and embracing "contingency, flexibility, and an openness to the new") ERIK VERMEULEN, MARK FENWICK & WULF A. KAAL, REGULATION TOMORROW: WHAT HAPPENS WHEN TECHNOLOGY IS FASTER THAN THE LAW? (2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2834531 (last visited Mar 15, 2017).

[208] RANCHORDÁS, *supra* note 14 at 212; see also Roberta Romano, *Regulating in the Dark*, *in* REGULATORY BREAKDOWN: THE CRISIS OF CONFIDENCE IN US REGULATION 86 (Cary Coglianese ed., 2012), https://books.google.com/books?hl=en&lr=&id=VH6u3fgTZUEC&oi=fnd&pg=PA86&ots=BGHYiOSFBF&sig=krfhQUq5cZJPX0M3lVHNhUE0gdw (last visited Mar 16, 2017) (discussing regulation of financial markets).

[209] VERMEULEN, FENWICK, AND KAAL, *supra* note 207.

[210] Kaal, *supra* note 14.

[211] WULF A. KAAL & ERIK P. M. VERMEULEN, HOW TO REGULATE DISRUPTIVE INNOVATION - FROM FACTS TO DATA (2016), https://papers.ssrn.com/abstract=2808044 (last visited Mar 16, 2017).

[212] *Id.*; Gerard Roe & Jason Potts, *Detecting new industry emergence using government data: a new analytic approach to regional innovation policy*, 18 INNOVATION 373–388 (2016); KAAL AND VERMEULEN, *supra* note 211.

[213] Scherer, *supra* note 16.

[214] Graeme Laurie, Shawn HE Harmon & Fabiana Arzuaga, *Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty*, 4 LAW INNOV. TECHNOL. 1–33 (2012).

[215] Gregory N Mandel, *Regulating Emerging Technologies*, 1 LAW INNOV. TECHNOL. 75–92 (2009).

problems is that while there are common regulatory challenges across different areas of innovation and technology policy, there are also highly context-specific challenges.[216] Ensuring regulatory approaches are closely connected with their context requires individual responses to different technologies in different locations at different times. As Brownsword points out, this means that inevitably, 'the details of the regulatory regime will always reflect a tension between the need for flexibility (if regulation is to move with the technology) and the demand for predictability and consistency (if regulatees are to know where they stand)'.[217] Brownsword concludes that 'while we should try to develop stock (tried and trusted) responses to challenges that we know to be generic, simple transplantation of a particular regulatory response from one technology to another is not always appropriate'.[218]

This spectrum of regulatory approaches from command and control to self-regulation or peer regulation presents a quandary for those trying to regulate in this area. There is no quick fix that can be implemented to resolve the problems we have outlined. In the next Part, we consider some innovative approaches to developing a governance approach to the development of AI that includes considering a number of tools from within self-regulation, and risk regulation theories. We conclude that, while these theories may eventually influence the regulation of AI, there is a moment in time now where all of the stakeholders may be able to influence the development and regulation of AI through cooperation and collaboration in the nascent stages of development. In this way, all stakeholders can have a role and a stake in the way that regulation develops. This may take the form of overt self-imposed industry codes of practice or conduct from the participants,[219] and involve less intrusive and direct guidance from public regulators – what might be termed a nudge. The alternative to this cooperative approach is command and control regulation by the State that could stifle innovation and hobble development of this promising technology.

## 5. Innovation in regulation

Risk based frameworks usually entail the following sequence: firstly the regulator sets the level and type of risks it will tolerate; secondly the regulator conducts some form of risk assessment and assesses the likelihood of the risk eventuating; thirdly, regulators will evaluate the risk and rank the regulated entities on their level of risk – high, medium or low and fourthly, will allocate resources according to the level of risk that they have assessed.[220] Van Asselt and Ren emphasise the need for communication and inclusion when assessing risk. They argue that 'various actors are included, [and] play a key role in framing the risk'. This inclusion includes 'roundtables, open forums, negotiated rule-making exercises, mediation, or mixed advisory committees, including scientists and stakeholders.'[221] 'It is important to know what the various actors label as risk problems. In that view, inclusion is a means to an end: integration of all relevant knowledge and inclusion of all relevant concerns'.[222] The participants, they argue, should include 'a range of actors which have complementary roles and diverging interests'.[223] Hutter concludes that to achieve regulatory

[216] Roger Brownsword & Karen Yeung, *Tools, Targets and Thematics*, *in* REGULATING TECHNOLOGIES 3–22, 6 (Roger Brownsword & Karen Yeung eds., 2008).

[217] Brownsword, *supra* note 12 at 27.

[218] *Id.* at 32.; Brownsword and Yeung, *supra* note 216 at 6.

[219] See the contributions from industry participants outlined in Part 2 above.

[220] Black and Baldwin, *supra* note 201 at 184–185.

[221] van Asselt and Renn, *supra* note 150 at 440.

[222] *Id.* at 441.

[223] *Id.* at 441.

excellence, 'regulators must have access to accurate information so that they have a clear idea of the risks they are regulating'.[224]

A similar requirement for fuller information underpins behavioural insights into how to modify group behaviours. Much has been made of nudge theory in recent years.[225] This has tended to focus on nudging individual behaviours. However, there has been some recent work on how behavioural economics approaches might influence a broader spectrum of decision-makers. For example Weber argues in relation to environmental policy-making that 'decisions could be reframed in ways that might affect choices is by changing the focus of such decisions from individuals to groups'.[226] She argues that 'cultures that emphasize the importance of affiliation and social goals over autonomy and individual goals have been shown to influence the way in which decisions under risk and uncertainty get made'.[227] Weber states that 'policy interventions should be designed to prime social roles that will induce people to use rule-based processes.' She argues that 'the goal of environmental policy is to change behaviour of companies, governing boards and committees, and members of the general public in the direction of more sustainable, long-term, and socially and environmentally responsible actions.'[228] Weber concludes that 'conventional policy interventions are not using the full range of goals that motivate behaviour and changes in behaviour … [and] do not utilize the full range of processes that people use to decide on a course of action.'[229]

Taking this a step further, Tyler argues that 'the behaviour of the people within groups shapes the viability of those groups. The suggestion that the thoughts, feelings, and behaviour of the people within groups are linked with group viability and functioning is widely supported by studies within law, … and public policy and government.'[230] Examples from some Australian regulatory agencies in this regard are apposite.

The Australian Taxation Office formulated behavioural nudges to deter company directors from engaging in illegal behaviour within the ATO's regulatory mandate such as payment of taxation and employee superannuation entitlements.[231] These nudges are part of an early

[224] Hutter, *supra* note 111 at 104.

[225] THALER AND SUNSTEIN, *supra* note 172.

[226] Elke U Weber, *Doing the right thing willingly: Using the insights of behavioral decision research for better environmental decisions*, *in* THE BEHAVIORAL FOUNDATIONS OF PUBLIC POLICY 380–397, 388 (Eldar Shafir ed., 2013).

[227] *Id.* at 388.

[228] *Id.* at 391.

[229] *Id.* at 391.

[230] Tom Tyler, *The psychology of cooperation: Implications for public policy*, *in* THE BEHAVIORAL FOUNDATIONS OF PUBLIC POLICY 77–90, 84 (Eldar Shafir ed., 2013) (citations removed).

[231] These nudges are being used to specifically target illegal phoenix activity. See Australian Taxation Office, *Our approach to phoenixing*, 28 AUST. RESTRUCT. INSOLV. TURNAROUND ASSOC. J. 43 (2016); Illegal phoenix activity is the deliberate and often cyclical misuse of a company to evade creditors. The business or assets may be stripped from the existing company and transferred to a new or related company for this illegal purpose. , The existing company is then abandoned or wound up as an assetless shell, with the inevitable and intended result of leaving creditors such as the ATO unpaid. The new or related company continues the business, now debt free. Fundamental doctrines of corporate law including limited liability and the doctrine of the separate legal entity shield the new company from the creditors of the previous company. For a full discussion see Anne Matthew, *The conundrum of phoenix activity: Is further reform necessary?*, 23 INSOLV. LAW J. 116–135, 117 (2015).

intervention system.[232] Potential offenders are identified using a risk model.[233] A dedicated risk and intelligence team then determines whether early intervention is appropriate, and if so what action should be taken. Early intervention strategies include a field audit or nudging via communication with the potential offender encouraging compliance.[234]

ASIC is Australia's corporate watchdog, regulating markets, corporations and financial services.[235] ASIC is systematically deploying insights from behavioural economics research across its regulatory business.[236] In a move toward 'smarter regulation', ASIC's development of behavioural strategies focuses on key drivers of behaviour: culture, deterrence and incentives.[237] In its policy development, ASIC is seeking to better understand regulatory problems and identify poor architecture or methods that may amplify biases or lead to poor market outcomes.[238] This includes reconsideration of timing, the messenger and the context of the regulatory response.[239]

---

[232] AUSTRALIAN TAXATION OFFICE, SUBMISSION TO THE SENATE ECONOMICS REFERENCES COMMITTEE - INSOLVENCY IN THE CONSTRUCTION INDUSTRY (2015), http://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Economics/Insolvency_construction/Submissions (last visited Mar 9, 2017).

[233] Australian Taxation Office, "Submission to the Senate Economics References Committee", *Id.* at 20.; Australian Taxation Office, "Our Approach to Phoenixing", , *supra* note 232.

[234] Australian Taxation Office, "Submission to the Senate Economic References Committee", , *supra* note 232 at 23; Australian Taxation Office, "Our Approach to Phoenixing", , *supra* note 231; A full account of the ATO's approach, including templates of nudge correspondence is contained in a major report prepared as part of an Australian Research Discovery Grant: HELEN ANDERSON ET AL., QUANTIFYING PHOENIX ACTIVITY: INCIDENCE, COST, ENFORCEMENT 1–160 53–54, 110–113 (2015), http://law.unimelb.edu.au/centres/cclsr/research/major-research-projects/regulating-fraudulent-phoenix-activity (last visited Mar 10, 2017).

[235] ASIC is a national government regulator established under the AUSTRALIAN SECURITIES AND INVESTMENTS COMMISSION ACT 2001 (CTH), .

[236] This is being accomplished with assistance from a dedicated behavioural economics team within ASIC's Strategic Intelligence Unit. The Strategic Intelligence Unit plays an important role in initiatives targeting the future-proofing of ASIC's capabilities and positioning ASIC to keep pace with rapid change in its regulatory environment. The Behavioural Economics team is responsible for developing ASIC's understanding of when behavioural economics is the most appropriate regulatory tool in the kit to address regulatory issues. See Peter Kell, ASIC AND BEHAVIOURAL ECONOMICS: REGULATING FOR REAL PEOPLE 4 (2016), http://asic.gov.au/about-asic/media-centre/speeches/asic-and-behavioural-economics-regulating-for-real-people/ (last visited Mar 8, 2017); Australian Government Treasury, FIT FOR THE FUTURE: A CAPABILITY REVIEW OF THE AUSTRALIAN SECURITIES AND INVESTMENTS COMMISSION TREASURY, http://www.treasury.gov.au/PublicationsAndMedia/Publications/2016/ASIC-capability-review/Fit-for-the-Future/Foreword/Executive-summary (last visited Mar 9, 2017).

[237] Greg Medcraft, Chairman, Australian Securities and Investments Commission, CREATING GROWTH THROUGH OUR MARKETS: USING THE RIGHT NUDGE (2015), http://asic.gov.au/about-asic/media-centre/speeches/creating-growth-through-our-markets-using-the-right-nudge/ (last visited Mar 9, 2017); For a further discussion regarding the regulation of corporate culture in Australia see John H C Colvin & James Argent, *Corporate and personal liability for "culture" in corporations*, 34 CORP. SECUR. LAW J. 30–47 (2016); Robert Baxt, *What is the fuss about culture all about?*, 90 AUST. LAW J. 621–624 (2016).

[238] AUSTRALIAN SECURITIES AND INVESTMENTS COMMISSION, ASIC'S STRATEGIC OUTLOOK 13 (2014); Kell, *supra* note 236 at 4; Greg Medcraft, OPENING STATEMENT (2015), http://asic.gov.au/about-asic/media-centre/speeches/parliamentary-joint-committee-opening-statement-20-march-2015/ (last visited Mar 7, 2017).

[239] Kell, *supra* note 236 at 4.

ASIC has suggested that other regulators seeking to engage with behavioural economics in policy development should be prepared to work iteratively overtime, by testing, trialling and adapting interventions.[240] A smarter approach to regulation informed by behavioural economics will benefit from creating networks to collaborate widely with a range of stakeholders and experts including other regulators and end-users.[241] ASIC recommends an evidence based approach to behavioural interventions, and suggests that a considered approach necessarily involves fighting the urge to rush in with regulatory intervention before fulling assessing the problem or area the targeted intervention is intended to address. [242]

### 5.1. The current state of play

The approach of the United States Government in attempting to shape behaviours in the development of AI is in its infancy. However, as an early indicator, the government has shown that it is prepared to consult with groups of stakeholders and in 2016 its Office of Science and Technology Policy conducted a series of public workshops held at Washington University, Stanford University, Carnegie Mellon University and New York University. The OSTP also participated in various industry conferences and sought public comment in the form of a Request for Information.[243] This might be seen in a number of different ways. Firstly, the government is seen to be consultative and is attempting to engage with stakeholders in the area. Abbott noted that 'modern regulatory policy, including risk regulation policy, views public communication, inpute and participation as essential'. He cites the 2012 OECD recommendations on regulatory policy that 'call for "open government", including transparency and communication, stakeholder engagement throughout the regulatory process and open and balanced public consultations'.[244] Secondly, this could also be seen as an information gathering exercise – something that is necessary in the risk regulation literature as well as in behavioural economics theories. Thirdly, government could be seen to be signposting its intentions to regulate if necessary. By making itself heard, it sends a signal to stakeholders that it is aware of the industry and should not be forgotten.

The United States government has also published the *National Artificial Intelligence Research and Development Strategic Plan*. The Strategic Plan states:

> AI presents some risks in several areas, from jobs and the economy to safety, ethical and legal questions. Thus, as AI science and technology develops, the federal government must also invest in research to better understand what the implications are for AI for all of these rooms, and to address these implications by developing AI systems that align with ethical, legal and societal goals.[245]

This may be seen as the government seeking to influence decision makers in the AI industry to shape behaviours. Whether this is a nudge or is simply a mechanism that is being implemented early on in the regulation cycle is not clear. However, its emphasis on beneficial development clearly states the government's intentions and focus.

---

[240] *Id.* at 7.

[241] Kell, *supra* note 236.

[242] *Id.* at 4.

[243] EXECUTIVE OFFICE OF THE PRESIDENT NATIONAL SCIENCE AND TECHNOLOGY COUNCIL COMMITTEE ON TECHNOLOGY, *supra* note 166 at 12.

[244] Abbott, *supra* note 60 at 10.

[245] NATIONAL SCIENCE AND TECHNOLOGY COUNCIL AND NETWORKING AND INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT SUBCOMMITTEE, *supra* note 166. Again, this report no longer appears in searches on the whitehouse.gov website.

We Robot Conference at Yale University March 2017. Please do not cite without the authors' consent.

In a way, Asimov's three 'laws' of robotics have acted as a nudge, or have been in the consciousness of roboticists for over half a century. Similarly, the golem stories have nudged people's consciences as they toy with developing potentially life threatening inventions.

## 6. Conclusion

On 21 May 1946, as scientists were still experimenting with the new power of nuclear energy, Louis Slotin, a Canadian physicist who had worked on the Manhattan project to develop nuclear weapons during World War II, was preparing to conduct an experiment in a lab in the New Mexico desert. Slotin was slowly lowering a hemispherical beryllium tamper over a piece of plutonium to excite the neutrons that were emitting from the plutonium core. This process would create a small nuclear reaction so that the scientists could measure the results. The process was aptly referred to as 'tickling the dragon's tail'. On 21 May, Slotin slipped and dropped the beryllium tamper directly onto the core causing a momentary but powerful reaction that irradiated the whole room. Slotin bore the brunt of the reaction. He died a painful death nine days later from radiation poisoning.[246]

Seventy years later, scientists, engineers and technicians are experimenting with a new scientific development with potentially destructive capabilities. If we are to heed the allegory in the golem stories or the metaphor of the dragon's tail, we must come to the conclusion that any such danger, no matter its potential, should be carefully handled. We do not suggest a draconian, command and control type of regulation, and do not even think it would work. However, we do suggest a new and more nuanced, responsive, and adaptive regulation developed to foster innovation and minimise the risks of AI. This approach, as with the approach in relation to the treatment of nuclear weapons, needs a global solution.

Since 2007, the face of technology, the institutions involved, and therefore the AI regulatory space has changed dramatically. The last decade has seen the rise of some of the biggest technology companies including Microsoft, Apple, Facebook and Google as major leaders in AI. It is arguable that in terms of new technology development including AI, these companies hold the lion's share of regulatory resources.[247] Public regulators, by contrast, appear to be increasingly in the difficult position of needing to find mechanisms to regulate technology they have only limited capabilities to understand by influencing firms that are very well resourced and connected and can exercise substantial choice about the jurisdictions in which they operate.

There are encouraging signs from recent publications – certainly the emphasis on more research to pay attention to social impacts of AI from both the United States government and from private coalitions is encouraging. Still, largely, the rhetoric of avoiding over-regulation is worrying – even the biggest government regulators are hesitant and probably will not be particularly well equipped to deal with this any time soon. For smaller regulators – including

[246] See Alex Wellerstein, 'The Demon Core and the Strange Death of Louis Slotin', 21 May 2016, *The New Yorker* retrieved from http://www.newyorker.com/tech/elements/demon-core-the-strange-death-of-louis-slotin.

[247] HOOD AND MARGETTS, *supra* note 184 (discussing resources of nodality, authority, treasure, and organisation).

those outside of the US, there is almost no chance of successfully intervening in current technological development.

There are benefits to self-regulation particularly where public regulators lack the requisite knowledge to under the problem that needs regulating. Self-regulation has in its favour that it involves iterative and cooperative development of standards with input from various stakeholders at the coalface of the problem. The downside to self-regulation is that it works best where there is some imminent threat of state based penalty for non-compliance. As discussed, governments are at a disadvantage, probably for the first time in history at this scale, against the major corporate stakeholders in AI.

The United States government is perhaps best able to shape the development of AI. It has recently set about the task of informing itself about AI and has set an R&D policy that seeks to influence beneficial development of AI. Recent studies in behavioural policy making suggest that the attitudinal settings of people within groups shape the development of the group. The government has set a positive benchmark that seeks to sway participants in the field. Whether this can be called nudging or not is moot, but the intention is clear.

Because regulators do not have the expertise, if we are ever to ensure AI is developed in a way that is beneficial for humanity, developers must acknowledge both their social obligation to share information (be transparent and accountable) with others, and critical importance of collaborations with thinkers from other disciplines. This is where we can go back to DeepMind – this is such a great example of developers building accountability into the system – we need to encourage this. This cannot be done only by regulators, but must be multidisciplinary and multi-stakeholder: often, developers themselves don't know the right questions to ask. We need to empower civil society and researchers to raise new questions. We also need to empower research and black-box testing[248] for the times when we know we are not going to get straight answers from developers for a variety of commercial and other reasons. We can and probably should also get better at regulating in the absence of perfect information. Risk-based approaches can help regulators identify where to spend their energy.

[248] Christian Sandvig et al., *An Algorithm Audit*, *in* DATA AND DISCRIMINATION: SELECTED ESSAYS 6–10 (Seeta Peña Gangadharan, Virginia Eubanks, & Solon Barocas eds., 2014), https://www.newamerica.org/oti/policy-papers/data-and-discrimination/; Gillespie, *supra* note 72.