Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence

Michael Guihot,[†] Anne F. Matthew[‡] and Nicolas P. Suzor[*]

There is a pervading sense of unease that artificially intelligent machines will soon radically alter our lives in ways that are still unknown. Advances in AI technology are developing at an extremely rapid rate as computational power continues to grow exponentially. Even if existential concerns about AI do not materialise, there are enough concrete examples of problems associated with current applications of artificial intelligence to warrant concern about the level of control that exists over developments in AI. Some form of regulation is likely necessary to protect society from risks of harm. However, advances in regulatory capacity have not kept pace with developments in new technologies including AI. This is partly because regulation has become decentered; that is, the traditional role of public regulators such as governments commanding regulation has been dissipated and other participants including those from within the industry have taken the lead. Other contributing factors are the dwindling of resources in governments on the one hand and the increased power of technology companies on the other. These factors have left the field of AI development relatively unregulated. Whatever the reason, it is now more difficult for traditional public regulatory bodies to control the development of AI. In the vacuum, industry participants have begun to self-regulate by promoting soft law options such as codes of practice and standards. We argue that, despite the reduced authority of public regulatory agencies, the risks associated with runaway AI require regulators to begin to participate in what is largely an unregulated field. In an environment where resources are scarce, governments or public regulators must develop new ways of regulating. This paper proposes solutions to regulating the development of AI ex ante. We suggest a two-step process: first, governments can set expectations and send signals to influence participants in AI development. We adopt the term nudging to refer to this type of influencing. Second, public regulators must participate in and interact with the relevant industries. By doing this, they can gather information and knowledge about the industries, begin to assess risks and then be in a position to regulate those areas that pose

[†] Senior Lecturer, Commercial and Property Law Research Centre, Queensland University Technology (QUT) Faculty of Law, michael.guihot@qut.edu.au. The authors wish to thank participants in the We Robot 2017 conference at Yale University, especially Kate Crawford, Dustin Lewis, and Matthew Scherer for their helpful suggestions and comments.

[‡] Lecturer, Commercial and Property Law Research Centre, Queensland University of Technology (QUT) Faculty of Law, a.matthew@qut.edu.au

[*] Associate Professor, Queensland University of Technology (QUT) Faculty of Law, n.suzor@qut.edu.au. Associate Professor Suzor is the recipient of an Australian Research Council DECRA Fellowship (Project Number DE160101542).

most risk first. To conduct a proper risk analysis, regulators must have sufficient knowledge and understanding about the target of regulation to be able to classify various risk categories. We have proposed an initial classification based on the literature that can help to direct pressing issues for further research and a deeper understanding of the various applications of AI and the relative risks they pose.

I. 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.[4] This insistence betrayed concerns about AI development from within the industry. Google apparently agreed to set up the ethics board, 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.[5] Google 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'.[6] Working at 'Google scale' presumably means using Google's worldwide infrastructure to test its AI systems – the opportunities for which appear to be limitless. Google has already expanded its testing using DeepMind in other areas such as to reduce global warming,[7] and to improve diagnosis and treatment in healthcare.[8] If the results of the application of AI in Google's data centres can be replicated more broadly so as to reduce the world's energy consumption, avert global warming, or enable affordable, accessible health care, then humanity will reap great benefits.[9] However, while the results of the tests appear laudable, some questions linger such as what checks and balances were in place to govern the application of AI here? Were any risks of its application considered and ameliorated in the tests? What governance is in place to control companies testing beta versions of AI applications on a large scale? Conversely, if regulation is put in place prematurely or without proper thought and consultation, would the potential benefits that might result from the general application of these programs in other areas be retarded or lost? In short, would regulation have a chilling effect on innovation that is harmful for the long-term public interest? We argue that, with

[4] 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).

[5] 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).

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

[7] Sam Shead, DEEPMIND IS FUNDING CLIMATE CHANGE RESEARCH AT CAMBRIDGE AS IT LOOKS TO USE AI TO SLOW DOWN GLOBAL WARMING BUSINESS INSIDER AUSTRALIA (2017), https://www.businessinsider.com.au/deepmind-is-funding-climate-change-research-at-cambridge-university-2017-6 (last visited Jul 26, 2017).

[8] DeepMind, WORKING WITH THE NHS TO BUILD LIFE SAVING TECHNOLOGY DEEPMIND, https://deepmind.com/ (last visited Mar 19, 2017).

[9] Here, we have concentrated on the work of Google but it is only one of the major innovators in this area. Similar work on developing AI is also being carried out by Facebook, Microsoft and Apple, to name a few – see the discussion in Part II below.

these questions in mind, AI should be more actively regulated because the benefits that can be achieved through controlled or regulated application outweigh the potential negative impacts of regulating. This paper addresses these and some of the many other issues that must be addressed by potential regulators when seeking to regulate new technologies such as AI.

In Part II, we outline the range of threats posed by different applications of AI and introduce the case for regulating its development. While some argue that developing AI poses an existential threat to humanity, others point to the benefits attained by relatively controlled development and application of more benign systems. We argue that these arguments are at cross purposes and distract from a more pressing need: government ought not only play a part in guiding the development of AI for the broader benefit of humankind, but must also regulate to address the very real and present problems associated with current applications of AI today. These include bias and safety concerns as well as the pressing effect on employment and the inherent intrusion into our privacy caused by AI interrogating the data we generate as part of our everyday lives. Before we contemplate regulating AI though, we must more precisely define and classify the different technologies that are often referred to as AI. This classification exercise, we argue, is vital to understanding the different types of risks that regulation might seek to address. This spectrum of risks posed by different classes of AI provides the basis upon which we ultimately argue for a stratified approach to regulation. This is developed further in Part V.

In Part III, we set out the challenges of regulating AI. 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.[10] In these situations, regulation lags behind or in some circumstances 'decouples' from the technology it seeks to address.[11] Another core challenge regulatory agencies face lies in the difficulty in understanding the social impacts of AI on a systems level, and engaging with these impacts at every (or any) stage of development.[12] A 'social systems analysis' will allow regulators to understand the operation of AI in a broad social context.[13] 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,[14] and sometimes even

[10] 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).

[11] 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).

[12] Kate Crawford & Ryan Calo, *There is a blind spot in AI research*, 538 NATURE 311–313 (2016).

[13] *Id.*

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

unknowable.[15] Research and development in AI is carried out in many different locations, at different times, and in ways that are not highly visible. The scale of research also varies and can be carried out by a single person on a home computer or at a scale that only large multinational companies such as Google can attain.

There is no shortage of advice given to regulators about how to respond to technological change. We review the challenges that current and future developments in AI are likely to pose for regulators, and the different and sometimes conflicting advice that commentators have urged regulators to follow. We consider the urgency of developing effective mechanisms of regulation, and explain how the challenges of regulating AI are different in kind to challenges of regulating in other domains. We argue that as many public regulators now find themselves without the resources to adequately understand or intervene in the range of complex issues that rapid developments in AI present, some regulatory innovation is required. In order to meet these challenges, we suggest that regulators will need to be adaptable, develop new strategies to learn about risks, and identify opportunities to influence technological developers. We show that recent developments in how regulation is conceived go some way to identifying potential future strategies for public regulators, but that more work is needed.

In Part IV, we consider how 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. While the regulation of other emerging technologies is not directly applicable to AI, there is much that can be learned from innovations in regulation in other fields.[16] Current regulatory mechanisms, including laws governing tort, copyright, privacy, and patent, and regulations that govern other emerging technologies are either unsuitable or, for other reasons, cannot easily be applied to novel technological developments in areas such as the regulation of AI.[17] The challenge in regulating this field is magnified by the fundamental uncertainty about how AI will develop and how that development may impact on the other challenges we will face in the future.[18]

The effect of the size and power of the multinational companies that develop most of the applications of AI in the world, such as Google, Facebook and Microsoft, raises fundamental issues about the ability of governments to regulate in this area at all. Far fewer of the traditional tools of regulation once available to governments seeking to regulate AI remain viable or available. We highlight the concerns being expressed about the rampant research and development into AI by

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

[16] 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).

[17] Allenby, *supra* note 11 at 20–21.

[18] 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).

some of the world's biggest companies, ostensibly ungoverned,[19] and propose some innovative solutions to counterbalance the power disparity. We review the range of proposals and suggestions for regulating AI, and consider how regulatory theory provides guidance.

In Part V we argue that in the context of highly constrained governance resources, some regulatory innovation is required. Some 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.[20] We argue that there is a role for government to play in shaping the regulatory environment at a very broad policy level by nudging or influencing beneficial development.[21] By using its influence in this way, government can seek to guide the development of AI by framing the agenda in positive ways without wholly relinquishing its traditional regulatory role. This will also allow governments to develop a fuller regulatory response over time. The multitude of different applications of AI would make it improbable that nudging would have an effect at the micro level of individual applications. At this micro level, we suggest that other more concrete regulatory approaches need to be employed. 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.[22] When the focus shifts to regulation within individual industries or of particular types of AI applications, regulatory agencies must move beyond nudging and adopt more focussed, nuanced and adaptive approaches to regulation.[23] Other theorists have proposed greater roles for regulatory agencies with specific expertise.[24] Still others have suggested that public regulators may be able to experiment with more

[19] 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 Crawford and Calo, *supra* note 12 at 311.

[20] 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).

[21] 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).

[22] 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).

[23] See Richard S. Whitt, *Adaptive Policymaking: Evolving and Applying Emergent Solutions for U.S. Communications Policy*, 61 FED. COMMUN. LAW J. 483, 487 (2008). for example who proposed applying his version of 'adaptive policymaking', where regulators 'tinker' with 'inputs, connectivity, incentives, and feedback' to encourage firms to act in ways that further the public good.

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

rapid, temporary laws,[25] although the potential lack of legal certainty that results may create further problems for investors and other participants in the field. We identify some of these opportunities for innovation in the work of public regulators.

A regulatory intervention in the development of AI technology must consider the spectrum of risks that different AI applications pose. In Part V, we introduce a risk-based regulation framework to help regulators work through the different forms of AI and to identify where scarce regulatory resources should be concentrated. Our initial typology presents three discrete categories: low, medium and high risk applications of AI. Of these, we suggest that the most productive area for regulators to focus on at the moment is medium to high risk categories but that the potential for low risk AI to quickly develop into high risk should mean that these areas must not be completely discounted.

In Part VI we conclude with a suggestion for greater cooperation and information sharing between regulators and the potentially regulated. We argue that, with the increase in societal concerns about the risk inherent in developing AI, regulation of AI is an inevitable and responsible approach to governance.

II. ARTIFICIAL INTELLIGENCE: WHAT DOES THE FUTURE HOLD?

To be able to regulate AI, regulatory bodies must understand both the thing that they seek to regulate and the potential risks that it poses. We must cast our gaze both back and into the future to see how AI has been defined and what it might become, and what risks AI has posed and might pose in the future. We outline some attempts that have been made to define AI and see that there is no concrete definition. This has led to an informal classification system based upon the 'strength' of the underlying algorithm or its ultimate effect. Traditional classifications of AI differentiate between 'narrow' and 'strong' AI. This is unsatisfactory in terms of defining AI because one measures breadth while the other measures strength. We propose a different classification; one based upon the risks that each AI application poses. In this way, we can begin to sort various classes of AI based on whether the AI poses a low, medium, or high risk to either society or to human safety or wellbeing. This classification is crucial to understanding how regulatory strategies can be tailored to the relevant AI risk profile. Regulatory bodies need to perform this risk analysis before they develop laws that affect a class of AI. It is important then to distinguish the various meanings given to the term 'artificial intelligence' and the different forms AI may take. This allows us to identify a subset or range of applications of AI most suitable for governments or regulatory bodies to initially regulate.

A. Defining AI

Before defining artificial intelligence we need first to define intelligence. Intelligence in human terms has been described as a set of factors that include

[25] 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).

'consciousness, self-awareness, language use, the ability to learn, the ability to abstract, the ability to adapt, and the ability to reason'.[26] Once intelligence is defined, estimations or approximations of those qualities should form the benchmark of attempts to create or simulate it – hence artificial intelligence. But a simulation of which of those characteristics of intelligence can be called artificial intelligence? Must it replicate all aspects of intelligence? John McCarthy did not limit intelligence in AI to a replication of human intelligence but argued that machines could display other intelligences that involve 'much more computing than people can do'.[27] He defined artificial intelligence as 'the science and engineering of making intelligent machines, especially intelligent computer programs.'[28] Omohundro adopted an external agency requirement and defined AI as a system that 'has goals which it tries to accomplish by acting in the world'.[29] Russell and Norvig summarised eight definitions of AI differentiated by how they reflected expectations of human thinking and behavior or (machine) rational thinking and behavior.[30] Ultimately, Russell and Norvig preferred the rational agent approach in which machine agents 'operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue [the best expected outcome]'.[31] To be able to display these characteristics, AI also needs to be actuated in machinery, whether that is a computer system or a robot. Typically, though, these machine behaviors have been compared against human abilities to process language, to reason, and to perceive and manipulate objects in the environment to attain pre-determined goals.[32]

All of these definitions set a fairly high bar for an algorithm to attain before it meets the definition of AI. AI then can be differentiated from machine learning systems or even machine learning that learns from examining large data sets,

[26] Scherer, *supra* note 24 at 360. Consciousness on its own has proved notoriously difficult to define, a difficulty amplified when attempting to define artificial consciousness. See Christof Koch et al., *Neural correlates of consciousness: progress and problems*, 17 NAT. REV. NEUROSCI. 307–321 (2016); GERALD M EDELMAN, THE REMEMBERED PRESENT: A BIOLOGICAL THEORY OF CONSCIOUSNESS (2000); Francis Crick & Christof Koch, *Towards a neurobiological theory of consciousness*, 2 *in* SEMINARS IN THE NEUROSCIENCES 263–275 (1990), http://authors.library.caltech.edu/40352/ (last visited Jul 17, 2017); Francis Crick & J. Clark, *The astonishing hypothesis*, 1 J. CONSCIOUS. STUD. 10–16 (1994); Stanislas Dehaene & Jean-Pierre Changeux, *Experimental and Theoretical Approaches to Conscious Processing*, 70 NEURON 200–227 (2011); Steve Torrance, *Ethics and consciousness in artificial agents*, 22 AI SOC. 495–521 (2008); Wendell Wallach, Colin Allen & Stan Franklin, *Consciousness and ethics: artificially conscious moral agents*, 03 INT. J. MACH. CONSCIOUS. 177–192 (2011); Paul FMJ Verschure, *Synthetic consciousness: the distributed adaptive control perspective*, 371 PHIL TRANS R SOC B 20150448 (2016).

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

[28] *Id.*

[29] Stephen M. Omohundro, *The Basic AI Drives*, *in* ARTIFICIAL GENERAL INTELLIGENCE 2008 483–493 (Pei Wang, Ben Goertzel, & Stan Franklin eds., 2008).

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

[31] McCarthy, *supra* note 27 at 4–5. This combination of perception, adaptability, creativity and autonomous operation reflects what would be required of an agent to pass the Turing test.

[32] RUSSELL AND NORVIG, *supra* note 30 at 2–3.

sometimes using neural networks to make deep connections among the data. If these computations do not display the other characteristics of AI such as operating autonomously, adapting to change, creating and pursuing their own goals,[33] then they cannot be AI. However, while they cannot be AI based on the definitions above, machine learning systems are also, perhaps erroneously, often referred to as being AI.

The lack of definitional clarity means that the broad label, AI, has become the vernacular term for a range of programs, algorithms and networks that are used in a multitude of applications. For example AI is used to refer to the programs underlying chess and other game playing programs and Roomba vacuum cleaners,[34] but also to the coordinated systems controlling autonomous vehicles and the personal agents developed by Microsoft, Apple and Google among others. Some of these uses of the term AI are differentiated by using descriptors such as 'narrow AI' to distinguish their limited application to a single set task. When AI is developed so as to apply more broadly or with greater effectiveness it is often referred to as becoming stronger,[35] rather than the opposite of narrow: broader.

Complicating this definitional problem further, research by mathematicians and engineers who seek to develop self-replicating and self-aware algorithms is said also to be work 'in AI'.[36] There has been some attempt to distinguish this work from narrow or even stronger AI and algorithms that display these characteristics by referring to it as 'strong AI'. A more common reference is artificial general intelligence (AGI). As opposed to narrow AI, AGI is said to possess 'a reasonable degree of self-understanding and autonomous self-control, [has] the ability to solve a variety of complex problems in a variety of contexts, and [can] learn to solve new problems that they didn't know about at the time of their creation'.[37] AGI is 'subject to a variety of "drives" including self-protection, resource acquisition, replication, goal preservation, efficiency, and self-improvement'.[38] It is generally recognised that AGI does not yet exist but it is AGI that causes most concern to those who believe that AI creates an existential threat to humanity. We discuss this further in Part II C.

[33] Liza Daly, AI LITERACY: THE BASICS OF MACHINE LEARNING WORLD WRITABLE (2017), https://worldwritable.com/ai-literacy-the-basics-of-machine-learning-2e20f93e34b4 (last visited Apr 13, 2017).

[34] These single task applications are often classified as narrow AI - see ARTIFICIAL GENERAL INTELLIGENCE, VI (Ben Goertzel & Cassio Pennachin eds., 2007). See also RAY KURZWEIL, THE SINGULARITY IS NEAR 264 (2010). The bulk of AI research and development today is conducted into this narrow type of AI – see ARTIFICIAL GENERAL INTELLIGENCE, *supra* note at 1.

[35] KURZWEIL, *supra* note 34 at 289 and 409. This classification system refers to narrow AI as opposed to strong AI. Perhaps a clearer dichotomy would be to refer to weak AI and strong AI but we retain the traditional classification in this paper.

[36] Laurent Orseau, *Asymptotic non-learnability of universal agents with computable horizon functions*, 473 THEOR. COMPUT. SCI. 149–156 (2013).

[37] ARTIFICIAL GENERAL INTELLIGENCE, *supra* note 34 at VI.

[38] 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).

The range of applications of AI sits on a spectrum from those applications that are not strictly AI,[39] through to narrow applications of AI (as found in chess games etc) to AGI. When referring to AI then, we must bear in mind this vast array of uses and misuses of the term. It is neither possible nor even desirable to govern all of these diverse uses of AI using one regulatory approach. However, the risks associated with these different applications of AI will arguably drive different regulatory responses and must therefore be treated differently. This is why we argue for a classification based on the risk that various AI applications pose. For public regulators that have limited resources and information, classifying AI can inform their decisions about which applications or class of AI to regulate first, and at what level.

B. Introducing risk as a defining point

We propose that risk should be considered as a quality that differentiates classes of AI. We develop this idea further in Part III but here we argue that once applications of AI are classified according to the potential risk each poses to society or to the people or environment in which they are applied, then public regulators can more efficiently and effectively direct their regulatory responses. Without that knowledge, they will be grasping in the dark at even understanding the regulatory problem.[40]

Even categorising risk in relation to AI is complicated by a lack of clarity on AI's potential. On one hand, underpinning the risk analysis is the pervasive fear that AI will develop rapidly to the point at which it will annihilate humans as a species either through some miscalculation in replicating software or because humans are suboptimal to the machine's set goals. When talking about risk in relation to AI, it is these risks that linger just below the surface of each argument. On the other hand, others argue that the development of AI is benign and beneficial to society. However, these arguments may be at cross-purposes due, we argue, to a lack of a sufficient and agreed-upon definition for AI. We suggest that classifying AI based upon potential risk factors as suggested in this paper may clarify some of these arguments so that regulation may be used where required to minimise risks, while at the same time allowing development of less risky AI with only minimal

[39] See for example Adi Prakash, "DOING AI": WHAT LEGAL SHOULD REMEMBER ABOUT BIG DATA LEGALTECH NEWS (2017), http://m.legaltechnews.com/?slreturn=20170726234213/#/article/1202792798132/Doing-AI-What-Legal-Should-Remember-About-Big-Data?utm_content=buffer01b0a&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer&_almReferrer=https:%2F%2Ft.co%2Fwki9DruH9A (last visited Jul 27, 2017).

[40] The United States government has recognised this. See recommendation 5 in its EXECUTIVE OFFICE OF THE PRESIDENT NATIONAL SCIENCE AND TECHNOLOGY COUNCIL COMMITTEE ON TECHNOLOGY, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (2016) which states "Agencies should draw on appropriate technical expertise at the senior level when setting regulatory policy for AI-enabled products. Effective regulation of AI-enabled products requires collaboration between agency leadership, staff knowledgeable about the existing regulatory framework and regulatory practices generally, and technical experts with knowledge of AI. Agency leadership should take steps to recruit the necessary technical talent, or identify it in existing agency staff, and should ensure that there are sufficient technical 'seats at the table' in regulatory policy discussions".

regulatory intervention. In this way, we can avoid suggesting the same regulatory response to the AI in a Roomba, for example, as we would to regulate autonomous weapons systems or the comparatively simple algorithms that regulate critical environmental or energy systems.

In Part II C, we discuss the arguments made in relation to the existential risks posed by some in relation to AI. As discussed, these arguments are often raised as reasons to regulate the development of AI and must be addressed. Then in Part II D, we outline concrete examples of problems associated with current applications of AI in use today that, we argue, also require a regulatory response but for more concrete reasons.

C. Reports of the Singularity and the End of Humanity May be Greatly Exaggerated

Perhaps the most visceral fear about the development of AI is the existential threat to humanity that is said will be caused by the rise of super-intelligent machines.[41] These concerns pervade the collective conscious in relation to AI. We argue that this fear may be overstated given the current state of development in AI but it is such a pervading idea that it informs every level of discussion. It is also addressed in every code of conduct, standard, or values statement that has been developed by those in the industry[42] and should be addressed in any regulatory intervention.

In 1965 Good proposed that society would be transformed by the invention of a machine with ultra-intelligence. It would surpass human intelligence and be able to design even more intelligent machines.[43] It would, he argued, be the last machine that humans would ever need to make for themselves[44] and would save humanity.[45] Good argued that this *ex machina* in human image would be designed from an understanding of human intellect.[46] Good's 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

[41] See for example, 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 (1993), https://ntrs.nasa.gov/search.jsp?R=19940022855 (last visited Mar 14, 2017); Writing 15 years after Vinge, Kurzweil appears most optimistic about the outcome of the singularity but maintains an element of caution. See KURZWEIL, *supra* note 34; NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES (First edition ed. 2014); Nick Bostrom, *When machines outsmart humans,* 35 FUTURES 759–764 (2003); JOHN VON NEUMANN & RAY KURZWEIL, THE COMPUTER & THE BRAIN (2012).

[42] See the discussion of these soft law approaches in Part IV B 2 below.

[43] Irving John Good, *Speculations Concerning the First Ultraintelligent Machine,* 6 *in* ADVANCES IN COMPUTERS 31–88, 33 (Franz L. Alt and Morris Rubinoff ed., 1966), http://www.sciencedirect.com/science/article/pii/S0065245808604180 (last visited Mar 14, 2017).

[44] *Id.* at 31–32.

[45] *Id.* at 31.

[46] *Id.* at 78.

could become aware of its own superior intelligence. This event, which he described as the singularity, would spell the end of humanity.[47]

These fears are not new and are not confined to fears of AI. Age-old concerns in human mythology about humans playing god-the-creator form the basis of stories such as the Frankenstein and golem stories. These stories have parallels with, and lessons for, the development of AGI. 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'[48] — comparable to the algorithm in AI. In some golem stories, the golem obtains superhuman strength and, uncontrolled, causes destruction and mayhem. The parallels to the creation of AGI with super human intelligence are apt. A further parallel might be drawn with the desire to regulate or control these fears. For example, golems were bound by Jewish law. They were programmed not to kill unless necessary and could not lie.[49] We see here the birth of the idea of embedding legal codes within technical code.[50]

Existential concerns have stimulated the minds of ethicists and philosophers since soon after work began on AI.[51] However, discussions about the legal ramifications of AI were typically slower to develop and early considerations of AI and the law only appear in the early 1980s.[52] Even then, the dangers associated with the inability to understand and control AI were apparent.[53] This problem has not diminished and, if anything, has probably increased in the nearly 40 years since 1981. Researchers in AI recognise that there is a potential risk that if autonomous AGI is developed, it will be difficult for a human operator to maintain control.[54]

Some of the risks seem remote, or are, at this stage, only potential problems, 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.[55] People can be particularly risk averse when they stand to lose something,[56] and governments respond to the desires and concerns of the societies

[47] 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).

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

[49] *Id.* at 205.

[50] See LAWRENCE LESSIG, CODE: VERSION 2.0 (2nd edition ed. 2006).

[51] See NORBERT WIENER, CYBERNETICS (2nd ed. 1961).

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

[53] *Id.* at 446. citing; WIENER, *supra* note 51.

[54] See Omohundro, *supra* note 38.

[55] 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).

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

they govern. One aim of the law is to predict what might go wrong and to design laws to prevent or avoid it.

It is characteristic of exponential growth that all of the significant effects of the growth occur in the last short timeframe at the end of the growth set. AI has had a long gestation period. There have been many failed predictions about the imminent explosion of AI over the last sixty years, but, far from dissipating, the questions about AI's impact will only become more urgent as we draw nearer to the exponential inflection point and its growth takes a sudden and dramatic vertical trajectory. The question is are we, after sixty years of growth, now approaching that inflection point or are we still in the slower gradual development phase? The answer must be that as we approach the point where AI begins to develop more quickly, we should begin to prepare for and guide the development of AI in ways that will benefit society while still avoiding existential threats as best we can. This should be the role of the law, but lawmaking processes are often criticised as being overly responsive or reactive rather than sufficiently proactive.

Those within the AI industry have already taken steps to counter the concerns about AGI autonomously self-replicating out of human control. For example, Orseau and Armstrong, an engineer at DeepMind and a researcher into systemic risk respectively, acknowledged that 'reinforcement learning agents … are unlikely to behave optimally all the time'.[57] They recognised 'concerns that a "superintelligent" agent may resist being shut down, because this would lead to a decrease of its expected reward',[58] and detailed how DeepMind's engineers have developed a 'big red button', or an off switch for such an artificially intelligent reinforcement learning agent. Any regulation of AI might consider compulsory adoption of this program in all research and development into AGI.

However, not everyone shares these concerns. The panel that contributed to the Stanford Report into AI titled *Artificial Intelligence and Life in 2030* noted that:

> 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.[59]

While threats to humankind posed by AI may yet still be some way off, it is important to listen to those in the industry who are calling for controls to be put in place now to prepare for the future. If AI ever does develop to a point where it becomes a threat to humanity, they argue, it may well be too late to do anything about it. Far from ignoring these fears and threats, any regulatory response to AI must address the risks they pose in some manner. The more recent warnings of technology entrepreneurs like Elon Musk and scientists like Stephen Hawking about the risks of runaway AI should at least cause regulators to pause and consider

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

[58] *Id.* at 2.

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

whether they have appropriate risk identification and mitigation strategies in place.[60]

The call to regulate comes not only from the deep human fears of the singularity but also because of the more concrete problems associated with the narrow AI that currently exists, has already been implemented, and pervades our everyday lives. In Part II D, we analyse some of these unforeseen problems that are occurring now in current applications of AI. These issues highlight the potential for unforeseen errors to occur. These types of demonstrable errors and unforeseen problems are the canary in the coalmine of AI development. They provide warning about how things can go wrong when society and governments allow AI systems to be developed and deployed without appropriate regulation in place. Any regulatory response needs to ensure that AI systems are designed and deployed so that they do not pose any harm (in its broadest sense) to people or society.[61]

D. Problems Associated With Current Applications of AI

For the moment, the dystopian ramifications of rampant, uncontrollable AI are still the imaginings of science fiction writers.[62] The current challenge for regulating AI is the proliferation in the capabilities of relatively narrow AI systems tasked with performing specific functions.[63] Developments in AI technology have been smouldering since research on it began shortly after World War II.[64] Today, AI is at the forefront of technological development and 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. For example, IBM's cognitive computing platform, Watson, has developed from its initial challenge of winning the gameshow Jeopardy to being

[60] See e.g. AI Open Letter, FUTURE OF LIFE INSTITUTE, https://futureoflife.org/ai-open-letter/ (last visited Mar 13, 2017).

[61] Crawford and Calo refer to this as the blind spot in AI research. See Crawford and Calo, *supra* note 12.

[62] 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."

[63] KURZWEIL, *supra* note 37 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; See also 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.

[64] McCarthy, *supra* note 27.

applied to provide real solutions to problems in commerce, law, and health.[65] DeepMind's AlphaGo recently defeated the human master of the complex Chinese board game Go and Google also used DeepMind's AI to reduce the electricity consumption in Google's data centres;[66] Microsoft has incorporated AI into its personal agents such as Cortana and Zo which can perform a dizzying array of tasks and answer seemingly unlimited questions using a mellifluous (female by design) computer generated voice;[67] Microsoft's algorithm, DeeperCoder, is capable of writing code to solve simple problems;[68] Facebook uses AI in its face recognition, language translation, and camera effects and its research arm, Facebook Artificial Intelligence Research (FAIR) is said to be 'committed to advancing the field of machine intelligence'.[69] Joaquin Candela, Director of Engineering for Facebook's Applied Machine Learning (AML) group has stated

[65] 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).

[66] See this and other examples of DeepMind's application in the introduction to this paper.

[67] 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).

[68] 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).

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

that Facebook is working towards 'generalization of AI'[70] which will, it is argued, be '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'.[71] Advances in AI technology are vaulting toward the exponential as computer capacity and speed double every two years.[72] The Stanford Report predicts that as driverless cars fall into common use, they will form the first public impressions of AI in a corporeal form.[73] This experience will be an important one for AI, since we are on the cusp of a surge of AI with a physical embodiment. The Stanford Report also predicts that the typical North American city will by 2030 feature personal robots, driverless trucks and flying cars.[74]

These AI systems present a spectrum of immediate issues that may require a regulatory response. Some are likely to be dealt with by developers as they come to their attention, and end users of the system may deal with others as they refine their use of the system and work with developers in overcoming issues as and when they arise. In this Part, we outline several of the issues that may require a regulatory response including biases that appear in law enforcement decisions made by AI systems, safety, particularly in relation to driverless cars, the lack of a human 'heart' when relying on AI in judicial decision making, privacy in relation to a vast number of applications, and the pressing problems associated with unemployment caused by increasing rates of automation supported by AI.

1. Bias

The coalescing of AI and big data opens significant possibilities for the synthesis and analysis of that data, but it also stands to compound problems that presently exist in that process. These include unintended racism, sexism and discrimination in the outcomes of data analysis.[75] Ajunwa, Crawford and Ford have proposed a model to regulate big data to address privacy concerns and to allow a pathway to correct erroneous assumptions made from an assemblage of

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

[71] *Id.*

[72] 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).

[73] STONE ET AL., *supra* note 59 at 18–25.

[74] STONE ET AL., *supra* note 59, 18-23 (automated vehicles), 24-25 (home robots), 7, 18, 20 (flying vehicles).

[75] 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).

that data.[76] Bias can be difficult to detect, and if care is not taken can 'become part of the logic of everyday algorithmic systems'.[77] 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.[78] Conversely, risk scores were mistakenly low for white defendants.[79] Bias also arises in the work of private platforms that filter, index, and sort online content and mediate communications.[80] 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.[81] Bias issues such as these are unlikely to provoke a regulatory response if they are dealt with in AI system design. However, these issues can be ameliorated with regulation that requires either careful design or prompt troubleshooting when the issues are identified.

2. Safety

AI is being touted as a solution to a number of social problems. However, when it is implemented in a social context, it also presents a range of safety issues.[82] For example, autonomous vehicles such as cars and trucks have the potential to improve safety on roads if they succeed in reducing accidents caused by driver error such as inattention, impairment, slow reaction times and inappropriate risk-taking. Social benefits potentially include improving mobility for those unable to drive, or those that live in heavily traffic congested urban areas.[83] Hence, there is an urgency to deploy autonomous vehicles and developers

[76] 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).

[77] Crawford, supra note 78.

[78] 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).

[79] *Id.*

[80] 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).

[81] Crawford, *supra* note 78; 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).

[82] Patrick Lin, Keith Abney & George Bekey, *Robot ethics: Mapping the issues for a mechanized world*, 175 ARTIF. INTELL. 942–949, 945–946 (2011); Drew Simshaw et al., *Regulating Healthcare Robots: Maximizing Opportunities While Minimizing Risks*, 22 RICHMOND J. LAW TECHNOL. 1–38 (2015); Eliezer Yudkowsky, *Cognitive biases potentially affecting judgment of global risks*, 1 GLOB. CATASTROPHIC RISKS 13 (2008).

[83] JAMES MANYIKA ET AL., DISRUPTIVE TECHNOLOGIES: ADVANCES THAT WILL TRANSFORM LIFE, BUSINESS, AND THE GLOBAL ECONOMY 78–83 (2013), http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/disruptive-technologies (last visited Jun 29, 2017). In this report, the McKinsey Global Institute

have already been testing autonomous vehicles on public roads. Indeed, the authors of the Stanford Report expect that ‘transportation is likely to be one of the first domains in which the general public will be asked to trust the reliability and safety of an AI system for a critical task’.[84]

However the safety risks present in autonomous vehicles include the risk of accidents that may not otherwise have occurred; accidents created by even minor software or hardware errors, flawed or deficient programming of software, or unethical decision-making in the face of a high-risk, multi-risk scenario.[85] Regulation is key to providing an environment that will give the technology a chance to develop to its full potential, while protecting the public from unacceptable risks.[86] Public regulators are already developing regulatory frameworks for safety assurance during the development and testing phases.[87] These frameworks extend to design standards, vehicle modification, and the

identifies government regulation as potentially both an enabler and barrier to the socio-economic benefits of autonomous vehicles.

[84] STONE ET AL., *supra* note 57, 18.

[85] Lin, Abney, and Bekey, *supra* note 82 at 945. Examples of programming issues present with autonomous vehicles include for example, the trolley-car problem. Programming issues may be highly specific and unique to certain cultures, geographical terrain, or indigenous fauna. See for example, reports that Volvo is working on difficulties arising with the animal detection system in its autonomous vehicles when confronted with the unusual way in which kangaroos move. The system had previously been tested on moose in Sweden: Jake Evans, THE VERY AUSTRALIAN PROBLEM OF ROOS AND DRIVERLESS CARS ABC NEWS (2017), http://www.abc.net.au/news/2017-06-24/driverless-cars-in-australia-face-challenge-of-roo-problem/8574816 (last visited Jun 29, 2017).

[86] Upon the introduction of the Federal Automated Vehicles Policy in the US, President Obama noted, ‘the quickest way to slam the brakes on innovation is for the public to lose confidence in the safety of new technologies’: Barack Obama, SELF-DRIVING, YES, BUT ALSO SAFE PITTSBURGH POST-GAZETTE (2016), http://www.post-gazette.com/opinion/Op-Ed/2016/09/19/Barack-Obama-Self-driving-yes-but-also-safe/stories/201609200027 (last visited Jun 29, 2017); US DEPARTMENT OF TRANSPORTATION, FEDERAL AUTOMATED VEHICLES POLICY: ACCELERATING THE NEXT REVOLUTION IN ROADWAY SAFETY (2016), https://www.transportation.gov/AV (last visited Jun 29, 2017).

[87] See for example, US DEPARTMENT OF TRANSPORTATION, *supra* note 86; DEPARTMENT OF TRANSPORT (UK), THE PATHWAY TO DRIVERLESS CARS: A CODE OF PRACTICE FOR TESTING - MOVING BRITAIN AHEAD (2015), https://www.gov.uk/government/publications/automated-vehicle-technologies-testing-code-of-practice (last visited Jun 29, 2017); (SVG), STRASSENVERKEHRSGESTZ, (Road Traffic Act, Germany), amended in June 2017 to allow for automated vehicles on public roads; NATIONAL TRANSPORT COMMISSION AUSTRALIA, REGULATORY REFORMS FOR AUTOMATED ROAD VEHICLES (2016); PILOT PROJECT AUTOMATED VEHICLES, ONTARIO REGULATION 305/15 UNDER HIGHWAY TRAFFIC ACT, RSO 1990, C H 8, (2014). The UNITED NATIONS CONVENTION ON ROAD TRAFFIC, VIENNA, (1968), (Vienna Convention on Road Traffic), Articles 8 (5bis) and 39(1) were amended to facilitate use of autonomous vehicles on public roads, while ensuring the driver of the vehicle maintained its position in a superior role. The justifications for the amendment to the *Vienna Convention on Road Traffic* are included as an appendix to UNITED NATIONS, ECONOMIC AND SOCIAL COUNCIL, ECONOMIC COMMISSION FOR EUROPE, INLAND TRANSPORT COMMITTEE, WORKING PARTY ON ROAD TRAFFIC SAFETY, REPORT OF THE SIXTY-EIGHTH SESSION OF THE WORKING PARTY ON ROAD TRAFFIC SAFETY 11 (2014).

development of safety principles, criteria and assurance standards that are 'efficient, affordable and create a minimal administrative burden'.[88]

The success of AI in solving social problems will ultimately lie in public and regulatory confidence in its use, and much of this confidence will turn upon trust in safety assurance.[89] Safety, in this sense, ought not be confined to physical safety, but should extend to concern for non-physical harm[90] such as privacy, security, and the dehumanization of care for people at their most vulnerable.[91] For example, the benefit of AI-enabled healthcare robots could be impeded by lack of regulation to assure public trust and confidence across a range of safety issues including these types of non-physical harm.

These risks are most acute with personal care robots. Trust and confidence in AI assisted robots may be hard-won in personal care situations given that they have traditionally involved human-to-human interaction.[92] Also, to be effective and efficient, personal care robots must be able to access personal and medical information, 'know… and possibly shar[e] the location of medication, objects and people',[93] connect with hospital or other healthcare networks, and connect with networked technology such as personal devices including phones, wearable devices, or mobile applications.[94] The unprecedented amount of personal and medical information that could potentially be accessed, used, processed and stored by personal care robots is vulnerable to the same privacy and security concerns raised in relation to the Internet of Things.[95] Aside from these security and privacy concerns, the health care context may raise unique safety concerns, for example, if

[88] NATIONAL TRANSPORT COMMISSION AUSTRALIA, *supra* note 87 at 14; Current projects: Automated vehicle trial guidelines, NATIONAL TRANSPORT COMMISSION, http://www.ntc.gov.au/current-projects/automated-vehicle-trial-guidelines/ (last visited Jun 29, 2017).

[89] STONE ET AL., *supra* note 59 at 35–36; Simshaw et al., *supra* note 82 at 8–10.

[90] In the healthcare context see Bernd Carsten Stahl & Mark Coeckelbergh, *Ethics of healthcare robotics: Towards responsible research and innovation*, 86 ROBOT. AUTON. SYST. 152–161 (2016); Simshaw et al., *supra* note 82; David D. Luxton, Susan Leigh Anderson & Michael Anderson, *Chapter 11 - Ethical Issues and Artificial Intelligence Technologies in Behavioral and Mental Health Care*, *in* ARTIFICIAL INTELLIGENCE IN BEHAVIORAL AND MENTAL HEALTH CARE 255–276 (2016), http://www.sciencedirect.com/science/article/pii/B9780124202481000118 (last visited Jun 28, 2017).

[91] Healthcare robots include surgical, routine-task and personal care robots. See Simshaw et al., *supra* note 82 at 9–10; Stahl and Coeckelbergh, *supra* note 90 at 154, 157; Luxton, Anderson, and Anderson, *supra* note 90.

[92] Laurel D. Riek, *Chapter 8 - Robotics Technology in Mental Health Care*, *in* ARTIFICIAL INTELLIGENCE IN BEHAVIORAL AND MENTAL HEALTH CARE 185–203, 194 (David D. Luxton ed., 2016), http://www.sciencedirect.com/science/article/pii/B9780124202481000088 (last visited Jun 28, 2017).

[93] Simshaw et al., *supra* note 82 at 11–12.

[94] *Id.* at 13–15.

[95] Lin, Abney, and Bekey, *supra* note 82 at 945; Simshaw et al., *supra* note 82; Similarly complex safety issues arise with the non-commercial or recreational use of drones. See Roger Clarke & Lyria Bennett Moses, *The regulation of civilian drones' impacts on public safety*, 30 COMPUT. LAW SECUR. REV. 263–285 (2014).

an external party can hack medical devices such as pacemakers.[96] These risks escalate with unsophisticated home users.[97] Again, we should give careful consideration to regulation that can address these concerns.

3. Legal Decision-Making

AI has been applied in highly specific legal tasks such as sentencing and judicial interpretation in an effort to improve transparency and consistency in judicial decisions.[98] However, these systems have been criticized as lacking capacity to exercise discretion and make situational value judgments. Concerns have been raised about mechanistic reliance upon these applications of AI and their capacity to influence and shape the behavior of people involved in the decision-making process.

Decision-making in the application of legal principles necessarily involves discretion. Decision-making in sentencing relies on 'induction and intuition as well as the capacity to assess the social impact of the decision'.[99] These have not yet proven to be among AI's greatest strengths. There is a significant body of scholarship that argues against using AI in making definitive legal decisions,[100] and cautions against even a narrowly limited role for AI in informing human decisions.[101] As Simpson argued, even if AI is able to approximate human discretion in sentencing decision-making, the question that remains is the extent to

[96] David D. Luxton et al., *Chapter 6 - Intelligent Mobile, Wearable, and Ambient Technologies for Behavioral Health Care*, *in* ARTIFICIAL INTELLIGENCE IN BEHAVIORAL AND MENTAL HEALTH CARE 137–162 (2016), http://www.sciencedirect.com/science/article/pii/B9780124202481000064 (last visited Jun 28, 2017); Simshaw et al., *supra* note 82 at 22.

[97] Simshaw et al., *supra* note 82 at 22.

[98] 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); 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); Hall et al., *supra* note.

[99] Hall et al., *supra* note 98 at 9.

[100] CASS R. SUNSTEIN, OF ARTIFICIAL INTELLIGENCE AND LEGAL REASONING (2001), https://papers.ssrn.com/abstract=289789 (last visited Mar 13, 2017); Philip Leith, *The Judge and the Computer: How Best 'Decision Support'?*, 6 ARTIF. INTELL. LAW 289–309 (1998); Philip Leith, *The Emperor's New Expert System*, 50 MOD. LAW REV. 128–132 (1987); Philip Leith, *The rise and fall of the legal expert system*, 30 INT. REV. LAW COMPUT. TECHNOL. 94–106 (2016); Cooper, *supra* note 65 at 97–99; John O. McGinnis & Russell G. Pearce, *The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services*, (2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2436937 (last visited Mar 10, 2017).

[101] URI J SCHILD, EXPERT SYSTEMS AND CASE LAW (1992); John Zeleznikow, *Building decision support systems in discretionary legal domains*, 14 INT. REV. LAW COMPUT. TECHNOL. 341–356 (2000); 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); Abdul Paliwala, *Rediscovering artificial intelligence and law: an inadequate jurisprudence?*, 30 INT. REV. LAW COMPUT. TECHNOL. 107–114 (2016).

which 'an algorithm can have a heart'.[102] Simpson questions whether 'such algorithms [can] deal with the unexpected, quirky or unique individual that may require appeals to a sense of justice?'[103] Lippe et al propose that an optimal combination of AI and humans is required to provide balance.[104]

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'.[105] The implication, at least in Europe, is that from 2018 a human must somehow be involved in making the decisions. How effective this is likely to be remains to be seen. Public regulators in all jurisdictions similarly ought to consider the risks of allowing the involvement of AI in making automated final legal decisions.

Even where the decision is not automated, but say AI is used to support human decision-making, public regulators ought to be wary of undesirable risks and consequences. Reliance upon AI systems in judicial decision-making enlivens long-standing fears that reducing human processes to their most mechanistic may have an unintended regulatory effect.[106] That is, once a process is reduced to its most mechanistic, it may make the humans involved in the decision-making process more compliant or programmable to the process.[107] Even where the goal and purpose of involving AI in legal decision-making is to increase consistency, there are still risks that it will lead to standardization,[108] which in automated legal decision-making processes can have a regulatory effect on people involved.[109] This regulatory impact may extend to an unintended chilling effect on individualization, even where the legislature intended there to be some flexibility.[110] People involved in the decision-making process may have difficulty deviating from the standardization in order for example to 'have a heart',[111] to 'introduce an element of humanity in special circumstances'[112] or to consider whether the decision is in the best interests of society.[113]

[102] 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).

[103] *Id.* at 56.

[104] Lippe, Katz, and Jackson, *supra* note 101 at 20.

[105] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL OF 27 APRIL 2016 ON THE PROTECTION OF NATURAL PERSONS WITH REGARD TO THE PROCESSING OF PERSONAL DATA AND ON THE FREE MOVEMENT OF SUCH DATA, AND REPEALING DIRECTIVE 95/46/EC (GENERAL DATA PROTECTION REGULATION), 2016/679 (2016), http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 (last visited Jul 26, 2017) to take effect from 25 May 2018.

[106] 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).

[107] *Id.* at 498–499.

[108] Hall et al., *supra* note 98.

[109] Anja Oskamp & Maaike W. Tragter, *Automated Legal Decision Systems in Practice: The Mirror of Reality*, 5 ARTIF. INTELL. LAW 291–322, 293 (1997).

[110] *Id.* at 293.

[111] Simpson, *supra* note 102 at 56.

[112] Hall et al., supra note 60 at 33.

[113] Oskamp and Tragter, *supra* note 109; Paliwala, *supra* note 101 at 112–113.

The array of concerns surrounding the use of AI systems in judicial decision-making is likely to be managed by the continual refinement of how AI systems are deployed by people in the decision-making process: the judiciary and their administrators and should ultimately be regulated.

4. Privacy

The leaps in advancement that are the promise of AI will sometimes turn on the quality and quantity of information available to it to inform AI learning. Public regulators will need to regulate to protect the privacy of individuals if large data sets are disclosed to tech companies with AI capabilities. For example, maintaining patient privacy should be paramount where data sets held by public health services are shared with technology companies. This should be so even where data is disclosed for a specific purpose and is technically compliant with current regulatory disclosure models. Even so, sensitivities surrounding well-intentioned disclosures should result in a regulatory response, even where the disclosure technically complies with existing regulatory processes.[114] Such a regulatory response may be because the existing regulatory compliance process did not contemplate the scale of the disclosure, the use to which the data is put 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.[115] Such a regulatory response may involve the

[114] 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, *supra* note 8; 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).

[115] See Sam Shead, The UK data regulator has ruled that Google DeepMind's first deal with the NHS was illegal Business Insider Australia (2017),

imposition of a command and control model heavily restricting future access to such data sets.

5. Unemployment

The socio-economic and socio-political impact of AI is a serious risk for public regulators. The deployment of AI in workplaces via algorithms, robotics or automation targeting increased speed, efficiency, or safety is expected to radically change the workforce.[116] These concerns speak to a fundamental issue beyond the economics of increased productivity. The sheer scale of the disruptive impact on wages and employment is unlikely to be matched by increased productivity and may instead 'exacerbate inequality rather than promote greater opportunity and shared prosperity'.[117] Regulators must consider issues such as the benefits that society can attain from AI, and how regulators can support workers through the expected job displacement if the scale of that displacement is anything approaching the levels anticipated. Public regulators need to consider the socio-economic and socio-political dis-equilibrium that might be caused if the AI revolution causes widespread unemployment. Ultimately, regulators must consider if society will require a living wage paid for by taxes on robots.[118]

Adverse impacts on employment will not be confined to manufacturing or blue-collar work where robots are already used.[119] While unskilled routine tasks that lend themselves to automation are at high risk, jobs that are highly skilled involving high levels of dexterity, creativity, social intelligence, collaboration, negotiation, and problem solving will also be at risk with further advances in technology.[120] Every robot introduced into the workplace is estimated to have a

https://www.businessinsider.com.au/ico-deepmind-first-nhs-deal-illegal-2017-6 (last visited Jul 26, 2017).

[116] Carl Benedikt Frey & Michael A. Osborne, *The future of employment: how susceptible are jobs to computerisation?*, 114 TECHNOL. FORECAST. SOC. CHANGE 254–280, 291 (2017). McKinsey's occupational study estimates that 51% of jobs in the US ($2.7 trillion of wages) could be automated by 2055, or decades earlier depending on the pace of technological development: see James Manyika et al., HARNESSING AUTOMATION FOR A FUTURE THAT WORKS 8, 49, http://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works (last visited Jul 19, 2017). The World Bank estimates that 57% of jobs in the OECD nations could be displaced: see World Development Report 2016: Digital Dividends, 129, http://www.worldbank.org/en/publication/wdr2016 (last visited Jul 19, 2017).

[117] World Development Report 2016: Digital Dividends, *supra* note 116 at 275; DARON ACEMOGLU & PASCUAL RESTREPO, ROBOTS AND JOBS: EVIDENCE FROM US LABOR MARKETS 7–9 (2017), https://papers.ssrn.com/abstract=2940245 (last visited Jul 19, 2017).

[118] Bill Gates, THE ROBOT THAT TAKES YOUR JOB SHOULD PAY TAXES (2017), https://qz.com/911968/bill-gates-the-robot-that-takes-your-job-should-pay-taxes/ (last visited Jul 24, 2017); Why taxing robots is not a good idea, THE ECONOMIST (2017), https://www.economist.com/news/finance-and-economics/21717374-bill-gatess-proposal-revealing-about-challenge-automation-poses-why-taxing (last visited Jul 24, 2017).

[119] ACEMOGLU AND RESTREPO, *supra* note 117 at 1.

[120] Frey and Osborne, *supra* note 116 at 291–294; Manyika et al., *supra* note 116; World Development Report 2016: Digital Dividends, *supra* note 116 at 108–109.

sizable impact on wages and employment/population ratios.[121] As the use of robots in workplaces increases, the aggregate effect on employment and wages is expected to increase.[122]

The pace of change and the sheer extent of displacement caused by the effects of automation, robots and AI on wages and employment will be unprecedented. Workers will be marginalized and forced to upskill to find work.[123] The World Bank has cautioned that public regulators are 'in a race between skills and technology', and for many skills 'people are losing the race'.[124] At least part of the answer is to reform education and training, but as the World Bank has observed, these types of reforms have such a long lag time until they can prove effective that targeted educational reforms must begin early, and in youth.[125] Therefore, a regulatory response needs to consider support for education, training, and the process of transitioning displaced workers through the process of job disruption and reemployment.[126] Public regulators ought to influence education and training agendas now to ensure the development of resilient, transferrable skills, not easily automated, that lend themselves to a lifetime of working with and adapting to technological change.[127] Longer working lifetimes and the pace of technological development may see low-skilled workers experience this type of job disruption more than once.

In this Part we have outlined examples of problems associated with current applications of AI systems that will provoke a regulatory response. The examples provided illustrate concrete problems and the possibility of far greater, even existential, problems if the development of AI is left unattended. 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. However, we argue for a considered, principled and consultative approach.

III. THE DIFFICULTY IN REGULATING AI

Even in the simplest of industries, 'regulation is extraordinarily difficult'.[128] When considering the regulation of new technologies, former justice of the High Court of Australia, Michael Kirby noted that 'the normal organs of legal regulation

[121] ACEMOGLU AND RESTREPO, *supra* note 110 at 36 where it is observed that this impact will only be marginally correlated with the more usual effects of imports, other technologies, and the natural attrition of 'routine jobs'.

[122] *Id.* at 35–36.

[123] World Development Report 2016: Digital Dividends, *supra* note 116 at 130.

[124] *Id.* at 131.

[125] *Id.* at 131.

[126] Manyika et al., *supra* note 116 at 1.

[127] World Development Report 2016: Digital Dividends, *supra* note 116 at Chapter 2.

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

often appear powerless'.[129] Further along that continuum, regulating the development of AI may yet be the hardest task for regulators to tackle.

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.[130] However, the real risks of AI may yet be unknown and perhaps are unknowable. This necessarily makes them difficult to evaluate for the purposes of risk assessment which involves balancing a range of social attitudes and will often reflect the culture and values of the society in which it is deployed. However, it is clear that the variety of applications of AI in operation today poses a range or a spectrum of risks.

A. The Range of Risks Associated With AI and AGI

In Part II we outlined a range or spectrum of classes of AI – from narrow AI through to AGI. However, the level of risk associated with the applications within each class does not directly correlate to the class. The applications of AI within each class could pose a range of risks that might range from low to moderate to high. Further, an application of AI in the narrow class may have the potential to become stronger as the AI learns or develops. Whether that AI could then develop into AGI and thus pose a greater risk is often unknowable. Still further, the *type* of risk posed by each application may not be the same within each class of AI. For example, with a particular application of AI there might be a low risk to safety or to human life, but a high risk of a breach of privacy, or a high risk of causing unemployment. Therefore, it is too simplistic to merely take a class of AI such as narrow AI and to seek to regulate it based upon a presumed level of risk. A still further complicating factor is that similar types of application will be used differently in different industries or areas. For example, the same narrow AI application used in a product in the aviation industry may be applied to a different product in an agricultural setting. This will very likely mean that different regulatory agencies will be required to regulate the same AI but in different applications. Taking this complication one step further, the risk posed by the application's use in the agricultural setting may be lower than when the same AI is applied in the aviation industry. Therefore, the same AI application will have to be treated differently by two separate agencies.

Public regulators must become informed about the AI used in their field, assess the risks posed by the AI application as they are used in the industry in which they operate, and regulate appropriately. Earlier research has acknowledged that the reliability and fidelity of organisations involved ought to be evaluated based on factors including the intended use to which the technology might be put.[131] Armed with a deeper understanding of the industry and the intended use of

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

[130] See Deryck Beyleveld & Roger Brownsword, *Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning*, 4 LAW INNOV. TECHNOL. 35–65 (2012).

[131] 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 Jessica Carlo, Kalle Lyytinen & Richard Boland,

the AI, stakeholders involved in informing the regulatory approach will be better placed to ask the right questions to assuage, or at least contextualise, their concerns about levels of risk. Iterative and cooperative involvement of all stakeholders including public regulators is the key to avoiding the necessity to hastily adopt command and control regulatory action and its unintended consequences.[132] We therefore must consider the type of risk that different classes and types of AI pose – starting with a look at the systemic nature of AI risk that exists even now.

B. Systemic Risk

Not all applications of AI will eventuate in a 'singularity' scale event.[133] However, immediate systemic risk issues are present with existing AI applications.[134] Systemic risk is the embedded risk 'to human health and the environment in a larger context of social, financial and economic risks and opportunities'.[135] Systemic risks exist in an atmosphere of uncertainty and they are not restrained by sector, domain or geography.[136] Assessed at its height, strong AI or AGI presents inherent systemic risk.[137] However, the integrated nature and

Systemic Risk, Information Technology Artifacts, and High Reliability Organizations: A Case of Constructing a Radical Architecture, *in* ICIS 2004 CONFERENCE PROCEEDINGS (2004), http://aisel.aisnet.org/icis2004/56 (last visited Mar 16, 2017).

[132] 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: Cass R Sunstein, *Regulatory Paradoxes*, 57 UNIV. CHIC. LAW REV. 407–441, 413, 418, 441. (1990).

[133] For a fuller 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: Carlo, Lyytinen, and Boland, *supra* note 136 at 686; 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).

[134] See Omohundro, *supra* note 29 at 483. who argues that even a chess-playing robot will be 'dangerous unless it is designed very carefully. Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety'.

[135] Ortwin Renn & Andreas Klinke, *Systemic risks: a new challenge for risk management*, 5 EMBO REP. S41–S46 (2004).

[136] These characteristics of systemic risk are identified and discussed in van Asselt and Renn, *supra* note 133 at 435.

[137] 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).

embeddedness of even narrow AI's deployment into complex, interdependent social, financial and economic systems or networks amplifies the potential for risk impact, particularly where it is deployed in a pervasive way.[138] The more complex and non-linear these networks are, the easier it is for the impacts of an AI 'incident' to proliferate rapidly throughout the network affecting multiple stakeholders.[139] Systemic risks are problematic for regulation. While systemic risks are not unknown to public regulators,[140] the potential size of the connectedness of the network that AI can access is unprecedented. For these reasons, it is unlikely that command and control models of regulation would be effective to regulate systemic risk.[141]

According to van Asselt and Renn, systemic risk should 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'.[142] Then the public regulator, business and society can ensure that there are 'early warning systems' are in place to detect risk if it

[138] van Asselt and Renn, *supra* note 133 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).

[139] van Asselt and Renn, *supra* note 133 at 436; Carlo, Lyytinen, and Boland, *supra* note 131 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 137 at 234.

[140] Carlo et al consider for example management of risk hazards associated with nuclear facilities. See: Carlo, Lyytinen, and Boland, *supra* note 131 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 137; 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).

[141] Neil Gunningham & Darren Sinclair, *Smart regulation*, *in* REGULATORY THEORY: FOUNDATIONS AND APPLICATIONS 133–148, 142 (Peter Drahos ed., 2017).

[142] van Asselt and Renn, *supra* note 133 at 438.

eventuates.[143] Then public regulators could initially develop agreed principles that synthesise those things that need to be considered before formulating the processes necessary to govern those risks.[144]

In the regulation of AI, the mix and interplay of stakeholders will be important in the formulation of 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 in this area. The necessarily diverse mix of stakeholders and heterogeneous interests may make unified agreement on principles difficult.[145] Those charged with developing principles will need to consider not only the technological and scientific concerns but also a range of societal norms and social and economic considerations.[146] Settling on a set of principles will involve an element of trust in the science and technology. Creating a culture of iterative and cooperative development could engender this trust. Progress could be smoothed by a culture of fidelity and transparency from those with technical knowledge and scientific expertise in AI. Even if fuller information is available to public regulators, it will still be difficult to know everything necessary to regulate effectively because of the opacity of the algorithms that are not transparent on their face and are said to reside in an impenetrable black box.[147]

C. The Risks of Failing to Regulate must be evaluated against the Risks Involved in Regulating AI

Some academics have proposed that to avoid legal uncertainty and to avoid the difficulties associated with regulating AI we should merely adapt existing liability regimes.[148] 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.[149] So, for example, a judicial

[143] *Id.* at 439.

[144] van Asselt and Renn, *supra* note 133; Principles suggested include communication and inclusion, integration, and reflection. See: ORTWIN RENN, RISK GOVERNANCE: COPING WITH UNCERTAINTY IN A COMPLEX WORLD (2008); Bridget M Hutter, *Risk, Regulation and Management*, *in* RISK IN SOCIAL SCIENCE 202–227, 214–215 (2006); Carlo, Lyytinen, and Boland, *supra* note 131 at 686.

[145] Note Carlo et al, have made similar observations in their research on high reliability organizations: Carlo, Lyytinen, and Boland, *supra* note 131 at 694.

[146] Carlo, Lyytinen, and Boland, *supra* note 131.

[147] Crawford has observed that the "algorithmic black box" is compounded by the fact that "algorithms do not always behave in predictable ways". See: Crawford, *supra* note 80 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, political and institutional contexts". See: Karppi and Crawford, *supra* note 139 at 74.

[148] Allenby, *supra* note 11 at 20–21.

[149] See, e.g., Diana M. Bowman, *The hare and the tortoise: an Australian perspective on regulating new technologies and their products and processes*, *in* INNOVATIVE GOVERNANCE MODELS FOR EMERGING TECHNOLOGIES (Gary E. Marchant, Kenneth W Abbott, & Braden R. Allenby eds., 2013), http://ebookcentral.proquest.com/lib/qut/detail.action?docID=1569419; Kaal, *supra* note 25.

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' – may adequately deal with concerns about potential harm caused by autonomous cars. 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.[150] 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. Those implementing and enforcing the laws could avoid much of the uncertainty surrounding new regulatory regimes. We do not subscribe to this light touch method and argue that AI requires a *sui generis* approach such as is outlined in this paper.

Entrepreneurs and technological innovators maintain a healthy fear of regulation, which is often seen as red tape that hinders or stymies development.[151] Thierer, for example, argues for what he terms 'permissionless innovation'; that 'unless a compelling case can be made that a new invention will bring serious harm to society, innovation should be allowed to continue unabated'.[152] Kurzweil too argues against regulation – preferring a free-market system to deal with problems if and when they arise. However, he does this while simultaneously urging caution.[153] Technology-rich industries have a long history of seeking to avoid the impulse to regulate that often accompanies widespread social fears about new technologies.[154] 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. 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.'[155] Risk analysis too generally involves striking a balance, and the promise of AI may make taking some risk worthwhile.

However, the calls for innovation without any regulation must be viewed critically. In Part II of this paper we provided a number of concrete examples of potential and existing problems and risks that current applications of AI developed without regulation pose for society. We also argued that there is at least the potential for AI development to cause harm to humanity and society. Arguing that regulation necessarily stymies innovation is a syllogistic fallacy; not all regulation stymies innovation. There are enough problems already with relatively narrow AI to persuade regulators that some regulation may indeed be necessary. While regulation may be difficult and may meet resistance from the industry, it is important that we as a society begin to consider the regulation of this vital area. We take up the challenge of contributing to AI research from a legal and regulatory

[150] Ibid.

[151] Adam D. Thierer, *Technopanics, threat inflation, and the danger of an information technology precautionary principle*, 14 MINN. J. LAW SCI. TECHNOL. 309, 317, 339, 375 (2012).

[152] 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).

[153] KURZWEIL, *supra* note 34 at 420.

[154] Thierer, *supra* note 151.

[155] Gurkaynak, Yilmaz, and Haksever, *supra* note 18 at 753.

perspective in this paper. Next in this Part we detail some of the specific problems that regulators face when attempting to regulate in this area.

D. The Problems Posed by AI for Regulators

Scherer set 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'.[156]

These broad categories succinctly capture some of the major problems facing those seeking to regulate AI. Certainly, AI is being developed and deployed in many parts of the world and it is difficult to predict what problems might arise when even two of these powerful technologies are combined. However, while there is the potential for AI development to occur without the need for large scale institutional frameworks such as government agencies, most of the investment in research and development is currently being made by large private companies such as Google, Facebook, Microsoft, Apple and Amazon[157] and this is where major innovations and developments will be likely to occur. This also exacerbates the opacity problem because private companies are apt to maintain secrecy, are not required to share information and in fact are benefitting from the law of patent to protect their legitimate interests in new technology from other developers. However, these broad problems represent only the top layer of issues and much more specific and deeper issues are at play through a deeper analysis.

Scherer also proposed a system under which an agency would be set up to certify AI systems as safe,[158] and where such certified systems would 'enjoy limited tort liability'[159] 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*. In this Part, we outline the potential difficulties associated with this process. As well as the general problems with regulating new technologies outlined above, there are a number of specific problems associated with regulating AI.

1. The Pacing Problem

A particular problem that regulators face is that developments in the technology outpace any attempt at regulating it.[160] In the face of the continuously

[156] Scherer, *supra* note 24 at 369.

[157] We discuss this in more detail in Part IV.

[158] Scherer, *supra* note 24 at 394.

[159] *Id.* at 394.

[160] THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT: THE PACING PROBLEM, *supra* note 10.

increasing speed of innovation, legal and ethical oversight has lagged.[161] This pacing problem plagues the regulation of technology generally and often leads to the technology disengaging or decoupling from the regulation that seeks to regulate it. Because the technology is at the forefront of scientific discovery and is developing so quickly, AI 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 to provide meaningful guidance regarding any specific technology.[162] 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.

2. Information Asymmetry and the Collingridge Dilemma

Private companies are investing heavily in AI research and development. The result is information asymmetries between those companies and public regulators seeking to understand those developments.[163] 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.[164] 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. It is an especial problem for regulators who need to fully understand the subject of regulation.

This information problem forms the first half of the Collingridge Dilemma on the control of technology, 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 the technology is so entrenched in our daily lives that there is a

[161] *Id.*; 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); ROGER BROWNSWORD, RIGHTS, REGULATION, AND THE TECHNOLOGICAL REVOLUTION (2008), http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199276806.001.0001/acprof-9780199276806 (last visited Mar 15, 2017); 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).

[162] THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT: THE PACING PROBLEM, *supra* note 10; VERMEULEN, FENWICK, AND KAAL, *supra* note 161; BROWNSWORD, *supra* note 161; Ray Purdy, *Legal and Regulatory Anticipation and "Beaming" Presence Technologies*, 6 LAW INNOV. TECHNOL. 147–192 (2014).

[163] BROWNSWORD, *supra* note 161; Laurie, Harmon, and Arzuaga, *supra* note 161.

[164] MC Stephenson, *Information Acquisition and Institutional Design*, 124 HARV. LAW REV. 1422 (2011); H BAKHSHI, A FREEDMAN & PJ HEBLICH, STATE OF UNCERTAINTY: INNOVATION POLICY THROUGH EXPERIMENTATION (2011); Gregory N Mandel, *Regulating Emerging Technologies*, 1 LAW INNOV. TECHNOL. 75–92 (2009).

resistance to regulatory change from users, developers and investors.[165] AI has already been deployed in society in a wide variety of fields, from medical diagnostics to criminal sentencing to social media, rendering the need to address this issue even more urgent.[166]

3. Little Established Ethical Guidance, Normative Agreement, or Regulatory Precedent

The ethical and social implications of introducing robots into mainstream society is a very weighty issue that remains largely unresolved, even as the consequences of this interaction are already unfolding.[167] Other areas in which ethical issues arise include the use of military robots, human-robot relationships, such as the use of robots as sex partners, caregivers, and servants.[168] Lin et al. argue that robot ethics issues can be classified in terms of safety and errors; law and ethics; and social impact, and consider the possibility and desirability of programming ethics into AI systems.[169] A regulatory regime for the design and deployment of robots and AI in society must consider the need to include ethics rules in the code that underpins their operation. That system of ethics must reflect a broad normative consensus on what ethical values robots and AI systems should include.

4. Regulatory Delay and Uncertainty

Regulatory delay[170] occurs as regulators consider if and when they will approve the implementation of a new development. For example legislators may pre-emptively ban the commercialisation of new products in response to public concerns, acting even before enough research can be conducted to ascertain whether the concerns are valid or well founded. This delay causes uncertainty for

[165] DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY 11 ff (1980).; Laurie, Harmon, and Arzuaga, *supra* note 161.

[166] ANDREAS MARGELISCH, A STATE OF THE ART REPORT ON LEGAL KNOWLEDGE-BASED SYSTEMS (1999), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.6315&rep=rep1&type=pdf (last visited Nov 17, 2016); Jason Borenstein & Yvette Pearson, *Companion Robots and the Emotional Development of Children*, 5 LAW INNOV. TECHNOL. 172–189 (2013); Angwin et al., *supra* note 78; B.M. Dickens & R.J. Cook, *Legal and ethical issues in telemedicine and robotics*, 94 INT. J. GYNECOL. OBSTET. 73–78 (2006).

[167] Miles Brundage, *Limitations and risks of machine ethics*, 26 J. EXP. THEOR. ARTIF. INTELL. 355–372 (2014); Gordana Dodig Crnkovic & Baran Çürüklü, *Robots: ethical by design*, 14 ETHICS INF. TECHNOL. 61–71 (2012); Perri 6, *Ethics, regulation and the new artificial intelligence, part II: autonomy and liability*, 4 INF. COMMUN. SOC. 406–434 (2001); BERT-JAAP KOOPS, THE CONCEPTS, APPROACHES, AND APPLICATIONS OF RESPONSIBLE INNOVATION. AN INTRODUCTION (2015), https://papers.ssrn.com/abstract=2673753 (last visited Nov 19, 2016).

[168] Lin, Abney, and Bekey, *supra* note 82.

[169] *Id.*

[170] See BAKHSHI, FREEDMAN, AND HEBLICH, *supra* note 164; Mandel, *supra* note 164; Ronald R. Braeutigam, *The effect of uncertainty in regulatory delay on the rate of innovation*, 43 LAW CONTEMP. PROBL. 98–111 (1979); Stephenson, *supra* note 164.

developers.[171] 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 are no longer able or willing to bear the risk.[172] This effect adds to the concerns of developers about regulators seeking to regulate the development of AI.

Of course, some social benefits that may come from innovation and development of AI may well be lost or limited if regulation is implemented prematurely.[173] Sunstein in particular has adverted to the problems associated with adopting what is known as the 'precautionary principle' to regulate risk.[174] People, he argues, are nothing if not 'predictably irrational';[175] and tend to be overly concerned with losses rather than the gains that might be made from, for example, new technology. He argues that regulators should therefore avoid regulating purely on the threat of unknown future risks.[176] However, at the same time, he warns against regulatory inaction 'because a probability of harm is, under many circumstances, a sufficient reason to act'.[177] Ultimately, Sunstein urges that 'a wide variety of adverse effects may come from inaction, regulation, and everything in between', noting the need to 'attempt to consider all of those adverse effects and not simply a subset'.[178] This measured approach should find some favour. After over 60 years of developments in AI, regulation now could not be criticised as being overly reactive or precautionary. As we argued in Part II, as AI development continues apace, some caution in this area is warranted.

5. Coordination across Regulatory Bodies

Coordinating the many regulatory bodies involved in a new technology is a problem that plagues every innovating industry.[179] Already, many groups and industry bodies have developed codes of conduct and standards to regulate the development of AI.[180] Given the increasingly interdisciplinary nature of AI

[171] Braeutigam, *supra* note 170.

[172] Mandel, *supra* note 164; Roberta Romano, *Regulating in the Dark*, *in* REGULATORY BREAKDOWN: THE CRISIS OF CONFIDENCE IN US REGULATION 86–117 (Cary Coglianese ed., 2012), http://www.jstor.org.ezp01.library.qut.edu.au/stable/j.ctt3fhzfx.8.

[173] See THIERER, *supra* note 152; KURZWEIL, *supra* note 34.

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

[175] 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).

[176] Sunstein, *supra* note 125 at 1009.

[177] *Id.* at 1055.

[178] *Id.* at 1056.

[179] JASON POTTS, THE NATIONAL ORIGINS OF GLOBAL INNOVATION POLICY AND THE CASE FOR A WORLD INNOVATION ORGANIZATION (2015), https://papers.ssrn.com/abstract=2705906 (last visited Jul 26, 2017); Lyria Bennett Moses, *Agents of Change: How the Law 'Copes' with Technological Change*, 20 GRIFFITH LAW REV. 763–794 (2011); Stuart Benjamin & Arti Rai, *Fixing Innovation Policy: A Structural Perspective*, 77 GEORGE WASH. LAW REV. 1–88 (2008); Mandel, *supra* note 164.

[180] For example, see the codes of conduct produced by the Partnership on AI and the standards prepared by the Institute of Electrical and Electronics Engineers (IEEE) discussed in Part IV B below.

research, coordinating these industry bodies is no less a challenge for public regulators in this field.[181] An AI regulatory regime would need to account for existing laws, other governmental regulatory bodies, and self-regulatory industry bodies that develop professional codes of ethics, and it needs to do this across many different fields such as neuroscience; neurobiology; mechanical, electrical and software engineering; psychology; innovation studies; and economics and finance.[182] Soft law developments such as industry codes of practice, principles and standards developed by groups of industry participants vary and can often be at cross-purposes. Marchant and Wallach propose that this multiplicity of perspectives and approaches requires an 'issues manager' to oversee and coordinate the various principles, codes and other approaches.[183] Marchant and Wallach have proposed to form a Governance Coordination Committee to 'provide oversight, cultivate public debate, and evaluate the ethical, legal, social, and economic ramifications of … important new technologies'.[184] The current efforts to attempt to govern using these industry-led soft law approaches is discussed further in Part IV.

6. Agency Capture

Regulatory failure due to agency capture occurs where regulators become sympathetic 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.[185] While each of these problems is relatively common throughout innovating industries, the AI industry is particularly at risk of the revolving door issue.[186] The information asymmetry issue where AI companies hold all the relevant information about the technology 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 (and if) they can.

[181] Bennett Moses, *supra* note 179.

[182] Christopher-Paul Milne & Joyce Tait, *Evolution Along the Government - Governance Continuum: FDA's Orphan Products and Fast Track Programs as Exemplars of "What Works" for Innovation and Regulation*, 64 FOOD DRUG LAW J. 733–754 (2009); PETER W. B. PHILLIPS, GOVERNING TRANSFORMATIVE TECHNOLOGICAL INNOVATION: WHO'S IN CHARGE? (2007); POTTS, *supra* note 179.

[183] Gary E. Marchant & Wendell Wallach, *Governing the Governance of Emerging Technologies*, *in* INNOVATIVE GOVERNANCE MODELS FOR EMERGING TECHNOLOGIES 136–152, 142 (Kenneth W Abbott, Gary E. Marchant, & Braden R. Allenby eds., 2014).

[184] See Gary E. Marchant & Wendell Wallach, *Coordinating Technology Governance*, XXXI ISSUES IN SCIENCE AND TECHNOLOGY, 2015, at 43–50.

[185] Thomas O. McGarity, *MTBE: a precautionary tale*, 28 HARV. ENVIRON. LAW REV. 281 (2004); BEN GOLDACRE, BAD PHARMA: HOW DRUG COMPANIES MISLEAD DOCTORS AND HARM PATIENTS (2013).

[186] See Anne Weismann, SILICON VALLEY COMPANIES LOBBY TO REMAIN UNREGULATED THE NEW YORK TIMES (2016), https://www.nytimes.com/roomfordebate/2016/10/24/silicon-valley-goes-to-washington/silicon-valley-companies-lobby-to-remain-unregulated?referer=https://t.co/86PwxwASfJ&nytmobile=0 (last visited Jul 26, 2017).

7. Limited Enforcement Mechanisms and Jurisdiction Shopping

Added to the complexities outlined above, the major players in the development of AI 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 have described 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, highlights the difficulties that might be faced by a regulator in trying to regulate these companies.[187] This idea is further explored in Part IV below.

The fact that the technology is relatively opaque[188] 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 used 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 created a version of their app specifically designed to identify users likely to be regulators and prevent them from accessing the system to investigate concerns or collect evidence.[189]

In Part III, we have outlined various risks associated with AI and broadly grouped applications of AI into 3 classes based upon the risks that each poses. We also highlighted the general and specific difficulties that regulators face when attempting to regulate new technologies, and particularly, AI. In Part IV, we outline how public regulators will need to adopt new strategies to begin to regulate AI as the old strategies lose effectiveness. One strategy will be to regulate based upon the relative risks associated with particular applications of AI.

IV. THE NEED FOR REGULATORY INNOVATION

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; require, but 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. Traditionally, governments had the information and resources that put them in the best position to regulate in most instances. We have seen a long period where government held legislative control of the state. In the field of new technology, at least, the machinery of control is drifting away from government and is becoming decentred. In this Part, we review this decentring of

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

[188] PASQUALE, *supra* note 15.

[189] 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).

regulation and outline some examples of peer or self-regulation that has begun to proliferate in the vacuum of government control.

We review 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 but may still provide some guidance to regulators seeking to approach regulating AI. 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 include. They are also, by definition, 'responsive' and hence presuppose the existence of a regulatory framework. As such, they are best used to guide interactions between regulators and the regulated when regulatory systems are already in place. Further, responsive regulation is limited in its ability to regulate new technologies that exhibit the kinds of characteristics set out in Part III above – it lacks the flexibility required to react quickly enough in such a dynamic field. Further, while much can be learned from regulation of other emerging technologies, the regulation of AI must be sui generis. While in its nascent stages, it will require a more nuanced set of regulatory approaches.

A. 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.[190] It is now widely recognised that there are far more techniques in the regulation toolbox than 'command and control' style rules backed by sanctions.[191] Ayres and Braithwaite's concept of 'responsive regulation' for example sets out a graduated pyramid of interventions by the state in policing behaviour, in order to encourage and direct an optimal mix of regulatory work by private and public entities.[192] The responsive element of responsive regulation is that, as the regulatory response moves up the pyramid, 'escalating forms of government intervention will reinforce and help constitute less intrusive and delegated forms of market regulation'.[193] That is, responsive regulation still requires government to assert a 'willingness to regulate more intrusively' and by so doing can guide the regulation where it is most effective, mostly through 'less intrusive and less centralised forms of government intervention'.[194] 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. It is still government that maintains the ability and responsibility to ultimately regulate if required.[195] The threat relies on government's ability to inflict varying degrees of discretionary punishment or other forms of persuasion within the pyramidal structure if there is the regulated entity fails to comply with initial regulatory

[190] THOMAS HOBBES, LEVIATHAN (2006); JOHN AUSTIN, THE PROVINCE OF JURISPRUDENCE DETERMINED (2nd ed. 1861).

[191] Julia Black, *Critical reflections on regulation*, 27 AUSTL J LEG PHIL 1, 4 (2002).

[192] IAN AYRES & JOHN BRAITHWAITE, RESPONSIVE REGULATION (1994).

[193] *Id.*

[194] *Id.* at 4.

[195] *Id.* at 5.

attempts — this is referred to as tit-for-tat approach.[196] The critical effect of responsive regulation was to highlight developments in alternative means of regulating other than command and control – and therefore avoid some of the more problematic effects of blunt regulatory tools. It appears, however, to be a tool that is still too blunt to hone new technologies such as AI. Part of the reason for this is that the traditional role of government has diminished over time.

1. The Traditional Role of Government in Regulation

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, referred to the government's central position as a receiver and distributor of information that allows it access to and control of the full range of information.[197] Governments held a strategic position with nearly full information about the area and topic of regulation.[198] Authority refers to the authority of the government to determine what is legal[199] and to 'demand, forbid, guarantee, [and] adjudicate'.[200] Treasure refers to the government's assets both in money and other tangible assets such as buildings and equipment which gives it the power to control development at the time and place of its choosing.[201] Organisation refers to the resources in people employed by government with the knowledge and skills to be able to carry out any required task. This includes '(soldiers, workers, bureaucrats), land, buildings, materials, computers and equipment'.[202] The interaction of these roles traditionally held by government simplifies analysis of the role of government in regulation.[203] When these theories are applied to the difficult tasks of regulating AI, the challenges that regulators face are clearly visible. In these contexts, the government's nodality, authority, treasure and organisation have been depleted or usurped and are not always sufficient to match that of the major technology companies such as Google, Facebook, Microsoft and Apple. 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.

Similarly, in the context of regulating AI, the emphasis of responsive regulation in 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, for example, but we argue, that responsive regulation 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. The notion of government as at the apex of power

[196] See generally *Id.* at 38–41.

[197] CHRISTOPHER C HOOD & HELEN Z MARGETTS, THE TOOLS OF GOVERNMENT IN THE DIGITAL AGE 5 (2007).

[198] *Id.* at 6.

[199] *Id.* at 6.

[200] HOOD AND MARGETTS, *supra* note 197.

[201] *Id.*

[202] *Id.*

[203] *Id.* at 12.

structures is arguably no longer applicable, if it truly ever was solely the case. This is especially so when considering the power, global reach and diffuse company structures of companies operating at ‘Google scale’.

2. De-Centred Regulation

Over the last three decades, regulatory scholars and regulatory agencies have been grappling with the ‘decentring’ of regulation, and with it, a recognition that regulation is not the exclusive work of states, and state power to command obedience.[204] As Black contends, a ‘decentred understanding’ of regulation involves ‘complexity, fragmentation of knowledge and of the exercise of power and control, autonomy, interactions and interdependencies, and the collapse of the public/private distinction’.[205] The hallmarks of ‘decentred’ regulation she argues are that it is ‘hybrid (combining governmental and non-governmental actors), multi-faceted (using a number of different strategies simultaneously or sequentially), and indirect’.[206] The current environment surrounding the development of AI shows that if regulation of it is to succeed, that regulation must evolve in an environment that displays these characteristics. Those regulating in this field need to understand and work within these parameters. 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.[207]

Regulators must address the challenges of regulating with limited resources. These resource constraints have curbed the impact of regulatory bodies in general. However, they are particularly debilitating in the context of new technologies that involve a steep learning curve and require regulatory bodies to engage deeply in the industry. Regulatory agencies that seek to regulate AI in this environment should first seek to engage with and work with the relevant actors to learn about and grapple with the complexities in the field. By doing this, they can begin to understand the motivations of the relevant players so that they might start to influence the direction AI development will take. This process, as Black recommends, will involve recurring loops of discussion and feedback where effective ideas are fostered, and redundant notions are jettisoned. Public regulators faced with resource constraints must do this while also managing a shifting regulatory environment where regulators are subject to pressure from interest groups and citizens to pursue conflicting agenda and must also consider how regulation of AI might affect regulatory work in other fields and industries. Regulators must also be able to reflect on the effectiveness of their strategies, often

[204] Black, *supra* note 22.
[205] Black, *supra* note 191 at 8.
[206] Black, *supra* note 22 at 111.
[207] *Id.* at 111.

in an information vacuum, and be able to change strategies when one approach does not work, is ineffective, or even retrograde.[208]

B. Self-Regulation and Peer Regulation

One result of decentred regulation is that governments that once held a central position of power and influence have ceded some of that influence and power to a dissipated group of regulatory participants that would ordinarily be the subject of regulation. Where political influence and power exists in those industries, self-regulation evolves and becomes the default position. In recent years prominent figures from within the AI industry have begun to warn about the need to ensure that the development and deployment of AI technology is effectively regulated.[209] In the absence of government led intervention in the industry, those within the industry are regulating themselves. This is not typical self-regulation under the auspices of a formal government agency, but is self-regulation in a vacuum of input from government.

Some academics have described a kind of self-regulation where the influence of corporate peers guides the behaviour of industry participants.[210] Jessop describes a system of governance that limits the role of regulatory bodies and emphasises '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'.[211] This appears to describe what is happening in practice in relation to AI. 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.[212]

This bottom up development is happening now in the development of AI. Prominent industry participants have developed several codes of conduct and practice, and standards already and the next phase of coordinating these strategies has begun, all outside the auspices of government control.

The challenges of regulating fast moving technology are so great that industry self-regulatory approaches are often presented as the most effective mechanism to

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

[209] See Omohundro, *supra* note 38; Stuart Russell, Daniel Dewey & Max Tegmark, *Research Priorities for Robust and Beneficial Artificial Intelligence*, AI MAGAZINE, 2015, at 105–114.

[210] 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).

[211] *Id.*

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

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 what they might describe as 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',[213] 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'.[214] 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.

Another industry-led initiative to attempt to regulate AI was developed at the Future of Life Institute's Asilomar conference in January 2017. The 23 Asilomar principles as they are known are grouped under three headings: research issues, ethics and values, and longer-term issues. Principles falling within the longer term issues include principle 22 titled 'Importance'. It states that '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'.[215] These principles reflect concerns that even those within the industry hold about the development of, what from its description is, AGI. The Asilomar principles contain a similar basket of issues that are reflected in other industry codes or values statements in relation to AI. While they express well-meaning principles of behavior, quite who will enforce these control measures and what sanctions may be levied for their breach is uncertain.

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'.[216] 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

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

[214] *Id.*

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

[216] 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).

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.[217]

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.[218] 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.*[219] 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. In a classic statement of self-regulation that usurps the traditional role of public regulators, the Partnership on AI has stated that it will continue to pursue 'ongoing engagement … to bring stakeholders together, create best practices, share findings and insights, and to contribute to charting a path forward.'[220] Perhaps, given the government's retreat from regulating in this area, the Partnership on AI may be best placed to continue this work for the time being.

It may well be that self-regulation will be effective in mitigating the most important risks of the development and deployment of AI systems. However, there is also a risk that self-regulation may not be sufficient.[221] First, industry codes or

[217] *Id.* at 5.

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

[219] 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, *supra* note 40; 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).

[220] Partnership on AI, Ibid.

[221] 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*

principles are not obligatory. The principles or codes are often drafted broadly as vision or values statements that do not contain any mandatory requirements but are guides to practice that may be ignored. Second, they lack effective enforcement regimes. Even if they do contain some element of obligation, participants may lack the will to enforce those obligations. Third, many different suggested approaches such as the IEEE standards or the principles proposed by the Partnership on AI or the Asilomar principles vary in their content and focus and lack a central governing body that will coordinate direction and compliance.[222]

Certainly, it will be important to avoid regulation that is ineffective or unduly stymies research and development. we suggest that governments need to consider and engage with the concerns and risks associated with AI now in order to protect public interests that industry-led regulation is not well suited to addressing.[223]

C. The Evolving Nature of Regulation

Despite the efforts of those within the industry to self-regulate, the task of regulating the development and deployment of AI is 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

the Regulator, 24 J. 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: Ayres and Braithwaite, *supra* note 180; 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).

[222] See Marchant and Wallach, *supra* note 184.

[223] Black, *supra* note 22 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 Britain: The Changing Face of Self-Regulation*, 67 Public Adm. 435–454 (1989); Baldwin, Cave, and Lodge, *supra* note 221 at 259–280.

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 and promote interdisciplinary AI research initiatives that look at AI systems' impact from multiple perspectives, combining the computational, social scientific, and humanistic.[224]

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'.[225] The report calls on lawyers to be 'part of the discussions on regulation, governance, and domestic and international legislation in these areas.'[226]

Russell, Dewey and Tegmark set out two policy questions they argue need to be addressed by regulators, academics and those in the industry: '(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?'[227] They proposed that the qualities that these policies should 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'.[228] It appears inevitable that there will eventually be some form of regulation of AI. The European Union has begun to develop Civil Law Rules on Robotics[229] that will ultimately govern the development of robotics and AI in Europe. We discuss this further in Part V.

In Table 1 below we set out some of the major theories of regulation that have evolved over the last two decades. Regulatory theory has developed from the prominent but increasingly less influential command and control style to more and more nuanced and adaptive approaches as increasingly complex situations have

[224] 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).

[225] INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, STANDARDS ASSOCIATION, *supra* note 216 at 89.

[226] *Id.* at 89.

[227] Russell, Dewey, and Tegmark, *supra* note 209 at 107.

[228] *Id.* at 107.; Other principles of good governance that might be added to this list include that they should be "participatory, consensus oriented, accountable, transparent, responsive, effective and efficient, equitable and inclusive and follows the rule of law": see ECONOMIC AND SOCIAL COMMISSION FOR ASIA AND THE PACIFIC, WHAT IS GOOD GOVERNANCE? (2009), http://www.unescap.org/sites/default/files/good-governance.pdf (last visited Jul 27, 2017).

[229] See REPORT OF THE EUROPEAN PARLIAMENT PLENARY SITTING JAN 21 2017 WITH RECOMMENDATIONS TO THE COMMISSION ON CIVIL LAW RULES ON ROBOTICS, (2017), http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0005+0+DOC+PDF+V0//EN (last visited Jul 27, 2017).

demanded. 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':[230] a 'tit-for-tat' approach to enforce compliance by persuasion and education before escalating up a 'pyramid' of more punitive sanctions.	Braithwaite, 2011: 1. Think in context 2. Listen actively (build commitment with stakeholders) 3. Engage with fairness 4. Praise those who show commitment 5. Signal a preference for support and education 6. Signal a range of escalating sanctions that may be used if necessary 7. Engage a wider network of partners as regulatory responses increase in severity 8. Elicit active responsibility from stakeholders where possible 9. Evaluate regulations and improve practices.[231]
'Smart regulation'	• Prefer a mix of regulatory instruments while avoiding 'smorgasboardism'; • Prefer less interventionist measures • Escalate up a pyramid of sanctions when required (responsive regulation) • Empower third parties to act as surrogate regulators • Maximise opportunities for win-win outcomes by encouraging businesses to go 'beyond compliance'[232]
'Risk-based regulation'	Hampton Review:[233] • Regulators, and the regulatory system as a whole, should use comprehensive risk assessment to concentrate resources on the areas that need them most;

[230] Ayres and Braithwaite, *supra* note 192.

[231] John Braithwaite, *The Essence of Responsive Regulation*, 44 UBC Law Rev. 475–520 (2011).

[232] 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).

[233] 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).

	• Regulators should be accountable for the efficiency and effectiveness of their activities, while remaining independent in the decisions they take; • 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; • No inspection should take place without a reason; • Businesses should not have to give unnecessary information, nor give the same piece of information twice; • The few businesses that persistently break regulations should be identified quickly, and face proportionate and meaningful sanctions; • Regulators should provide authoritative, accessible advice easily and cheaply; • 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; • 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 • 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 2. Define the problem precisely 3. Determine how to measure impact 4. Develop solutions or interventions 5. Implement the plan with periodic monitoring, review, and adjustment 6. Close project, allowing for long-term monitoring and maintenance.[234]
‘Really Responsive regulation’	Regulators should be responsive to: • firms’ compliance responses (Responsive regulation); but also • the ‘attitudinal settings’ (operating and

[234] Malcolm K. Sparrow, The regulatory craft: controlling risks, solving problems, and managing compliance 142 (2011).

	cognitive framework of the target of regulation) • the institutional environment • the 'logics of different regulatory strategies and tools' • the regulatory regime's own performance and effects • change in priorities, circumstances, and objectives.[235]
'Really Responsive Risk-based regulation'	In applying risk-based regulation, regulators should: • be responsive to regulated firms' behavior, attitude, and culture; institutional environments; interactions of controls; regulatory performance; and change; • take attitudinal matters on board • identify how attitudes vary across regulatory tasks; • 'be clear about the degree to which any particular regulatory task can and should be guided by a risk-scoring system'[236] Risk-based regulation must focus on: • '*detecting* undesirable or non-compliant behaviour, • *responding* to that behaviour by developing tools and strategies, • *enforcing* those tools and strategies on the ground, • *assessing* their success or failure, and *modifying* them accordingly'.[237]

Table 1 outlines a number of theories that describe traditional methods of regulation. While no one theory would apply as a whole to the regulation of AI, a risk-based approach in combination with several of the elements of Really Responsive Regulation and Smart Regulation may ultimately prove effective. Black and Baldwin's Really Responsive Risk-Based Regulation is perhaps the closest to this approach. It requires regulators to be responsive to 'regulated firms' behavior, attitude, and culture; institutional environments; interactions of controls; regulatory performance; and change'. However, Black and Baldwin could not have

[235] Baldwin and Black, *supra* note 208.

[236] Julia Black & Robert Baldwin, *Really Responsive Risk-Based Regulation*, 32 LAW POLICY 181–213, 193 (2010).

[237] *Id.* at 187. (emphasis in original).

foreseen the changes in the AI environment that would occur subsequent to 2010 when their article was written. The levels and speed at which responses to all of these elements in the AI environment would make it difficult for these responsive regulatory theories to adequately respond in good time. The proliferation of AI into our daily lives has been fast and furtive. The consequence is that public regulators will need to be even more responsive in the forms Black and Baldwin suggest. Perhaps a Really *Really* Responsive Risk-Based Regulation would be required. However, we suggest a more nuanced approach is required in Part V.

The regulation of AI requires a theory of regulation that is not bound by the normative straits in which the theories above evolved. Those theories detail a normative approach to regulation. They presuppose a regulatory environment already being in place and the ability of the government to impose its control if ultimately required to. This, as we have argued, is no longer the case. However, the main problem that each theory faces when it comes to new technologies such as AI is that the mechanisms to respond to change are too slow. They require the machinery of the state to respond to changes in the regulatory environment, but that machinery is not easily engaged and, when engaged, responds too slowly.

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: • Cautious • Macroscopic • Incremental • Experimental • Contextual • Flexible • Provisional • Accountable • Sustainable[238]
One Hundred Year Study on AI	Government should: • Accrue greater technical expertise in AI • Remove impediments to research on the social impacts of AI • Increase public and private funding for research on the social impacts of AI • Resist pressure for 'more' and 'tougher' regulation that stifles innovation or forces innovators to leave the jurisdiction • Encourage a 'virtuous cycle' of

[238] Whitt, *supra* note 23.

	accountability, transparency, and professionalization among AI developers • Continually re-evaluate policies in the context of research on social impacts[239]
Whitehouse report – *Preparing for the Future of Artificial Intelligence*	Regulatory agencies should: • Recruit and develop technical expertise in AI • Develop a workforce with 'more diverse perspectives on the current state of technology' • Use risk-assessment to identify regulatory needs • Avoid increasing compliance costs or slowing development or adoption of beneficial innovations where possible • Avoid premature regulation that could stifle innovation and growth[240]
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: • Embed diagnostic monitoring and evaluation into regulatory programmes at the outset • Collaborate closely with private firms and non-governmental actors • Share and compare results of policy experimentation with other jurisdictions[241]

Table 2 lists a set of strategies rather than broad theories. They are more practically applicable than theoretical. In that vein, many scholars have suggested

[239] Stone et al., *supra* note 59.

[240] 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 40.

[241] 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).

specific regulatory tools that may be useful in regulating new technologies such as AI which include:

- Enhancing flexibility through 'temporary regulation' by using 'experimental legislation'[242] and sunset clauses to 'define adaptable goals and enable the adjustment of laws and regulations according to the evolution of the circumstances'.[243]
- 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'.[244]
- Developing 'anticipatory rulemaking'[245] 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.'[246]
- Making increased use of data analysis to identify what, when, and how to regulate;[247]
- 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;[248]
- 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;[249]
- Creating new multi-stakeholder fora to help overcome information and uncertainty issues that stifle innovation or inhibit effective regulation.[250]

While there is no shortage of suggested regulatory responses, it is hard to distil a clear set of concrete recommendations from the wide and varied literature. This may partly be due to the disparate nature of AI including the definitional problems outlined in Part II. Ultimately, one of the key problems is that while there are common regulatory challenges across different areas of innovation and technology policy, there are also highly context-specific challenges.[251] Ensuring regulatory approaches are closely connected with their context requires individual

[242] (Recommending engaging in policy and regulatory experiments to compare different regulatory regimes and embracing 'contingency, flexibility, and an openness to the new') VERMEULEN, FENWICK, AND KAAL, *supra* note 161.

[243] RANCHORDÁS, *supra* note 25 at 212; see also Romano, *supra* note 172 (discussing regulation of financial markets).

[244] VERMEULEN, FENWICK, AND KAAL, *supra* note 161.

[245] Kaal, *supra* note 25.

[246] 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).

[247] *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 246.

[248] Scherer, *supra* note 24.

[249] Laurie, Harmon, and Arzuaga, *supra* note 161.

[250] Mandel, *supra* note 164.

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

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)'.[252] Brownsword concluded 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'.[253]

The 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 practical and innovative means to begin the process of regulating 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,[254] and involve less intrusive and direct guidance from public regulators – what might be termed a nudge.

V. STRATEGIES TO REGULATE AI

In Part IV we outlined a number of theories of regulation and detailed some of their deficiencies when it comes to regulating AI. We also outlined some theories of regulation that may not be applicable to regulating AI. In this Part, we argue that, in the lag time it takes to properly devise an appropriate regulatory structure to address AI, public regulatory bodies should begin to exert their influence on the nascent development of AI so as to broadly guide its development in beneficial ways. We then suggest that public regulators should begin to develop risk-based strategies to most effectively target their limited regulatory resources. Regulating the risk profile for AI outlined in Part III requires a staggered approach where the highest risks, as assessed by public regulators, are addressed first. At the very least, regulators should be taking steps now to establish what risks pertain to what class and type of AI and be in a position to regulate if eventually required. However, as governments have so far shown an inability to engage with regulation in this area, we suggest that there is a broader and more immediate role for the state in influencing the development of AI systems, but that doing so well will require some innovation in regulatory practices. We suggest that this can be done immediately while the harder and more onerous task of preparing risk profiles can happen over a longer term. The recent Stanford Report recommended a 'vigorous and informed debate' to 'steer AI in ways that enrich our lives and our society'.[255]

[252] Brownsword, *supra* note 16 at 27.

[253] *Id.* at 32.; Brownsword and Yeung, *supra* note 251 at 6.

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

[255] STONE ET AL., *supra* note 59 at 49.

How government regulators may actually be able to steer AI development, however, is a crucial and, as yet, unanswered question. In this Part, we consider how public regulatory agencies may be able to adopt strategies to 'nudge'[256] the development of AI. In this way, 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.

A. The Influence of Regulators, or Nudging

Much has been made of nudge theory in recent years.[257] Psychological observations as applied in behavioural economics reveal that normative human behavior can be skewed or distorted by inherent human biases. Nudge theory proposes that by exploiting these biases, human behavior can be nudged to behave in a way so as to achieve an outcome desired by the nudger. The theory 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.[258] In an example used in a study of environmental policy-making, Weber argued that 'decisions could be reframed in ways that might affect choices by changing the focus of such decisions from individuals to groups'.[259] She argued 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'.[260] Weber argued further that 'the goal of environmental policy is to change the 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.'[261] Weber concluded 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.'[262] These regulatory interventions apply the idea of nudging in its broadest sense. It is not only the behavior of the individual that can be the target of behavioural policy-making. The theory can be used to influence those who govern companies such as boards of directors. In this way regulatory policy can shape the behaviours of companies and, even more broadly, groups of companies within industries.

In Weber's example, policies are directed to influence the environmental responsibility of companies. We argue that similar broad policies directed at

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

[257] *Id.*

[258] Saurabh Bhargava & George Loewenstein, *Behavioral Economics and Public Policy 102: Beyond Nudging*, 105 AM. ECON. REV. 396–401 (2015); Brigitte C. Madrian, *Applying Insights from Behavioral Economics to Policy Design*, 6 ANNU. REV. ECON. 663–688 (2014).

[259] 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).

[260] *Id.* at 388.

[261] *Id.* at 391.

[262] *Id.* at 391.

companies developing AI would begin to influence or guide beneficial behaviours in those companies. If governments are unable yet to fully participate in gathering information because of resource constraints or because of the diffuse nature of AI development, it can begin to shape the behavioural environment by proposing policy statements that foster beneficial and benign development of AI. This approach has several immediate benefits for public regulators. It is relatively inexpensive; it does not require a great deal of investment to be able to set broad policy indicators that outline the regulators' attitude to AI development. It also would buy the regulator time to take on the task of fully engaging with the regulatory environment as outlined in this paper.

B. Examples of Influence as Regulation

The approach of the United States government in attempting to shape behaviours of those developing AI is in its infancy. As an early indicator, the government has shown that it was, until recently, prepared to consult with groups of stakeholders. In 2016 its Office of Science and Technology Policy (OSTP) 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.[263] As a further signal of its policy intention that is designed to shape behaviours, the United States government also published two documents: a *National Artificial Intelligence Research and Development Strategic Plan*, and a *National Artificial Intelligence Research and Development Strategic Plan*.[264] 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.[265]

The *Preparing for the Future of Artificial Intelligence* report made 23 recommendations on what government agencies, schools and universities, and AI professionals could do to prepare for the future of AI. This document on its own had the effect of engaging with and shaping or influencing the development of AI. Evidence for the immediate impact that the report had includes that it was adopted by Partnership on AI.

These strategies 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, input and participation as essential'. He cited the 2012 OECD recommendations on regulatory policy that

[263] EXECUTIVE OFFICE OF THE PRESIDENT NATIONAL SCIENCE AND TECHNOLOGY COUNCIL COMMITTEE ON TECHNOLOGY, *supra* note 40 at 12.

[264] NATIONAL SCIENCE AND TECHNOLOGY COUNCIL AND NETWORKING AND INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT SUBCOMMITTEE, *supra* note 219. Again, this report no longer appears in searches on the whitehouse.gov website.

[265] *Id.*. Again, this report no longer appears in searches on the whitehouse.gov website.

'call for "open government", including transparency and communication, stakeholder engagement throughout the regulatory process and open and balanced public consultations'.[266] Secondly, it could be seen as an information gathering exercise – something that is noted as being a necessary first step in the risk regulation literature as well as in behavioural economics theories. Thirdly, the government could be seen to be signposting its intention to regulate if necessary.

The United States government, by engaging with AI and those responsible for developing it and publishing its stated intentions, sent a clear signal to all those involved in the developing field of AI. It showed that the government was engaged in the conversations and was prepared to stake a claim in game. This also may be seen as the government seeking to influence or nudging decision makers in the AI industry and to shape behaviours within that field. The government's emphasis on beneficial development clearly articulates its intentions and focus and sends a clear signal to the entire industry in the United States and more broadly in the western world. Because many of the companies that develop AI are based in the United States, such a clear policy signal from the United States government would obviously have an influential effect on the behaviours of the major AI companies and the people who work within them.

However, in a worrying development, the United States Government appears to have retreated from its laudable approach to participate in the development of AI. The *Preparing for the Future of Artificial Intelligence* report has been removed from the government website and archived. Similarly, the government's *National Artificial Intelligence Research and Development Strategic Plan*[267] is no longer available online and is presumably no longer government policy. Retreating from its former position sends an altogether different and equally strong message: that regulation is not a priority, is not going to happen in the near future, and the government is uninterested in the development of AI, at least for now. If AI is to be regulated in any meaningful way then, it may well, in the absence of government direction, be up to those developing the AI to control its development. However, this is hardly the ideal solution. It is unfortunate that the planned policy no longer can have the influential effect that it once had.

While the United States government has retreated from its role as influencing the development of AI, the European Parliament has taken positive steps. In February 2017 it passed a resolution to recommend to the European Union Commission to develop Civil Law Rules on Robotics (and included AI).[268] The resolution recommends that the EU adopt rules on liability for issues arising from robots and AI,[269] and also recommends that the EC designate a European Agency for Robotics and Artificial Intelligence to govern robotics and AI. The Agency would:

[266] Abbott, *supra* note 65 at 10.

[267] NATIONAL SCIENCE AND TECHNOLOGY COUNCIL AND NETWORKING AND INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT SUBCOMMITTEE, *supra* note 219. Again, this report no longer appears in searches on the whitehouse.gov website.

[268] See REPORT OF THE EUROPEAN PARLIAMENT PLENARY SITTING JAN 21 2017 WITH RECOMMENDATIONS TO THE COMMISSION ON CIVIL LAW RULES ON ROBOTICS, *supra* note 229.

[269] *Id.* at 20.

> provide the technical, ethical and regulatory expertise needed to support the relevant public actors, at both Union and Member State level, in their efforts to ensure a timely, ethical and well-informed response to the new opportunities and challenges, in particular those of a cross-border nature, arising from technological developments in robotics, such as in the transport sector.[270]

The resolution recommends a system of registration of so-called 'smart robots', the definition of which is wide enough to capture AI. The registration would apply across EU.[271] The resolution also recommends developing a Code of Ethical Conduct for researchers and designers in robotics and AI to 'act responsibly and with absolute consideration for the need to respect the dignity, privacy and safety of humans.'[272] This move by the European Parliament and Commission sends a clear signal to the industry intended to influence the research, development and design of robots and AI, at least in Europe. Once set up, the Agency for Robotics and Artificial Intelligence will begin to gather the technical, ethical and regulatory expertise so needed to begin the regulatory process. This initiative represents the most advanced work towards regulation of AI today and should be lauded as a model for the rest of the world.

There is therefore a place for government policy to shape the behavior of those in the AI industry. At the same time though, more needs to be done to begin the process of developing regulation. As discussed in Part IV, we propose that a Really *Really* Responsive Risk-Based Regulatory framework will be most effective. The risk based regulatory approach will allow regulatory bodies to target their intervention to the most pressing elements of AI development based upon a risk analysis.

C. Risk-based Regulation of AI

In Parts III and IV we outlined a number of risk profiles for various classes of AI. Given the risks posed by AI, it is appropriate that regulation responds to those risks. 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.[273] These tasks are usually carried out by a regulatory agency after consultation with those within the industry.

In the regulation of AI then, public regulators must undertake a risk analysis of current applications of AI. After the regulator has assessed and set the level of risk that it might tolerate, it must gather as much information about the state of affairs as is possible. This can be done by consulting with those already in the industry and participating in or organising information sessions such as roundtables that involve all relevant stakeholders. The risks can only be properly assessed with all relevant information. Only when public regulatory agencies or governments are

[270] *Id.* at 10.
[271] *Id.* at 20.
[272] *Id.* at 21.
[273] See Black and Baldwin, *supra* note 236 at 184–185.

aware of the issues will they be in a position to properly rank the risks that meet or exceed their tolerance levels and to allocate the necessary resources to regulate the risks involved. Our initial triage of risks posed by various applications of AI in Part III could then be refined and further developed in a feedback loop after multiple consultation processes.

Van Asselt and Ren emphasised the need for communication and inclusion when assessing risk. They argued 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.'[274] They emphasised that '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'.[275] The participants, they argue, should include 'a range of actors which have complementary roles and diverging interests'.[276] Hutter also noted that to achieve regulatory excellence, 'regulators must have access to accurate information so that they have a clear idea of the risks they are regulating'.[277] As outlined in Part IV, relevant industry parties are forming industry level associations and groups to share information and agree on principles and shared values. As discussed, this has already resulted in a range of principles and proposed standards by which many in the industry have agreed to be bound. However, government and regulatory bodies must now engage in the process. The United States government in particular, up until recently, had shown that it was willing to take the lead in this information gathering and sharing phase of the regulatory process. It is essential for the government to continue this level of involvement if it is to put itself in a position to be able to regulate effectively. Without such involvement, it will continue to have little influence on the direction that AI development takes. At the same time, regulatory bodies need to begin to assess and rank the various risks associated with AI applications.

D. Classifying the Risks

The high costs of, and challenges to, effective regulatory intervention requires that the attention of regulators should be carefully focused on the areas posing greatest risk. We argued in Part III that different claims to AI can be refined into at the very least three broad subcategories based upon whether it is (a) narrow and single use AI, (b) it displays some characteristics of operating autonomously or may pursue its own goals, or (c) are or display some of the characteristics of AGI. Each of these classes poses different risks and those risks vary within classes depending on the application. Within each category there are many sub-categories of application. Relatively benign applications of AI such as in Roomba, Pandora or simple game applications can be placed within the low risk category. On the next level, we include more robust applications such as the AI in AlphaGo, the more experimental aspects of AI work carried out by Google, Facebook Microsoft, Apple, Amazon, and any other large player experimenting with AI as referred to in Part II above. The third class includes the more concerning aspects associated with

[274] van Asselt and Renn, *supra* note 133 at 440.
[275] *Id.* at 441.
[276] *Id.* at 441.
[277] Hutter, *supra* note 128 at 104.

experiments seeking to attain AGI. This group would include research conducted by mathematicians and engineers who seek to create either a self-replicating AI or AGI without concern or knowledge aforethought for its ultimate capabilities.[278] The problem with regulating AI identified in Part II is that each of these applications could plausibly lay claim to being, applying or using AI. However, each category does not and cannot justify or require the same regulatory response, and some applications may not even require a regulatory response at this stage.[279] It is only when the risk profile of an AI application increases that a regulatory response may be required. For example, more and less risky applications of AI will exist within a single class of AI (for example, narrow AI). However, without a risk analysis, the level of risk of each application within a class is as yet unascertained.

The class of AI that poses the greatest risk to humanity as a systemic risk is AGI. We discuss AGI even though it does not currently exist because it requires an immediate regulatory response, if indeed it is not already too late to regulate its research and development. AI professionals are already experimenting with self-replication and AI autonomy. While these experiments do not yet reach the level of AGI, they remain a very high potential and perhaps imminent risk. If one of these experiments, through accident or serendipity, creates a form of AGI, then the concerns expressed by many in the industry become reality and the chance to control its behaviour may well be lost. Lethal autonomous weapons also pose an extremely high risk to human wellbeing but this sub-category of AI application is subject to its own unique regulatory environment and is outside the scope of this paper.[280]

A further complication in regulating AI using a risk-based strategy arises because none of the risks or classes of AI is static. The level of risk posed by applications within in classes may increase or decrease. Various push and pull factors will move the applications in each class up and down depending on features that either ameliorate or accentuate the risks associated with its use. The risks posed by narrow applications may become stronger and hence may ultimately become AGI. The question for regulators is at what point they should intervene. Should they begin to regulate as soon as AI poses some risk or should they wait until an imminent risk is apparent? A further complication is that, at this stage, relevant regulators are not even in a position to discern which application of AI fits within which class. No clear system of classification currently exists. Our

[278] See James Babcock, Janos Kramar & Roman Yampolskiy, *The AGI Containment Problem*, *in* ARTIFICIAL GENERAL INTELLIGENCE: 9TH INTERNATIONAL CONFERENCE (2016). Paul Christiano, CRYPTOGRAPHIC BOXES FOR UNFRIENDLY AI LESSWRONG (2010), http://lesswrong.com/lw/3cz/cryptographic_boxes_for_unfriendly_ai/ (last visited Apr 28, 2017). Eliezer Yudkowsky & Marcello Herreshoff, *Tiling Agents for Self-Modifying AI, and the Lobian Obstacle. Early Draft. Machine Intelligence Research Institute, Berkeley, CA*, (2013). Orseau and Armstrong, *supra* note 57.

[279] There is some kudos too in being able to advertise to the public that your product uses AI. So while the product may not strictly use AI, as a marketing ploy, manufacturers sometimes claim that their product uses AI: see for example Prakash, *supra* note 39; see also Omohundro, *supra* note 38; Omohundro, *supra* note 29.

[280] For a thorough investigation of this topic including suggestions on possible regulation in the area see Dustin A. Lewis, Gabriella Blum, and Naz K. Modirzadeh, *War-Algorithm Accountability*, (August 31, 2016). Available at https://pilac.law.harvard.edu/war-algorithm-accountability-report/#_ftnref515.

suggestion is to begin classifying based on the level of risk each application currently poses.

Yet a further complication arises because the same public regulator or regulatory agency will not regulate all (or even more than one) of the applications within each class. The classes we have identified are separated broadly by risk factors and not by application type. So, even though they may be on the same class and the same level of risk for the purposes of our classification, the regulators who might respond to concerns raised by the use of AI in autonomous vehicles will not be required to consider Google's use of AI to reduce electricity consumption in its data centres for example.

We suggest that it is the role of governments and regulatory bodies to begin to influence the direction that AI is to take at a broad level and to attempt to intercede now in its development. We have provided suggestions on the role of public regulators as to how this might be done. In the meantime, our initial proposal is for governments or public regulators to take steps towards regulating AI by obtaining information, joining and commencing conversations with stakeholders in the industries that use AI, and influencing the development of AI in ways beneficial to society. We contend that the most dangerous (and as yet unattained) class of AI, AGI, should be regulated now and serious questions about its development should be considered and discussed among AI professionals now.

VI. 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.[281]

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 and will not be easy.

[281] See Alex Wellerstein, THE DEMON CORE AND THE STRANGE DEATH OF LOUIS SLOTIN THE NEW YORKER (2016), http://www.newyorker.com/tech/elements/demon-core-the-strange-death-of-louis-slotin (last visited Jul 27, 2017).

In the last two decades, the face of technology, the institutions involved, and therefore the AI regulatory space has changed dramatically. This period 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.[282] 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 the social impacts of AI from both the United States government and from private coalitions is encouraging. Still, the rhetoric of avoiding over-regulation is worrying – even the biggest and most well-resourced government regulators are hesitant and probably will not be particularly well equipped to deal with this any time soon. For smaller regulators – including those outside of the United States, there is almost no chance of successfully intervening in current technological development. Governments are left to try to influence or nudge the development of AI at the broad policy level. This remains one of the only roles that might remain available to government given the changing power dynamics between government and these large companies.

There are benefits to self-regulation, particularly where public regulators lack the requisite knowledge to understand 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 because many of the major AI companies are based in the United States. Recent studies in behavioural policy making suggest that the attitudinal settings of people within groups shape the development of the group. The government recently set about the task of informing itself about AI and has set out both a strategic and a research and development policy that seeks to influence beneficial development of AI. By setting out its agenda as it has and by investing in collaboration with industry participants, the United States Government had set a positive benchmark that sought to sway participants in the field. Whether this can be called nudging or not is moot, but the intention was clear. However, the government has more recently retreated from this stance; this is regrettable. The latest retrograde steps send an equal and opposite message to AI developers. In a positive sign though, the European Union has taken positive steps toward regulation of robots and AI and other countries might do well to replicate its example.

Because regulators do not yet have the expertise or even enough information to create expertise, if we are ever to ensure AI is developed in a way that is

[282] HOOD AND MARGETTS, *supra* note 197 (discussing resources of nodality, authority, treasure, and organisation).

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. The ethics board set up by DeepMind and Google, and the Partnership on AI are great examples of this. However, the problems that face potential regulators attempting to regulate such a dynamic field illustrate that more collaboration and information sharing between all relevant parties is required if we are to safely reap the benefits of AI.

The risks that different classes of AI pose lie along a spectrum. Similarly, the different applications of AI pose different and variable risks within the field in which they are applied. Public regulators must begin to engage with researchers and professionals in the area to gain the necessary information required to be able to identify and regulate in relation to the greatest risks that AI poses. By adopting a risk-based approach, public regulators will be able to target their approaches to achieve the most efficient and effective regulatory outcomes.