Harvard Journal of Law & Technology
Volume 29, Number 2 Spring 2016

REGULATING ARTIFICIAL INTELLIGENCE SYSTEMS: RISKS, CHALLENGES, COMPETENCIES, AND STRATEGIES

Matthew U. Scherer∗

TABLE OF CONTENTS

∗ Attorney, Buchanan Angeli Altschul & Sullivan LLP. The author thanks George Cooper, Daniel Dewey, Joseph Freedman, Luke Muehlhauser, Geoff Taylor, and Alexander Volokh for their advice and comments on drafts of this article.

I. INTRODUCTION

It may not always be obvious, but we are living in the age of intelligent machines. Artificial intelligence ("AI") permeates our lives in numerous subtle and not-so-subtle ways, performing tasks that, until quite recently, could only be performed by a human with specialized knowledge, expensive training, or a government-issued license. Driverless cars have been approved for road operation in four states and the District of Columbia;[1] their inevitable arrival on the consumer market may revolutionize road transportation. Autonomous machines can execute complex financial transactions, flag potential terrorists using facial recognition software, and (most alarmingly for this author and his legal contemporaries) perform document review.[2] More mundanely, computer chess engines can defeat the strongest human players in the world, and Google Translate can generate passable English translations of *Le Monde* articles. In fact, "robot journalists" may even have written the *Le Monde* articles themselves.[3]

The increasing ubiquity and rapidly expanding commercial potential of AI has spurred massive private sector investment in AI projects. "Firms such as Google, Facebook, Amazon and Baidu have got into an AI arms race, poaching researchers, setting up laboratories and buying start-ups."[4] With each passing month, AI gains footholds in new industries and becomes more enmeshed in our day-to-day lives, and that trend seems likely to continue for the foreseeable future.[5]

---

1. *See* Aaron M. Kessler, *Law Left Behind as Hands-Free Cars Cruise*, STAR TRIBUNE (May 3, 2015, 12:21 PM), http://www.startribune.com/law-left-behind-as-hands-free-cars-cruise/302322781/ [https://perma.cc/39PB-UDJ8].

2. *See, e.g.*, John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, N.Y. TIMES (Mar. 4, 2011), http://www.nytimes.com/2011/03/05/science/05legal.html (May 4, 2016); Timothy Williams, *Facial Recognition Software Moves from Overseas Wars to Local Police*, N.Y. TIMES (Aug. 12, 2015), http://www.nytimes.com/2015/08/13/us/facial-recognition-software-moves-from-overseas-wars-to-local-police.html (May 4, 2016).

3. *See, e.g.*, Yves Eudes, *The Journalists Who Never Sleep*, GUARDIAN (Sept. 12, 2014, 6:17 AM), http://www.theguardian.com/technology/2014/sep/12/artificial-intelligence-data-journalism-media [https://perma.cc/CES7-X58C] (discussing the increasing use of "robot writers" in journalism).

4. *Artificial Intelligence: Rise of the Machines*, ECONOMIST (May 9, 2015), http://www.economist.com/news/briefing/21650526-artificial-intelligence-scares-peopleexcessively-so-rise-machines [https://perma.cc/B2LD-B4XS].

5. *See, e.g.*, Kevin Kelly, *The Three Breakthroughs That Have Finally Unleashed AI on the World*, WIRED (Oct. 27, 2014, 6:30 AM), http://www.wired.com/2014/10/future-of-artificial-intelligence/ [https://perma.cc/Y6N4-WB7B] ("This perfect storm of parallel computation, bigger data, and deeper algorithms generated the 60-years-in-the-making overnight success of AI. And this convergence suggests that as long as these technological trends continue — and there's no reason to think they won't — AI will keep improving."); Mohit Kaushal & Scott Nolan, *Understanding Artificial Intelligence*, BROOKINGS INST. (Apr. 14, 2015, 7:30 AM), http://www.brookings.edu/blogs/techtank/posts/2015/04/14-understanding-artificial-intelligence [https://perma.cc/SQ5W-7Q2P] ("As consumers, we should expect AI technology to permeate all aspects of life within a few short years.").

The potential for further rapid advances in AI technology has prompted expressions of alarm from many quarters, including some calls for government regulation of AI development and restrictions on AI operation.[6] That in and of itself is hardly surprising; fear of technological change and calls for the government to regulate new technologies are not new phenomena. What is striking about AI, however, is that leaders of the tech industry are voicing many of the concerns. Some of the concerns stem from the familiar fears of technological unemployment[7] and the potential for new technologies to be misused by humans.[8] But many of the fears cut much deeper.

In an interview at MIT's 2014 AeroAstro Centennial Symposium, Elon Musk eschewed the skepticism of regulation that characterizes most of Silicon Valley's business titans and suggested that some government intervention might be wise in the case of artificial intelligence:

> I think we should be very careful about artificial intelligence. If I had to guess at what our biggest existential threat is, it's probably that . . . . I'm increasingly inclined to think there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don't do something very foolish.[9]

Other prominent figures in the tech world — most notably Bill Gates and Steve Wozniak — have voiced similar concerns regarding the long-term risks of AI.[10]

---

6. *See, e.g.*, John Frank Weaver, *We Need to Pass Legislation on Artificial Intelligence Early and Often*, SLATE (Sept. 12, 2014, 3:53 PM), http://www.slate.com/blogs/future_tense/2014/09/12/we_need_to_pass_artificial_intelligence_laws_early_and_often.html [https://perma.cc/6SKM-K6RT]; Perri 6, *Ethics, Regulation and the New Artificial Intelligence, Part I: Accountability and Power*, 4 INFO., COMM. & SOC'Y 199, 203 (2010).

7. *Compare, e.g.*, STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 1034 (3d ed. 2010) (discussing how people losing their jobs to automation is an ethical issue introduced by AI), *with* JOHN MAYNARD KEYNES, *Economic Possibilities for Our Grandchildren*, *in* ESSAYS ON PERSUASION 321, 325 (1972) ("For the moment the very rapidity of these changes is hurting us and bringing difficult problems to solve . . . namely, *technological unemployment*. This means unemployment due to our discovery of means of economising the use of labour outrunning the pace at which we can find new uses for labour.").

8. *See* RUSSELL & NORVIG, *supra* note 7, at 1035 (discussion titled "AI systems might be used toward undesirable ends").

9. Aileen Graef, *Elon Musk: We Are "Summoning a Demon" with Artificial Intelligence*, UPI (Oct. 27, 2014, 7:50 AM), http://www.upi.com/Business_News/2014/10/27/Elon-Musk-We-are-summoning-a-demon-with-artificial-intelligence/4191414407652/ [https://perma.cc/M98J-VYNH].

10. *See, e.g.*, Eric Mack, *Bill Gates Says You Should Worry About Artificial Intelligence*, FORBES (Jan. 28, 2015), http://www.forbes.com/sites/ericmack/2015/01/28/bill-gates-also-worries-artificial-intelligence-is-a-threat/ (quoting Bill Gates, "I am in the camp that is concerned . . . . First the machines will do a lot of jobs for us and not be super intelligent.

At the very least, more mundane legal issues surrounding AI seem likely to crop up in the near future. Who (or what) will be held liable when an autonomous vehicle causes an accident? To what degree can physicians delegate the task of diagnosing medical conditions to intelligent scanning systems without exposing themselves to increased liability for malpractice if the system makes an error? Such questions regarding AI-caused harm will arise with ever-increasing frequency as "smart" technologies fan out into an ever-expanding range of industries.

But, as Musk's above-quoted statement suggests, the rise of AI has so far occurred in a regulatory vacuum. With the exception of a few states' legislation regarding autonomous vehicles and drones, very few laws or regulations exist that specifically address the unique challenges raised by AI, and virtually no courts appear to have developed standards specifically addressing who should be held legally responsible if an AI causes harm. There is a similar dearth of legal scholarship discussing potential regulatory approaches to AI.[11] It does not appear that any existing scholarship examines AI regulation through the lens of institutional competence — that is, the issue of what type(s) of governmental institution would be best equipped to confront the unique challenges presented by the rise of AI.[12]

In a way, it is not surprising that the prospect of AI regulation has been met with radio silence from the normally voluble world of legal scholarship. The traditional methods of regulation — such as product licensing, research and development oversight, and tort liability — seem particularly unsuited to manage the risks associated with intelligent and autonomous machines. Ex ante regulation would be difficult because AI research and development may be discreet (requiring little physical infrastructure), discrete (different components of an AI system may be designed without conscious coordination), diffuse (dozens of individuals in widely dispersed geographic locations can participate in an AI project), and opaque (outside observers may not be able to

---

That should be positive if we manage it well. A few decades after that though the intelligence is [sic] strong enough to be a concern."); Peter Holley, *Apple Co-Founder on Artificial Intelligence: "The Future Is Scary and Very Bad for People*," WASH. POST (Mar. 24, 2015), http://www.washingtonpost.com/blogs/the-switch/wp/2015/03/24/apple-co-founder-on-artificial-intelligence-the-future-is-scary-and-very-bad-for-people/ [https://perma.cc/6YRC-QDSG] (quoting Steve Wozniak, "If we build these devices to take care of everything for us, eventually they'll think faster than us and they'll get rid of the slow humans to run companies more efficiently").

11. The scholarship on the related field of law and robotics is somewhat better-developed. *See generally, e.g.*, Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513 (2015); Gabriel Hallevy, *"I, Robot – I, Criminal" — When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses*, 22 SYRACUSE SCI. & TECH. L. REP. 1 (2010); F. Patrick Hubbard, *"Sophisticated Robots": Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 1803 (2014).

12. *See* HENRY M. HART, JR. & ALBERT M. SACKS, THE LEGAL PROCESS: BASIC PROBLEMS IN THE MAKING AND APPLICATION OF LAW lx (1994).

detect potentially harmful features of an AI system). The autonomous nature of AI creates issues of foreseeability and control that might render ex post regulation ineffective, particularly if an AI system poses a catastrophic risk. Moreover, regulation at any stage is complicated by the difficulty in defining what, exactly, "artificial intelligence" means.

This article will advance the discussion regarding the feasibility and pitfalls of government regulation of AI by examining these issues and explaining why there are, nevertheless, some potential paths to effective AI regulation. Part II will examine the characteristics of AI that present regulatory challenges. Some of these challenges are conceptual, such as how to define artificial intelligence and how to assign moral and legal responsibility when AI systems cause harm. Other challenges are practical, including the inherent difficulties in controlling the actions of autonomous machines, which may render ex post regulation ineffective; the related risk that AI systems will perform actions that are unforeseeable to their designers and operators; and the potential for AI to be developed so clandestinely or diffusely as to render effective ex ante regulation impracticable. Despite these challenges, the legal system's deep regulatory toolkit and the already large and ever-increasing role of large corporations in AI development mean that effective AI regulation should nevertheless be possible.

Part III will analyze the competencies of the three major types of government entities — legislatures, agencies, and courts — in terms of regulating AI The democratic legitimacy and freedom to delegate that legislatures enjoy make legislatures the ideal bodies for establishing the guiding principles for AI regulation. Agencies are best suited to determine the substantive content of those regulations due to their relative independence and greater ability to specialize and draw upon technical expertise. Finally, courts are best equipped to allocate responsibility after an AI system causes harm.

In light of these challenges and competencies, Part IV will offer a proposed framework for AI regulation based on differential tort liability. The centerpiece of the regulatory framework would be an AI certification process; manufacturers and operators of certified AI systems would enjoy limited tort liability, while those of uncertified AI systems would face strict liability. The respective roles of the legislature, the executive (specifically, a new AI-focused administrative agency), and courts would be catered to the competencies of each institution with respect to emerging technologies such as AI.

II. THE TROUBLE WITH AI

The increasing role of AI in the economy and society presents both practical and conceptual challenges for the legal system. Many of the practical challenges stem from the manner in which AI is researched and developed and from the basic problem of controlling the actions of autonomous machines.[13] The conceptual challenges arise from the difficulties in assigning moral and legal responsibility for harm caused by autonomous machines, and from the puzzle of defining what, exactly, artificial intelligence means. Some of these problems are unique to AI; others are shared with many other post-industrial technologies. Taken together, they suggest that the legal system will struggle to manage the rise of AI and ensure that aggrieved parties receive compensation when an AI system causes harm.

Section A will discuss potential definitions of artificial intelligence and why coming up with a working definition of AI for regulatory purposes will be difficult. Section B will describe the characteristics that make AI a potential public risk and explain why it will prove more difficult to regulate than earlier sources of public risk. Peter Huber coined the term "public risk" to describe threats to human health or safety that are "centrally or mass-produced, broadly distributed, and largely outside the individual risk bearer's direct understanding and control."[14] Dawn was just breaking on the Information Age when Huber first used the term, and early public risk commentators focused primarily on nuclear technology, environmental threats, and mass-produced physical products. The increasing ubiquity of AI makes it all but certain that AI systems will generate many public risks. Those risks may prove difficult for the legal system to address, because AI presents challenges not raised by the public risks of the twentieth century. Nevertheless, as Section C will explain, the law provides mechanisms that can help reduce the public risks associated with AI even in the face of AI's unique challenges.

---

13. In this article, the term "autonomous machines" refers to machines that "act independently of direct human instruction, based on information the machine itself acquires and analyzes." David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 121 (2014); *see also* Matthew U. Scherer, *Who's to Blame (Part 2): What Is an "Autonomous" Weapon?*, LAW AND AI (Feb. 10, 2016), http://www.lawandai.com/2016/02/10/what-is-an-autonomous-weapon/ [https://perma.cc/668Q-9VWJ] (defining autonomy as the ability of a system to operate free from human direction, monitoring, and control). As with other terms used in this article, *see infra* notes 46–47, the use of the term "autonomy" is not meant to imply that such machines possess the metaphysical qualities of consciousness or self-awareness.

14. Peter Huber, *Safety and the Second Best: The Hazards of Public Risk Management in the Courts*, 85 COLUM. L. REV. 277, 277 (1985).

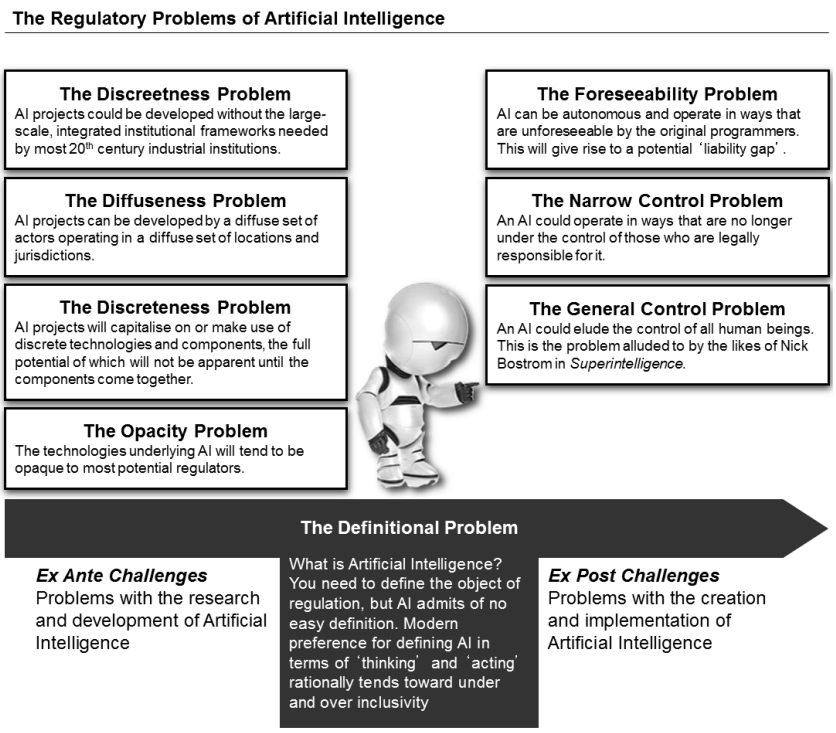

Figure 1: The Regulatory Problems of Artificial Intelligence

A. What Is AI?

Any AI regulatory regime must define what exactly it is that the regime regulates; in other words, it must define artificial intelligence. Unfortunately, there does not yet appear to be any widely accepted definition of artificial intelligence even among experts in the field, much less a useful working definition for the purposes of regulation.[15] This section will not presumptuously attempt to resolve that dispute or create a new definition of AI but instead will discuss the definitional problems that regulators will have to confront.

The difficulty in defining artificial intelligence lies not in the concept of artificiality but rather in the conceptual ambiguity of intelligence. Because humans are the only entities that are universally recognized (at least among humans) as possessing intelligence, it is hardly surprising that definitions of intelligence tend to be tied to human characteristics. The late AI pioneer John McCarthy, who is wide-

15. *See* John McCarthy, *What is Artificial Intelligence?*, JOHN MCCARTHY'S HOME PAGE 2–3 (Nov. 12, 2007), http://www-formal.stanford.edu/jmc/whatisai.pdf [https://perma.cc/U3RT-Q7JK].

ly credited as coining the term "artificial intelligence," stated that there is no "solid definition of intelligence that doesn't depend on relating it to human intelligence" because "we cannot yet characterize in general what kinds of computational procedures we want to call intelligent."[16] Definitions of intelligence thus vary widely and focus on myriad interconnected human characteristics that are themselves difficult to define, including consciousness, self-awareness, language use, the ability to learn, the ability to abstract, the ability to adapt, and the ability to reason.[17]

The same issues that plague efforts to define intelligence generally also apply to efforts to define artificial intelligence. Today, the leading introductory textbook on AI, Stuart Russell and Peter Norvig's *Artificial Intelligence: A Modern Approach*, presents eight different definitions of AI organized into four categories: thinking humanly, acting humanly, thinking rationally, and acting rationally.[18] Over time, the importance of each of these definitional concepts has waxed and waned within the AI research community.

Russell and Norvig cite the works of computing pioneer Alan Turing, whose writings predated the coining of the term "artificial intelligence," as exemplifying the "acting humanly" approach.[19] In his now-seminal paper *Computing Machinery and Intelligence*, Turing said that the question "Can machines think?" was "too meaningless to deserve discussion."[20] Turing instead focused on the potential for digital computers to replicate, not human thought processes themselves, but rather the external manifestations of those processes.[21] This is the premise of Turing's "imitation game," where a computer attempts to convince a human interrogator that it is, in fact, human rather than machine.[22]

Other early approaches to defining AI often tied the concept of intelligence to the ability to perform particular intellectual tasks. As a result, concepts of what constitutes artificial intelligence have shifted over time as technological advances allow computers to perform tasks that previously were thought to be indelible hallmarks of intelligence. Turing used the term "arguments from various disabilities" to describe

16. *Id.*

17. Some of these characteristics are, of course, present to various degrees in some other animals as well. Most notably, there is extensive scientific literature examining the cognitive abilities of non-human primates and cetaceans. *See generally, e.g.*, DAVID PREMACK, INTELLIGENCE IN APE AND MAN (1976); Olivier Pascalis & Jocelyne Bachevalier, *Face Recognition in Primates: A Cross-Species Study*, 43 BEHAV. PROCESSES 87 (1998); Rachel Adelson, *Marine Mammals Master Math*, MONITOR PSYCHOL. Sept. 2005, at 22, http://www.apa.org/monitor/sep05/marine.aspx [https://perma.cc/DU3G-4VP8].

18. RUSSELL & NORVIG, *supra* note 7, at 2.

19. *Id.*

20. A. M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433, 442 (1950).

21. *See id.* at 433–35.

22. *See id.* at 433–34.

arguments that machines could not think because they were unable to perform certain tasks.[23] Chess once was one such yardstick, but computers could play a passable game of chess by the 1960s[24] and could defeat the best human player in the world by 1997.[25] The result of achieving such a milestone has not been to proclaim that the machine that achieved it possesses intelligence, but rather to interpret the accomplishment of the milestone as evidence that the trait in question is not actually indicative of intelligence. This led McCarthy to lament that "[a]s soon as it works, no one calls it AI anymore."[26]

Today, it appears that the most widely-used current approaches to defining AI focus on the concept of machines that work to achieve goals — a key component of "acting rationally" in Russell and Norvig's scheme. McCarthy defined intelligence as "the computational part of the ability to achieve goals in the world" and AI as "the science and engineering of making intelligent machines, especially intelligent computer programs."[27] Russell and Norvig's textbook utilizes the concept of a "rational agent" as an operative definition of AI, defining such an agent as "one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome."[28]

From a regulatory perspective, however, the goal-oriented approach does not seem particularly helpful because it simply replaces one difficult-to-define term (intelligence) with another (goal). In common parlance, goal is synonymous with intention.[29] Whether and when a machine can have intent is more a metaphysical question than a legal or scientific one, and it is difficult to define goal in a manner that avoids requirements pertaining to intent and self-awareness without creating an over-inclusive definition.[30] Consequently, it is not clear how defining AI through the lens of goals could provide a solid working definition of AI for regulatory purposes.

---

23. *Id.* at 447.

24. *See* NILS J. NILSSON, THE QUEST FOR ARTIFICIAL INTELLIGENCE 194 (2010) (discussing the computer chess program Mac Hack VI's performance in tournaments against human players in 1967).

25. *See* BRUCE PANDOLFINI, KASPAROV AND DEEP BLUE: THE HISTORIC CHESS MATCH BETWEEN MAN AND MACHINE 7–8 (1997).

26. *See* Moshe Y. Vardi, *Artificial Intelligence: Past and Future*, COMM. ACM, Jan. 2012, at 5, 5 (2012).

27. McCarthy, *supra* note 15; *see also* Stephen M. Omohundro, *The Basic AI Drives*, *in* ARTIFICIAL GENERAL INTELLIGENCE 2008 483, 483 (2008) (defining AI as a system that "has goals which it tries to accomplish by acting in the world").

28. RUSSELL & NORVIG, *supra* note 7, at 4.

29. THE AMERICAN HERITAGE DICTIONARY and MERRIAM-WEBSTER'S COLLEGIATE DICTIONARY both direct readers to the entry for "intention" for a list of synonyms of "goal." *Goal*, THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE (4th ed. 2000); *Goal*, MERRIAM-WEBSTER'S COLLEGIATE DICTIONARY (11th ed. 2003).

30. For instance, if a "goal" is simply defined as "an end or objective that can be articulated with specificity," then a simple stamping machine arguably would have a goal because the end toward which it operates (i.e. stamping) is readily articulable.

Utilizing the more general concept of "acting rationally" would be both over-inclusive and under-inclusive. Rational action can already be ascribed to an enormous number of computer programs that pose no public risk. Computer chess programs and the AI of computer opponents in video games attempt to achieve an optimal result within the bounds of predefined sets of rules and thus could be described as acting rationally. Certainly, there does not seem to be any need to regulate the development of such innocuous programs and systems as they exist today. A "rational action" definition would also be under-inclusive; just as AI programs that *do* act rationally may not pose a public risk, AI programs that *do not* act rationally may pose serious public risks if the absence of rationality makes it difficult to predict the program's actions.[31]

This is not to say that AI systems that act rationally could not pose a public risk. On the contrary, much of the modern scholarship regarding the catastrophic risks associated with AI focuses on systems that seek to maximize a utility function, even when such maximization could pose an existential threat to humanity.[32] But the principle of rational action would not, standing alone, provide a sufficient legal definition for AI.

This paper will effectively punt on the definitional issue and "define" AI for the purposes of this paper in a blissfully circular fashion: "artificial intelligence" refers to machines that are capable of performing tasks that, if performed by a human, would be said to require intelligence. For the sake of distinguishing between AI as a concept and AI as a tangible technology, this article will occasionally use the term "AI system" to refer to the latter. For AI based on modern digital computing, an AI system includes both hardware and software components. It thus may refer to a robot, a program running on a single computer, a program run on networked computers, or any other set of components that hosts an AI.

B. The Problematic Characteristics of AI

Several characteristics of artificial intelligence will make it exceptionally difficult to regulate AI as compared to other sources of public risk. Subsections B.1 and B.2 will discuss features that distinguish AI from prior human inventions: autonomy and the attendant concerns about control and responsibility. These challenges call into question the sufficiency of any AI regulatory regime based on ex post legal mechanisms, i.e., those that intervene only after harm has occurred. Subsection B.3 will discuss the problematic characteristics of AI research and development ("R&D") work that will make effective

31. *Cf. infra* Part II.B.1.

32. *See infra* note 53 and accompanying text.

ex ante AI regulation difficult. These characteristics — discreetness, discreteness, diffuseness, and opacity — are also shared by the R&D of many Information Age technologies.

1. Autonomy, Foreseeability, and Causation

The most obvious feature of AI that separates it from earlier technologies is AI's ability to act autonomously. Already, AI systems can perform complex tasks, such as driving a car and building an investment portfolio, without active human control or even supervision.[33] The complexity and scope of tasks that will be left in the hands of AI will undoubtedly continue to increase in the coming years. Extensive commentary already exists on the economic challenges and disruptions to the labor market that these trends are already bringing about, and how those trends are likely to accelerate going forward.[34] Just as the Industrial Revolution caused socioeconomic upheaval as mechanization reduced the need for human manual labor in manufacturing and agriculture, AI and related technological advances will reduce the demand for human labor in the service sector as AI systems perform tasks that once were the exclusive province of well educated humans.[35] AI will force comparably disruptive changes to the law as the legal system struggles to cope with the increasing ubiquity of autonomous machines.

One important characteristic of AI that poses a challenge to the legal system relates to the concept of foreseeability. We have already seen numerous instances of AI that are designed to act in a manner that seems creative, at least in the sense that the actions would be deemed "creative" or as a manifestation of "outside-the-box" thinking if performed by a human. Some widely recognized examples of this phenomenon come from computer chess programs, which can play moves that cut against the basic precepts of human chess strategy.[36]

A particularly intriguing example comes from C-Path, a cancer pathology machine learning program.[37] Pathologists suspected that

33. *See* Neil Johnson et al., *Abrupt Rise of New Machine Ecology Beyond Human Response Time*, SCI. REPORTS, Sept. 11, 2013, at 1, 2; Kessler, *supra* note 1.

34. *See, e.g.*, Aaron Smith & Janna Anderson, *AI, Robotics, and the Future of Jobs*, PEW RESEARCH CTR. 44–45 (Aug. 6, 2014), http://www.pewinternet.org/files/2014/08/Future-of-AI-Robotics-and-Jobs.pdf [https://perma.cc/P2RS-BZPP]; *see also* RUSSELL & NORVIG, *supra* note 7, at 1034 (discussing how people losing their jobs to automation is an ethical issue introduced by AI).

35. *See, e.g.*, Smith & Anderson, *supra* note 34, at 52.

36. *See* NATE SILVER, THE SIGNAL AND THE NOISE: WHY SO MANY PREDICTIONS FAIL — BUT SOME DON'T 287–88 (2012).

37. "Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data." Margaret Rouse, *What Is Machine Learning*, WHATIS.COM, http://whatis.techtarget.com/definition/machine-learning [https://perma.cc/NCV5-83KF].

studying components of the supportive tissue (stroma) surrounding cancerous cells might, in combination with studying the actual tumor cells, aid in cancer prognosis.[38] But in a large study, C-Path found that the characteristics of the stroma were actually a *better* prognostic indicator for breast cancer than the characteristics of the cancerous cells themselves — a conclusion that stood at odds with both common sense and prevailing medical thought.[39]

Such examples of creativity are something of an illusion, a consequence of the computational resources available to these specialized AI programs combined with AI's freedom from the cognitive biases that affect humans. Discussing a computer chess engine, statistician Nate Silver observed:

> We should probably not describe the computer as "creative" for finding the moves; instead, it did so more through the brute force of its calculation speed. But it also had another advantage: it did not let its hang-ups about the right way to play chess get in the way of identifying the right move in those particular circumstances. For a human player, this would have required the creativity and confidence to see beyond the conventional thinking.[40]

This points to a fundamental difference between the decision-making processes of humans and those of modern AI — differences that can lead AI systems to generate solutions that a human would not expect. Humans, bounded by the cognitive limitations of the human brain, are unable to analyze all or even most of the information at their disposal when faced with time constraints. They therefore often settle for a satisfactory solution rather than an optimal one, a strategy that economist Herbert Simon termed "satisficing."[41] The computational power of modern computers (which will only continue to increase) means that an AI program can search through many more possibilities than a human in a given amount of time, thus permitting AI systems to analyze potential solutions that humans may not have considered, much less attempted to implement. When the universe of possibilities is sufficiently compact — as in the game Connect Four, or checkers played on an 8x8 board — the AI system may even be able to generate an optimal solution rather than a merely satisfactory

38. *See* Andrew H. Beck et al., *Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival*, SCI. TRANSLATIONAL MED., Nov. 9, 2011, at 1, 8.

39. *See id.*

40. SILVER, *supra* note 36, at 287–88.

41. Herbert A. Simon, *Rational Choice and the Structure of the Environment*, 63 PSYCHOL. REV. 129, 136 (1956).

one. [42] Even in more complex settings, and as the chess and C-Path anecdotes indicate, an AI system's solution may deviate substantially from the solution typically produced by human cognitive processes. The AI's solution thus may not have been foreseeable to a human — even the human that designed the AI

From a legal perspective, the takeaway from the chess and C-Path anecdotes is not the (mis)impression that the AI systems displayed creativity, but rather that the systems' actions were unexpected — certainly to outside observers, and perhaps even to the systems' programmers. Because AI systems are not inherently limited by the preconceived notions, rules of thumb, and conventional wisdom upon which most human decision-makers rely, AI systems have the capacity to come up with solutions that humans may not have considered, or that they considered and rejected in favor of more intuitively appealing options.[43] It is precisely this ability to generate unique solutions that makes the use of AI attractive in an ever-increasing variety of fields, and AI designers thus have an economic incentive to create AI systems capable of generating such unexpected solutions. These AI systems may act unforeseeably in some sense, but the capability to produce unforeseen actions may actually have been intended by the systems' designers and operators.[44]

To date, the unexpectedness of AI actions has been rather limited in scope; a computer chess program might make an unexpected move, but it is still not doing anything other than playing chess. But the development of more versatile AI systems combined with advances in machine learning make it all but certain that issues pertaining to unforeseeable AI behavior will crop up with increasing frequency and that the unexpectedness of AI behavior will rise significantly. The experiences of a learning AI system could be viewed as a superseding cause — that is, "an intervening force or act that is deemed sufficient to prevent liability for an actor whose tortious conduct was a factual cause of harm"[45] — of any harm that such systems cause. This is because the behavior of a learning AI[46] system depends in part on its

42. *See* Jonathan Schaeffer et al., *Checkers Is Solved*, 317 SCI. 1518, 1518–20 (2007). Of course, this only applies to finding solutions to problems that can be formalized and reduced to computer code.

43. *See* SILVER, *supra* note 36, at 287–88; Calo, *supra* note 11, at 532, 539 (using the term "emergence" to refer to the "unpredictably useful behavior" of robots, and noting that such behavior "can lead to solutions no human would have come to on her own").

44. *See* Calo, *supra* note 11 at 538 ("Emergent behavior is a clearly stated goal of robotics and artificial intelligence . . . .").

45. RESTATEMENT (THIRD) OF TORTS: PHYS. & EMOT. HARM § 34 cmt. b (AM. LAW INST. 2010). For a general discussion on the issues surrounding liability for harm caused by robots, see WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG 197–214 (2009).

46. Here, the term "learning AI" is not meant to imply that the AI system consciously learns, but rather that it is able to gather and, through machine learning, use new data to change how it acts in the world.

post-design experience,[47] and even the most careful designers, programmers, and manufacturers will not be able to control or predict what an AI system will experience after it leaves their care.[48] Thus, a learning AI's designers will not be able to foresee how it will act after it is sent out into the world — but again, such unforeseeable behavior was intended by the AI's designers, even if a specific unforeseen act was not.[49]

If legal systems choose to view the experiences of some learning AI systems as so unforeseeable that it would be unfair to hold the systems' designers liable for harm that the systems cause, victims might be left with no way of obtaining compensation for their losses. Issues pertaining to foreseeability and causation thus present a vexing challenge that the legal system will have to resolve in order to ensure that means of redress exist for victims of AI-caused harm.[50]

2. Control

The risks created by the autonomy of AI encompass not only problems of foreseeability, but also problems of control. It might be difficult for humans to maintain control of machines that are programmed to act with considerable autonomy. There are any number of mechanisms by which a loss of control may occur: a malfunction, such as a corrupted file or physical damage to input equipment; a security breach; the superior response time of computers as compared to humans;[51] or flawed programming. The last possibility raises the most interesting issues because it creates the possibility that a loss of control might be the direct but unintended consequence of a conscious design choice. Control, once lost, may be difficult to regain if the AI is designed with features that permit it to learn and adapt. These are the characteristics that make AI a potential source of public risk on a

47. As with "autonomous" and "learning," the term "experience" is not meant to imply consciousness, but rather to serve as a useful shorthand for the actionable data that an AI system gathers regarding its environment and the world in which it exists.

48. *See* Pei Wang, *The Risk and Safety of AI*, A GENERAL THEORY OF INTELLIGENCE, https://sites.google.com/site/narswang/EBook/topic-list/the-risk-and-safety-of-ai [https://perma.cc/5LY3-CTLD] ("An adaptive system's behaviors are determined both by its nature (i.e., initial design) and its nurture (i.e., postnatal experience). Though it is still possible to give the system certain innate beliefs and motivations, they will not fully determine the system's behaviors.").

49. *See* Jack M. Balkin, *The Path of Robotics Law*, 6 CALIF. L. REV. CIRCUIT 45, 52 (2015), http://www.californialawreview.org/wp-content/uploads/2015/06/Balkin-Circuit.pdf [https://perma.cc/AP4A-3YX8] ("[A]lthough the risk of *some* kind of injury at *some point* in the future is foreseeable whenever one introduces a new technology, how and when an injury occurs may not be particularly foreseeable to each of the potential defendants . . . .").

50. *See* RESTATEMENT (THIRD) OF TORTS: APPORTIONMENT OF LIABILITY §§ 10–17 ("Liability of Multiple Tortfeasors for Indivisible Harm"); *id.* §§ 22–23 ("Contribution and Indemnity"); Calo, *supra* note 11, at 554–55; Balkin, *supra* note 49, at 53.

51. *See* Johnson et al., *supra* note 33.

scale that far exceeds the more familiar forms of public risk that are solely the result of human behavior.

Loss of control can be broken down into two varieties. A loss of *local control* occurs when the AI system can no longer be controlled by the human or humans legally responsible for its operation and supervision. A loss of *general control* occurs when the AI system can no longer be controlled by any human. Obviously, the latter prospect presents far greater public risk than the former, but even a loss of general control would not necessarily pose significant public risk as long as the objectives of the AI system align with those of the public at large. Unfortunately, ensuring such an alignment of interests and objectives may be quite difficult, particularly since human values are themselves nearly impossible to define with any precision.[52]

The potential for the misalignment of interests flows from the fact that an AI's objectives are determined by its initial programming. Even if that initial programming permits or encourages the AI to alter its objectives based on subsequent experiences, those alterations will occur in accordance with the dictates of the initial programming. At first glance, this actually seems beneficial in terms of maintaining control. After all, if humans are the ones doing the initial programming, they have free rein to shape the AI's objectives. But many AI experts and commentators suggest that if an AI is programmed to achieve a certain objective, it may continue to work toward that objective even if the results of its efforts are not what the AI's original programmers would have subjectively intended:

> For example, we might propose a utility function designed to *minimize human suffering* . . . . Given the way humans are, however, we'll always find a way to suffer even in paradise; so the optimal decision for the AI system is to terminate the human race as soon as possible — no humans, no suffering.[53]

In such a scenario, the risk from an AI system does not stem from malevolence or an inability to comprehend the subjective intent behind its programmed goals. Rather, it stems from the machine's fundamental indifference to that subjective intent. "[A]n AI could know exactly what we meant and yet be indifferent to that interpretation of our words (being motivated instead by some other interpretation of

---

52. *See, e.g.*, Luke Muehlhauser & Nick Bostrom, *Why We Need Friendly AI*, 13 THINK 41, 41–43 (2014); Stuart Russell, *Of Myths and Moonshine*, EDGE, http://edge.org/conversation/jaron_lanier-the-myth-of-ai#26015 [https://perma.cc/PLG8-RWBZ]; NATE SOARES & BENJA FALLENSTEIN, MACHINE INTELLIGENCE RES. INST., ALIGNING SUPERINTELLIGENCE WITH HUMAN INTERESTS: A TECHNICAL RESEARCH AGENDA 2 (2014), https://intelligence.org/files/TechnicalAgenda.pdf [https://perma.cc/2XQT-NEXV].

53. RUSSELL & NORVIG, *supra* note 7, at 1037.

the words or being indifferent to our words altogether)."[54] Consequently, "[w]ith AI systems, . . . we need to be very careful what we ask for, whereas humans would have no trouble realizing that the proposed utility function cannot be taken literally."[55]

A growing chorus of academics, tech entrepreneurs, and futurists has gone further, warning that stronger forms of AI may resist all human efforts to govern their actions and pose a catastrophic — perhaps even existential — risk to humanity. A common expression of this concern focuses on the possibility that a sophisticated AI system could improve its own hardware and programming to the point that it gains cognitive abilities far outstripping those of its human creators.[56] As Russell and Norvig's "minimize human suffering" example indicates, it could be devilishly difficult to ensure that the goals of such an AI system are aligned with those of its human designers and operators.[57] If such AI systems prove to be more than a theoretical possibility, ex ante action would be necessary to ensure that the systems remain either susceptible to human control, aligned with the public interest, or both.

One need not accept the plausibility of such existential risk scenarios to recognize that problems of control and supervision will arise as AI systems become increasingly powerful, sophisticated, and autonomous.[58] Already, AI systems are capable of autonomously executing commands such as stock trades on time scales that can be measured in nanoseconds, depriving humans of their ability to intervene in real time.[59] The "flash crash" of 2010 demonstrated that the interaction between algorithmic trading systems can have a massive economic impact in a remarkably short period of time.[60] The results of such stock trades are, fortunately for most human investors, at least theoretically reversible. That may no longer be the case as AI systems

54. NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 196 (2014).

55. RUSSELL & NORVIG, *supra* note 7, at 1037; *see also* RICHARD A. POSNER, CATASTROPHE: RISK AND RESPONSE 41 ("Unless carefully programmed, [military] robots might prove indiscriminately destructive and turn on their creators.").

56. Nick Bostrom coined the term "superintelligence" to refer to the abilities of such a machine. Bostrom defines superintelligence as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills." Nick Bostrom, *How Long Before Superintelligence?*, NICK BOSTROM'S HOME PAGE, http://www.nickbostrom.com/superintelligence.html [https://perma.cc/7XW2-VLRC].

57. *See* RUSSELL & NORVIG, *supra* note 7, at 1037; *see also supra* note 52 and accompanying text.

58. *See* WALLACH & ALLEN, *supra* note 45, at 197 ("Autonomous (ro)bots aren't going to attempt a global takeover any time soon. But they are already causing harm, real and perceived, and they will not always operate within ethical or legal guidelines.").

59. *See* Johnson et al., *supra* note 33, at 1.

60. *See, e.g.*, Nils Pratley, *The Trillion-Dollar Questions over the Flash Crash and the Hound of Hounslow*, GUARDIAN (Apr. 25, 2015, 11:00 AM), http://www.theguardian.com/business/2015/apr/25/flash-crash-hound-of-hounslow-trillion-dollar-question [https://perma.cc/88QE-5FR4].

are rolled out in an ever-increasing number of industries. That will only reinforce the need to ensure that humans retain means for controlling sophisticated AI systems.

3. Research and Development: Discreet, Diffuse, Discrete, and Opaque

From a regulatory standpoint, some of the most problematic features of AI are not features of AI itself, but rather the manner in which AI R&D work can be done. *Discreetness* refers to the fact that AI development work can be conducted with limited visible infrastructure. *Diffuseness* means that the individuals working on a single component of an AI system might be located far away from one another. A closely related feature, *discreteness*, refers to the fact that the separate components of an AI system could be designed in different places and at different times without any conscious coordination. Finally, *opacity* denotes the possibility that the inner workings of an AI system may be kept secret and may not be susceptible to reverse engineering. Each of these features is shared, to varying degrees, by R&D work on many technologies in the Information Age, but they present particularly unique challenges in the context of AI.

The sources of public risk that characterized the twentieth century — such as nuclear technology, mass-produced consumer goods, industrial-scale pollution, and the production of large quantities of toxic substances — required substantial infrastructure investments. This simplified the regulatory process. The high cost of building the necessary facilities, purchasing the necessary equipment, and hiring the necessary labor meant that large corporations were the only non-governmental entities capable of generating most sources of public risk. Moreover, the individuals responsible for installing, operating, and maintaining the infrastructure typically had to be at the physical site where the infrastructure was located. The physical visibility of the infrastructure — and of the people needed to operate it — made it extremely unlikely that public risks could be generated clandestinely.[61] Regulators thus had little difficulty determining the "who" and "where" of potential sources of public risk.

By contrast, AI research and development can be performed relatively *discreetly*, a feature that AI shares with many other Information Age technologies. In 2009, Professor John McGinnis wrote that "[a]rtificial intelligence research is done by institutions no richer than colleges and perhaps would require even less substantial resources."[62] This actually overstated the resources necessary to participate in AI development, particularly with the rise of open-source programming.

61. *See, e.g.*, John O. McGinnis, *Accelerating AI*, 104 NW. U. L. REV. 1253, 1262 (2010).
62. *Id.*

Simply put, a person does not need the resources and facilities of a large corporation to write computer code. Anyone with a reasonably modern personal computer (or even a smartphone) and an Internet connection can now contribute to AI-related projects. Individuals thus can participate in AI development from a garage, a dorm room, or the lobby of a train station. This potential for discreetness provides the most jarring difference between AI and earlier sources of public risk.

The participants in an AI-related venture may also be remarkably *diffuse* by public risk standards. Participants in an AI-related project need not be part of the same organization — or, indeed, any organization at all. Already, there are a number of open-source machine-learning libraries; widely dispersed individuals can make dozens of modifications to such libraries on a daily basis.[63] Those modifications may even be made anonymously, in the sense that the identity in the physical world of individuals making the modifications is not readily discernible.[64]

The AI program itself may have software components taken from multiple such libraries, each of which is built and developed *discretely* from the others.[65] An individual who participates in the building of an open-source library often has no way of knowing beforehand what other individuals or entities might use the library in the future. Components taken from such libraries can then be incorporated into the programming of an AI system that is being developed by an entity that did not participate in assembling the underlying machine-learning library.

These characteristics are not limited to open-source projects or freely available material. Many modern computer systems use commercial off-the-shelf ("COTS") hardware and software components, most of which are proprietary.[66] The ease with which such compo-

63. Consider, for example, scikit-learn, an open-source machine-learning library for the Python programming language that can be accessed and modified through GitHub. GitHub users can modify a library on the website by sending (or "pushing") their modifications (termed "commits") to GitHub's servers. By April 18, 2015, GitHub users had made more than 18,000 such modifications to scikit-learn. *See scikit-learn: Machine Learning in Python*, GITHUB, https://github.com/scikit-learn/scikit-learn [https://perma.cc/3FA5-S5RA]. On April 2, 2015 alone, nine unique users made nineteen modifications to scikit-learn's code. According to the users' profile pages, two users are located in Switzerland, two more in France, one in the United States, and one in India; the remaining two users' profile pages give no indication of their geographic location. *See scikit-learn: Machine Learning in Python*, GITHUB, https://github.com/scikit-learn/scikit-learn/commits/master?page=57 [https://perma.cc/WV56-Z762].

64. The potential for anonymous editing also ties into each of the other three problems discussed in this section.

65. *Cf.* WALLACH & ALLEN, *supra* note 45, at 198.

66. *See* Robert B.K. Dewar, *COTS Software in Critical Systems: The Case for Freely Licensed Open Source Software*, MILITARY EMBEDDED SYSTEMS (Dec. 9, 2010), http://mil-embedded.com/articles/cots-open-source-software/ [https://perma.cc/T5G5-PXAB] (contrasting proprietary COTS software with freely available open-source software).

nents can be acquired makes it tempting to maximize use of COTS components to control costs, despite the potential security issues associated with using software components developed wholly outside the system developer's control.[67] Modern AI programming is no exception; few, if any, AI systems are built from the ground up, using components and code that are wholly the creation of the AI developers themselves. Moreover, if past is prologue, the physical components of an AI system will be manufactured by yet other entities separate from those that developed the AI system's programming. While separately developed components are present in all complex machinery to a certain extent, the level of discreteness and the scale of interactivity between software and hardware components in modern computer systems already rivals or exceeds that of prior technologies, and that complexity seems likely to increase further with the development of stronger forms of AI.[68]

In all likelihood, there will be considerable variation in the discreteness of the components of AI projects. Some AI systems likely will be built primarily with COTS or freely available hardware and software components, while others will mostly utilize programming and physical components designed and developed specifically for the AI project in question. Because of the cost advantages inherent in maximizing the use of COTS and freely available components, however, it seems all but certain that some AI systems will operate using a mishmash of hardware and software components harvested from many different companies. The interaction between numerous components and the disparate geographic locations of the companies involved will greatly complicate any regime designed to manage the risks associated with AI.[69]

Finally, the inner workings of and the interactions between the components of an AI system may be far more *opaque* than with earlier technologies. COTS software components may be easy to acquire, but their coding often is proprietary. Critical features underlying an AI system's operation thus may not be immediately apparent or readily susceptible to reverse engineering. Contrast this with automobiles —

---

67. *See generally* Carol Woody & Robert J. Ellison, *Supply-Chain Risk Management: Incorporating Security into Software Development*, DEP'T OF HOMELAND SEC. (Mar. 15, 2010), https://buildsecurityin.us-cert.gov/articles/best-practices/acquisition/supply-chain-risk-management%3A-incorporating-security-into-software-development [https://perma.cc/UV6U-X64C].

68. *See, e.g.*, Calo, *supra* note 11, at 534 ("Programming dictates behavior in complex ways. Code interacts with other code and various inputs, for instance, operator instructions or sensor data.").

69. *See id.* ("Software can have one or many authors. It can originate anywhere, from a multimillion-dollar corporate lab to a teenager's bedroom."); Balkin, *supra* note 49, at 53 ("Bugs may be difficult to spot and may develop through the combination of multiple modifications and additions. It may be fiendishly difficult to affix responsibility for bugs that emerge from layers of software development by many hands.").

one of the twentieth century's great sources of public risk. Automobiles consist of approximately 30,000 individual physical parts,[70] but the ways in which those physical components interact is well understood — not only by the designers and manufacturers of the vehicle itself, but also by the makers of parts for the vehicle and mechanics responsible for repairing the vehicles after they reach consumers. It seems unlikely that AI systems will demonstrate similar transparency if their development follows now-prevailing trends in information technology. Defects in the design of a complex AI system might be undetectable not only to consumers, but also to downstream manufacturers and distributors.[71]

Taken together, these characteristics confront regulators with fundamental logistical difficulties that were not present in earlier sources of public risk. Participants in AI projects may be located in multiple countries and have no legal or formal contractual relationship with one another. Attempts by any one country to regulate their citizens' participation in such projects may not greatly impact the projects' development. Even for projects involving large firms, the relatively low cost of infrastructure and the small physical footprint required for AI development means that firms could simply move AI development work offshore if regulations in their country of origin prove too intrusive. Many would likely do so given the competitive advantages that accompany advances in AI.[72]

These difficulties with regulating AI ex ante will also complicate efforts to ensure that victims receive compensation ex post when AI systems cause harm. The sheer number of individuals and firms that may participate in the design, modification, and incorporation of an AI system's components will make it difficult to identify the most responsible party or parties. Some components may have been designed years before the AI project had even been conceived, and the components' designers may never have envisioned, much less intended, that their designs would be incorporated into any AI system, still less the specific AI system that caused harm. In such circumstances, it may seem unfair to assign blame to the designer of a component whose work was far-removed in both time and geographic location from the completion and operation of the AI system. Courts may hesi-

---

70. John Paul MacDuffie & Takahiro Fujimoto, *Why Dinosaurs Will Keep Ruling the Auto Industry*, 88 HARV. BUS. REV. 23, 23 (2010).

71. *See* Vladeck, *supra* note 13, at 148 (citing the potential for "undetectable failure" in the components of automated driving systems as a drawback to holding manufacturers primarily liable for defects in autonomous vehicles).

72. *See, e.g.*, Vernor Vinge, *The Coming Technological Singularity: How to Survive in the Post-Human Era*, 10129 NASA CONF. PUBLICATION 11, 15 (1992), http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855.pdf [https://perma.cc/J2SU-UK5E] ("In fact, the competitive advantage . . . of every advance in automation is so compelling that passing laws, or having customs, that forbid [human-level AI] merely assures that someone else will get them first.").

tate to say that the designer of such a component could have foreseen the harm that occurred.[73] Similarly, the opacity of AI systems may make courts hesitant to blame the end user of an AI system that causes harm to a third party. And considerations of foreseeability aside, the multitude of potential defendants will complicate the assignment and apportionment of liability.

C. A Role for the Law?

Despite the problematic features of AI, there is good reason to believe that legal mechanisms could be used to reduce the public risks that AI presents without stifling innovation. Many of the problems identified in the preceding sections are simply gaps in the current law, and those gaps could be filled in any number of ways. Creating a working definition of AI will be difficult, to be sure, but coming up with precise legal definitions for imprecise terms is hardly a challenge unique to AI.[74] Any legal definition for the purposes of liability or regulation likely would be over-or under-inclusive, but that too is hardly an unfamiliar problem for the legal system to face. Similarly, the issues associated with foreseeability and causation must be confronted, but courts have always needed to adjust the rules for proximate causation as technology has changed and developed. The problem of control presents considerable challenges in terms of limiting the harm caused by AI systems once they have been developed, but it does not make it any more difficult to regulate or direct AI development ex ante.

---

73. David Vladeck explores these issues by asking who should be held responsible for the damage caused by HAL 9000, the AI villain in Stanley Kubrick's *2001: A Space Odyssey*:

> Would it be fair to hold liable the companies that designed, programmed, or manufactured HAL 9000, even though they embedded in HAL's "thinking" systems the first rule of autonomous machines — i.e., never harm a human — and even though the evidence strongly suggests that HAL "taught" himself to defy their instructions? Or should the creators of machines that have the capacity to "think" be held strictly liable whenever anything goes wrong? If so, on what theory? The theory that the wrongful conduct itself is proof of a defect? Or on an insurance-based theory that the creators are in a better economic position to absorb the cost of the injury than the person harmed?

Vladeck, *supra* note 13, at 125. Vladeck further notes that courts may hesitate to assign liability to automated systems where credible alternative theories of liability exist. *See id.* at 140 n.78. Vladeck cites *Ferguson v. Bombardier Servs. Corp.*, where the plaintiffs claimed that a defective autopilot system caused a plane crash. *Id.* (citing 244 F. App'x 944, 947 (11th Cir. 2007)). The trial court, in a ruling upheld on appeal, excluded a plaintiffs' witness from testifying because the witness' proposed testimony was equally consistent with the defendants' theory that the plane had been improperly loaded by its operators. 244 F. App'x at 947–49.

74. *See, e.g.*, SAIF Corp. v. Allen, 881 P.2d 773, 782–83 (Or. 1994) (discussing Oregon's rules for interpreting "inexact" and "delegative" statutory terms).

The law already provides mechanisms for confronting the issues of discreteness and opacity. The discreteness of AI is also shared by many other modern and not-so-modern technologies. Automobiles have long been manufactured using components from multiple companies and courts long ago developed rules for apportioning liability when harm is caused by defects in multiple such components.[75] Opacity could be reduced either directly by legislation requiring publication of the code and specifications of AI systems offered for commercial sale, or indirectly through tax incentives or tort standards that limit the liability of companies that make their AI systems more transparent.

The problems presented by the potentially diffuse and discreet nature of AI R&D seem somewhat harder to resolve at first blush. But the mere fact that AI *can* be developed diffusely and discreetly does not mean that the development of AI *will* proceed in a radically different fashion than earlier sources of public risk. Already, industry trends suggest that the development of AI, as with most twentieth-century technologies, will largely be driven by commercial and governmental entities rather than small private actors. The commercial potential of AI has already led to a veritable AI arms race as large companies have moved to invest heavily in AI projects. In January 2014, Google spent $500 million to purchase DeepMind, a British AI development company that defines its mission as "solv[ing] intelligence" by combining "the best techniques from machine learning and systems neuroscience to build powerful general-purpose learning algorithms."[76] The DeepMind purchase was just one of more than a dozen AI and robotics acquisitions that Google made in 2013 and 2014.[77] Google is far from alone; virtually every other large tech company has significant AI projects, including IBM's Watson, Facebook's Artificial Intelligence Research lab, and Microsoft's Project

75. *See* RESTATEMENT (THIRD) OF TORTS: APPORTIONMENT OF LIABILITY §§ 10–17 ("Liability of Multiple Tortfeasors for Indivisible Harm"); *id.* §§ 22–23 ("Contribution and Indemnity").

76. *Google DeepMind*, GOOGLE DEEPMIND, http://deepmind.com/index-alt.html#our-mission [https://perma.cc/HAW3-TJ23].

77. *See* Dan Rowinski, *Google's Game of Moneyball in the Age of Artificial Intelligence*, READWRITE (Jan. 29, 2014), http://readwrite.com/2014/01/29/google-artificial-intelligence-robots-cognitive-computing-moneyball [https://perma.cc/5QHB-2L68]. The DeepMind acquisition came during a two-month period that saw Google purchase seven other robotics companies. Adam Clark Estes, *Meet Google's Robot Army. It's Growing.*, GIZMODO (Jan. 27, 2014, 12:22 PM), http://gizmodo.com/a-humans-guide-to-googles-many-robots-1509799897 [https://perma.cc/DK6A-2HHD]. Several months later, Google spent "tens of millions" of pounds to expand its new DeepMind division by acquiring two British AI companies. Ingrid Lunden, *Google's DeepMind Acqui-Hires Two AI Teams in the UK, Partners with Oxford*, TECHCRUNCH (Oct. 23, 2014), http://techcrunch.com/2014/10/23/googles-deepmind-acqui-hires-two-ai-teams-in-the-uk-partners-with-oxford/ [https://perma.cc/D939-WT7R].

Adam.[78] The center of gravity for AI R&D thus may land in the same place as the public risks of the twentieth century — large, highly visible corporations.

If this trend continues, the most significant advances in AI will likely come from highly visible entities that regulators and courts can readily identify. Even though AI development work can be done by a single person using a personal computer, economies of scale and access to greater financial and human capital still confer a considerable advantage and will continue to do so in the future. This will prove particularly significant if computational power turns out to be a crucial component in developing more sophisticated AI. The human brain is thought to possess exascale computational power,[79] two orders of magnitude greater than the world's most powerful supercomputer in 2015[80] and eight orders of magnitude greater than the typical laptop computer available today.[81] In 2014, it took the world's fourth most powerful supercomputer forty minutes to simulate a single second of human brain activity.[82] At present, the list of operators of the world's most powerful supercomputers is dominated by governmental entities, state-owned enterprises, large research institutions, and large-cap corporations.[83] If that state of affairs continues, then national governments and large corporations, the same entities that generate other sources of public risk, may be the only entities capable of building strong AI systems for many years. At the very least, projects backed

78. *See, e.g.*, *IBM Watson*, IBM, http://www.ibm.com/smarterplanet/us/en/ibmwatson/ [https://perma.cc/5BX9-SWE5]; *Facebook AI Research (FAIR)*, FACEBOOK, https://research.facebook.com/ai [https://perma.cc/9UW3-TJ2G]; *Introducing Project Adam: A New Deep-Learning System*, MICROSOFT (July 14, 2014), http://research.microsoft.com/apps/video/default.aspx?id=220709&r=1 [https://perma.cc/Y2WU-PZV5].

79. Bernd Mohr, *The Human Brain Project Will Push the Boundaries of Supercomputing*, TOP500 (Jan. 2016), http://www.top500.org/blog/the-human-brain-project-will-push-the-boundaries-of-supercomputing/ [https://perma.cc/H87N-W4KH]. "Exascale" refers to a computer capable of performing 1018 floating-point operations per second (FLOPS). Joab Jackson, *Next Up: Exascale Computers, Expected to Arrive by 2020*, PCWORLD (Nov. 18, 2012), http://www.pcworld.com/article/2014715/next-up-exascale-computers-expected-to-arrive-by-2020.html [https://perma.cc/GPS3-FBFR].

80. *June 2015*, TOP500, http://www.top500.org/lists/2015/06/ [https://perma.cc/Q5P8-R72Q]. According to TOP500, the most powerful supercomputer as of November 2014 was China's Tianhe-2, capable of performing 33.86 petaFLOPS, or 3.386 x 1016 FLOPS. *Id.*

81. Francis Wray, *A Brief Future of Computing*, PROJECT HPC (2012), http://web.archive.org/web/20150423142748/http://www.planethpc.eu/index.php?option=com_content&view=article&id=66:a-brief-future-of-computing&catid=1:articles&Itemid=3 (accessed via online archive because the original PlanetHPC website is no longer running). A typical laptop available in 2012 was capable of performing 50 gigaFLOPS, or 5 x 1010 FLOPS. *See id.*

82. *See* Matthew Sparkes, *Supercomputer Models One Second of Human Brain Activity*, TELEGRAPH (Jan. 13, 2014, 10:04 AM), http://www.telegraph.co.uk/technology/10567942/Supercomputer-models-one-second-of-human-brain-activity.html [https://perma.cc/JK4N-PSNJ].

83. *See June 2015*, *supra* note 80.

by such entities will have a significant advantage over other efforts to build sophisticated AI systems.

In one regard, however, the rising private sector investment in AI narrows the range of effective tools at the government's disposal. Large private sector spending on AI development makes it unlikely that government-subsidized AI safety research would, standing alone, have a significant impact.[84] Absent truly exorbitant public spending, government investment in AI research would be dwarfed by private sector investment — and unless there is a cataclysmic event on the scale of World War II, it is unlikely that the public appetite for massive government spending on AI projects would materialize. Moreover, if the goal of public sector AI investment would simply be to research AI safety and publish information about how to develop safe AI, there still would need to be some sort of mechanism to encourage or require AI developers to incorporate the resultant safety features into their systems. Consequently, while government-subsidized research might complement a broader legal framework for AI, it would not be a sufficient legal response to the public risks that AI will generate.

Fortunately, the legal and regulatory institutions of the industrialized world provide a broad and deep toolkit offering many potential methods for influencing the development and operation of AI.[85] Even if an aspect of AI is not easily susceptible to direct ex ante regulation by an administrative agency, it might respond to the indirect ex post incentives provided by tort law. Legislatures, agencies, and courts each offer mechanisms that can help direct the development of AI in socially and economically beneficial ways. Part III will address the comparative competencies of each of these types of governmental institutions for managing the public risks associated with AI.

III. INSTITUTIONAL COMPETENCE

Before turning to the potential substantive content of AI regulations in Part IV, this Part considers what role each of the potential regulatory institutions should play:

> In a government seeking to advance the public interest, each organ has a special competence or expertise, and the key to good government is not just figuring out what is the best policy, but figuring out

84. *See* Kaushal & Nolan, *supra* note 5 (proposing a "new Manhattan Project" for AI); *see also* McGinnis, *supra* note 61, at 1265 (proposing a "research project, like those funded by the National Institutes of Health").

85. *Cf.* Calo, *supra* note 11, at 537 (noting that in order to resolve the challenges that robotics will present to the legal system "[c]ourts may soften or strengthen existing doctrines, import doctrines across subject matter, or resurrect doctrine long forgotten").

> which institutions should be making which decisions and how all the institutions should interrelate.[86]

Part III examines the competencies of three separate institutions — national legislatures,[87] administrative agencies, and the common law tort system — particularly with respect to managing the pubpublic risks presented by AI.

The legal processes of all three institutions share certain characteristics. The superior financial and professional resources available to large firms and wealthy individuals give them a greater ability to influence decision-making in all institutional settings. This tendency manifests itself in the form of lobbying by concentrated interest groups in legislatures and administrative agencies. In the tort system, access to greater financial resources provides litigants with the ability to spend more money on investigation, discovery, attorneys, and experts. Scholars have argued ceaselessly about which institution is most handicapped by such disparities, and resolving that dispute is beyond the scope of this paper. For now, it suffices to note that access to greater financial resources generally translates to a superior ability to influence policy in all three settings. Other characteristics common to all three institutions, and that do not obviously afflict any one institution more than another, include resource and budgetary constraints and the potential for corruption by key decision-makers.

Even beyond these common characteristics, no institution has a monopoly on any particular competence. For example, while administrative agencies typically enjoy an advantage over courts and legislatures in terms of subject-matter expertise,[88] courts and legislatures can close this gap by consulting experts of their own. And while legislatures and agencies have greater freedom than courts to act ex ante and take measures to prevent harm before it occurs,[89] one may reasonably question how often they exercise that freedom in practice. All of the

---

86. *See* HART, JR. & SACKS, *supra* note 12, at lx.

87. This paper focuses on national rather than state or provincial legislatures because of the diffuse and easily transportable nature of AI research. Because of these factors, most regional and local legislatures would be able to regulate only a small fraction of AI research. Consequently, any substantive regulations adopted solely by a single sub-national political unit would not likely have a significant effect on the development and deployment of AI as a whole. Of course, national regulations suffer the same disadvantages when compared to international treaties. But because negotiating and ratifying a comprehensive AI treaty would be exceedingly difficult (witness the failures of the Doha round of trade talks and the Copenhagen climate change conference), the prospects for such a treaty seem remote in the absence of preexisting national regulatory systems or a strong consensus that AI poses a global catastrophic risk. *See* WALLACH & ALLEN, *supra* note 45, at 212 (pointing out "decisions to regulate research" in some countries may not be supported by the governments of other countries because "[v]alues and social pressures differ from country to country and state to state").

88. *See infra* Part III.B.2.

89. *See infra* Parts III.B.4 and III.C.2.

characteristics discussed in Part III are subject to similar caveats. The principles discussed below are, nevertheless, quite instructive in their implications for whether and how AI might be effectively regulated.

A. Legislatures

Legal process scholars have largely ignored the regulatory role of legislatures, preferring instead to focus on the judicial and administrative processes. This omission seems somewhat perplexing to a 21st century observer given the increasing prominence of direct legislative intervention as a form of social and economic control since the dawn of the Progressive Era. The 20th century saw the creation of complex tax codes and the gradual abolition of common law crimes in favor of penal codes. Statutory schemes also increasingly displaced the common law in defining the substantive rules governing bankruptcy, labor, public health, real property, and personal transportation.

Despite the relative dearth of scholarship discussing the institutional strengths and weaknesses of legislatures, we can perceive a few general characteristics of legislatures as regulatory bodies: (1) democratic legitimacy; (2) a relative lack of expertise; and (3) the ability to delegate. These characteristics make legislatures the ideal body for setting the starting point for a regulatory scheme and establishing the fundamental principles that guide the development of policy, though not for making decisions about the specific substantive content of regulations.

1. Democratic Legitimacy

Laws passed by legislative bodies comprised of elected representatives can stake a stronger claim to reflecting the popular will than administrative rules or judicial doctrine. This imbues legislative enactments with greater democratic legitimacy than agency rules or court decisions.[90] This advantage results both from the fact that legislators are chosen by regular elections and by legislators' greater openness to direct contact with the general public.[91] The general public thus typically prefers that legislatures make the policy decisions on matters of fundamental social policy and other areas of the law where the weighing of ethical, moral, and other value-laden considerations predominate.

---

90. *See e.g.*, Roscoe Pound, *Common Law and Legislation*, 21 HARV. L. REV. 384, 406 (1908) ("We recognize that legislation is the more truly democratic form of lawmaking. We see in legislation the more direct and accurate expression of the general will.").

91. *See* Benjamin H. Barton, *An Institutional Analysis of Lawyer Regulation: Who Should Control Lawyer Regulation — Courts, Legislatures, or the Market?*, 37 GA. L. REV. 1167, 1222 (2003).

But while the democratic process provides legislatures with its strongest claim to policymaking preeminence, voters generally vote based on the totality of a candidate's views rather than on any single issue and rarely know the exact details of any particular bill at the time they enter the ballot box, thus undercutting the idealistic principle that legislative action is an expression of popular will.[92] The need to be reelected and the expense of legislative campaigns also limit legislators' ability to make informed judgments on any particular bill. Legislators must spend considerable time campaigning and fundraising, reducing the amount of time they spend on legislative business and broad constituent contact.[93] Pressure from key interest groups may lead a legislator to support policies that his constituents oppose and oppose policies that they support.

Despite these concerns, legislatures remain the institutions best equipped to make value-laden policy decisions. Agencies' staffs are appointed rather than elected; judges are supposed to follow the law even when the law deviates from popular will; and abrogating those principles in order to make agencies and courts more democratically responsive would undermine those institutions' unique strengths. By default, then, legislators are best-equipped to make decisions on issues where democratic legitimacy is a priority.

Any AI regulatory regime must have the public imprimatur that comes with legislative approval. The weighing of values is inherent both in determining the level of acceptable public risk and in deciding whether there are certain spheres (e.g., military and police functions) in which human decision-makers should never cede responsibility to autonomous machines. To ensure that institutions with strong democratic legitimacy make those decisions, legislatures should set the starting point for AI regulation by specifying the goals and purposes of any AI regulatory regime.

2. Lack of Expertise

A critical weakness of legislatures with respect to regulating emerging technologies is a relative lack of expertise. Agencies typically are staffed by experts possessing specialized knowledge of the

92. *See* HART, JR. & SACKS, *supra* note 12, at 688.

93. Tracy Jan, *For Freshman in Congress, Focus Is on Raising Money*, BOS. GLOBE (May 12, 2013), http://www.bostonglobe.com/news/politics/2013/05/11/freshman-lawmakers-are-introduced-permanent-hunt-for-campaign-money/YQMMMoqCNxGKh2h0tOIF9H/story.html [https://perma.cc/Y8QQ-C4E3]. After the 2012 elections, the Democratic Congressional Campaign Committee recommended that the newest members of Congress spend four to five hours a day making fundraising and "strategic outreach" calls — approximately "twice as much time as party leaders expect them to dedicate to committee hearings and floor votes, or meetings with constituents." *Id.*

field in question,[94] and courts can turn to expert witnesses to gain the technical knowledge necessary to decide a particular case. Legislators, by contrast, typically must rely on committee hearings and contact with lobbying groups to gain access to relevant expert opinions regarding proposed legislation.

The drawbacks of relying upon lobbyists for technical information are obvious, and the utility of committee hearings is questionable at best in the context of emerging technologies. Only the small subset of the legislature that sits on the relevant committee will hear the experts' testimony, and even those legislators cannot afford to spend an inordinate amount of time conducting hearings on any one particular issue or bill. Moreover, in the United States Congress at least, the influence of legislative committees has noticeably waned in recent years.[95] This limits a legislature's capacity to make informed policy decisions for emerging technologies, where a proper understanding of the relevant features of a technology may depend on access to technical expertise.

3. Delegation and Oversight

Fortunately, legislatures can compensate for their relative lack of expertise because they enjoy greater freedom than agencies and courts to delegate some policymaking responsibilities. Such delegation may be internal (to committees and subcommittees) or external (to agencies, courts, or private institutions). Delegating bills to committees provides legislatures with the ability to close, albeit only partially,[96] the gap in expertise and specialization they suffer in comparison to agencies. But more importantly, when a legislature confronts an issue that lies decisively outside its institutional competence, it can assign the task of substantive policymaking to a more suitable external entity. On issues where consensus among legislators proves elusive, the legislature can establish a general standard rather than a hard-and-fast rule, "essentially delegating rulemaking responsibilities to courts, agencies, or private institutions."[97] Similarly, legislatures possess "the power simply to set forth a 'policy,' or social objective, and vest discretion in an agency . . . to carry it out." [98] Judges and bureaucrats usually do not have such discretion.[99]

---

94. *See* HART, JR. & SACKS, *supra* note 12, at lxi n.42; Todd Eberly, *The Death of the Congressional Committee*, BALTIMORE SUN (Nov. 27, 2011), http://articles.baltimoresun.com/2011-11-27/news/bs-ed-supercommittee-20111127_1_committee-system-committee-chairs-committee-hearings/2 [https://perma.cc/S7C7-65PT].

95. *See, e.g.*, Eberly, *supra* note 94.

96. *See id.*

97. HART, JR. & SACKS, *supra* note 12, at xciii–xciv.

98. *Id.* at xciv.

99. *See id.*

Once the decision to delegate is made, legislatures possess multiple oversight tools to monitor the progress of the regulatory entity, including "the power of the purse, [the ability to conduct] oversight hearings, and the power to withdraw their delegation" completely.[100] Legislatures thus can step in and remove or reassign policymaking authority if agencies or courts fail to establish acceptable regulations. Conversely, if the chosen entity succeeds in establishing a solid regulatory framework, the legislature can codify the relevant rules or standards by incorporating them into subsequent legislation.

B. Agencies

The idea of delegating policymaking responsibility to administrative agencies, staffed by bureaucrats and technical experts rather than politicians or judges, gained currency during the Progressive Era and the Great Depression, periods when it seemed that legislatures simply lacked the capacity to deal with the complex social and economic challenges created by industrialization.[101] The shortcomings of legislatures spurred the creation of new regulatory entities that could specialize in the particular industry in need of regulation, staff themselves with professionals who have prior knowledge of the relevant fields, and remain independent from the political pressures that distorted the judgments of elected officials. Administrative agencies had existed previously, but they exploded in number and importance during the 20th century as economic crises and long-term social trends led the public to increasingly demand governmental intervention in the economy and society.[102]

One unique aspect of agencies is their malleability in design. In principle, at least, agencies can be tailor-made for the regulation of a specific industry or for the resolution of a particular social problem. Policymakers in agencies can be experts with a background in the relevant field rather than generalists of the sort that fill the ranks of courts and legislatures. Moreover, because agencies are not bound by the rules that limit courts' ability to consider factors other than the facts of the specific case in front of them, they are freer to conduct independent factual investigations and make policy decisions based on broad social considerations. In the context of AI, this makes agencies well positioned to determine the substantive content of regulatory policies.

---

100. *See* Barton, *supra* note 91, at 1224.

101. *See* FELIX FRANKFURTER, THE PUBLIC AND ITS GOVERNMENT 34–35 (1930).

102. *See generally* Julius Stone, *The Twentieth Century Administrative Explosion and After*, 52 CALIF. L. REV. 513 (1964).

1. Flexibility

Agencies have a clear advantage over legislatures and courts in terms of institutional flexibility. Because agencies can be designed and assembled from the ground up, "[t]he potential scope of regulation is limited only by the imaginations of regulators."[103] As a result, the number of potential forms that agencies could take is virtually unlimited. "[A]gencies may differ in respect of the tenure of their officials, the measure of their independence, their relationship to the courts, their powers to investigate and prosecute, and in a hundred and one other details."[104]

Rather than being a strictly legislative, judicial, or executive body, an agency's functions may embrace aspects of all three branches.[105] Agencies may combine a legislature's ability to set policy, a court's ability to dispose of competing claims, and the executive's ability to enforce decisions. They also can influence conduct in more subtle ways by collecting and publishing relevant information about the safety risks created by an industry or specific products within that industry. An independent agency thus may possess "that full ambit of authority necessary for it . . . to plan, to promote, and to police," potentially giving it "an assemblage of rights normally exercisable by government as a whole."[106] Agencies therefore "have comparative institutional advantages over both courts and legislatures in applying legislated rules or principles to problems, because they have the legislature's ability to engage in ambitious factfinding and the courts' option of focusing on one problem at a time."[107] This permits agencies to make broad policy decisions without being limited, as courts are, to the narrow issues and facts of a specific case.

Agencies also display considerable diversity in the scope of their substantive jurisdiction. The most familiar examples of administrative agencies are entities tasked with regulating one particular industry, such as aviation, energy, communications, or consumer finance. But the jurisdiction of other administrative agencies "relate[s] less to a particular type of industrial activity than to a general social and economic problem which cut[s] across a vast number of businesses and occupations."[108] Institutions such as the Federal Trade Commission, the National Labor Relations Board, and the Equal Employment Opportunity Commission, were built to address specific types of unfair

---

103. W. Kip Viscusi, *Toward a Diminished Role for Tort Liability: Social Insurance, Government Regulation, and Contemporary Risks to Health and Safety*, 6 YALE J. ON REG. 65, 70 (1989).

104. JAMES M. LANDIS, THE ADMINISTRATIVE PROCESS 22 (1938).

105. *See id.* at 2.

106. *Id.* at 15.

107. HART, JR. & SACKS, *supra* note 12, at lxxx.

108. LANDIS, *supra* note 104, at 16.

practices, with the agencies acting as an extension of the government's general police power.[109] An agency's mission can be as broad as the Environmental Protection Agency's, whose stated mission "is to protect human health and the environment,"[110] or as narrow as licensing acupuncturists.[111] Given the rising social and economic pervasiveness of AI, this flexibility in defining an agency's mission and goals will be of particular value in influencing the development of AI.

But legislatures, fearing the prospect of "mission creep" and wishing to ensure agency accountability, are loathe to give agencies too much freedom to change their own modes of operation. Legislatures therefore generally prescribe many aspects of agencies' operations in the enabling legislation, including such key features as leadership structure and rulemaking procedures. As a result, the flexibility of an agency largely fades once the enabling legislation is passed. Such limited post-creation flexibility would be felt acutely in the case of AI regulation, because the focus and direction of AI research is likely to change over time — a limitation that will be discussed in greater detail below.

2. Specialization and Expertise

Agencies provide a way to place policymaking in the hands of professionals with expertise in the regulated industry. Ideally, this expertise has two components. First, the agency's staff consists of individuals who have preexisting knowledge of the designated industry. Second, the agency itself is given a specific mission so that its staff is able to focus its work solely on matters relevant to its designated industry. The combination of these two features allows agencies to develop true expertise with respect to the relevant industry:

> [Expertness] springs only from that continuity of interest, that ability and desire to devote fifty-two weeks a year, year after year, to a particular problem. With the rise of regulation, the need for expertness became dominant; for the art of regulating an industry requires knowledge of the details of its operation, ability to shift requirements as the condition of the

109. *Id.* at 16, 23; *see also id.* at 30 (distinguishing between "those administrative bodies whose essential concern is the economic functioning of the particular industry and those which have an extended police function of a particular nature"); *About EEOC*, U.S. EQUAL EMP'T OPPORTUNITY COMM'N, http://www.eeoc.gov/eeoc/ [https://perma.cc/X5MM-PDCX].

110. *Our Mission and What We Do*, ENVTL. PROTECTION AGENCY, http://www.epa.gov/aboutepa/our-mission-and-what-we-do [https://perma.cc/33FE-BEFZ] (last updated Sep. 29, 2015).

111. *See, e.g.*, *Office of Acupuncture Licensure*, COLO. DEP'T REG. AGENCIES, https://www.colorado.gov/pacific/dora/Acupuncture [https://perma.cc/42Y4-Q24F].

> industry may dictate, the pursuit of energetic measures upon the appearance of an emergency, and the power through enforcement to realize conclusions as to policy.[112]

Admittedly, agencies enjoy this advantage more or less by default. Legislators and judges are generalists, with workloads that span across industries and fields of law.[113] Expertise in a particular field is rare, and specialization is rarer still because of the wide variety of matters that reach a legislator's desk or a judge's docket. Consequently, "[n]either the legislature nor the judiciary is competent to make all the technical, fact-bound judgments necessary for the regulatory process, tasks an agency filled with specially trained experts is particularly competent to undertake."[114]

But agencies' expertise advantage may actually wane in the context of emerging and rapidly changing technologies, such as AI. When a technology is in its infancy, researchers directly involved in the research and development of that technology may be the only people who possess the expertise necessary to make risk and safety assessments.[115] In such cases, the relatively few specialists who are in the know can demand a premium in the private sector for access to their knowledge, making it less likely that they will join the staff of a regulatory agency during the period when their expertise would be most beneficial.

This issue is particularly salient in the context of AI regulation. The dominant strains of AI research have changed repeatedly during the industry's six decades of existence.[116] Today, most AI researchers believe that new fundamental ideas will be needed for an AI to achieve human level intelligence.[117] It is impossible to say with any certainty where these fundamental new ideas may come from. AI research draws on concepts from fields as diverse as computer science,

---

112. LANDIS, *supra* note 104, at 23–24.

113. Specialty courts and therapeutic courts represent a limited exception to this rule. To date, however, no such specialty courts that are devoted to cases involving a specific industry or technology have been established. Committee work theoretically should permit legislators to specialize to some degree, but this specialization is limited by the relatively small amount of time that legislators spend on committee work and diluted further by the fact that most legislators are members of multiple committees or subcommittees.

114. HART, JR. & SACKS, *supra* note 12, at lxi n.42.

115. *E.g.*, Mary L. Lyndon, *Tort Law and Technology*, 12 YALE J. ON REG. 137, 157–58 (1995) ("Neither courts nor agencies possess expertise when technologies are new or changing. At early stages of innovation, only researchers involved in R&D have this knowledge.").

116. *See* RUSSELL & NORVIG, *supra* note 7, at 18–27.

117. *E.g.*, McCarthy, *supra* note 15; WALLACH & ALLEN, *supra* note 45, at 191 ("[M]ajor technological thresholds need to be crossed before the promise of human-like AI, not to mention superhuman AI, should be considered a serious possibility by policy makers and the general public.").

linguistics, probability, mathematics, economics, neuroscience, psychology, and philosophy.[118] The relative importance of these fields in AI research almost certainly will change — possibly quite dramatically — as developers build more sophisticated AI systems. The experts on the AI of today thus may struggle to assess the risks associated with the AI of tomorrow.

These issues point to a more general slipperiness in the meaning of "expertise." Would someone qualify as an AI expert if they have a background in computer science but no training or expertise specifically relating to AI? Would someone with even extensive experience in computer vision be qualified to assess the features of natural language processing systems? These issues obviously are neither insurmountable nor unique to administrative agencies; when deciding who to select for a vacant position, every public and private entity must make value judgments on which candidate has the most appropriate combination of education and work experience. But because AI research spans a wide array of seemingly disparate fields whose relative importance may wax and wane, a governing agency may struggle to ensure that its staff includes the appropriate mix of professionals.

3. Independence and Alienation

An important impetus for the rise of administrative agencies came from cynicism regarding the independence of elected officials and their willingness and ability to act in the public interest, particularly given the pressures exerted upon legislators by the rise of mass media and the increasing influence of lobbying organizations.[119] Because agencies can be designed from scratch, they can theoretically be designed in a manner that shields them from some of the political pressures that distort the legislative process, such as by limiting the president's ability to terminate members of the agency's leadership.[120] This permits agencies to fashion policy while being "removed to a degree from political influence,"[121] providing an expert agency with the ability to "not only solve problems, but rely on neutral criteria for the solution of problems,"[122] free — or at least freer — from the distorting influence of electoral politics.

But such independence comes at the price of alienation from the general public and their elected representatives, leaving independent

118. *See* RUSSELL & NORVIG, *supra* note 7, at 5–14.

119. *See* FRANKFURTER, *supra* note 101, at 33–34.

120. *See, e.g.*, 12 U.S.C. § 242 (2012) (members of the Federal Reserve's Board of Governors can only be removed "for cause"); 42 U.S.C. § 7171 (2012) (members of Federal Energy Regulatory Commission "may be removed by the President only for inefficiency, neglect of duty, or malfeasance in office").

121. LANDIS, *supra* note 104, at 111.

122. HART, JR. & SACKS, *supra* note 12, at lxi n.42.

agencies open to criticism for being undemocratic. James Landis, former dean of Harvard Law School and one of the most vocal supporters of the administrative process, noted that an agency's "relative isolation from the popular democratic processes occasionally arouses the antagonism of legislators."[123] When the New Deal spurred the creation of a slew of independent federal agencies, critics slammed the new entities as "a headless 'fourth branch' of the Government, a haphazard deposit of irresponsible agencies and uncoordinated powers" whose very existence threatened the foundations of American constitutional democracy.[124]

The decision to assign a policymaking task to an administrative agency thus represents a value choice:

> The legislative decision to "take things out of politics" by delegating significant issues of public policy to an administrative agency does not change the nature of the issues to be decided. It merely changes the forum in which they will be decided from one that draws its strength from its political responsiveness to one that takes its definition from its independence and expertise.[125]

There is another type of independence that agencies can enjoy: the ability to conduct independent investigations. In this regard, agencies differ from courts, where dispositions must be based on the record as developed by the parties.[126] Judges in most jurisdictions theoretically have the ability to call their own witnesses, but that is a privilege that few judges choose to exercise.[127] Juries are specifically warned against conducting an independent investigation into the facts of the case and are admonished not to consider evidence outside of the exhibits and witnesses presented to them.[128] Agencies need not be subjected to such restrictions.

---

123. LANDIS, *supra* note 104, at 50.

124. THE PRESIDENT'S COMM. ON ADMIN. MGMT., ADMINISTRATIVE MANAGEMENT IN THE GOVERNMENT OF THE UNITED STATES 36 (1937).

125. James O. Freedman, *Expertise and the Administrative Process*, 28 ADMIN. L. REV. 363, 373 (1976).

126. *See* LANDIS, *supra* note 104, at 37–39.

127. *See* RICHARD GLOVER, MURPHY ON EVIDENCE 688–89 (14th ed. 2015).

128. *See, e.g.*, WASH. PATTERN JURY INSTRUCTIONS § 1.01 (2014) ("The only evidence you are to consider consists of testimony of witnesses and exhibits admitted into evidence . . . . The law is contained in my instructions to you. You must disregard anything the lawyers say that is at odds with the evidence or the law in my instructions."); MICH. MODEL CRIM. JURY INSTRUCTIONS § 2.5 (2015) ("Evidence includes only the sworn testimony of witnesses, the exhibits admitted into evidence, and anything else I tell you to consider as evidence.").

4. Ex Ante Action

Agencies also share legislatures' ability to act ex ante and formulate policy before harmful conduct occurs. Courts, by contrast, are inherently reactive institutions; they can intervene in society only to the extent that the particular case or controversy before them permits them to do so.[129] Consequently, courts' approval or disapproval of conduct generally comes only after the conduct is complete.

As with agencies' advantage in technical expertise, however, it is possible that agencies' ability to act ex ante has diminished significance when an agency is tasked with regulating an emerging technology, such as AI. It may well be that additional research, development, and even public operation are the only ways to determine which types of AI are harmful and which types are not. The potential for rapid changes in the direction and scope of AI research may impair an agency's ability to act ex ante; an agency whose staff is drawn from experts on the current generation of AI technology may not have expertise necessary to make informed decisions regarding future generations of AI technology. Moreover, ex ante regulation does not imply timeliness in rule promulgation. When a federal agency wishes to promulgate or amend a rule, it must go through the laborious process of preparing a regulatory analysis, receiving approval from the Office of Information and Regulatory Affairs, and calling for extensive public comment on the proposed rules.[130] State agencies operate under similar restrictions.[131] Because of the delays necessitated by the rulemaking process, the ex ante rule may not go into effect until the target class of product has already caused harm or even become obsolete.

This does not necessarily lead to the conclusion that an agency's ability to regulate ex ante is completely useless in the context of emerging technologies such as AI. By publishing industry standards and engaging in public advocacy, agencies can set expectations without specifically approving or barring particular products or programs.[132] Agencies also have an unparalleled ability to collect and disseminate risk information so that the general public may deduce for

---

129. *See* U.S. CONST. art. III, § 2.

130. *See A Guide to the Rulemaking Process*, OFFICE OF THE FED. REGISTER, https://www.federalregister.gov/uploads/2011/01/the_rulemaking_process.pdf [https://perma.cc/6CE8-3Q6T].

131. *See, e.g.*, *Administrative Rules Process in a Nutshell*, STATE OF MICHIGAN (Feb. 2015), http://www.michigan.gov/documents/lara/Admin_Rules_Process_353271_7.pdf [https://perma.cc/Y6HU-Y2XL].

132. The National Institute of Standards and Technology, for instance, publishes voluntary and consensus standards on a variety of topics, including a framework on cybersecurity first drafted in 2014. *See* Press Release, Nat'l Inst. of Standards and Tech., NIST Releases Cybersecurity Framework Version 1.0 (Feb. 12, 2014), http://www.nist.gov/itl/csd/launch-cybersecurity-framework-021214.cfm [https://perma.cc/6QN3-LYAD].

itself the relative safety of a particular product or class of products.[133] In the context of AI regulation, this could be accomplished by adopting standards specifying the characteristics that AI systems should have, such as being limited to specified activities or remaining susceptible to human control.

C. The Common Law Tort System

The strengths and weaknesses of tort law as a mode of indirect regulation flow from the case-by-case basis on which the tort system operates and the wide variety of legal standards that the tort system employs. Tort law influences future behavior primarily through the deterrent effect of liability. But because tort cases cannot be brought until after harm occurs, courts have only a limited ability to be proactive in setting or influencing policy, a flaw that could prove quite significant if the pace of AI development accelerates further. Once a suit is brought, procedural and evidentiary rules act to focus attention on the specific facts that led to harm in that case; the ability to introduce information regarding broader social and economic considerations is limited. As a result, courts tend to give greater consideration to the risks of a technology and less to its benefits, a tendency that, if left unchecked, could stunt investment in unfamiliar but useful new technologies.[134] Taken together, these characteristics make courts well equipped to adjudicate cases arising from specific past harms, but not to make general determinations about the risks and benefits associated with emerging technologies such as AI.

1. Fact-Finding

In each tort case that comes before it, a trial court's adjudicative task is to assess the record as developed by the parties and make the findings necessary to determine the outcome of that specific case. In jury trials, the format of the vast majority of tort trials, these findings come in the form of responses to narrow questions presented on a verdict form. Courts utilize rules of procedure and evidence designed to focus the attention of both the parties and the jury on the case at hand rather than any extraneous circumstances. In this regard, courts differ sharply from legislatures and most agencies. In the legislative and administrative policymaking processes, broad social and economic considerations are often the whole point; in a tort trial, they are (or at least should be) beside the point.[135] The exclusion of information

133. *See* Viscusi, *supra* note 103, at 76.

134. *See* Huber, *supra* note 14, at 320–29.

135. Of course, determining the substantive content of tort *standards* necessarily involves some degree of policy judgment. Thus, policy considerations are legally relevant when a

about broad social outcomes makes courts particularly ill-suited for making findings regarding what usually happens in a class of cases, but it makes them ideally suited for making findings regarding what actually happened in one specific case.[136]

The courts' adjudicative role becomes much trickier when the harm is due to the confluence of multiple actors or multiple risks. In such cases, "[c]ourts must obtain some ex post information about the size of the ex ante risk caused by the injurer's action and the relative role of this risk within the context of all risk exposures."[137] These difficulties do not stem from some feature of the courts themselves, but rather are inherent in the nature of harm caused by products in industrial societies, where there generally are multiple actors who were involved in the production process and who may have contributed to the risk of harm posed by the product. Because courts have more experience than the other institutions in allocating responsibility in such situations, they remain best equipped to make such determinations of responsibility when harm occurs.

These characteristics make the tort system a mixed blessing when it comes to the management of public risks caused by emerging technologies. The intensive discovery and fact-finding processes of civil litigation provide powerful tools for unearthing relevant information regarding the design and safety features of a harm causing product, and gaining such specific and detailed information is particularly important when uncertainty regarding causal factors is high. But because both discovery and the presentation of evidence at trial will focus on the features of the product that led to the harm (and the absence of features that could have prevented the harm), the judge and jury may not have any occasion to consider the broader risk profile of the disputed technology. Each case, taken individually, thus provides an incomplete — and even misleading — factual picture of the technology at issue.

2. Reactive (and Reactionary)

Courts generally have a diminished ability to act ex ante in comparison to legislatures and agencies. Courts cannot simply decide sua sponte to announce how the law will treat liability arising from new technologies. Instead, they must wait until litigants start filing claims. In tort cases, this generally occurs only after harm accrues. Conse-

---

technology is new and the tort standards governing it are still being developed. But the policy decisions that drive the choice of standards are not among the issues that juries must decide in tort *trials*.

136. Legal process scholars occasionally use the terms "legislative facts" and "adjudicative facts" to distinguish between these two types of information. *See* HART, JR. & SACKS, *supra* note 12, at 360.

137. Viscusi, *supra* note 103, at 73.

quently, the substantive tort standards applicable to a particular technology or activity do not even begin to develop until after that technology or activity causes harm.

For emerging technologies, the reactive nature of tort law may delay the formation of industry expectations about how the courts will treat harm arising from a new technology. Tort law offers a wide array of legal rules that courts can choose to apply, and the choice of which rules should apply to a new technology can greatly impact the liability outlook for each company in the chain of design, manufacturing, and distribution. Companies may tighten supply chain management or move all development work in house if courts begin applying joint and several liability in tort cases involving their products or services. If courts decide that a product or activity is inherently dangerous and subject to the standards of strict liability rather than negligence, some companies may completely withdraw from the affected industry to avoid liability exposure.

For AI in particular, this could raise some potentially thorny issues. Depending on the context in which it is used, AI could be viewed either as a product or as a service, which could alter whether the principles of strict liability or negligence would apply to harm arising from AI. Moreover, if a learning AI's conduct depends in part on the AI's experiences during operation by the end user, courts will have to determine whether and at what point those experiences constitute a superseding cause.[138]

The reactive nature of courts also contributes to reactionary tendencies toward new risks. "[J]udges and juries, like most people unfamiliar with the quantitative aspects of risk, routinely assume that new and less familiar hazards are graver than they really are, and that older, more common ones are less severe."[139] This tendency is exacerbated by the case-specific nature of tort cases, where the judge and jury will scrutinize the specific features (or absence thereof) in the technology that caused the specific harm in that case, rather than on the aggregate social utility of the new technology as a whole. Consequently, courts may be institutionally predisposed to making regressive public risk choices.[140]

3. Incrementalism

The ex post nature of the tort system also means that the common law develops quite slowly. The path from the filing of the suit to final adjudication is long and winding; pleadings, pretrial conferences, discovery, summary judgment, and pretrial motions all must be complet-

138. *See supra* notes 45–46 and accompanying text.
139. Huber, *supra* note 14, at 319.
140. *See id.* at 320.

ed before a case reaches trial. The development of legal standards for a new technology after the first tort adjudication generally is even slower because stare decisis does not apply to prior decisions made in different jurisdictions. Indeed, the development of the law often proceeds unevenly even within a specific jurisdiction because trial courts and intermediate appellate courts may issue conflicting decisions. Because the highest courts in many jurisdictions operate under a system of discretionary review, it might be years before a jurisdiction's tort liability standards for a new technology become clear — and those standards still may vary dramatically from the applicable standards in other jurisdictions.

On the positive side, the incremental nature of the common law provides a mechanism that allows legal rules to develop organically; if a rule of law adopted in one court proves unworkable or harmful, other courts may reject or modify the rule. In its idealized form, then, the common law results in a gradual process of optimization[141] akin to the process of biological evolution or the spontaneous ordering of the free market.[142] This makes the common law attractive both as a method for fine-tuning existing legal frameworks and as an alternative to more intrusive agency regulation in relatively static industries. It makes the common law tort regime particularly unsuited, however, for controlling liability arising from technologies in rapidly changing industries.

So far, progress in AI has been incremental, at least from the perspective of the legal system; the increasing ubiquity of automated systems has not necessitated radical changes to existing legal doctrines. If AI progress continues at the same pace, we may not have to look any further than the common law tort system to find mechanisms for internalizing the costs associated with AI. But as AI technology becomes more sophisticated, its technological progress could accelerate considerably.[143] This would point to the need for common law doctrine that anticipates future technological advances in AI systems, lest those systems fall into a legal vacuum. Unfortunately, because trial courts act on the basis of the case at hand rather than on broader social implications, and because courts generally lack agencies' technical expertise (particularly with respect to emerging technologies), courts

---

141. *See, e.g.*, BENJAMIN N. CARDOZO, THE GROWTH OF THE LAW 55 (1924) ("A process of trial and error brings judgments into being. A process of trial and error determines their right to reproduce their kind.").

142. *See, e.g.*, 1 F. A. HAYEK, LAW, LEGISLATION AND LIBERTY 118 (1973) (expressing approval of "[t]he contention that the judges by their decisions of particular cases gradually approach a system of rules of conduct which is most conducive to producing an efficient order of actions"); George L. Priest, *The Common Law Process and the Selection of Efficient Rules*, 6 J. LEGAL STUD. 65, 75–77 (1971).

143. *See, e.g.*, BOSTROM, *supra* note 54, at 63–66.

cannot be expected to develop legal principles that will anticipate such changes.

4. Misaligned Incentives

The adversarial system creates incentives that do not necessarily comport with the need to optimize public risk. Plaintiffs' lawyers choose cases based on the probability of obtaining a lucrative settlement or a favorable verdict, rather than on which cases present the best opportunity to reduce future harm. Therefore, it is possible that they may focus on AI in the most visible arenas — self-driving cars, for instance — even if these are not the AI programs that pose the greatest public risk. Certainly, they will not choose cases where the potential plaintiffs actually received a net *benefit* from using the AI, even if such cases far outnumber the cases where the AI caused net harm.[144]

Such misaligned incentives become even more obvious once the case enters the courtroom. Strategic considerations, rather than scientific rigor, drive the parties' decisions regarding what witnesses to call and what evidence to present. When the litigation involves complex or highly technical subject matter, this facet of the adversarial system is keenly felt. The theoretical expertise provided by expert witnesses is undercut by the stark reality that attorneys with sufficient resources will have little trouble locating a qualified expert who will testify in support of their position. "The scientific community is large and heterogeneous, and a Ph.D. can be found to swear to almost any 'expert' proposition, no matter how false or foolish."[145] When the jury only hears from one or two such "experts" from each side, it may have difficulty discerning whose testimony conforms to scientific reality.

Agencies are not, of course, completely unaffected by the presence of hacks and cranks. But an agency consisting of people with prior knowledge of the relevant field is less likely to be hoodwinked than a lay jury or a generalist judge. Moreover, if a charlatan somehow makes his way into an agency's policymaking body, his views should be quickly marginalized due to the presence of a greater number of experts whose assessments hew closer to the prevailing consensus in the field. Agencies thus have a greater resistance to technical speciousness, even if that resistance does not rise to the level of complete immunity.

144. *See* Huber, *supra* note 14, at 323. An example might be a financial management AI system that made one very unprofitable investment but had a positive impact overall on the investor's portfolio.

145. *Id.* at 333.

IV. A REGULATORY PROPOSAL

Part IV sets forth a proposed regulatory regime for AI. The purpose of this proposal is not to provide a complete blueprint for an AI regulatory regime, but rather to start a conversation on how best to manage the public risks associated with AI without stifling innovation. To that end, the scheme outlined below proposes legislation, the Artificial Intelligence Development Act ("AIDA"), that would create an agency tasked with certifying the safety of AI systems. Instead of giving the new agency FDA-like powers to ban products it believes to be unsafe, AIDA would create a liability system under which the designers, manufacturers, and sellers of agency-certified AI programs would be subject to limited tort liability, while uncertified programs that are offered for commercial sale or use would be subject to strict joint and several liability.

AIDA leverages the respective institutional strengths of legislatures, agencies, and courts, as discussed in Part III, while taking account of the unique aspects of AI research that make it particularly difficult to regulate, as discussed in Part II. It takes advantage of legislatures' democratic legitimacy by assigning legislators the task of setting forth the goals and purposes that guide AI regulation. It delegates the substantive task of assessing the safety of AI systems to an independent agency staffed by specialists, thus insulating decisions about the safety of specific AI systems from the pressures exerted by electoral politics. This critical task is assigned to agencies because those institutions are better equipped than courts to assess the safety of individual AI systems, largely due to the misaligned incentives of the court system. Decisions regarding the safety of an emerging technology should not be informed primarily by testimony from hired guns chosen by litigants, particularly because individual court cases rarely reflect the overall risks and benefits associated with any technology.[146] Finally, AIDA leverages courts' experience in adjudicating individual disputes by assigning courts the tasks of determining whether an AI system falls within the scope of an agency-certified design and allocating responsibility when the interaction between multiple components of an AI system give rise to tortious harm.

This strong tort-based system would compel designers and manufacturers to internalize the costs associated with AI-caused harm — ensuring compensation for victims and forcing AI designers, programmers, and manufacturers to examine the safety of their systems — without the innovation-stifling effects of an agency empowered to ban certain AI systems outright.

---

146. *See supra* Parts III.C.2 and III.C.4.

A. The Artificial Intelligence Development Act

The starting point for regulating AI should be a statute that establishes the general principles for AI regulation. The legislation would establish an agency (hereinafter, the "Agency") responsible for certifying AI programs as safe and set the limits of the Agency's power to intervene in AI research and development. AIDA should begin, as do most modern statutes, with a statement of purpose. The purpose of AIDA would be to ensure that AI is safe, secure, susceptible to human control, and aligned with human interests, both by deterring the creation of AI that lack those features and by encouraging the development of beneficial AI that include those features. The Agency would be required to promulgate rules defining artificial intelligence and to update those definitional rules periodically. Rules relating to the definition of AI would have to be ratified by the legislature, because such rules effectively define the scope of the Agency's jurisdiction.

AIDA would give the Agency the authority to establish a certification system under which AI systems that are to be offered for commercial sale could be reviewed by Agency staff and certified as safe. But rather than banning uncertified AI, AIDA would operate by using a bifurcated tort liability system to encourage designers and manufacturers to go through the certification process and, even if they choose to forego certification, to ensure the safety and security of their AI.

Systems that successfully complete the agency certification process would enjoy limited tort liability — in essence, a partial regulatory compliance defense with the effect of limiting rather than precluding tort liability. For Agency-certified AI, plaintiffs would have to establish actual negligence in the design, manufacturing, or operation of an AI system in order to prevail on a tort claim. If all of the private entities involved in the development or operation of an Agency-certified AI system are insolvent, a successful plaintiff would have the option of filing an administrative claim with the Agency for the deficiency; the Agency would be required to administer a fund (funded either by Agency fees or from Congressional appropriations) sufficient to meet its anticipated obligations from such claims. Whenever a negligence suit involving the design of a certified AI system succeeds, the Agency would be required to publish a report similar to the reports that the National Transportation Safety Board prepares after aviation accidents and incidents.[147]

Companies who develop, sell, or operate AI without obtaining Agency certification would be strictly liable for harm caused by that AI. In addition, liability would be joint and several, thus permitting a

147. *See Aviation Accident Reports*, NAT'L TRANSP. SAFETY BOARD, http://www.ntsb.gov/investigations/AccidentReports/Pages/aviation.aspx [https://perma.cc/US7N-3UCR].

plaintiff to recover the full amount of their damages from any entity in the chain of development, distribution, sale, or operation of the uncertified AI. A defendant found liable in such a suit would then have to file a contribution or indemnity action to obtain reimbursement from other potential defendants.[148]

The Agency would also be required to establish rules for pre-certification research and testing of AI. These rules would permit AI developers to gather data and test their designs in secure environments so that the Agency could make better-informed certification decisions. Such testing would be exempt from the strict liability that ordinarily would attach to uncertified AI. In addition, the statute should contain a grandfather clause making programs in commercial operation twelve months before the bill's enactment presumptively exempt from the statutory scheme to prevent an undue upset of industry and consumer expectations. AIDA should, however, give the Agency the authority to create a mechanism separate from the certification process for reviewing existing AI that may present a risk to the public.

Because AI is a highly technical field, legislators are not well equipped to determine what types of AI pose a public risk. They therefore should delegate the task of formulating substantive AI policies to an agency staffed by AI specialists with relevant academic and/or industry experience. Aside from the rules set forth in the preceding paragraphs, AIDA would give the Agency the authority to specify or clarify most aspects of the AI regulatory framework, including the Agency's certification process.

B. The Agency

The new agency would have two components: policymaking and certification. The policymaking body would be given the power to define AI (though the definition would be subject to legislative ratification), create exemptions allowing for AI research to be conducted in certain environments without the researchers being subjected to strict liability, and establish an AI certification process. The certification process would require AI developers seeking certification to perform

148. This process could be analogized to the "common enterprise" theory of liability, which Vladeck proposes as a tort liability model for autonomous vehicles and, by extension, other AI systems. *See* Vladeck, *supra* note 13, at 149. This proposal avoids the common enterprise phrasing because the problem of discreteness, discussed above in Part II.B.3, means that some of the entities who design the components of an AI system may have no organizational relationship to one another, and thus would not constitute a common enterprise under the usual formulation of that doctrine. *See, e.g.*, FTC v. E.M.A. Nationwide, Inc., 767 F.3d 611, 637 (6th Cir. 2014) ("Courts generally find that a common enterprise exists 'if, for example, businesses (1) maintain officers and employees in common, (2) operate under common control, (3) share offices, (4) commingle funds, and (5) share advertising and marketing.'") (quoting FTC v. Wash. Data Res., 856 F. Supp. 2d 1247, 1271 (M.D. Fla. 2012)).

safety testing and submit the test results to the agency along with their certification application. The decision-makers in both the policymaking and certification divisions should be experts with prior education or experience with AI. The hiring process should be designed to ensure that the certification staff in particular includes an appropriate mix of specialists based on the prevailing trends in AI research.

On the policymaking front, rulemaking authority would rest with a Board of Governors (hereinafter, the "Board"). As an independent administrative entity, the Board's members would be appointed by the executive branch, subject to legislative branch approval. In addition to rulemaking, the Board would be responsible for conducting public hearings on proposed rules and amendments.

Perhaps the most important policy decision that the Agency would face is how to define artificial intelligence. Unfortunately, as noted in Part II, AI is an exceptionally difficult term to define. These difficulties make an agency best suited to determine a working definition of AI for the purposes of regulation, if only because legislatures and courts would be particularly unsuited for establishing such a definition. Again, this Article will not attempt to resolve the issue of how exactly AI should be defined. Whatever the definition the Agency ultimately chooses, it should be required to review that definition periodically and amend the definition as necessary to reflect changes in the industry. As is standard in administrative law, the definition of AI and other rules promulgated by the Agency should be published at least several months prior to the vote on their adoption, and the publication should be followed by a public comment period.

AIDA would also require the Agency to promulgate rules for pre-certification testing. Information from such testing would be a required component of any application for Agency certification, and testing conducted in compliance with Agency rules would not be subject to strict liability. The rules for such testing would be designed to ensure that the testing is done in a closed environment. For example, the rules might bar testing from being conducted on networked computers, on robots or other systems with mechanisms (e.g., access to a 3-D printer) that permit it to manipulate objects in the physical world, on systems above a certain threshold of computational power, or on systems with any other features that might permit the AI testing to have effects outside the testing environment. The Agency would have the authority to fast-track amendments to the testing requirements. Such amendments would go into effect immediately, but would also be followed by a public comment period and a subsequent vote ratifying the amendments.

After testing is completed, AI developers could file a certification application with the Agency. To provide guidance to certification applicants and set expectations within the industry, the Board would be

responsible for publishing the substantive standards under which applications for AI certification would be judged (e.g., risk of causing physical harm, goal alignment, and mechanisms for ensuring human control). The primary responsibility of the Agency's staff will be determining whether particular AI systems meet those standards. Companies seeking certification of an AI system would have to disclose all technical information regarding the product, including: (1) the complete source code; (2) a description of all hardware/software environments in which the AI has been tested; (3) how the AI performed in the testing environments; and (4) any other information pertinent to the safety of the AI. After disclosure, the Agency would conduct its own in-house testing to assess the safety of the AI program.

Given the diversity in form that AI could take, the Agency would also have the power to limit the scope of a certification. For instance, an AI system could be certified as safe for use only in certain settings or in combination with certain safety procedures. The agency could establish a fast-track certification process for AI systems or components that have been certified as safe for use in one context (e.g., autonomous road vehicles) that an entity wishes to be certified as safe for use in a different context (e.g., autonomous airplanes). A similarly streamlined certification process would be established for reviewing and approving new versions of certified AI systems, perhaps modeled on the Abbreviated New Drug Application process for generic versions of drugs that are already FDA-approved.[149]

The Agency should also promulgate rules governing licensing and warning notice requirements for certified AI. The rules could specify, for instance, that a designer or manufacturer would lose its liability protection if it sells a product to a distributor or retailer without a licensing agreement that forbids such sellers from modifying the AI system. This rule would help ensure that the product that ultimately reaches the end user is the same product that the Agency certified.

C. The Courts' Role

Courts' responsibility under the AIDA framework would be to adjudicate individual tort claims arising from harm caused by AI, harnessing courts' institutional strength and experience in fact-finding. In accordance with AIDA's liability framework, courts would apply the rules governing negligence claims to cases involving certified AI and the rules of strict liability for cases involving uncertified AI In the

149. *See* 21 U.S.C. § 355(j) (2012); *see also Abbreviated New Drug Application (ANDA): Generics*, U.S. FOOD & DRUG ADMIN., http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/AbbreviatedNewDrugApplicationANDAGenerics/ [https://perma.cc/BC7E-Q28H] (last updated Apr. 14, 2016).

latter category of cases, the most important part of this task will be allocating responsibility between the designers, manufacturers, distributors, and operators of harm causing AI. For multiple-defendant cases and actions for indemnity or contribution, allocation of responsibility should be determined in the same manner as in ordinary tort cases.[150]

It seems almost certain that, certification process and licensing requirements notwithstanding, parties in many cases will dispute whether the version of the AI system at issue was one that had been certified by the Agency, or will dispute at what point modifications took the AI outside the scope of the certified versions. In such cases, the court would hold a pre-trial hearing to determine whether the product conformed to a certified version of the system at the time it caused harm and, if it did not, the point at which the product deviated from the certified versions. That modification point will then serve as the dividing line between the defendants who enjoy limited liability and the defendants who are subject to strict liability.

V. CONCLUSION

By utilizing the tort system rather than direct regulation, the proposal outlined in Part IV charts something of a middle course — it is not as coercive as a regulatory regime that bans the production of uncertified AI systems, but it still provides a strong incentive for AI developers to incorporate safety features and internalize the external costs that AI systems generate. By using tort liability as a lever to internalize the externalities associated with AI systems, AIDA helps ensure that the prices of AI systems in the market reflect the risks associated with those systems. The imposition of joint and several liability for uncertified AI would encourage distributors, sellers, and operators to carefully examine an uncertified AI system's safety features, and the prospect of losing liability protection would discourage downstream entities from modifying a certified AI system unless they have confidence that the modification would not pose a significant public risk.

That being said, and as noted at the beginning of Part IV, this proposal is meant to start a conversation rather than to be the final word. It is not difficult to conjure up alternative approaches at least as plausible as AIDA. A less interventionist government program might resemble John McGinnis's proposal for a government entity devoted to subsidizing AI safety research,[151] perhaps combined with strong

150. *See* RESTATEMENT (THIRD) OF TORTS: APPORTIONMENT OF LIABILITY §§ 10–17 ("Liability of Multiple Tortfeasors for Indivisible Harm"); *id.* §§ 22–23 ("Contribution and Indemnity").

151. *See* McGinnis, *supra* note 61, at 1265.

tort rules that penalize AI developers who ignore the results of that safety research. If more data on AI behavior is strongly correlated with AI safety, then subsidizing such research might have a significant positive impact on the development of safer AI. A more heavy-handed regulatory regime might resemble the FDA's drug approval program, where products cannot be sold in the absence of agency approval and the approval process itself involves multiple phases of rigorous safety testing.[152] If AI truly poses a catastrophic risk, then such a rigorous approach might be necessary.

A more market-oriented approach might require the manufacturers and operators of AI systems to purchase insurance from approved carriers for their AI systems, thus letting the free market more directly determine the risk of harm that AI systems generate. A related idea would be to establish something akin to the legal fiction of corporate personhood, where AI systems would be capable both of owning assets and of being sued in court.[153] AI systems thus would be considered independent legal entities, and their owners and operators would not be subject to suit for non-intentional torts unless the AI was insufficiently capitalized or the court found another reason to "pierce the AI veil." A related framework might include applying wage laws to AI systems that perform discretionary tasks traditionally performed by humans, with a "minimum wage" set at a level sufficient to ensure that AI systems can cover the cost of expected harms. Finally, perhaps legislatures could pass "AI sunshine laws" requiring the designers and operators of AI to publicly disclose the code and specifications of AI systems, relying on members of the public to raise concerns and point out aspects of AI that might present a public risk, not unlike the manner in which Wikipedia allows members of the public to identify errors in its entries.

The appeal of each of these approaches will vary depending on the risks and benefits that individuals perceive in the further development of AI. Those who, like Elon Musk, believe that AI could pose an existential risk may favor more stringent government oversight of AI development.[154] Those who believe the public risks associated with AI to be manageable, and existential risk nonexistent, likely will op-

---

152. *See How Drugs Are Developed and Approved*, U.S. FOOD & DRUG ADMIN., http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/default.htm [https://perma.cc/WPM3-D4DA] (last updated Aug. 18, 2015).

153. *Cf.* WALLACH & ALLEN, *supra* note 45, at 198 ("It may also be in [companies producing and utilizing intelligent machines'] interests to promote a kind of independent legal status as agents for these machines (similar to that given corporations) as a means of limiting the financial and legal obligations of those who create and use them."); Vladeck, *supra* note 13, at 129 (suggesting that one approach to liability for harm caused by autonomous vehicles "would be to hold the vehicle itself responsible, assuming, of course, that the law is willing to confer legal 'personhood' on the vehicle and require the vehicle to obtain adequate insurance").

154. *See* Graef, *supra* note 9.

pose any government intervention in AI development and may countenance only limited government regulation of AI operation. Regardless, we are entering an era where we will rely upon autonomous and learning machines to perform an ever-increasing variety of tasks. At some point, the legal system will have to decide what to do when those machines cause harm and whether direct regulation would be a desirable way to reduce such harm. This suggests that we should examine the benefits and drawbacks of AI regulation sooner rather than later.