AN FDA FOR ALGORITHMS

ANDREW TUTT*

The rise of increasingly complex algorithms calls for critical thought about how best to prevent, deter, and compensate for the harms that they cause. This Article argues that the criminal law and tort regulatory systems will prove no match for the difficult regulatory puzzles algorithms pose. Algorithmic regulation will require federal uniformity, expert judgment, political independence, and pre-market review to prevent—without stifling innovation—the introduction of unacceptably dangerous algorithms into the market. This Article proposes that certain classes of new algorithms should not be permitted to be distributed or sold without approval from a government agency designed along the lines of the FDA. This "FDA for Algorithms" would approve certain complex and dangerous algorithms when it could be shown that they would be safe and effective for their intended use and that satisfactory measures would be taken to prevent their harmful misuse. Lastly, this Article proposes that the agency should serve as a centralized expert regulator that develops guidance, standards, and expertise in partnership with industry to strike a balance between innovation and safety.

TABLE OF CONTENTS

* Attorney-Adviser, Office of Legal Counsel, Department of Justice. The views expressed in this essay are the author's only and do not necessarily reflect the views of the Department of Justice or the Office of Legal Counsel. The author wishes to thank the participants in the 2016 "Unlocking the Black Box" conference at Yale Law School. Special thanks are owed to Jack Balkin, Frank Pasquale, and Jonathan Manes. The author also wishes to thank the editors of the *Administrative Law Review*, especially Ross Handler and Kimberly Koruba, for their dogged editing and inexhaustible patience.

INTRODUCTION

Algorithms play an increasingly important role in the world.[1] In the form of software programs and applications, algorithms power personal computers and smart phones. In the form of search engines, social media websites, and online stores, algorithms help to sift, filter, and

1. Algorithms are "procedure[s] for solving a given type of mathematical problem." Diamond v. Diehr, 450 U.S. 175, 186 (1981) (citing Gottschalk v. Benson, 409 U.S. 63 (1972)); *see also* WEBSTER'S NEW WORLD DICTIONARY OF COMPUTER TERMS 17 (5th ed. 1994) (An algorithm is "a mathematical or logical procedure for solving a problem. An algorithm is a recipe for finding the right answer to a difficult problem by breaking down the problem into simple steps."); PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD 1 (2015) ("An algorithm is a sequence of instructions telling a computer what to do."). Every aspect of what a computer does is determined by an algorithm. When you stream a movie, algorithms help you figure out what to watch, work together to route the movie across the Internet, and compress and decompress the data in the video. DOMINGOS, *supra* at 1.

organize the world's information. Most algorithms are no cause for concern. They are carefully crafted with detailed instructions at every step to solve narrow well-defined problems. But a new family of algorithms, "Machine Learning" algorithms, has arrived.[2] These algorithms are not programmed to solve particular problems. Instead, they are programmed to learn to solve problems.[3] To be sure, even most machine-learning algorithms are no cause for concern. Machine learning algorithms that try to predict what movies people will want to watch, or what brand of soap they will want to buy, for example, are not necessarily dangerous if they fail.

But machine-learning algorithms will soon be used to solve problems that ordinary algorithms have never solved before, or never solved nearly as well before, and, in many of those applications, they stand to pose significant risks to individuals and society if they fail or are misused.[4] Self-driving cars rely on interlocking, machine-learning

2. *See* PETER FLACH, MACHINE LEARNING: THE ART AND SCIENCE OF ALGORITHMS THAT MAKE SENSE OF DATA 3 (2012); Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 88 (2014). There is a terminological divide in legal scholarship at the moment. Some of the most prominent authors in the field prefer to conceive of algorithmic regulation as the problem of regulating robots. At times these scholars have seemed to suggest that robots present issues that are qualitatively distinct from the issues that complex algorithms present. For example, in an important white paper to which this Article is indebted, Ryan Calo argued that a "Federal Robotics Commission" should be developed to distinctly regulate robots. *See, e.g.*, RYAN CALO, THE CASE FOR A FEDERAL ROBOTICS COMMISSION 14 (2014), https://www.brookings.edu/research/the-case-for-a-federal-robotics-com miss ion/ ("Robots . . . pose unique challenges to law and to legal institutions that computers and the Internet did not."). Many scholars that write about robotics, however, appear to present robot regulation as another way of addressing the problems inherent in the regulation of complex algorithms. *See, e.g.*, Jack M. Balkin, *The Path of Robotics Law*, 6 CAL. L. REV. CIR. 45, 50 (2015) ("A robot's ability to cause physical injury is not really an 'essential' characteristic of robotic technology. It is a particularly *salient* feature of robotics for lawyers . . . ."). This Article emphasizes that algorithms are the appropriate unit of regulation because it is the changing nature of algorithms that has sparked the need to begin to contemplate a new regulatory approach.

3. DOMINGOS, *supra* note 1, at 6 (2015) ("Every algorithm has an input and an output: the data goes into the computer, the algorithm does what it will with it, and out comes the result. Machine learning turns this around: in goes the data and the desired result and out comes the algorithm that turns one into the other. Learning algorithms—also known as learners—are algorithms that make other algorithms.").

4. *See* ERIC SIEGEL, PREDICTIVE ANALYTICS 12 (2013) ("In the future—and sooner than we may think—many aspects of our world will be augmented or replaced by computer systems that today are the sole purview of human judgment.").

algorithms to make driving decisions and to "see" obstacles in the road.[5] Machine-learning algorithms will soon be consulted to make medical diagnoses, assist in surgeries, and optimize the power grid.[6] They may even be called upon to design products while other machine-learning algorithms manage the factories and warehouses that produce and distribute them. In many of these applications, people's lives may depend on the safety and efficacy of these algorithms. Yet, owing to their enormous potential complexity, it may be almost impossible to know in advance when and how they will fail.

Machine learning is not the stuff of science fiction or a far-off future. Sophisticated machine-learning algorithms are already here.[7] The ancient game of Go was long "viewed as the most challenging of classic games for artificial intelligence [AI] owing to its enormous search space and the difficulty of evaluating board positions and moves."[8] But a machine-learning algorithm—AlphaGo—is now likely the world's greatest Go player.[9] The game of *Jeopardy!* was thought to represent "a

5. At least one self-driving car algorithm has been developed, however, that can take the raw pixel input from the road and use it to solve the self-driving problem "end-to-end." *See* Mariusz Bojarski et al., *End to End Learning for Self-Driving Cars*, ARXIV 2 (2016), https://arxiv.org/pdf/1604.07316v1.pdf ("The primary motivation for this work is to avoid the need to recognize specific human-designated features, such as lane markings, guard rails, or other cars, and to avoid having to create a collection of 'if, then, else' rules, based on observation of these features."). However, in a sense even a vehicle that solves the problem "end-to-end" is combining a machine-vision algorithm with a driving algorithm—the two tasks are simply being learned in parallel.

6. *See, e.g.*, Cynthia Rudin et al., *Machine Learning for the New York City Power Grid*, 34 IEEE TRANSACTIONS ON PATTERN ANALYSIS & MACHINE INTELLIGENCE 328 (2011) (presenting machine learning methods for enhancing electrical grid reliability); Yohannes Kassahun, *Surgical Robotics Beyond Enhanced Dexterity Instrumentation: A Survey of Machine Learning Techniques and their Role in Intelligent and Autonomous Surgical Actions*, 11 INT'L J. COMPUTER ASSISTED RADIOLOGY & SURGERY 553 (2016) (reviewing "the current role of machine learning (ML) techniques in the context of surgery with a focus on surgical robotics"); Igor Kononenko, *Machine Learning for Medical Diagnosis: History, State of the Art and Perspective*, 23 ARTIFICIAL INTELLIGENCE IN MEDICINE 89 (2001) (providing "an overview of the development of intelligent data analysis in medicine from a machine learning perspective").

7. *See infra* notes 8–11 (describing sophisticated machine learning algorithms); *see also* Volodymyr Mnih et al., *Playing Atari with Deep Reinforcement Learning*, ARXIV 1 (Dec. 19, 2013), http://arxiv.org/pdf/1312.5602v1.pdf (describing "breakthroughs in computer vision and speech recognition").

8. David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 NATURE 484, 484 (2016).

9. Choe Sang-Hun, *Google's Computer Program Beats Lee Se-dol in Go*

unique and compelling AI question" because, to "compete at the human champion level," a computer "would need to produce exact answers to often complex natural language questions with high precision and speed and have a reliable confidence in its answers . . . ."[10] Yet a machine-learning algorithm—Watson—is now indisputably the world's greatest *Jeopardy!* player.[11] Only a few years ago, a myriad of other problems—from accurate speech recognition to image recognition to self-driving cars—seemed far from reality. Machine-learning algorithms have made solutions imminent.

This new family of algorithms holds enormous promise, but also poses new and unusual dangers. Machine-learning algorithms will solve problems that ordinary predictive programming never could.[12] But machine-learning algorithms are unpredictable, almost by definition. They are programmed to learn to solve problems, then taught to solve those problems, and then asked to solve those problems in extreme situations in the real world. But how machine-learning algorithms learn—and how they reason from experience to practice—is almost entirely alien.[13] Machine-learning algorithms do not learn nor reason like humans do, and that can make their outputs difficult to predict and difficult to explain.[14] The result is that in some of the most important

---

Tournament, N.Y. TIMES, Mar. 15, 2016, http://www.nytimes.com/2016/03/16/world/asia/korea-alphago-vs-lee-sedol-go.html?pagewanted=all; Cade Metz, *Google's AI Wins Fifth and Final Game Against Go Genius Lee Sedol*, WIRED (Mar. 15, 2016, 5:01 AM), https://www.wired.com/2016/03/ googles-ai-wins-fifth-final-game-go-genius-lee-sedol. Zheping Huang, *Google's Aplha Go AI Secretively Won More than 50 Straight Games Against the World's Top Go Players*, QUARTZ (Jan. 4, 2017), https://qz.com/877721/the-ai-master-bested-the-worlds-top-go-players-and-then-revealed-itself-as-googles-alphago-in-disguise/.

10. David Ferrucci et al., *Building Watson: An Overview of the Deep QA Project*, 31 ASS'N FOR ADVANCEMENT ARTIFICIAL INTELLIGENCE 59, 60 (2010).

11. John Markoff, *Computer Wins on 'Jeopardy!': Trivial, It's Not,* N.Y. TIMES (Feb. 16, 2011), http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html?pagewanted=all.

12. *See, e.g.*, DOMINGOS, *supra* note 1, at 6 (2015) ("As of today people can write many programs that computers can't learn. But, more surprisingly, computers can learn programs that people can't write.").

13. *See, e.g.*, Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, GLOBAL CATASTROPHIC RISKS 308, 313 (Nick Bostrom & Milan M. Ćirković eds., 2008) ("Any two [Artificial Intelligence] AI designs might be less similar to one another than you are to a petunia. The term 'Artificial Intelligence' refers to a vastly greater *space of possibilities* than does the term 'Homo sapiens.'").

14. *See* Cade Metz, *AI Is Transforming Google Search. The Rest of the Web Is Next*,

applications to which they might one day be placed, we will be entrusting our fates to machines we do not, and perhaps even cannot, understand.

There are already examples of machine-learning algorithms failing in ways we hardly could have predicted. Google's image recognition algorithm, an algorithm taught to label photos, labeled photos of black people as containing "Gorillas."[15] IBM's *Jeopardy!* Supercomputer Watson, in the course of winning *Jeopardy!* against two of the greatest ever human players, made an error even the worst human *Jeopardy!* player never would have made.[16] On the second day of the Final *Jeopardy!* man versus machine tournament, the category was "U.S. Cities" and the clue: "Its largest airport is named for a World War II hero; its second largest for a World War II battle."[17] The humans both answered correctly: "Chicago."[18] Watson answered "Toronto."[19]

WIRED, (Feb. 4, 2016, 7:00 AM), https://www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches/ ("Edmond Lau, who worked on Google's search team and is the author of the book *The Effective Engineer*, wrote in a Quora post that Singhal carried a philosophical bias against machine learning. With machine learning, he wrote, the trouble was that 'it's hard to explain and ascertain why a particular search result ranks more highly than another result for a given query.' And, he added: 'It's difficult to directly tweak a machine learning-based system to boost the importance of certain signals over others.' Other ex-Googlers agreed with this characterization."); *see also* ERIK BRYNJOLFSSON & ANDREW MCAFFE, THE SECOND MACHINE AGE 255 (2014) ("Because they're examples of digital technologies doing human-like things, they can lead us to conclude that the technologies themselves are becoming human-like. But they're not—yet. We humans build machines to do things that we see being done in the world by animals and people, but we typically don't build them the same way that nature built us. As AI trailblazer Frederick Jelinek put it beautifully, 'Airplanes don't flap their wings.'").

15. *See* Jessica Guynn, *Google Photos Labeled Black People "Gorillas*," USA TODAY (July 1, 2015), http://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465; Alistair Barr, *Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms*, WALL ST. J. (Jul. 1, 2015 3:40 PM), http://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/.

16. *See* WENDELL WALLACH, A DANGEROUS MASTER 226–27 (2015); Betsy Cooper, *Judges in Jeopardy!: Could IBM's Watson Beat Courts at Their Own Game?*, 121 YALE L.J. ONLINE 87, 98 (2011).

17. *See* Steve Hamm, *Watson on Jeopardy! Day Two: The Confusion Over an Airport Clue*, BUILDING A SMARTER PLANET (Feb. 15, 2011, 7:30 PM), http://web.archive.org/web/20160422135346/http://asmarterplanet.com/blog/2011/02/watson-on-jeopardy-day-two-the-confusion-over-an-airport-clue.html.

18. *See id.*

19. *See id.*

Additionally, Tesla's Autopilot system, which relies on computer vision powered by machine learning to detect obstacles on the roadway and take appropriate action, may have caused a fatal accident by failing to apply the brakes when a tractor-trailer made a left turn in front of one driver's car.[20]

No one knows precisely why these algorithms failed as they did and, in the Tesla case, it is not entirely clear the algorithms failed at all.[21] Watson's engineers at IBM thought Watson might have malfunctioned because Watson does not approach problems like humans do. In creating Watson, "[t]he IBM team paid little attention to the human brain . . . . Any parallels to the brain are superficial, and only the result of chance."[22] Watson might have said Toronto—when the category was "U.S. Cities"—because Watson knows that "categories only weakly suggest the kind of answer that is expected" and "downgrades their significance."[23] Watson may have been confused because "there are cities named Toronto in the United States and the Toronto in Canada has an American League baseball team."[24] Watson may have had trouble linking the names of Chicago's airports to World War II.[25] Maybe it was any of these explanations—or all of them. Watson's programmers did not really know, nor did they have a ready-made way to "teach" Watson not to make the same mistake again.[26]

---

20. Anjali Singhvi & Karl Russell, *Inside the Self-Driving Tesla Fatal Accident*, N.Y. TIMES, July 12, 2016, http://www.nytimes.com/interactive/2016/07/01/business/inside-tesla-accident.html; Bill Vlasic & Neal E. Boudette, *Self-Driving Tesla Was Involved in Fatal Crash, U.S. Says*, N.Y. TIMES (,June 30, 2016), http://www.nytimes.com/2016/07/01/business/self-driving-tesla-fatal-crash-investigation.html; Steve Lohr, *A Lesson of Tesla Crashes? Computer Vision Can't Do It All Yet*, N.Y. TIMES, Sept. 19, 2016, http://www.nytimes.com/ 2016/09/20/science/computer-vision-tesla-driverless-cars.html.

21. Tesla has continuously offered competing explanations and it has never been clear that a definitive source for the crash was ever discovered. *See, e.g.*, David Shepardson, *Tesla Mulling Two Theories to Explain 'Autopilot' Crash: Source*, REUTERS, July 29, 2016, http://www.reuters.com/article/us-tesla-autopilot-congress-idUSKCN10928F. The explanation of Watson's answer was hedged, using words like "probably" indicating they did not really know exactly what happened. *See* Hamm, *supra* note 17.

22. BRYNJOLFSSON & MCAFFE, *supra* note 14, at 255 (quoting Gareth Cook, *Watson, the Computer Jeopardy! Champion, and the Future of Artificial Intelligence*, SCI. AM. (Mar. 1, 2011), http://www.scientificamerican.com/article/watson-the-computer-jeopa).

23. Hamm, *supra* note 17.

24. *Id.*

25. *Id.*

26. *Id.* Watson's programmers were nonetheless delighted that Watson at least had

Tesla also remains unsure precisely what led to the fatal crash involving its autopilot system.[27] Tesla has suggested that the failure may have been purely technical—that the car's radar and camera systems may simply have failed to detect the tractor-trailer.[28] It has also been suggested that the image-recognition system may not have been able to distinguish "the white side of the tractor-trailer against a brightly lit sky . . . ."[29] Or it may have known there was an object in its path but misidentified the truck as an overpass or overhead road sign and therefore disregarded it.[30] It may also be the case that perhaps the car did know it had to stop but was not able to stop or execute another safety maneuver in time to avert the crash.[31]

One purpose of this Article is to explain why the difficulties IBM and Tesla confront in predicting and explaining the sources of failure in their algorithms are not unique; that in fact our inability to understand, explain, or predict algorithmic errors is not only unsurprising, but destined to become commonplace. What Watson's blunder and the Tesla accident both show is that when and why machine-learning algorithms fail is difficult to predict and explain because what they do is probabilistic and emergent by design. What makes them valuable is what makes them uniquely hazardous. The other purpose of this Article is to argue that a federal regulatory agency would be an effective means of dealing with the challenges posed by these kinds of complex algorithms in the future. Making those points will require cutting through diverse legal and technological fields, ranging from the cutting edge of algorithm design, to the legal-policy literature that analyzes the merits of centralized federal regulation, to the history of the FDA. The goal is to show, once the foundation is laid, that a dedicated agency charged with the mission of supervising the development, deployment, and use of algorithms will soon be highly desirable, if not necessary.

This Article is divided into four parts. Part I is a basic primer on machine-learning algorithms. The primer is meant to bring the reader up to speed on the current trajectory of algorithmic development. It endeavors to explain how machine-learning algorithms work, how they

---

very little confidence in its answer, because it showed that the algorithm was aware that it was guessing. *Id.*

27. Neal E. Boudette, *Tesla Faults Brakes, but Not Autopilot, in Fatal Crash*, N.Y. TIMES, July 29, 2016, http://www.nytimes.com/2016/07/30/business/tesla-faults-teslas-brakes-but-not-autopilot-in-fatal-crash.html.

28. *Id.*

29. Vlasic & Boudette, *supra* note 20.

30. Boudette, *supra* note 27.

31. *Id.*

differ from other algorithms, and the unique regulatory challenges they pose.

Part II builds on the explanation in Part I to explain what a regulatory agency could do to address unique challenges posed by machine-learning algorithms. An agency could provide a comprehensive means of organizing and classifying algorithms into regulatory categories by their design, complexity, and potential for harm (in both ordinary use and through misuse). The agency could prevent the introduction of certain algorithms into the market until their safety and efficacy have been proven through evidence-based pre-market trials. Such an agency could also impose disclosure requirements and usage restrictions to prevent certain algorithms' harmful misuse.

Part III addresses the legal-policy arguments for regulating algorithms through a centralized federal regulatory agency, rather than leaving such regulation to the states or to an amalgam of other federal agencies. Ultimately, the argument is that centralized federal regulation is likelier to be responsive and appropriately tailored. For consumers, tort and criminal law are unlikely to effectively counter the harms from algorithms. For innovators, the availability of federal preemption from local and *ex post* liability is likely to be desired. Thus, when compared to other approaches, regulation through a centralized agency would strike an acceptable balance between regulation and innovation.

Finally, Part IV turns from algorithms to pharmaceuticals to highlight the analogy between complex algorithms and complex drugs. With respect to the operation of many drugs, the precise mechanisms by which they produce their benefits and harms are not well understood. The same will soon be true of the most important (and potentially dangerous) future algorithms. Drawing on lessons from the fitful growth and development of the FDA, this Article proposes that the FDA's regulatory scheme is an appropriate model from which to design an agency charged with algorithmic regulation. Anticipating some objections, it offers a brief history of the FDA, to show that objections to that agency—registered throughout its century-long life—have been overcome by the public's desire to prevent major public health crises. Analogous safety concerns are likely to create pressure to regulate algorithms as well.

This Article concludes that, regardless of the path we take, there is now a need to think seriously about the future of algorithms and the unique threats they pose. A piecemeal approach may be incapable of addressing the problems presented by future algorithms.

I. WHAT "ALGORITHMS" ARE AND SOON WILL BE

A. The Basics

At their most basic level, algorithms are simply instructions that can be executed by a computer.[32] Software programs are algorithms running atop algorithms.[33] The computers we interact with each day have a set of extremely basic algorithms known as the BIOS (the Basic Input/Output System that carries out the gnomish task of telling the mechanical parts in the computer what to do.[34] Atop those algorithms runs the OS (the Operating System) that can start other software programs and shut them down.[35] And all the programs we use, from web browsers to word processors, are simply algorithms bundled together to accomplish specific tasks.

Most algorithms are extremely straightforward. The instructions are

---

32. *See* DOMINGOS, *supra* note 1, at 1 (2015) ("An algorithm is a sequence of instructions telling a computer what to do."); DONALD E. KNUTH, THE ART OF COMPUTER PROGRAMMING 1–9 (2d ed., 1973). In the 1930s, Alan Turing, Kurt Gödel, and Alonzo Church formalized what it means for a problem to be computable by an algorithm. *See id.* For more on the basics of algorithms, see Jennifer Golbeck, *How to Teach Yourself About Algorithms*, SLATE (Feb. 9, 2016, 9:45 AM), http://www.slate.com/articles/technology/future_tense/2016/02/how_to_teach_yourself_about_algorithms.single.html, and Jacob Brogan, *What's the Deal with Algorithms?*, SLATE (Feb. 2, 2016, 10:29 AM), http://www.slate.com/articles/technology/future_tense/2016/02/what_is_an_algorithm_an_explainer.html.

33. DOMINGOS, *supra* note 1, at 5 (2015) ("Algorithms combine with other algorithms to use the results of other algorithms, in turn producing results for still more algorithms . . . . Algorithms form a new kind of ecosystem . . . .").

34. *See, e.g.*, J. Dianne Brinson, *Copyrighted Software: Separating the Protected Expression from Unprotected Ideas, A Starting Point*, 29 B.C. L. REV. 803, 853 (1988) ("For example, to make an IBM-compatible computer a developer must provide a basic input/output system (BIOS). The compatible computer's BIOS is the part of the operating system that interfaces between the user's applications programs and the hardware . . . .").

35. DANIEL B. GARRIE & FRANCIS M. ALLEGRA, PLUGGED IN: GUIDEBOOK TO SOFTWARE AND THE LAW § 2:5 (2015) ("Software can be categorized in many different ways; however, one distinct group in software is the operating system. An operating system (hereinafter 'OS') is software that provides a mechanism to manage a computer's hardware and applications. It serves as the primary user interface. The OS is the functional equivalent of a 'software platform,' since other software is constructed to operate within the parameters of the particular OS.").

relatively basic and the outcomes relatively deterministic.[36] The algorithm responds to specific inputs with specific outputs that the programmer anticipated in advance. If something goes wrong, the programmer can go back through the program's instructions to find out why the error occurred and correct it.

Many extremely impressive algorithms are basically not much more complicated than that. Take Google's "PageRank Algorithm," the algorithm that made Google the company to beat in search engines.[37] The algorithm is conceptually quite simple: it determines the rank of a page by determining how many other webpages link to that page, and then it determines how much to value those links by determining how many pages link to those pages.[38] The revolutionary thing about the PageRank algorithm was not necessarily or even primarily the idea that webpages should be ranked that way, but that Larry Page and Sergey Brin figured out how to write an algorithm that could rank the whole web, which was comprised of 26 million web pages at that time, "in a few hours on a medium size workstation" using a "simple iterative algorithm."[39] PageRank, brilliant as it is, is fairly easy to grasp.

Or consider another famous algorithm: Deep Blue, the supercomputer-driven software program that defeated chess champion Gary Kasparov in 1997.[40] Deep Blue is conceptually rather simple. On its turn, the computer tried its best to make the move that would maximize its chances of winning. To do that, it would hypothesize each of the moves it could make, each of the moves that could be made in response, and so on, out to as many as six to eight moves ahead, and then it would choose the next move based on what would give it the best position several moves down the road.[41] The tough part about programming Deep Blue was figuring out how to know how good a particular future board arrangement was without simulating moves and countermoves all the way to the end of the game (which would have

36. *See, e.g.*, DOMINGOS, *supra* note 1, at 9.

37. *See* U.S. Patent No. 6,285,999 (filed Jan. 9, 1998) ("Method for Node Ranking in a Linked Database").

38. *See id.*; *see also* Sergey Brin & Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, 30 COMPUTER NETWORKS & ISDN SYS. 107 (1998), http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf.

39. *See* Brin & Page, *supra* note 38, at 107.

40. *See generally* FENG-HSIUNG HSU, BEHIND DEEP BLUE: BUILDING THE COMPUTER THAT DEFEATED THE WORLD CHESS CHAMPION (2002).

41. *See* Nate Silver, *Rage Against the Machines*, FIVETHIRTYEIGHT (Oct. 23, 2014), http://fivethirtyeight.com/features/rage-against-the-machines/ ("Deep Blue was thought to be limited to a range of six to eight moves ahead in most cases.").

been technically infeasible).[42] To do that, Deep Blue's programmers came up with over eight thousand different parameters (known as "features") that might be used to determine whether a particular board position was good or bad.[43] Yet, remarkably, "the large majority of the features and weights in the Deep Blue evaluation function were created/tuned by hand . . . ."[44] Deep Blue was kind of like a Swiss watch. It ran extremely well, but to make it tell the time its designers had to decide that they were building a watch and then handcraft all the components.[45]

Increasingly, however, algorithms are not "programmed" in the way that PageRank and Deep Blue were programmed. Rather, it would be more apt to say that they are "trained." Put simply, rather than building an algorithm that plays chess very well, programmers are now developing algorithms that can learn to play chess well. That difference will have profound consequences.

B. Trained Algorithms

The future of algorithms is algorithms that learn. Such algorithms go by many names, but the most common are "Machine Learning,"[46] "Predictive Analytics,"[47] and "Artificial Intelligence,"[48] although the use of "intelligent" and its variants can be misleading because it is more important to distinguish between algorithms that learn and algorithms that do not, than it is to distinguish between algorithms that appear

42. *See* Murray Campbell et al., *Deep Blue*, 134 ARTIFICIAL INTELLIGENCE 57, 59, 61, 63 (2002), http://ac.els-cdn.com/S0004370201001291/1-s2.0-S0004370201001291-main.pdf?_tid=29ec99c2-b0f4-11e6-bbb5-00000aacb35e&acdnat=1479847520_ce14c0 d14d 4cb639a 6c4b4e9bd3e7cfc.

43. *Id.* at 73.

44. *Id.* at 76.

45. *See* Kunihito Hoki & Tomoyuki Kaneko, *Large-Scale Optimization for Evaluation Functions with Minimax Search*, 49 J. ARTIFICIAL INTELLIGENCE RES. 527, 527 (2014), http://www.jair.org/media/4217/live-4217-7792-jair.pdf ("Fully automated learning of the heuristic evaluation functions remains a challenging goal in chess variants. For example, developers have reported that the majority of the features and weights in Deep Blue were created/tuned by hand.").

46. *See* FLACH, *supra* note 2; Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 88 (2014).

47. *See* SIEGEL, *supra* note 4, at 3–4, 9 (2013) ("Each of the preceding accomplishments is powered by prediction, which is in turn a product of machine learning.").

48. DOMINGOS, *supra* note 1, at xix, 8.

intelligent and those that do not. Learning algorithms can be almost impossibly complex, while non-learning algorithms are often not as difficult to understand.[49] As one author put it, "as of today people can write many programs that computers can't learn," but "more surprisingly, computers can learn programs that people can't write."[50]

Basic machine-learning algorithms are already ubiquitous. How does Google guess whether a search query has been misspelled? Machine learning.[51] How do Amazon and Netflix choose which new products or videos a customer might want to watch? Machine learning.[52] How does Pandora pick songs? Machine learning.[53] How do Twitter and Facebook curate their feeds? Machine learning. How did President Obama win reelection in 2012? Machine learning.[54] Even online dating is guided by machine learning.[55] The list goes on and on.[56]

Algorithms that engage in Machine Learning differ fundamentally from other algorithms.[57] Machine-learning algorithms require the programmer to answer a question conceptually different from the question a programmer confronts when building other kinds of algorithms.[58] A programmer designing a typical algorithm for use in a particular task confronts the question: "How can I make this algorithm good at performing this task?"[59] A programmer designing a machine-learning algorithm confronts the question: "How can I make this

---

49. *See id.* at 3 (2015) (explaining that, to function, algorithmic instructions must be "precise and unambiguous"); *see also id.* at 4 (quoting Richard Feynman for the proposition: "What I cannot create, I do not understand"). *But see id.* (noting that all algorithms can become too complex to effectively predict).

50. *See id.* at 6.

51. *See* Shaz Ide, *How Does Google's 'Did You Mean' Algorithm Work?,* IT ENTERPRISE (Feb. 23, 2016), http://itenterprise.co.uk/how-does-googles-did-you-mean-algorithm-work/ (Google's 'did you mean…' algorithm "could, in effect, be seen as a hybridized algorithm which is constantly changing, evolving, and expanding.").

52. *See* SIEGEL, *supra* note 4, at 5–9, 142–43; DOMINGOS, *supra* note 1, at xi–xxv.

53. *See* SIEGEL, *supra* note 4, at 5–9; DOMINGOS, *supra* note 1, at xi–xxv.

54. *See* SIEGEL, *supra* note 4, at 6, 213–217; DOMINGOS, *supra* note 1, at 16–17 ("Machine learning was the kingmaker in the 2012 presidential election.").

55. *See* SIEGEL, *supra* note 4, at 5–9; DOMINGOS, *supra* note 1, at xi–xxv.

56. *See* SIEGEL, *supra* note 4, at 5–9 (listing dozens of examples of the real-world use of machine learning from predicting mortality and injury rates to decoding from MRI scans what people are thinking, to engaging in automated essay grading).

57. *See* DOMINGOS, *supra* note 1, at xi, 9.

58. *See id.* at 6–7, 23.

59. *See id.*

algorithm good at learning to perform this task?"[60]

Sometimes the two questions are essentially the same. Consider one of the most basic machine-learning algorithms: the Spam Filter. Unwanted e-mails containing malicious software programs, links to dangerous websites, and advertisements for Viagra are sent to inboxes by the thousands each day. A challenging task is to figure out how to distinguish "spammy" e-mails from good ones.[61] Even if we think we know what makes an e-mail likely to be spammy—say, it includes an executable attachment or the words "Nigerian Prince"—it would be extremely challenging for a human to figure out precisely how much the inclusion of those things should matter when trying to distinguish spam from legitimate e-mails. A machine-learning algorithm can automate that task by seeing which e-mails the humans consider spam, and being told what information in an e-mail might be relevant to deciding on its spamminess, and then calculating for itself the optimal weights to place on each factor that together most accurately determine how to separate spam from other email.[62] At its most abstract, a spam filter could simply be given all of the information in tens of millions of e-mails and be told at the outset which are spam and which are not. The algorithm could then decide not only how much *weight* to put on the information in an e-mail, but also *which* information in an e-mail is relevant in the first place. That's how, for example, a machine-learning algorithm can intuit that the inclusion of the word "via6a" is likely to mean an e-mail is spammy without a human needing to tell it so.[63]

That last bit is, in a nutshell, both the promise and peril of the future of machine-learning algorithms. In the AI community, extracting a "feature" requires knowing what information in a dataset might be relevant to solving a problem.[64] "In essence, *features* define a 'language' in which we describe the relevant objects in our domain, be they e-mails or complex organic molecules."[65] Traditionally, "this *feature construction* process [has proven] absolutely crucial for the success of a

60. *See id.*

61. *See* FLACH, *supra* note 2, at 1–6.

62. *See id.* at 1–12 (describing how a basic spam filter works).

63. *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 11 (2013).

64. *See* FLACH, *supra* note 2, at 13, 50; *see also* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 688 (2016) ("Through a process called 'feature selection,' organizations—and the data miners that work for them—make choices about what attributes they observe and subsequently fold into their analyses.").

65. FLACH, *supra* note 2, at 13.

machine-learning application."[66] For example, a programmer might select the features of an e-mail—the words in the body of the e-mail, the names of the attachments, and the words in the subject line—and leave it to the algorithm to figure out which words are spammy and how spammy they are. As noted earlier, Deep Blue had more than 8,000 features in its evaluation function, most of them handpicked and hand-weighted.[67]

Algorithms with even basic features can be hugely complex and powerful. Consider what Google was able to do with the H1N1 virus using an algorithm about as complex as a spam filter in combination with the massive Google search database.[68] In 2009, H1N1 was spreading rapidly, but because it took a while for ill patients to consult their doctors after an infection, the Centers for Disease Control and Prevention (CDC) was only able to track the spread of the disease with a two-week delay.[69] Google unleashed an algorithm that used search terms as the feature, simply looking for correlations between search terms and H1N1 infection rates.[70] The algorithm struck gold, discovering forty-five search terms that could be used to predict where H1N1 was in real time, without a two-week lag.[71]

Even Watson has fairly straightforward features. Watson uses the outputs of many other algorithms as its features.[72] When Watson is asked a question, it uses natural language processing algorithms to extract keywords, categories, and concepts from the question.[73] Watson combines the outputs of its natural language processing algorithms with

66. *Id.* at 41.

67. Campbell, *supra* note 43, at 59, 73, 76.

68. *See, e.g.*, MAYER-SCHÖNBERGER & CUKIER, *supra* note 63, at 1–3.

69. *See id.* at 1.

70. *See id.* at 1–3.

71. *See id.* at 2; *see also* Jeremy Ginsburg et al., *Detecting Influenza Epidemics Using Search Engine Query Data*, 457 NATURE 1012, 1012–14 (2009) (publishing the results of Google's algorithm).

72. *See* SIEGEL, *supra* note 4, at 165 ("Watson merges a massive amalgam of methodologies. It succeeds by fusing technologies.").

73. Those natural language algorithms have a certain hard-coded edge to them. For example, Watson had a dedicated algorithm designed to extract puns from questions by relying on at least a few *Jeopardy!*-specific quirks (such as the fact that on *Jeopardy!* puns are often set off by quotation marks). As Eric Brown, one of Watson's programmers, revealed in one Q&A session about the puns algorithm, "it was probably not done in as general a way as you would like." *See* Eric Brown, *How Jeopardy Champ IMB Watson Handles Puns*, YOUTUBE (Oct. 30, 2016), https://www.youtube.com /watch?v=gcmh XOR7LJQ.

information retrieval algorithms—similar to the algorithms that power search engines—applied to a massive database, which included the entire contents of Wikipedia.[74] Programmers then provided Watson with thousands of *Jeopardy!* questions, along with the correct answers, and told Watson to figure out which natural language algorithms, combined with which information retrieval algorithms, maximized the likelihood that Watson would guess a correct answer.[75] The more questions Watson saw, the better Watson got at predicting which combinations of search results were most likely to be right answers.[76]

But one can easily see that it is miserably difficult, if not impossible, to figure out what Watson will do with a question it has never been asked before. And in the grand scheme, Watson is the algorithmic equivalent of a single-celled organism. It will one day be regarded as little more than a curio. In the future, programmers will unleash ultra-sophisticated algorithms on huge amounts of data with only the vaguest of goals. Those sophisticated algorithms will decide for themselves, based on the data, both what in the data is relevant and how relevant it is.[77] They will be "algorithms that make other algorithms."[78] That is, they will determine the features and weight them. Indeed, it is that development—the development of algorithms that can "extract high-level features from raw sensory data"—that has led "to breakthroughs in computer vision and speech recognition."[79]

To see the difference between old-school machine-learning algorithms and the new-school algorithms, compare Deep Blue with "Giraffe," an algorithm that uses deep reinforcement learning to play chess.[80] Deep Blue, like nearly every other chess engine ever made,

---

74. *See* SIEGEL, *supra* note 4, at 153, 157, 168.

75. *See, e.g.*, Urvesh Bhowan & D.J. McCloskey, *Genetic Programming for Feature Selection and Question-Answer Ranking in IBM Watson*, 9025 LECTURE NOTES IN COMPUTER SCI. 153, 153 (Penousal Machado et al. eds., 2015) (explaining that Watson "uses ML [machine learning] to rank candidate answers generated by the system in response to an input question using a large extremely heterogeneous feature set derived from many distinct and independently developed NLP [natural language processing] and IR [information retrieval] algorithms").

76. *See* SIEGEL, *supra* note 4, at 175.

77. *See* Quoc V. Le et al., *Building High-Level Features Using Large Scale Unsupervised Learning*, ARXIV at 1 (July 12, 2012), http://arxiv.org/pdf/1112.6209v5.pdf.

78. DOMINGOS, *supra* note 1, at 6; SIEGEL, *supra* note 4, at 115 (portraying a "computer [that] is literally programming itself").

79. Mnih, *supra* note 7, at 1.

80. *See* Matthew Lai, *Giraffe: Using Deep Reinforcement Learning to Play Chess*,

relied on human chess experts to determine how to evaluate the relative strength or weakness of a particular board arrangement by tweaking the features of the evaluation function.[81] And as any computer programmer will tell you, "almost all improvements in playing strength among the top engines nowadays come from improvements in their respective evaluation functions"—often improvements made by hand.[82]

Giraffe improves its evaluation function by going "beyond weight tuning with hand-designed features," instead using "a learned system [to] perform feature extraction" through "a powerful and highly non-linear universal function approximator which can be tuned to approximate complex functions like the evaluation function in chess."[83] Giraffe's algorithm makes it capable of learning from self-play, and the result of training for "72 hours on a machine with 2x10-core Intel Xeon E5-2660 v2 CPU" was the development of an algorithm capable of playing chess "at least comparably to the best expert-designed counterparts in existence today, many of which have been fine tuned over the course of decades."[84]

The outputs of machine-learning algorithms that engage in their own feature extraction are sometimes almost indistinguishable from magic.[85] A team of researchers was able to use deep reinforcement learning to create a single super-algorithm that could be taught to play more than a half-dozen Atari games using information "it learned from nothing but the video input, the reward and terminal signals, and the set of possible actions—just as a human player would."[86] The trained algorithm surpassed the performance of previous game-specific AIs on six of the seven games and exceeded human expert performance on three of them.[87] Video of the expert algorithm playing the Atari games is stunning.[88]

ARXIV 2, 8–9, 12–13 (Sept. 14, 2015), http://arxiv.org/pdf/1509.01549v2.pdf.

81. Campbell, *supra* note 42, at 76–77.

82. *See* Lai, *supra* note 80, at 12.

83. *Id.* at 15.

84. *Id.* at 25, 32–33.

85. *See, e.g.*, DOMINGOS, *supra* note 1, at xv (calling them "seemingly magical technologies"); *see also* Andrej Karpathy, *The Unreasonable Effectiveness of Recurrent Neural Networks*, ANDREJ KARPATHY BLOG (May 21, 2015), http://karpathy .github.io/2015/05/21/rnn-effectiveness ("There's something magical about Recurrent Neural Networks (RNNs).").

86. *See* Mnih, *supra* note 7, at 2.

87. *Id.* at 2.

88. *See* Demis Hassabis, *Deepmind Artificial Intelligence @ FDOT14*, YOUTUBE (Oct. 30, 2016), https://www.youtube.com/watch?v=EfGD2qveGdQ.

The development of ever-more-abstract and sophisticated learning algorithms is happening at an accelerating pace. Only a few years ago, it was thought that problems like accurate speech recognition, image recognition, machine translation, and self-driving cars, were many years from satisfactory algorithmic solutions.[89] But it is now apparent that learning algorithms can apply extraordinary processing power to immense datasets to achieve results that come close to human-level performance.[90] And as "remarkable" as the growth in machine-learning algorithms is, "it's only a foretaste of what's to come."[91] When "algorithms now in the lab make it to the front lines, Bill Gates's remark that a breakthrough in machine learning would be worth ten Microsofts will seem conservative."[92]

Game-changing breakthroughs will involve combining learning algorithms with other learning algorithms and incredible amounts of data to create systems that meet or exceed human performance.[93] Self-driving cars, for example, may combine algorithms that can learn to distinguish objects based on sensory input with algorithms that can use that information to learn how to drive a car.[94] Better-than-human machine translation will come from scaling up the number of sentences used to teach the algorithm, from millions to hundreds of billions,[95] relying, for example, on complementary algorithms that can discern

89. *See* Mnih, *supra* note 7, at 2.

90. *See* BRYNJOLFSSON & MCAFFE, *supra* note 14, at 34 (describing the accelerating sophistication of algorithms in several areas once thought to be intractable for computers, predicting that "we're at an inflection point").

91. DOMINGOS, *supra* note 1, at 22.

92. *Id.*

93. *Id.* at 7, 15; *see also* Kate Allen, *How a Toronto Professor's Research Revolutionized Artificial Intelligence*, THE STAR (Apr. 17, 2015), http://www.thestar.com/news/world/2015/04/17/how-a-toronto-professors-research-revolutionized-artificial-intelligence.html ("The holy grail is a system that incorporates all these actions equally well: a generally intelligent algorithm. Such a system could understand what we are saying, what we mean by what we say, and then get what we want.").

94. *See* Alexis C. Madrigal, *The Trick That Makes Google's Self-Driving Cars Work*, THE ATLANTIC (May 15, 2014), http://www.theatlantic.com/technology/archive/2014/05/all-the-world-a-track-the-trick-that-makes-googles-self-driving-cars-work/370871.

95. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 63, at 37–39 (explaining how Google's decision to use "95 billion English sentences, albeit of dubious quality" to train its translation algorithm resulted in the most accurate and rich machine-translation algorithm available).

similarities between languages to greatly increase the available training data.[96]

The upshot is algorithms are becoming increasingly self-reliant or semi-autonomous. We will soon no longer need (or wish) to provide algorithms with hard-coded hints about how to solve problems. Instead, algorithms will be provided with some basic tools for solving problems, and then left to construct for themselves tools to solve intermediate problems, on the way to achieving abstract goals.[97]

Looking twenty to forty years ahead, a fear of many futurists is that we may develop an algorithm capable of recursive self-improvement, i.e. producing learning algorithms more efficient and effective than itself.[98] That development is popularly known in the AI community as the "singularity."[99] A learning algorithm capable of developing better learning algorithms could rapidly and exponentially improve itself beyond humanity's power to comprehend through methods humans could never hope to understand.[100] Again, however, that development is probably a long way off.

C. Predictability and Explainability

Looking to the more immediate future, we confront two especially salient difficulties as learning algorithms become more sophisticated. They are the problems of "predictability" and "explainability."[101] An

96. *See, e.g.*, Tomas Mikolov, et al., *Exploiting Similarities Among Languages for Machine Translation*, ARXIV 1 (Sept. 17, 2013), http://arxiv.org/pdf/1309.4168v1.pdf.

97. *See, e.g.*, MAYER-SCHÖNBERGER & CUKIER, *supra* note 63, at 55–56 (describing how algorithms do not need to be designed with a theory about how they are supposed to make predictions); DOMINGOS, *supra* note 1, at 23–26, 40–45.

98. *See* Yudkowsky, *supra* note 13, at 314.

99. This Article acknowledges the fact that what type of algorithm would be considered the singularity is disputed. *See Singularity*, LESSWRONG WIKI (last modified Feb. 10, 2014), https://wiki.lesswrong.com/wiki/Singularity ("Singularity can be broadly split into three 'major schools'—Accelerating Change (Ray Kurzweil), the Event Horizon (Vernor Vinge), and the Intelligence Explosion (I.J. Good)."). The majority view seems to be that the singularity would be the result of the development of an algorithm that could make itself smarter or otherwise engage in "recursive self-improvement." Initially, however, the concept of the "Singularity" was coined to describe the achievement of "intelligences greater than our own." *See* BRYNJOLFSSON & MCAFFE, *supra* note 14, at 254–55 (quoting Vernor Vinge).

100. Yudkowsky, *supra* note 13, at 313–14, 323–28.

101. *See* Mark G. Core et al., *Building Explainable Artificial Intelligence Systems*, AM. ASS'N FOR ARTIFICIAL INTELLIGENCE 1 (2006), https://www.aaai.org/Papers/AAAI/2006/AAAI06-293.pdf ("These new explanation systems are not modular and not

algorithm's predictability is a measure of how difficult its outputs are to predict, while its explainability is a measure of how difficult its outputs are to explain.[102] Those problems are familiar to the robotics community, which has long sought to grapple with the concern that robots might misinterpret commands by taking them too literally (e.g., instructed to darken a room, the robot destroys the lightbulbs).[103] Abstract learning algorithms run headlong into that difficulty. Even if we can fully describe what makes them work, the actual mechanisms by which they implement their solutions are likely to remain opaque: difficult to predict and sometimes difficult to explain.[104] And as they become more complex and more autonomous, that difficulty will increase.

Explainability and predictability are not new problems. Technologies that operate on extremely complex systems have long confronted them. Consider pharmaceutical drugs. When companies begin developing those drugs, their hypotheses about why they might prove effective are little better than smart guesses. And even if the drug proves effective for its intended use, it is hard to predict its side effects because the body's biochemistry is so complex. For example, Pfizer was developing Viagra as a treatment for heart disease when it discovered that the drug is a far more effective treatment for erectile dysfunction.[105] Rogaine first came to market as Loniten, a drug used to treat high blood pressure before it was discovered that it could regrow hair.[106] Sometimes, once

---

portable; they are tied to a particular AI system."); *see also* MAYER-SCHÖNBERGER & CUKIER, *supra* note 63, at 179 ("'Explainability,' as it is called in artificial intelligence circles, is important for us mortals, who tend to want to know why, not just what."); *see generally* Yu Zhang et al., *Plan Explainability and Predictability for Cobots*, ARXIV, at 1 (Nov. 25, 2015), http://arxiv.org/pdf/1511.08158v1.pdf; Conference Paper, Ryan Turner, *A Model Explanation System* 1 (2015), http://www.blackboxworkshop.org/pdf/Turner2015_MES.pdf; David Barbella et al., *Understanding Support Vector Machine Classifications Via a Recommender System-Like Approach* (2009), http://bret-jackson.com/papers/dmin09-svmzen.pdf.

102. *See* Zhang et al., *supra* note 101, at 1.

103. *See id.*

104. *See* Barbella et al., *supra* note 101, at 1 ("Because support vector machines are 'black-box' classifiers, the decisions they make are not always easily explainable. By this we mean that the model produced does not naturally provide any useful intuitive reasons about why a particular point is classified in one class rather than another.").

105. *See* Naveen Kashyap, *Why Pfizer Won in the United States but Lost in Canada, and the Challenges of Pharmaceutical Industry*, 16 T.M. COOLEY J. PRAC. & CLINICAL L. 189, 202–03 (2014).

106. *See* John N. Joseph et al., *Enforcement Related to Off-Label Marketing and Use*

a drug is discovered, its mechanisms (including the reasons for its side effects) can be easily explained, and sometimes they cannot. But efficacy and side effects can be very difficult to predict in advance.

Humans are another example of an often unpredictable and inexplicable system. We use legal rules, incentives, entitlements, and rights to change human behavior. Nevertheless, it is sometimes difficult to know in advance whether a given social intervention will be effective and, even if it is effective, whether it will produce unintended consequences.[107] But an important difference between machine-learning algorithms and humans is that humans have a built-in advantage when trying to predict and explain human behavior.[108] Namely, we evolved to understand each other.[109] Humans are social creatures whose brains have evolved the capacity to develop theories of mind about other human brains.[110] There is no similar natural edge to intuiting how algorithms will behave.[111]

Determining that an algorithm is sufficiently predictable and explainable to be "safe" is difficult, both from a technical perspective and a public policy perspective. If an algorithm is insufficiently predictable, it could be more dangerous than we know. If an algorithm is insufficiently explainable, it might be difficult to know how to correct its problematic outputs. Indeed, it may be extremely difficult even to know what kinds of outputs are "errors." For example, a few recent articles

---

of Drugs and Devices: Where Have We Been and Where Are We Going?, 2 J. HEALTH & LIFE SCI. L. 73, 100–01 (2009); *see also* W. Nicholson Price II, *Making Do in Making Drugs: Innovation Policy and Pharmaceutical Manufacturing*, 55 B.C. L. REV. 491, 525 n.229 (2014) (noting that Rogaine's first patented use was as a treatment for high blood pressure).

107. *See, e.g.*, Samuel Issacharoff & George Loewenstein, *Unintended Consequences of Mandatory Disclosure*, 73 TEX. L. REV. 753, 785–86 (1995) (explaining that changing a certain procedural rule to respond to a problem that emerged in a limited minority of actual cases was likely to have harmful unintended consequences).

108. Yudkowsky, *supra* note 13, at 309 ("Querying your own human brain works fine, as an adaptive instinct, if you need to predict other humans.").

109. *Id.* ("Humans evolved to model other humans—to compete against and cooperate with our own conspecifics. It was a reliable property of the ancestral environment that every powerful intelligence you met would be a fellow human.").

110. *Id.* ("We evolved to understand our fellow humans *empathically*, by placing ourselves in their shoes; for that which needed to be modelled was similar to the modeller. Not surprisingly, human beings often 'anthropomorphize'—expect humanlike properties of that which is not human.").

111. *See id.* at 308–14 (discussing anthropological biases that humans have when misunderstanding the evolution of AI).

have made the point that self-driving cars will need to be programmed to intentionally kill people (pedestrians or their occupants) in some situations to minimize overall harm and thereby implement utilitarian ethics.[112] Crash investigators deconstructing a future accident may want to know whether the accident was a result of the ethics function or a critical algorithmic error. The self-driving car algorithm's explainability will be crucial to that investigation.

What we know, and what can be known, about how an algorithm works will play vital roles in determining whether it is dangerous or discriminatory.[113] Algorithmic predictability and explainability are hard problems. And they are as much public policy and public safety problems as technical problems.[114] At the moment, however, there is no centralized standards-setting body that decides how much testing should be done, or what other minimum standards machine-learning algorithms should meet, before they are introduced into the broader world.[115] Not only are the methods by which many algorithms operate non-transparent, many are trade secrets.[116]

---

112. *See, e.g.*, Jean-François Bonnefon et al., *Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?*, ARXIV (Oct. 12, 2015), http://arxiv.org/pdf/1510.03346v1.pdf; *Why Self-Driving Cars Must Be Programmed to Kill*, MIT TECH. REV. (Oct. 22, 2015), https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be-programmed-to-kill.

113. *See generally* Barocas & Selbst, *supra* note 64 (discussing the results of not fully understanding algorithms and programmer error on discrimination law).

114. *See generally* CALO, *supra* note 2.

115. Recently, the National Highway Transportation Administration (NHTSA) released comprehensive guidance pertaining to self-driving cars. *See* Cecilia Kang, *Self-Driving Cars Gain Powerful Ally: The Government*, N.Y. TIMES, Sept. 19, 2016, http://www.nytimes.com/2016/09/20/technology/self-driving-cars-guidelines.html; Cecilia Kang, *The 15-Point Federal Checklist for Self-Driving Cars*, N.Y. TIMES, Sept. 20, 2016, http://www.nytimes.com/2016/09/21/technology/the-15-point-federal-checklist-for-self-driving-cars.html; Joan Lowy et al., *Innovation, Safety Sought in Self-Driving Car Guidelines*, ASSOC. PRESS NEWS (Sept. 20, 2016), http://bigstory.ap.org/article/921af0749a12495781606094d3984ccc/feds-preview-rules-road-self-driving-cars. The NHTSA guidance comes close to the kind of guidance that an algorithm-specific agency would issue. But it is limited to self-driving cars and it is unclear how NHTSA will acquire the expertise necessary to effectively ensure the safety and efficacy of the algorithms that automobile manufacturers develop. Reports indicate that "[t]he agency, for the first time in its history, may even seek authority from Congress to approve technology before it goes on the road," similar to the pre-market review proposed in this Article. *Id.*

116. *See generally* Frank Pasquale, *Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries*, 104 NW. U. L. REV. 105 (2010)

II. THINGS AN AGENCY COULD SORT OUT

The rising complexity and varied uses of machine-learning algorithms promise to raise a host of challenges when those algorithms harm people. Consider three: (1) algorithmic responsibility will be difficult to measure; (2) algorithmic responsibility will be difficult to trace; and (3) human responsibility will be difficult to assign.[117]

Consider the difficulty of measuring algorithmic responsibility. The problem is multi-faceted. Algorithms are likely to make decisions that no human would have made in a variety of circumstances no human has confronted or even could confront. Those decisions might be a "bug" or a "feature." Often it will be difficult to know which.[118] A self-driving car might intentionally cause an accident to prevent an even more catastrophic collision. A stock-trading algorithm may make a bad bet on the good faith belief (whatever that means to an algorithm) that a particular security should be bought or sold. The point is, we have a generally workable view of what it means for a person to act negligently or otherwise act in a legally culpable manner, but we have no similarly well-defined conception of what it means for an algorithm to do so.[119]

Next, consider the difficulty of tracing algorithmic harms. Even if algorithms were programmed with specific attention to well-defined legal norms, it could be extremely difficult to know whether the algorithm behaved according to the legal standard or not in any given circumstance. The stock trading algorithm that made the bad bet might have made its decision based solely upon the "signal" in its training data—i.e., the algorithm was right about the circumstance it was confronting, but the event it predicted did not come to pass. Or it might

---

(discussing the difficulties in regulating search engine functions because of information asymmetry—between the consumers and companies like Google—and trade secret protections).

117. For scholarly articles explaining and addressing some of the issues raised in the paragraphs that follow, see, for example, Jack Boeglin, *The Costs of Self-Driving Cars: Reconciling Freedom and Privacy with Tort Liability in Autonomous Vehicle Regulation*, 17 YALE J.L. & TECH. 171, 186 (2015); F. Patrick Hubbard, *"Sophisticated Robots": Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 1803 (2014); and David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117 (2014). Ryan Calo has also written about this issue with respect to robots and reached a similar conclusion that a federal agency is warranted. *See generally* CALO, *supra note* 2.

118. *See* CALO, *supra* note 2, at 7 (proposing that while driverless cars will reduce the number of accidents overall, they will create new kinds of accidents).

119. *See* WALLACH, *supra* note 16, at 239–43.

have made its decision based on "noise" in the training data—i.e., the algorithm looked for the wrong thing in the wrong place. Algorithms that engage in discrimination offer a good example. Suppose a company used a machine-learning algorithm to screen for promising job candidates. That algorithm could end up discriminating on the basis of race, gender, or sexual orientation—but tracing the discrimination to a problem with the algorithm could be nearly impossible. To be sure, the discrimination could be a result of a bug in the design of the training algorithm or a typo by the programmer, but it could also be because of a problem with the training data, a byproduct of latent society-wide discrimination accidentally channeled into the algorithm, or even no discrimination at all, but instead a low-probability event that just happened to be observed.[120]

Finally, consider the difficulty in fixing human responsibility. Algorithms can be sliced-and-diced in several ways that many other products are not.[121] A company can sell only an algorithm's code or even give it away. The algorithm could then be copied, modified, customized, and reused or used in a variety of applications its initial author never could have imagined. Figuring out how much responsibility the original developer bears when any harm arises down the road will be a difficult question. Or consider a second company that sells training data for use in developing one's own learning algorithms, but does not sell any algorithms itself. Depending on the algorithm the customer trains, and the use to which the purchaser wishes to put the data, the data's efficacy could be highly variable, and the responsibility of the data seller could be as well. Or imagine a third company that sells algorithmic services as a package, but the algorithm it offers relies partially or extensively on human interaction when determining its final decisions and outputs (e.g., a stock trading algorithm where a human must confirm all the proposed trades). Divvying up responsibility between the algorithm and the human is likely to prove complicated.

With those challenges in mind, the following subsections suggest the kinds of issues a federal agency could sort out.

120. *See* Barocas & Selbst, *supra* note 64.

121. Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 534 (2015).

A. Acting as a Standards-Setting Body

At its most basic, a federal agency could act as a standards-setting body that coordinates and develops classifications, design standards, and best practices.[122]

1. Classification

An agency could develop categories for classifying algorithms, varying the level of regulatory scrutiny based on the algorithm's complexity. Under a sufficiently nuanced rubric, the vast majority of algorithms could escape federal scrutiny altogether. For example, the agency could classify algorithms into types based on their predictability, explainability, and general intelligence, but only subject the most opaque, complex, and dangerous types to regulatory scrutiny—thereby leaving untouched the vast majority of algorithms with relatively deterministic and predictable outputs.

Table 1. A Possible Qualitative Scale of Algorithmic Complexity

Algorithm Type	Nickname	Description
Type 0	"White Box"	Algorithm is entirely deterministic (i.e., the algorithm is merely a pre-determined set of instructions).
Type 1	"Grey Box"	Algorithm is non-deterministic, but its non-deterministic characteristics are easily predicted and explained.
Type 2	"Black Box"	Algorithm exhibits emergent proprieties making it difficult or impossible to predict or explain its characteristics.
Type 3	"Sentient"	Algorithm can pass a Turing Test (i.e., has reached or exceeded human intelligence).
Type 4	"Singularity"	Algorithm is capable of recursive self-improvement (i.e., the algorithm has reached the "singularity").

122. *See* CALO, *supra* note 2, at 3–5, 11–12 ("Agencies, states, courts, and others are not in conversation with one another. Even the same government entities fail to draw links across similar technologies; drones come up little in discussions of driverless cars despite presenting similar issues of safety, privacy, and psychological unease. Much is lost in this patchwork approach.").

2. Performance Standards

An agency could also establish guidance for design, testing, and performance to ensure that algorithms are developed with adequate margins of safety. That guidance, in turn, could be based on knowledge of an algorithm's expected use, types of critical versus acceptable errors it might make, and the suggested predicted legal standard to apply to accidents involving that algorithm.

Table 2. Sample Possible Performance Standards

Algorithm	Performance Standard	Based On
Self-Driving Car (Autonomous)	With 95% statistical confidence, the algorithm must be involved in fewer than 1.13 fatal accidents per 100 million vehicle miles, and there must be fewer than 80 injuries per 100 million miles traveled.	Risk of death and injury per 100 million miles driven in 2012.[123]
Stock Trading Algorithm (Autonomous)	An algorithm's average return volatility must be predicted with 95% confidence based on historical data, and that volatility must be reported to investors.	Typical measure of the risk of a security (price volatility)
Job Applicant Screening Algorithm (Autonomous)	With 95% confidence, the pool of favored applicants drawn from a set of applicants must not underrepresent any protected class (based on EEOC guidance) by more than 20%.	The "80% rule" in the Uniform Guidelines on Employee Selection Procedures[124]

123. NHTSA, DOT, TRAFFIC SAFETY FACTS 2012 15 (2012), http://www-nrd.nhtsa.dot.gov/Pubs/812032.pdf.

124. *See Uniform Guidelines on Employee Selection Procedures*, 29 C.F.R. § 1607.4D (1997); *The Four-Fifths or Eighty Percent Rule*, [5 Emp. Practices] EMP. COORD. (RIA) § 23:28 (Oct. 2016) ("Under the Uniform Guidelines, a test or other selection procedure is generally regarded as having an adverse impact where its selection rate for any race, sex, or ethnic group is less than four-fifths (or 80%) of the rate for the identifiable group with the highest rate.").

3. Design Standards

An agency could also consider the knotty problem of establishing satisfactory measures of predictability and explainability and promulgate guidance for developing algorithms that meet those standards. Especially with respect to explainability, there is reason to believe that algorithm designers can design machine-learning algorithms with attention to ensuring explainability. For example, through testing, programmers might develop more transparent algorithms that match the performance of black-box algorithms by discovering the hidden features that make a particular black-box algorithm effective.[125] If explainability can be built into algorithmic design, the presence of a federal standard could nudge companies developing machine-learning algorithms into incorporating explainability from the outset.

4. Liability Standards

An agency could also make progress toward developing standards for distributing liability for harms among coders, implementers, distributors, and end-users. The development of such standards will prove complex and require careful consideration of many factors, including impacts on innovation, compensation for victims, and problems of justice and fairness. An agency could bring together diverse stakeholders—from the open source community to commercial firms, to customers, to potential victims—to develop flexible guidelines that do not unduly stifle innovation.

B. Acting as a Soft-Touch Regulator

A federal agency could also nudge algorithm designers through soft-touch regulations. That is, it could impose regulations that are low enough cost that they preserve freedom of choice and do not substantively limit the kinds of algorithms that can be developed or when or how they can be released.[126]

---

125. *See id.*

126. *Cf.* Cass R. Sunstein, *Nudges vs. Shoves*, 127 HARV. L. REV. F. 210, 211 (2014) (terming low-cost choice-preserving regulations "nudges"); Cass R. Sunstein, *The Storrs Lectures: Behavioral Economics and Paternalism*, 122 YALE L. J. 1826, 1830–31 (2013) (same assertion); Cass R. Sunstein, *The Ethics of Nudging*, 32 YALE J. ON REG. 413, 414 (2015) (same assertion). For a book-length treatment, see RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE (2008).

1. Transparency

Among the most meaningful soft-touch regulations an agency could impose would be requirements of openness, disclosure, and transparency.[127] There appears to be a growing consensus among scholars that the ability to require transparency should be one of the first tools used to regulate algorithmic safety.[128] Transparency can take many forms and can range from feather-light to brick-heavy.

Table 3. A Spectrum of Disclosure[129]

	Depth of Disclosure	Scope of Disclosure	Timing of Disclosure
Preserving Secrecy	Shallow and cursory	To small group of outside experts	Delayed for years or decades
Providing Transparency	Deep and thorough	To the public generally	Immediate

On the lighter end, an agency could require that certain aspects of certain machine-learning algorithms (their code or training data) be certified by third-party organizations, helping to preserve the trade secrecy of those algorithms and their training data. Intermediately, an agency could require that companies using certain machine-learning algorithms provide qualitative disclosures (analogous to SEC disclosures) that do not reveal trade secrets or other technical details about how their algorithms work but nonetheless provide meaningful notice about how the algorithm functions, how effective it is, and what errors it is most likely to make.

On the heavier end, in appropriate circumstances, the agency could

127. *But see* Eric A. Posner & E. Glen Weyl, *An FDA for Financial Innovation: Applying the Insurable Interest Doctrine to Twenty-First-Century Financial Markets*, 107 NW. U. L. REV. 1307, 1355 (2013) (explaining that although a disclosure requirement is a "less heavy-handed form of regulation" it is a "notoriously weak" form of regulation).

128. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 63, at 176–84 (describing recommended accountability mechanisms as disclosure and certifications); Pasquale, *supra* note 116, at 140–88 (offering a detailed account of the types of transparency that could be required and the public policy motivations that might drive particular disclosure solutions).

129. FRANK PASQUALE, THE BLACK BOX SOCIETY 142 (2015).

require that technical details be disclosed, potentially preempting state-level trade secret protections in the name of public safety. Frank Pasquale has discussed the pros and cons of requiring various kinds of transparency in depth in his book *The Black Box Society*.[130] Without addressing the benefits and drawbacks of striking any particular balance, it is worth emphasizing that the complex tradeoffs between innovation and safety will demand extensive and careful study. An agency could strike that difficult balance in a granular way by drawing together many stakeholders and mandating only those disclosures that are most appropriate to certain kinds of algorithms used in specific contexts.

C. Acting as a Hard-Edged Regulator

Finally, a federal agency could act as a hard-edged regulator that imposes substantive restrictions on the use of certain kinds of machine-learning algorithms, or even with sufficiently complex and mission-critical algorithms, act as a regulator that requires pre-market approval before algorithms can be deployed.

1. Pre-Market Approval

Among the most aggressive positions an agency could take would be to require that certain algorithms slated for use in certain applications receive approval from the agency before deployment. That pre-market approval process could provide an opportunity for the agency to require that companies substantiate the safety performance of their algorithms. For example, a self-driving car algorithm could be required to replicate the safety-per-mile of a typical vehicle driven in 2012. The agency could work with an applicant to develop studies that would prove to the agency's satisfaction that the algorithm meets that performance standard. Algorithms could also be conditionally approved subject to usage restrictions—for example, a self-driving car algorithm for cruise control could be approved subject to the condition that it is only approved for highway use. Off-label use of an algorithm, or marketing an unapproved algorithm, could then be subject to legal sanctions.

III. OTHER REGULATORY OPTIONS AND THEIR INADEQUACY

Although the regulation of complex algorithms is inevitable, there are at least two competing alternative regulatory paths that might be

130. *Id.* at 140–48.

pursued other than the creation of a centralized federal agency. One alternative to a federal agency would be regulation state-by-state.[131] In that scenario, most algorithmic regulation could be left to the tort and criminal law systems of the several states, or regulation could be performed by a combination of state-level agency, statutory, tort, and criminal regulation. A second alternative to a single federal agency would be regulation across several agencies regulating algorithms incident to their primary jurisdiction. In that scenario, the National Highway Transportation Safety Administration (NHTSA) would regulate self-driving cars, the FTC might regulate the Internet, the FAA would regulate drones, etc. Both of those alternatives seem, on balance, to be inferior to regulation through a centralized agency.

A. The Case for State Regulation

A weak case could be made that algorithm regulation should be left to the states to develop. The tort regulatory system has effectively, if imperfectly, dealt with transformational technological change in the past, adapting common-law tort precepts to the problems posed by modern industrial society, perhaps most notably the development of the automobile.[132] The states are famously laboratories of legal innovation, and competition between the states can sometimes produce a race to the top that tends toward optimal legal rules.[133] One could argue that, for those reasons, state-level regulation might prove agile, responsive, and effective.

Moreover, even if one were inclined to think that state-by-state regulation would not be particularly effective, one might nonetheless

---

131. The suggestion in the text that the two alternatives are state-level regulation or regulation by a federal agency assumes that Congress will not attempt to regulate algorithms through detailed and responsive legislation nor through a kind of federal tort regulatory system. In light of the long practice of Congress in these matters, that seems like a safe assumption.

132. *See* JOHN FABIAN WITT, THE ACCIDENTAL REPUBLIC 3–4 (2004); G. Edward White, *The Emergence and Doctrinal Development of Tort Law, 1870–1930*, 11 U. ST. THOMAS L.J. 463, 465–66 (2014) (describing the emergence of transportation accidents as central to the development of early twenty-first century tort law).

133. *See* Robert A. Schapiro, *Toward A Theory of Interactive Federalism*, 91 IOWA L. REV. 243, 267 (2005); Michael C. Dorf, *Foreword: The Limits of Socratic Deliberation*, 112 HARV. L. REV. 4, 60–61 (1998) (describing the states-as-laboratories theory); ROBERTA ROMANO, THE GENIUS OF AMERICAN CORPORATE LAW (1993) (arguing that jurisdictional competition between States for corporate charters produces efficient corporate law).

prefer it because it would be more effective than federal regulation. Federal agencies have been criticized for tending toward three forms of failure: (1) "tunnel vision," in which they do not engage in cost-justified regulation because they are unduly focused on carrying out their narrow mission without attention to broader side effects of regulatory choices;[134] (2) "random agenda selection," in which they tend to focus on high-salience political issues rather than the issues that pose the greatest threat to public safety;[135] and (3) "inconsistency," in which they treat similarly situated risks differently.[136] It might be argued that state-level regulation could better grapple with those sources of failure than a federal agency could because, for example, state legislatures are more attuned to competing priorities and stakeholders, and so will not as readily fall prey to tunnel vision and inconsistency.

But the case for state-level regulation is rather weak. An appropriately-structured federal agency is as capable of solving the tunnel vision, random agenda selection, and inconsistency problems as the states are.[137] Indeed, the solution offered by those who levy those criticisms of federal regulation is that federal agencies should place a greater premium on expertise and should be more politically insulated.[138] Moreover, the efforts of generalist state judges to adapt common law principles to rapidly evolving technological developments are likely to be fitful, imperfect, and slow. Further, state legislatures are as susceptible to regulatory capture as federal agencies—sometimes even more so.[139]

Algorithms also pose national problems, and such problems generally call for national solutions. The mobile nature of algorithms makes their regulation a national problem. Most of the technologies in which algorithms are embedded or extensively used are likely to be involved

---

134. *See* STEPHEN BREYER, BREAKING THE VICIOUS CIRCLE: TOWARD EFFECTIVE RISK REGULATION 10–19 (1993).

135. *See id.* at 19–21.

136. *See id.* at 21–29.

137. *See id.* at 59–63.

138. *See id.* at 55–81.

139. *See, e.g.*, Oren Bar-Gill & Elizabeth Warren, *Making Credit Safer*, 157 U. PA. L. REV. 1, 98–99 & nn.323, 325 (2008) (noting that "It is not clear that diffuse authority is less prone to regulatory capture than concentrated authority. For example, consumer groups find it difficult to oppose well-funded banking interests at multiple state legislatures, and they may be better able to serve as an effective counterweight at a single federal regulator."); Merrick B. Garland, *Antitrust and State Action: Economic Efficiency and the Political Process*, 96 YALE L.J. 486, 499 (1987) (mentioning that "special interests can capture state legislatures as well as regulatory bodies").

in national commerce—be it because they provide their services through the Internet, or because they are embedded in technologies like cars, planes, and drones. Absent a compelling case that algorithmic regulation would lead to a rapid race-to-the-top regulatory effort by the states, the most likely outcome of state-level regulation will be a checkerboard of regulatory efforts, with different standards of safety applicable in different geographic regions. That outcome is likelier to stifle innovation than to promote it.[140] Algorithm designers would probably prefer meeting a single national standard than attempting to figure out how to comply with the state-level standards of fifty jurisdictions.

B. The Case for Federal Regulation by Other Subject-Matter Agencies

It might also be argued that algorithms should not be treated as a single regulatory category but should instead be thought of as a kind of helper technology that should be regulated incident to the regulation of other technologies or fields, such as vehicles, aircraft, and the Internet. The argument would be that the bureaucratic burden of imposing double or overlapping regulatory jurisdiction would outweigh the benefit of obtaining the expertise of a single central agency.

The arguments for a central regulating agency are rather strong. Machine-learning algorithms will pose systematic, complex challenges that will transcend the technology with which they are associated. The same machine-learning algorithm could one day be deployed to drive a car and fly an airplane. Watson could be used to yield expert guidance in fields ranging from medicine to finance. Placing regulatory jurisdiction in multiple agencies would only make the problems of tunnel vision, random agenda selection, and inconsistency more acute. The same algorithm could be regulated two different ways depending on whether it is deployed in a car or a drone. In addition, lessons learned in developing regulatory solutions for one set of algorithms would not be readily available to other agencies developing solutions to identical or highly similar algorithms. Even if other agencies had overlapping jurisdiction, that would not necessarily undermine the case for a single

140. *Cf.* Bar-Gill & Warren, *supra* note 139, at 98–99 n.323 (explaining that "an optimally designed regulatory framework at the federal level is superior to state-level regulation" of consumer credit products because "not all states will be equally motivated to regulate consumer credit products," "not all states will be equally effective in regulating consumer credit products," and because "state-level regulation will potentially expose national lenders to fifty different regulatory regimes").

central expert agency. Often two or more agencies share regulatory jurisdiction and work jointly to develop comprehensive regulatory strategies.[141] Moreover, "shared responsibility may create a healthy competition between the two agencies, and it will be harder to capture two agencies instead of one."[142] Thus, a new federal agency in this space could add significant value—in the form of centralized expertise—even if other agencies retained primary jurisdiction over specific technologies.

Many of the foregoing arguments for a single-expert regulator are similar to those made by Oren Bar-Gill and Elizabeth Warren in their seminal article arguing for the creation of the Consumer Financial Protection Bureau (CFPB), *Making Credit Safer*.[143] As they explained in that article, certain features of the then-existing landscape of consumer financial protection regulation made it ineffectual and warranted replacement by a single agency.[144] They conceded that some agencies (bank regulators) already had the authority to protect consumers in the way a new agency would, and they admitted that at least one agency had the motivation to protect consumers (the FTC) in the way a new agency would.[145] But, they argued, the agencies with authority lacked the motivation to engage in consumer financial protection and the agency with motivation lacked the authority.[146] Bar-Gill and Warren concluded that the "litany of agencies [with overlapping authority], limits on rulemaking authority, and divided enforcement powers result[ed] in inaction."[147] Centralizing authority in a single agency with specialized expertise and a clear mission would unite authority and motivation and in that way improve the regulation of consumer financial protection.[148]

Bar-Gill and Warren's arguments apply with similar vigor here. Many

---

141. *See, e.g.*, Jody Freeman & Jim Rossi, *Agency Coordination in Shared Regulatory Space*, 125 HARV. L. REV. 1131, 1134 (2012) ("Many areas of regulation and administration are characterized by fragmented and overlapping delegations of power to administrative agencies.").

142. *See* Rachel E. Barkow, *Insulating Agencies: Avoiding Capture Through Institutional Design*, 89 TEX. L. REV. 15, 53 (2010).

143. *See generally* Bar-Gill & Warren, *supra* note 139, at 90, 98.

144. *Id.* at 86–97 (explaining that a new agency focused on protecting consumers is required because five separate agencies were designed "with a primary mission to protect . . . banks' profitability. Consumer protection is, at best, a lesser priority . . .").

145. *Id.*

146. *Id.* at 85–95 (describing the authority of various federal agencies along with their "lack of interest in exercising [that] power.").

147. *Id.* at 97.

148. *Id.* at 97–100.

safety agencies will have authority over a small slice of the algorithms that are developed, but they will lack the expertise and the motivation to regulate them consistently and effectively. A single highly-motivated regulator could develop comprehensive policy, could quickly respond to new products and practices, and could also ensure that consumers are adequately protected.

C. The Case for a Central Federal Agency

The case for regulation by a single expert agency outweighs the case for regulation by the states or jurisdiction distributed across multiple agencies because algorithms have qualities that make centralized federal regulation uniquely appealing. There are at least three qualities intrinsic to algorithms that make a national regulatory solution warranted—and, in particular, a national regulatory solution that may include pre-market approval requirements for some algorithms.[149]

1. Complexity

First, the kinds of algorithms that are most concerning are by their nature opaque, with benefits and harms that are difficult to quantify without extensive expertise. That feature of the market for algorithms contrasts sharply with the market for most products where individuals are able to easily assess the benefits and safety risks posed. Highly opaque and complex products benefit more from expert evaluation by a regulator than other products.[150]

2. Opacity

Second, the difficulties with assigning and tracing responsibility for harms to algorithms, and then associating that responsibility with human actors, further distinguish algorithms from other products. Algorithms could commit small but severe long-term harms or may commit grievous errors with low probability. Therefore, unlike many other products for which a combination of tort regulation and reputation will correct for accidents at an acceptable pace, the market and tort regulatory system are likely to prove too slow to respond to

149. *See* Posner & Weyl, *supra* note 127, at 1349–51 (naming the need for expertise in understanding product failures, delayed and uncertain feedback regarding when and how product defects occur, and the extent of potential danger arising from product failure as the three criteria that militate most strongly in favor of a safety agency).

150. *See id.* at 1349–50.

algorithmic harms.[151]

3. Dangerousness

Third, at least in some circumstances, algorithms are likely to be capable of inflicting unusually grave harm. When a machine-learning algorithm is responsible for keeping the power grid operational, assisting in a surgery, or driving a car, it can pose an immediate and severe threat to human health and welfare in a way many other products simply do not.

A central regulatory agency with pre-market review would be better able to contend with those problems than the states or an amalgam of subject-matter agencies working independently. Take expertise: to the degree significant subject matter expertise is required to understand the possible dangers an algorithm may pose, a single, central regulatory agency is more likely to be able to pool top talent than fifty jurisdictions or ten agencies. Take agility and nuance: a single federal regulator could grapple with the dangers algorithms pose holistically rather than piecemeal—effectively distinguishing between algorithms based on stakeholder feedback and expert judgment. A single national agency would be able to maximize the centralized expertise that can be brought to bear on the issue while offering the most agility and flexibility in responding to technological change and developing granular solutions.

D. But What Kind of Agency?

An agency with all the regulatory powers set out above may be warranted. But many structural and institutional questions remain: whether the agency should have a commission structure (like the SEC and the FTC) or a Director (like NHTSA and the CFPB); whether the agency should be independent, quasi-independent, or politically accountable; whether the agency's enforcement powers should be internal (by administrative law judges (ALJs)) or external (through the courts); and, whether the agency should be authorized to litigate on its own behalf or rather be required to rely on the Department of Justice to implement its enforcement authority.

Learning from recent analogous proposals to develop new safety regulators for financial products, any proposed regulatory framework should include three features.[152] First, the agency should be able to engage in *ex ante* regulation rather than relying on *ex post* judicial enforcement. Second, the agency should have a broad mandate to

151. *Id.* at 1350–51.

152. *See id.* at 1349–51; Bar-Gill & Warren, *supra* note 139, at 10.

ensure that unacceptably dangerous algorithms are not released onto the market, rather than charged with the enforcement of piecemeal legislation. Third, the agency should have ultimate authority over algorithmic safety regardless of the type or kinds of products in which those algorithms are embedded.

First, *ex ante* regulation is important because many of the types of harms that algorithms might cause can be mitigated through careful efforts at the design and development stage—including extensive pre-market testing and reliance on certain classes of explainable and predictable learning algorithms. Moreover, *ex post* judicial enforcement would likely be too blunt to effectively ensure unsafe algorithms will be kept off the market. Second, broad rulemaking and enforcement authority is important because of the high rate of innovation in the industry and the expertise necessary to understand algorithmic products. An agency charged with narrow authority to regulate only certain kinds of algorithms, or algorithms in only certain contexts, would be incapable of effectively responding to innovations that might prove unusually dangerous. Third, the agency must have ultimate authority over algorithms to eliminate regulatory gaps and contradictions and ensure that the states and other federal regulators do not undercut the agency's regulatory mission. Although concentrated broad authority runs the risk of regulatory capture, the alternative is likelier to be incoherence and inaction. At this juncture, the CFPB appears to be the state-of-the-art when it comes to consumer protection agency design. It combines the effectiveness of a single Director with the insulation traditionally afforded a commission-structured agency.[153] It has the full complement of conventional agency powers: rulemaking, enforcement, and adjudication. It can litigate on its own behalf and choose between prosecuting enforcement actions before its own ALJs or in the courts. The CFPB's design has made the agency remarkably nimble, powerful, scalable, and effective. At least at this early stage, the CFPB archetype seems like a good fit for an agency designed to make difficult tradeoffs between innovation and safety in a fast-paced industry.[154]

---

153. *See* PHH Corp. v. Consumer Fin. Prot. Bureau, No. 15-1177, 2016 WL 5898801, at *2–5 (D.C. Cir. 2016) (holding that the feature of the Consumer Financial Protection Bureau that makes it unique—a tenure protected single-director—is unconstitutional). To the extent that decision withstands further appeal, it would require any independent agency to be structured as a commission like the SEC and FTC.

154. *See id.* Again, in light of the D.C. Circuit's recent decision, an independent commission may be required because an agency cannot have a single for-cause

IV. THE FDA MODEL: THE ANALOGY BETWEEN DRUGS AND ALGORITHMS

Many will be skeptical that a new federal agency is warranted, and several arguments could be made against it. First, it might be argued that it is too soon to develop a regulator because algorithmic technology is still in its infancy. Second, it might be contended that algorithms are not a species of technology that calls for extensive regulation and oversight. Third, it might be offered that regulation is harmful in principle and that the public benefits most when there are fewer regulations and fewer obstacles to private-sector innovation.

Each of those arguments can be countered with logic. One might counter the first argument, for example, by asserting that the exponential pace at which algorithms develop means that we will likely progress from "too soon" to regulate algorithms to "too late" in the blink of an eye without much of a Goldilocks period in between. One might counter the second argument by pointing out that algorithms are precisely the kind of technology that calls for federal regulation: opaque, complex, and occasionally dangerous. Finally, one might answer the third argument—that less regulation is always better—by emphasizing that regulation in one form or another is inevitable, and the true choice is between regulation that is piecemeal, reactive, and slow or regulation that is comprehensive, anticipatory, and technically savvy.

But a better way to counter those arguments than a volume of logic is a page of history. Those types of objections have long been registered against what is perhaps the world's most popular, effective, and widely emulated regulatory agency: the FDA.[155] The products the FDA regulates, and particularly the complex pharmaceutical drugs it vets for safety and efficacy, are similar to black-box algorithms. And the crises the FDA has confronted throughout its more than one hundred years in existence are comparable to the kinds of crises one can easily imagine occurring because of dangerous algorithms. The FDA has faced steep

---

protected head. Interests in expertise and independence still militate in favor of agency independence, however.

155. *See* PHILIP J. HILTS, PROTECTING AMERICA'S HEALTH xiv (2003) ("The FDA . . . is the most known, watched, and imitated of regulatory bodies. Because of its influence outside the United States, it has also been described as the most important regulatory agency in the word."); DANIEL CARPENTER, REPUTATION AND POWER: ORGANIZATIONAL IMAGE AND PHARMACEUTICAL REGULATION AT THE FDA 9 (2010) (noting, *inter alia*, that "in a nation as purportedly anti-bureaucratic as the United States, the FDA's power in the national health system, in the scientific world, and in the therapeutic marketplace is odd and telling").

resistance at every stage, but its capacity to respond to, and prevent, major health crises has resulted in the agency becoming a fixture of the American institutional landscape.[156] We could draw on the FDA's history for lessons, and use those lessons as an opportunity to avoid repeating that history.

The FDA was born against a backdrop of public health crisis. Adulterated and misbranded foods and drugs were being sold nationwide, and people were becoming seriously ill.[157] In a political environment heavily resistant to federal regulation of any kind, overwhelming popular sentiment forced the issue to a vote, and the result was the creation of the Food and Drug Act, signed into law in 1906.[158] The law established minimum purity requirements and labeling requirements.[159]

But the law was very limited. All non-narcotic drugs could still be sold by anyone to anyone.[160] Homebrewed remedies were outside the Act's purview if they "didn't contain narcotics or one of a few listed poisons."[161] Public sentiment turned against this state-of-affairs when a severe public health crises rocked the nation.[162]

In the summer of 1937, a prominent Tennessee pharmaceutical manufacturer developed a new medicine by mixing a foul-tasting but effective antibacterial treatment (sulfanilamide) with a somewhat sweet-tasting liquid (diethylene glycol) to make the antibacterial more palatable to children.[163] Shockingly, the company "did not bother to test for toxicity, either in humans or animals."[164] But "diethylene glycol, a chemical customarily employed as an antifreeze, was a deadly poison and known to be such by the FDA."[165] With deaths mounting, the

156. *See* Ronald Hamowy, *Medical Disasters and the Growth of the FDA*, INDEP. POL'Y REPS. (2010), http://www.independent.org/pdf/policy_reports/2010-02-10-fda.pdf (describing how the FDA's mission has grown over the course of a number of successful responses to health crises).

157. *See* HILTS, *supra* note 155, at 21–22, 30 (describing how adulterating food was "easy and very profitable").

158. *See id.* at 52–55.

159. *See id.*

160. *See* Hamowy, *supra* note 156, at 5 (asserting that, "prior to 1938, all non-narcotic drugs could legally be sold without a prescription").

161. HILTS, *supra* note 155, at 75.

162. CARPENTER, *supra* note 155, at 73 ("By all accounts, the Federal Food, Drug, and Cosmetic Act of 1938 issued from crisis.").

163. *Id.* at 85–87; Hamowy, *supra* note 156, at 5–6; HILTS, *supra* note 155, at 89–90.

164. *See* Hamowy, *supra* note 156, at 6.

165. *See id.*

company attempted an informal recall of the product without informing anyone that it was poison.[166] Over one hundred people, most of them children, died before the FDA was able to track down and destroy the remainder of the medicine.[167] Shortly thereafter, Congress passed the Food, Drug, and Cosmetic Act of 1938, vastly expanding the powers of the FDA, including authorizing pre-market review.[168] In the years that followed, every "nation of the developed world would adopt its central principles."[169] "In image and in law, the sulfanilamide tragedy of 1937 became an instructive moment whose essential lesson was pre-market clearance authority over new drugs."[170]

The agency discovered, however, that even pre-market controls were not always enough. Another prominent health crisis led to further refinement of the FDA's regulatory mandate. In the fall of 1960, an American drug manufacturer applied for permission to market thalidomide in the United States.[171] Thalidomide was introduced internationally in the 1950s as a non-barbiturate sedative with supposedly few side effects and low toxicity.[172] Testing for new drugs was "still a matter of some discretion for companies, as the law did not specify what was needed,"[173] to prove that a drug was safe and effective. In the case of thalidomide, clinical testing had been almost comically slipshod, involving no controlled clinical trials or other systematic investigation into the drug's efficacy or side effects.[174]

Before it made its formal application, the American drug maker had already distributed tens of thousands of doses without any FDA oversight because "under the 1938 law, doctors could experiment on patients with new drugs, in any numbers and with any chemical, so long as they called the work an experiment."[175] The American manufacturer even began to sell the drug as a treatment for nausea for pregnant women, even though it had done zero testing to determine whether the drug was safe and effective for that use.[176]

---

166. *See id.*
167. *See id.*; HILTS, *supra* note 155, at 92.
168. Food, Drug, and Cosmetic Act, 21 U.S.C. § 301–399 (2012).
169. HILTS, *supra* note 155, at 93.
170. CARPENTER, *supra* note 155, at 73–74.
171. HILTS, *supra* note 155, at 152.
172. Hamowy, *supra* note 156, at 11.
173. HILTS, *supra* note 155, at 150.
174. *See id.* at 144–50.
175. *See id.* at 152.
176. *See id.* at 149–50.

To its credit, the FDA delayed the drug's approval for a year, apparently because the outlandish claims made on its behalf—that it was effective, had low toxicity, and few side effects—did not match the outcomes of even the most cursory animal trials.[177] And during that time the drug's show-stopping side effect emerged—it caused terrible birth defects.[178] Thousands of children were born with severe disabilities worldwide because of thalidomide, including a handful in the United States because of the drug's experimental use.[179] Shortly thereafter, Congress amended the FDA's statutes to put in place strict rules to ensure that drugs were not tested on humans without safeguards and that clinical trials were strictly controlled to ensure that they adequately determine a drug's efficacy and safety.[180]

The arc of the FDA's history shows a relatively stable pattern of public health crises causing the American public to expand the FDA's powers to ensure that drugs are proven safe and effective before they reach the marketplace.[181] Given the close analog between complex pharmaceuticals and sophisticated algorithms, leaving algorithms unregulated could lead to the same pattern of crisis and response. Consequently, we should learn from the FDA's history and decide to act before those crises occur. Some of the world's largest companies are hoping to transform the way people live and work through the power of algorithms.[182] The algorithms of the future may operate in ways that we can neither fully understand nor, without carefully controlled trials, reliably predict. We are poised to enter a world where algorithms can cause similarly outsized risks in similarly difficult-to-know ways as pharmaceutical drugs. Rather than wait for an algorithm to harm many people, we might take the FDA's history as a lesson and instead develop an agency now with the capacity to ensure that algorithms are safe and effective for their intended use before they are released.

---

177. *See id.* at 150–53; CARPENTER, *supra* note 155, at 243.

178. HILTS, *supra* note 155, at 154–55; CARPENTER, *supra* note 155, at 238–40.

179. HILTS, *supra* note 155, at 158.

180. *See id.* at 162–65.

181. *See also* Rebecca S. Eisenberg, *The Role of the FDA in Innovation Policy*, 13 MICH. TELECOMM. & TECH. L. REV. 345, 345–46 (2007).

182. *See The World's Biggest Companies*, FORBES, 2016, http://www.forbes.com/global2000/list/ (listing Apple, Microsoft, Alphabet (owner of Google), and IBM as among the world's one hundred largest companies); H. James Wilson et al., *Companies Are Reimagining Business Processes with Algorithms*, HARV. BUS. REV. (Feb. 8, 2016), https://hbr.org/2016/02/companies-are-reimagining-business-processes-with-algorithms (explaining how corporations across many different industries have successfully used learning algorithms to improve internal processes and interactions with customers).

CONCLUSION

The purpose of this Article is to make an early case for developing a new federal agency whose goal is to ensure that algorithms are safe and effective. Any proposal to introduce legal oversight into an uncharted domain merits careful scrutiny. There are legitimate concerns that regulation stifles innovation and impedes competition. Those who favor free markets may think a federal regulatory agency is too radical and more than is necessary at this early stage. However, given the pace of algorithmic progress, it may not be so early. The unique dangers algorithms pose, coupled with their complexity, make them similar to technologies we have closely regulated in the past. It may be that the future is here and that the time to treat algorithms as a mature technology, deserving of society's watchful eye, is now.