ROBOPHOBIA

ANDREW KEANE WOODS*

Robots—machines, algorithms, artificial intelligence—play an increasingly important role in society, often supplementing or even replacing human judgment. Scholars have rightly become concerned with the fairness, accuracy, and humanity of these systems. Indeed, anxiety about machine bias is at a fever pitch. While these concerns are important, they nearly all run in one direction: we worry about robot bias against humans; we rarely worry about human bias against robots.

This is a mistake. Not because robots deserve, in some deontological sense, to be treated fairly—although that may be true—but because our bias against nonhuman deciders is bad for us. For example, it would be a mistake to reject self-driving cars merely because they cause a single fatal accident. Yet all too often this is what we do. We tolerate enormous risk from our fellow humans but almost none from machines. A substantial literature—almost entirely ignored by legal scholars concerned with algorithmic bias—suggests that we routinely prefer worse-performing humans over better-performing robots. We do this on our roads, in our courthouses, in our military, and in our hospitals. Our bias against robots is costly, and it will only get more so as robots become more capable.

This Article catalogs the many different forms of antirobot bias and suggests some reforms to curtail the harmful effects of that bias. The Article's descriptive contribution is to develop a taxonomy of robophobia. Its normative contribution is to offer some reasons to be less biased against robots. The stakes

* Professor of Law, University of Arizona. The author thanks Jane Bambauer, Dan Rodriguez, Albertina Antognini, Tammi Walker, Alan Rozenshtein, Jess Findley, Mark Lemley, Kiel Brennan-Marquez, Dave Pozen, Shalev Roisman, Derek Bambauer, Marc Miller, Chris Robertson, and workshop participants at Harvard University, the University of Arizona, and Arizona State University. The author is grateful to the editors of the Colorado Law Review, especially Ming Lee Newcomb and Brendan Soane, for their diligence and care. Comments welcome: akwoods@arizona.edu.

could hardly be higher. We are entering an age when one of the most important policy questions will be how and where to deploy machine decision-makers.

INTRODUCTION

Robots—algorithms powered by sensors and networked with computers of increasing sophistication—are all around us.[1] They now drive cars,[2] determine whether a defendant should be granted bail,[3] perform life-or-death surgeries,[4] and more. This has rightly led to increased concern about the fairness, accuracy,

1. A note about terminology. The article is concerned with human judgment of automated decision-makers, which include "robots," "machines," "algorithms," or "AI." There are meaningful differences between these concepts and important line-drawing debates to be had about each one. However, this Article considers them together because they share a key feature: they are nonhuman deciders that play an increasingly prominent role in society. If a human judge were replaced by a machine, that machine could be a robot that walks into the courtroom on three legs or an algorithm run on a computer server in a faraway building remotely transmitting its decisions to the courthouse. For present purposes, what matters is that these scenarios represent a *human* decider being replaced by a *nonhuman* one.

This is consistent with the approach taken by several others. *See, e.g.*, Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135 (2019) (bundling artificial intelligence and physical robots under the same moniker, "robots"); Jack Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1219 (2017) ("When I talk of robots . . . I will include not only robots—embodied material objects that interact with their environment—but also artificial intelligence agents and machine learning algorithms."); Berkeley Dietvorst & Soaham Bharti, *People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error*, 31 PSYCH. SCI. 1302, 1314 n.1 (2020) ("We use the term algorithm to describe any tool that uses a fixed step-by-step decision-making process, including statistical models, actuarial tables, and calculators."). This grouping contrasts scholars who have focused explicitly on certain kinds of nonhuman deciders. *See, e.g.*, Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 529 (2015) (focusing on robots as physical, corporeal objects that satisfy the "sense-think-act" test as compared to, say, a "laptop with a camera").

2. *The Race for Self-Driving Cars*, N.Y. TIMES (June 6, 2017), https://www.nytimes.com/interactive/2016/12/14/technology/how-self-driving-cars-work.html [https://perma.cc/J25U-EXU6].

3. Dave Gershgorn, *California Just Replaced Cash Bail with Algorithms,* QUARTZ (Sept. 5, 2018), https://qz.com/1375820/california-just-replaced-cash-bail-with-algorithms/ [https://perma.cc/9AMF-KWH8].

4. Sandip S. Panesar, *The Surgical Singularity Is Approaching,* SCI. AM.: OBSERVATIONS (Dec. 27, 2018), https://blogs.scientificamerican.com/observations/the-surgical-singularity-is-approaching/ [https://perma.cc/B32Y-2SUW].

and safety of these systems.[5] Indeed, anxiety about algorithmic decision-making is at a fever pitch.[6] As Chief Justice Roberts recently admonished a group of high school students, "Beware the robots."[7] Or as thousands of British students put it in nationwide protests after England's university-sorting program erred, "Fuck the algorithm."[8] While concerns about algorithmic

5. *See, e.g.*, Sandra Mayson, *Bias in, Bias Out*, 128 YALE L.J. 2218, 2227–33 (2019) (surveying the literature and noting that "[a]s the use of criminal justice risk assessment has spread, concern over its potential racial impact has exploded"); Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer*, 43 HASTINGS L.J. 1321, 1323 (1992) (noting that "the law must pay attention to the structure of data processing," with a particular emphasis on transparency, rights, and accountability); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016) (describing how algorithms can be unintentionally biased in ways that challenge antidiscrimination law); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 5 (2014) (warning that there is "nothing unbiased about scoring systems" and urging human oversight of algorithmic scoring systems); Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245 (2016) (tracing automation bias and arguing for human deciders); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 636–38 (2017) (echoing the concerns about algorithmic bias and the need for oversight, and charting a path forward in algorithm design); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016) (alarming the public about the perils of large platforms's use of algorithms in hiring, credit, advertising, and more); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 18 (2015) ("The black boxes of reputation, search, and finance endanger all of us."); EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (2014), https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf [https://perma.cc/Y74A-85FJ] (describing the discrimination risks associated with big data).

6. Michael Kearns & Aaron Roth, *Ethical Algorithm Design Should Guide Technology Regulation*, BROOKINGS (Jan. 13, 2020), https://www.brookings.edu/research/ethical-algorithm-design-should-guide-technology-regulation/ [https://perma.cc/7HME-G9YE] ("Nearly every week, a new report of algorithmic misbehavior emerges.").

7. Deborah Cassens Weiss, *'Beware the Robots,' Chief Justice Tells High School Graduates*, ABA J. (June 8, 2018), https://www.abajournal.com/news/article/beware_the_robots_chief_justice_tells_high_school_graduates [https://perma.cc/GKQ2-MZKL]. This is hardly a unique view. *See also* European Parliament Comm. on Legal Affs., Draft Rep. with Recommendations to the Comm'n on Civ. L. Rules on Robotics, EUR. PARL. DOC. PE582.443 at 4 (May 31, 2016), https://www.europarl.europa.eu/doceo/document/JURI-PR-582443EN.pdf?redirect [https://perma.cc/76LZ-TUX4] ("[W]hereas ultimately there is a possibility that within the space of a few decades AI could surpass human intellectual capacity in a manner which, if not prepared for, could pose a challenge to humanity's capacity to control its own creation and, consequently, perhaps also to its capacity to be in charge of its own destiny and to ensure the survival of the species.").

8. Louise Amoore, *Why 'Ditch the Algorithm' Is the Future of Political Protest*, GUARDIAN (Aug. 19, 2020), https://www.theguardian.com/commentisfree/2020/aug/

decision-making are critical, they nearly all run in one direction: we worry about how algorithms judge humans; we rarely worry about how humans judge algorithms.[9] This is a mistake.

Deciding where to deploy machine decision-makers is one of the most important policy questions of our time. The crucial question is not whether an algorithm has *any* flaws, but whether it outperforms current methods used to accomplish a task. Yet this view runs counter to the prevailing reactions to the introduction of algorithms in public life and in legal scholarship.[10] Rather than engage in a rational calculation of who performs a task better, we place unreasonably high demands on robots. This is *robophobia*—a bias against robots, algorithms, and other nonhuman deciders.

Robophobia is pervasive. In healthcare, patients prefer human diagnoses to computerized diagnoses, even when they are

19/ditch-the-algorithm-generation-students-a-levels-politics [https://perma.cc/2SKP-FVXS].

9. The main exception is the small literature on algorithm aversion. *See* Berkeley J. Dietvorst, Joseph P. Simmons & Cade Massey, *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, 144 J. EXPERIMENTAL PSYCH. 114 (2015) [hereinafter Dietvorst et al., *Algorithm Aversion*] (showing in a series of experiments that people tend to have less confidence in algorithms than humans, even when they know the algorithm is more accurate); Berkeley J. Dietvorst, Joseph P. Simmons & Cade Massey, *Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms if They Can (Even Slightly) Modify Them*, 64 MGMT. SCI. 1155 (2018) [hereinafter Dietvorst et al., *Overcoming Aversion*] (showing that giving participants an opportunity to modify an algorithm increased their satisfaction and faith in the algorithm).

10. *See, e.g.*, Nada R. Sanders & Karl B. Manrodt, *The Efficacy of Using Judgmental Versus Quantitative Forecasting Methods in Practice*, 31 OMEGA 511 (2003) (showing that firms repeatedly rely on human forecasters instead of algorithmic forecasters); Robert Fildes & Paul Goodwin, *Good and Bad Judgment in Forecasting: Lessons from Four Companies*, 8 FORESIGHT 5 (2007) (showing that in four large companies where forecasting is essential to corporate success, people prefer human judgments to algorithmic predictions); Scott I. Vrieze & William M. Grove, *Survey on the Use of Clinical and Mechanical Prediction Methods in Clinical Psychology*, 40 PRO. PSYCH.: RSCH. & PRAC. 525 (2009) (showing that the overwhelming majority of clinical psychologists prefer their own clinical judgment over mechanical models); Reuben Binns et al., *'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions*, *in* CHI '18: PROCEEDINGS OF THE 2018 CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, Apr. 21–26, 2018, Montreal, Can., https://dl.acm.org/doi/pdf/10.1145/3173574.3173951 [https://perma.cc/GJ4W-TS8G] (finding considerable resistance to the idea of algorithmic justice). For a table summarizing the findings of research in the growing algorithmic-aversion literature, see Noah Castelo, Maarten W. Bos & Donald R. Lehmann, *Task-Dependent Algorithm Aversion*, 56 J. MKTG. RES. 809, 810 (2019).

told that the computer is more effective.[11] In litigation, lawyers are reluctant to rely on—and juries seem suspicious of—computer-generated discovery results, even when they have been proven to be more accurate than human discovery results.[12] In the military, autonomous weapons promise to reduce the risk of grave human errors, and yet there is a legal movement to ban what are tellingly referred to as "killer robots."[13] On the streets, robots are regularly physically assaulted.[14] In short, we are deeply biased against machines and in many different ways.

This is a problem. Not because of some deontological moral claim that robots deserve to be judged fairly—although that may be true[15]—but because human bias against robots is bad for humans. In many different domains, algorithms are simply better at performing a given task than people.[16] Algorithms outperform humans at discrete tasks in clinical health,[17] psychology,[18] hiring and admissions,[19] and much more. Yet in setting after setting, we regularly prefer worse-performing humans to a robot alternative, often at an extreme cost.

While robophobia is an urgent problem, it is not entirely new. In the 1950s, medical pioneer Paul Meehl crusaded for what he called "statistical prediction," or relying on algorithms and statistical tables to predict a patient's future behavior,

---

11. *See* Marianne Promberger & Jonathan Baron, *Do Patients Trust Computers?*, 19 J. BEHAV. DECISION MAKING 455 (2006) (finding that patients were more likely to follow medical advice from a physician than a computer and were less trustful of computers as providers of medical advice).

12. The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery,* 15 SEDONA CONF. J. 217 (2014).

13. *See infra* Section I.B.6.

14. *See infra* Section I.A.1.

15. I leave aside for now the very serious question of whether robots ought—in some deontological sense—to be treated fairly, justly, and so on. I also leave aside the question of whether robot mistreatment is an ethical failing on the part of humans. These are significant questions that deserve their own treatments. I do, however, ask whether robot mistreatment ought to be discouraged because of the beneficial effects for *humans* that might result.

16. Dietvorst & Bharti, *supra* note 1.

17. Stefania Ægisdóttir, et al., *The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction,* 34 COUNSELING PSYCH. 341–382 (2006).

18. William M. Grove et al., *Clinical Versus Mechanical Prediction: A Meta-Analysis*, 12 PSYCH. ASSESSMENT 19–30 (2000).

19. Nathan R. Kuncel et al., *Mechanical Versus Clinical Data Combination in Selection & Admissions Decisions: A Meta-Analysis*, 98 J. APPLIED PSYCH. 1060 (2013).

which was repeatedly shown to have advantages over "clinical prediction," or doctors' reliance on their training and intuitions.[20] Large meta-analyses across a range of domains have since shown that Meehl's claims about the benefits of using algorithms were valid.[21] But his public health campaign largely failed. Seventy years later, doctors continue to privilege their own intuitions over automated decision-making aids.[22] Since Meehl's time, a growing body of social psychology scholarship has offered an explanation: bias against nonhuman decision-makers.[23] Somehow, little of that empirical research has made it into law reviews.[24] As Jack Balkin notes, "When we talk about robots, or AI agents, or algorithms, we usually focus on whether they cause problems or threats. But in most cases, the problem isn't the robots. It's the humans."[25]

This Article is the first piece of legal scholarship to address our misjudgment of algorithms head-on. The Article catalogs different ways we misjudge algorithms and suggests some reforms to protect us from poor judgment. The descriptive contribution of the Article is to provide a taxonomy of different kinds of judgment errors. The evidence of our robophobia is overwhelming, but the research happens in silos—with some scholars working on human reluctance to trust algorithms as others work on automation bias—and little of it is seriously considered by legal scholars or policymakers. This Article brings these different literatures together and explores their implications for the law.

---

20. PAUL MEEHL, CLINICAL VS. STATISTICAL PREDICTION: A THEORETICAL ANALYSIS AND A REVIEW OF THE EVIDENCE (1954) (arguing that mechanical methods of prediction offer the promise of better decisions—more accurate and more reliable—than subjective, clinical assessments).

21. Ægisdóttir et al., *supra* note 17, at 341; Grove et al., *supra* note 18, at 19.

22. Promberger & Baron, *supra* note 11 (finding that patients were more likely to follow medical advice from a physician than a computer and were less trustful of computers as providers of medical advice).

23. Dietvorst et al., *Algorithm Aversion*, *supra* note 9, at 1 ("In a wide variety of forecasting domains, experts and laypeople remain resistant to using algorithms, often opting to use forecasts made by an inferior human rather than forecasts made by a superior algorithm.").

24. A search of law review and law journal articles regarding algorithmic bias produced over 500 results; only one of those discussed human bias against algorithms. *See* Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System,* 52 U.C. DAVIS L. REV. 1067, 1091 (2018) (discussing algorithm aversion as a potential barrier to public acceptance of algorithms in criminal justice).

25. Balkin, *supra* note 1, at 1223.

To be clear, the argument is not that robots are *good* and people are *bad*. Nor do I mean to downplay the risks that algorithms present, especially where they amplify discrimination and bias. Scholars have shown how poorly designed algorithms can exacerbate racial inequities in criminal justice;[26] undermine civil-rights protections in labor law;[27] and compound inequality in commerce, communications, and information.[28] These are serious concerns, and it would be a mistake to ignore them or to overrely on poorly designed machines. But humans also have a terrible track record of bias and discrimination. Given this, we must carefully assess not only whether a robot's decisions have distributional consequences but how they compare to the alternative. For example, some patients may have a general preference for humans over machines in healthcare because humans are warmer, more relatable, and so on; thus, privileging human decision-makers in medicine might be a net gain for people who feel that doctors treat them with dignity and respect. Yet recent work in healthcare suggests that doctors' implicit bias can make them worse than algorithms at diagnosing disease in underrepresented populations, despite all the well-founded concerns about algorithmic bias.[29] We must be attentive to the distributional consequences of how we implement robots in society, but that fact should not uniformly make one pro- or anti-algorithm.

The Article proceeds in six parts. Part I provides several examples of robophobia. This list is illustrative and far from exhaustive; tellingly, robophobia is so pervasive that there is simply not room in a single article to catalogue all of its instances. Part II distinguishes different *types* of robophobia. Part III interrogates potential explanations for robophobia and shows that none of these explanations is a sufficient justification for our pervasive robophobia. Part IV makes the strongest case for being wary of machine decision-makers, which includes concerns about equality, the current political economy of algorithms, and our own inclination to sometimes overrely on machines, even those we initially distrusted. Part V outlines the components of

26. Mayson, *supra* note 5.

27. Ifeoma Ajunwa, *An Auditing Imperative for Automated Hiring Systems*, 34 HARV. J.L. & TECH. 1 (2021).

28. PASQUALE, *supra* note 5.

29. *See* Emma Pierson et al., *An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations*, 27 NATURE MED. 136 (2021).

the normative case against robophobia. Finally, Part VI offers tentative policy prescriptions for encouraging rational thinking—and policy making—when it comes to nonhuman deciders.

I. EXAMPLES OF ROBOPHOBIA

Robophobia crops up in different domains and is shared by a wide range of people; this Part provides a number of examples of the phenomenon. These examples are divided into two broad categories. At the most basic level, we can distinguish between antirobot bias in general and antirobot bias that has been codified in laws and policies. While I am primarily concerned with laws and policies, generalized robophobia remains relevant because it will inevitably influence laws and policies.

One natural question about the following examples is whether they are meaningfully related to each other—and, indeed, whether they represent the same phenomenon. Is reluctance to trust an algorithmic decision the same as fear of killer robots? The answer, of course, is yes and no. They are distinct in the sense that they are likely motivated by different, if overlapping, sets of concerns. By lumping these very different examples together, I do not mean to suggest that they are motivated by the same things. Indeed, the diversity of the examples is partly the point. One aim of this Article is to show that there is a problem—human misuse of nonhuman actors—and that this problem comes in *many distinct forms*. Cataloguing the different forms and understanding their differences is key to doing something about the problem.

A. In General

1. On the Streets

Sometimes robophobia takes a violent form. When two university roboticists released HitchBOT, a hitchhiking robot capable of communicating with humans to request a ride in a given direction, their aim was merely to see if a robot could be trusted in the hands of the public.[30] The robot started its journey in

30. Dawn Chmielewski, *HitchBOT Gets Mugged in 'City of Brotherly Love' En Route to San Francisco*, VOX (Aug. 2, 2015, 2:14 PM), https://www.vox.com/

Salem, Massachusetts, with the goal of reaching San Francisco, California, but two weeks into the journey was found decapitated and dismembered in a Philadelphia alley.[31]

This might be dismissed as a one-off attack by hoodlums, but it is part of a larger pattern of abuse against robots. Self-driving cars being tested in Arizona have been attacked in over twenty incidents; the damage has included smashed windows, slashed tires, and attempts at even greater destruction.[32] In California, traffic safety investigators believe that a significant portion of accidents involving self-driving cars are caused by human drivers *intentionally* crashing into the self-driving cars.[33] In Osaka, a group of young boys were caught punching and kicking a robot that was attempting to navigate a mall.[34] In Moscow, a man attacked a teaching robot named Alantim with a baseball bat, despite the robot's repeated calls for help.[35] The list goes on.[36] There are now so many documented cases of human abuse of robots that journalists and psychologists have begun to ask, "Why do we hurt robots?"[37]

2015/8/2/11615286/hitchbot-gets-mugged-in-city-of-brotherly-love-en-route-to-san [https://perma.cc/2E2U-E9K5].

31. *Id.*

32. Jamie Court, *Arizona's Revolt Against Self-Driving Cars Should Be a Wake-Up Call to the Companies That Make Them*, L.A. TIMES (Jan. 11, 2019), https://www.latimes.com/opinion/op-ed/la-oe-court-self-driving-cars-20190111-story.html [https://perma.cc/32R7-SKYK].

33. Julia Carrie Wong, *Rage Against the Machine: Self-Driving Cars Attacked by Angry Californians*, GUARDIAN (Mar. 6, 2018, 2:25 PM), https://www.theguardian.com/technology/2018/mar/06/california-self-driving-cars-attacked [https://perma.cc/HKC8-PJDJ].

34. Nathan McAlone, *Japanese Researchers Watch a Gang of Children Beat Up Their Robot in a Shopping Mall*, BUS. INSIDER (Aug. 7, 2015, 10:22 AM), https://www.businessinsider.com/japanese-researchers-watch-as-gang-of-children-beats-up-their-robot-2015-8 [https://perma.cc/2K96-CG2E].

35. Becky Ferreira, *Watch a Robot Eulogize its 'Brother' at Moscow's New Cemetery for Dead Machines*, VICE (Nov. 8, 2017, 9:00 AM), https://www.vice.com/en_us/article/ywbqvk/watch-russian-robot-eulogize-brother-at-moscows-dead-machines-cemetery [https://perma.cc/LYB8-8L8F].

36. Jonah Engel Bromwich, *Why Do We Hurt Robots?*, N.Y. TIMES (Jan. 19, 2019), https://www.nytimes.com/2019/01/19/style/why-do-people-hurt-robots.html [https://perma.cc/J3GW-5R9Z].

37. *See id.*; Dražen Brščić et al., *Escaping from Children's Abuse of Social Robots*, *in* HRI '15: PROCEEDINGS OF THE TENTH ANNUAL ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION 59, Mar. 2–5, 2015, Portland, Or., https://dl.acm.org/doi/pdf/10.1145/2696454.2696468 [https://perma.cc/UR7D-YFYM].

2. In the Arts

If there is a central theme of human-robot stories, it is the idea that robots will inevitably turn on their creators. The English word "robot" comes from the Czech word "robota," meaning drudgery; it was used in a 1920 Czech play, *R.U.R. (Rossum's Universal Robots)*, which features artificially created workers who rise up and overthrow their creators.[38] This story echoes the old Hebrew tale of Gollum, a clay-formed creature that turns on its maker.[39] There is a good dramatic reason for this dark turn: it plays on our fascination with—and fear of—the uncanny machine. Another common trope in the literature about humans and robots is the morality tale about the risks of "playing god."[40] No robot story is likely better known than *Frankenstein* and its idea of a human creation turning on its maker.[41] Balkin describes the old literary trope as "the idea of the Frankenstein monster or the killer robot, which becomes evil or goes berserk."[42] In so many stories about robots, we reinforce the idea that robot-human interactions will end badly.

The idea of a "robot gone wild" was so pervasive in his time that the famous science-fiction author Isaac Asimov deliberately sought to counteract it with stories featuring robots as relatable protagonists; it is telling that in order to do this, he made them humanlike.[43] Indeed, Asimov wrote his robot characters in line with the most important of his "three laws of robotics,"[44] which reads, "a robot may not injure humanity, or, through inaction, allow a human being to come to harm."[45] Asimov was fighting

38. KAREL ČAPEK, R.U.R. (ROSSUM'S UNIVERSAL ROBOTS) (1920).

39. *See* Lee McCauley, *AI Armageddon and the Three Laws of Robotics*, 9 ETHICS & INFO. TECH. 153 (2007).

40. *Id.* at 154.

41. MARY SHELLEY, FRANKENSTEIN; OR, THE MODERN PROMETHEUS (1818). *See also* Jill Lepore*, The Strange and Twisted Life of "Frankenstein"*, NEW YORKER (Feb. 5, 2018), https://www.newyorker.com/magazine/2018/02/12/the-strange-and-twisted-life-of-frankenstein [https://perma.cc/V88A-WJWP] (noting that the book *Frankenstein* has been used for many purposes, including as a "catechism for designers of robots and inventors of artificial intelligences").

42. Balkin, *supra* note 1, at 1218.

43. *Id.* ("Asimov wrote his robot stories to fight against what he called the 'Frankenstein Complex,'—the idea that robots were inherently menacing or evil . . . .").

44. Isaac Asimov, *Runaround*, *in* I, ROBOT 37 (1950).

45. *Id.*

against what he called the "Frankenstein Complex,"[46] which has been described as the "fear of man broaching, through technology, into God's realm and being unable to control his own creations."[47]

The Frankenstein Complex is alive and well today. The central premise of one of the most popular shows on television, HBO's *Westworld*,[48] is that robots designed for human pleasure will inevitably rise up to kill their masters.[49] The central tension in the show is the difficulty telling humans from robots, and the unease builds as the robots slowly gain consciousness and start to collectively gather the courage to revolt.[50] The story is dark and incredibly popular because it speaks to a deep public fear about robot revenge.[51] Similar themes are explored in *Humans*,[52] *Ex Machina*,[53] *Blade Runner*,[54] *Battlestar Galactica*,[55] and many more popular films and shows.

Because Asimov sought to quell these fears, he is often seen as a corrective force to a robophobic society. But he also can be read to betray his own robophobia. Even in Asimov's effort to empathize and embrace robots, he adopts an absolutist position. Rather than say that robots may not do any more harm to humans than a similarly situated human might, he says a robot *may not* injure humanity, full stop. This absolute rule was surely designed to address what he knew to be growing public anxiety about a Frankenstein robot. But it is also consistent with the

46. Isaac Asimov, *The Machine and the Robot*., *in* SCIENCE FICTION: CONTEMPORARY MYTHOLOGY (P. S. Warrick et al. eds., 1978).

47. McCauley, *supra* note 39, at 154.

48. *Westworld* (HBO 2016).

49. Elizabeth Nolan Brown, *Westworld's Real Lesson? Screw or Kill All the Robots You Want*, REASON (Dec. 5, 2016, 4:45 PM), https://reason.com/2016/12/05/west-worlds-lesson/ [https://perma.cc/269G-GB3S].

50. Tim Surette, *Westworld: Who Is and Who Isn't a Robot?*, TV GUIDE (Oct. 27, 2016, 7:46 PM), https://www.tvguide.com/news/westworld-who-is-and-who-isnt-a-robot/ [https://perma.cc/JW4V-ASTW].

51. *See* Becky Ferreira, *Westworld's Female Hosts Signal a Shift in Our Fear of Robots*, MOTHERBOARD (Apr. 24, 2018, 7:00 AM), https://www.vice.com/en_us/article/bjpk43/westworlds-female-hosts-signal-a-shift-in-our-fear-of-robots [https://perma.cc/A5R5-CU9B].

52. *Humans* (Kudos, Channel 4 & AMC Studios 2015).

53. EX MACHINA (Film4 & DNA Films 2015).

54. BLADE RUNNER (The Ladd Co. & Shaw Bros. 1982).

55. *Battlestar Galactica* (R&D TV 2014).

kind of absolute rules that we have adopted in many areas of law and policy today.[56]

B. In Law & Public Policy

1. Transportation

Every year, around forty thousand people die on freeways in the United States.[57] Globally, the number of casualties is estimated at over one million, with road deaths being one of the ten most common causes of death around the world.[58] Traffic fatalities are so common they are not newsworthy—unless, that is, a self-driving car is involved. Indeed, when the first semi-autonomous cars were involved in fatal accidents, the news made international headlines.[59] Since autonomous vehicles have been on the road, they have been held to a near-impossible standard: perfection. And this phenomenon has occurred despite evidence that autonomous cars are involved in considerably fewer accidents than the average human driver.

Not all self-driving-car manufacturers publish accident-report numbers, so data comparing all robot-driven cars to all human-driven cars are not available. But Tesla—as a result of the overwhelming media coverage of its cars' accidents—now publishes a quarterly accident report.[60] Tesla's reports claim a single "accident or crash-like event" occurs for every 3.34-million

56. Michael Laris, *Uber Shutting Down Self-Driving Operations in Arizona*, WASH. POST (May 23, 2018, 5:50 PM), https://www.washingtonpost.com/news/dr-gridlock/wp/2018/05/23/uber-shutting-down-self-driving-operations-in-arizona/ [https://perma.cc/GX9T-HXWL] (noting that the governor of Arizona suspended Uber's self-driving tests because "safety was his top priority," despite the fact that the governor had previously taunted California for limiting innovation with "bureaucracy and more regulation").

57. Ryan Beene, *Traffic Deaths in the U.S. Exceed 40,000 for Third Straight Year*, BLOOMBERG (Feb. 12, 2019), https://www.bloomberg.com/news/articles/2019-02-13/traffic-deaths-in-u-s-exceed-40-000-for-third-straight-year [https://perma.cc/4XBY-FBZW].

58. WORLD HEALTH ORG., GLOBAL STATUS REPORT ON ROAD SAFETY 2018 (2018), https://www.who.int/publications/i/item/9789241565684 [https://perma.cc/JV4N-V8HV].

59. Bertel Schmitt, *Model S Crashes Make Headlines in Europe*, FORBES (Sept. 29, 2016, 3:03 PM), https://www.forbes.com/sites/bertelschmitt/2016/09/29/model-s-crashes-make-headlines-in-europe/#69f3b60656db [https://perma.cc/H2NG-6LGK].

60. *Tesla Q3 2018 Vehicle Safety Report*, TESLA (Oct. 4, 2018), https://www.tesla.com/blog/q3-2018-vehicle-safety-report [https://perma.cc/6PKT-WBZN].

miles driven with the car's semi-autonomous technology engaged.[61] According to the National Highway Traffic Safety Administration's most recent data, in the United States, there is a crash every 492,000 miles.[62] This means that Tesla's semi-autonomous cars are involved in *seven times* fewer crashes than the average car without autonomous features—that is, a car driven entirely by a human.[63] Yet the breathless press coverage of crashes involving autonomy would lead readers to believe that the semi-autonomous features are extremely dangerous.[64]

Other self-driving-car accidents have received similar treatment. When Uber's self-driving car crashed into a pedestrian in Phoenix, the press reported it as if a road fatality were a rare occurrence.[65] In fact, Arizona streets see an average of approximately three deaths a day by human drivers; 2018 saw 1,010 people killed in crashes with human drivers on Arizona roads.[66] In terms of pedestrian fatalities by human drivers, Arizona has the fourth highest rate in the country.[67] The vast majority of those crashes by human drivers received no media attention at all. Nationally, not a single human-caused car crash garners

---

61. *Id.*

62. *Id.*

63. If the statistics were available, the truest comparison would be between a Tesla on autopilot and a new car of similar value and performance, but I do not have those statistics.

64. *See, e.g.*, Jacob Bogage, *Tesla Driver Using Autopilot Killed in Crash*, WASH. POST (June 30, 2016), https://www.washingtonpost.com/news/the-switch/wp/2016/06/30/tesla-owner-killed-in-fatal-crash-while-car-was-on-autopilot/ [https://perma.cc/79Y9-BN58]; Tom Krisher, *3 Crashes, 3 Deaths Raise Questions About Tesla's Autopilot*, ASSOCIATED PRESS (Jan. 3, 2020), https://apnews.com/ca5e62255bb87bf1b151f9bf075aaadf [https://perma.cc/X5NE-CEDK].

65. For example, the *Arizona Republic*, the largest newspaper in the state, did an in-depth feature story one year after the Uber crash. The article discusses the dangers of self-driving cars on Phoenix streets yet makes no mention of how many fatalities occurred on Phoenix roads during the same time period. Ryan Randazzo, *Who Was Really at Fault in Fatal Uber Crash? Here's the Whole Story*, ARIZ. REPUBLIC (Mar. 17, 2019), https://www.azcentral.com/story/news/local/tempe/2019/03/17/one-year-after-self-driving-uber-rafaela-vasquez-behind-wheel-crash-death-elaine-herzberg-tempe/1296676002/ [https://perma.cc/D2XU-7LY6].

66. ARIZ. DEP'T OF TRANSP., ARIZONA MOTOR VEHICLE CRASH FACTS 2018 (2018), https://azdot.gov/sites/default/files/news/2018-Crash-Facts.pdf [https://perma.cc/3866-PR48].

67. *See* Perry Vandell, *Pedestrians in Arizona Are More Likely To Be Hit and Killed than Nearly Any Other State. Why?*, ARIZ. REPUBLIC (Sept. 30, 2020, 3:33 PM), https://www.azcentral.com/story/news/local/arizona-traffic/2020/09/28/arizona-has-4th-highest-pedestrian-death-rate-country-why/3511850001/ [https://perma.cc/YTJ2-5SX7].

anything close to the same level of media attention as crashes involving self-driving technology, which regularly make the front-page national and international news.[68] The news coverage of the Uber crash was all the more surprising because there was actually a human test driver behind the wheel who failed to intervene—so the crash involved, at a minimum, a mix of robot and human error—and the police said the accident would have been unavoidable for any driver, man or machine.[69]

News coverage may not be the best way to measure this effect because self-driving cars are novel, so while a car accident is not normally newsworthy, perhaps the novel technology makes it newsworthy. But the news coverage of crashes involving autonomous vehicles are not merely reporting on a novel event, a new kind of car crash; rather, they include demands to change policy. Uber's self-driving car crashing into a pedestrian in Phoenix led to public outcry,[70] and the Arizona governor decided to shut the program down.[71] In contrast, the Arizona governor made no special announcements or policy changes in reaction to the thousand-plus fatalities caused by human drivers the same year.

The public's robophobia and sensationalist news coverage are reinforcing phenomena. Psychologists have noted that the "[o]utsized media coverage of [self-driving car] crashes" amplifies preexisting fears.[72] Coupled with the fact that people tend to focus on sensational news stories of crashes rather than statistics about safety and given the general background fear of

68. *See* Dieter Bohn, *Elon Musk*: *Negative Media Coverage of Autonomous Vehicles Could Be 'Killing People'*, VERGE (Oct. 19, 2016, 9:16PM), https://www.theverge.com/2016/10/19/13341306/elon-musk-negative-media-autonomous-vehicles-killing-people [https://perma.cc/VA87-AQ5F].

69. Uriel J. Garcia & Ryan Randazzo, *Video Shows Moments Before Fatal Uber Crash in Tempe*, ARIZ. REPUBLIC (Mar. 21, 2018, 7:01 PM), https://www.azcentral.com/story/news/local/tempe-breaking/2018/03/21/video-shows-moments-before-fatal-uber-crash-tempe/447648002/ [https://perma.cc/4ZA3-VPAK].

70. *See* Ray Stern, *Ducey's Drive-By: How Arizona Governor Helped Cause Uber's Fatal Self-Driving Car Crash*, PHX. NEW TIMES (Apr. 12, 2018), https://www.phoenixnewtimes.com/news/arizona-governor-doug-ducey-shares-blame-fatal-uber-crash-10319379 [https://perma.cc/AJ5W-67N5].

71. Ryan Randazzo, *Arizona Gov. Doug Ducey Suspends Testing of Uber Self-Driving Cars*, ARIZ. REPUBLIC, https://www.azcentral.com/story/news/local/tempe-breaking/2018/03/26/doug-ducey-uber-self-driving-cars-program-suspended-arizona/460915002/ (Mar. 26, 2018, 6:59 PM) [https://perma.cc/EN5G-N3HK].

72. Azim Shariff, Jean-Francois Bonnefon & Iyad Rahwan, *Psychological Roadblocks to the Adoption of Self-Driving Vehicles*, 1 NATURE HUM. BEHAV. 694, 695 (2017).

algorithms, "the biggest roadblocks standing in the path of mass adoption may be psychological, not technological."[73] Indeed, over three-quarters of American drivers say they would be afraid to ride in a self-driving car, and 90 percent say they do not feel safer sharing the road with autonomous vehicles.[74] One study suggests that self-driving cars will need to be 90 percent safer than current human drivers to be accepted on the roads.[75] Others think the number is much higher.

2. Healthcare

We also find evidence of robophobia in healthcare, where both patients and doctors are reluctant to trust machine decision-makers. Algorithms are transforming nearly every aspect of healthcare, from reducing errors in diagnosis to improving accuracy in operations and shortening recovery times.[76] Algorithms can help fight prejudice in healthcare too. One recent study showed that algorithms spotted diseases on the x-rays of underserved populations when those same diseases were missed by doctors due to implicit bias.[77]

Yet patients are reluctant to trust this technology. A number of studies have shown that people generally prefer healthcare provided by humans over machines, even when that means the healthcare will be more costly and less effective. These patients appear to be willing to sacrifice the accuracy of their treatment in exchange for a human doctor. In fact, patients' bias against nonhuman decision-making manifests in many different ways throughout the healthcare system. Recent studies found that patients were less willing to schedule appointments for diagnosis by a robot; were willing to pay

---

73. *Id.* at 694.

74. Ellen Edmonds, *Americans Feel Unsafe Sharing the Road with Fully Self-Driving Cars*, AM. AUTO. ASS'N (Mar. 7, 2017), https://newsroom.aaa.com/2017/03/americans-feel-unsafe-sharing-road-fully-self-driving-cars/ [https://perma.cc/6BXJ-7X8B] (noting that the survey results were the same for 2016, suggesting that this fear was not decreasing over time).

75. DING ZHAO & HUEI PENG, UNIV. OF MICH, MCITY, FROM THE LAB TO THE STREET: SOLVING THE CHALLENGE OF ACCELERATING AUTOMATED VEHICLE TESTING (2017), https://mcity.umich.edu/wp-content/uploads/2017/05/Mcity-White-Paper_Accelerated-AV-Testing.pdf [https://perma.cc/A397-R649].

76. Harold Thimbleby, *Technology and the Future of Healthcare*, 2 J. PUB. HEALTH RES. 28 (2013).

77. Emma Pierson et al., *An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations*, 27 NATURE MED. 136 (2021).

significantly more money for a human provider; and preferred a human provider, even when told that the human provider was less effective than the robot.[78] Another study found that participants felt that doctors who relied on nonhuman decision aids had lower diagnostic ability than doctors who used their experience and intuition, despite evidence to the contrary.[79] Yet another study found that patients were more likely to follow the recommendations of a physician than of a computer program.[80] Finally, a 2018 Accenture survey of over 2,000 patients found that only 36 percent would consent to robot-assisted surgery, despite the fact that robot-assisted surgery is safer and leads to fewer complications when compared to more traditional surgery.[81]

Perhaps surprisingly, medical professionals also exhibit extreme suspicion of algorithmic decision aids. As one survey of the literature on doctors' resistance to artificial intelligence noted, "Although statistical models reliably outperform doctors, doctors generally prefer to rely on their own intuition rather than on statistical models, and are evaluated as less professional and competent if they do rely on computerized decision-aids."[82] These findings are consistent with the idea that medicine is as much "art" as science—a view many doctors hold.[83] And this distrust extends beyond decision aids: when a new robotic surgical device was developed that had many advantages over manual

78. Chiara Longoni, Andrea Bonezzi & Carey K. Morewedge, *Resistance to Medical Artificial Intelligence*, 46 J. CONSUMER RSCH. 629 (2021).

79. Hal R. Arkes, Victoria A. Shaffer & Mitchell A. Medow, *Patients Derogate Physicians Who Use a Computer-Assisted Diagnostic Aid*, 27 MED. DECISION MAKING 189 (2007).

80. Promberger & Baron, *supra* note 11 (finding that patients were more likely to follow medical advice from a physician than a computer and were less trustful of computers as providers of medical advice).

81. ACCENTURE CONSULTING, 2018 CONSUMER SURVEY ON DIGITAL HEALTH: US RESULTS (2019), https://www.accenture.com/t20180306t103559z__w__/us-en/_acnmedia/pdf-71/accenture-health-2018-consumer-survey-digital-health.pdf [https://perma.cc/ST9B-JE5A].

82. Chiara Longoni, Andrea Bonezzi & Carey K. Morewedge, *Resistance to Medical Artificial Intelligence*, 46 J. Consumer Rsch. 629, 630 (2019) (citations omitted).

83. *See* Robert Pearl, *Medicine Is an Art, Not a Science: Medical Myth or Reality?*, FORBES (June 12, 2014, 12:55 PM), https://www.forbes.com/sites/robertpearl/2014/06/12/medicine-is-an-art-not-a-science-medical-myth-or-reality/#532bd16f2071 [https://perma.cc/G62E-PDU4].

surgery, surgeons were almost universally opposed to its adoption.[84]

As an additional example, consider the 2014 campaign a national nurses' union launched to convince patients to demand human healthcare decision-making and reject automated aids.[85] Their campaign, produced by a Hollywood production studio, included nationwide radio and television advertisements.[86] In one dystopian video, a distraught patient is introduced to FRANK, his new robotic nurse.[87] The patient is seen saying in a panicked voice, "This thing isn't a nurse! It's not even a human!"[88] FRANK then misdiagnoses the patient—leading to the amputation of the patient's leg, to the patient's horror—and tells the patient he is pregnant.[89] All appears lost until a nurse shows up to save the day.[90] She tells the patient, "Don't worry, you're in the care of a registered nurse now." To the technician working on the computer, she then declares, "You and your computer are in over your heads. Get a doctor, and get out of my way!"[91] The message from America's healthcare providers is clear: *Robots are dangerous to your health; insist on a human.*[92]

Providers and patients have shown, at best, skepticism and, at worst, outright hostility toward the use of algorithms in the healthcare context. However, this bias against machines in healthcare is odd if one cares most about health outcomes. Machines have been shown to be better at detecting skin cancer,[93]

84. *See* D. T. Max, *Paging Dr. Robot*, NEW YORKER (Sept. 23, 2019), https://www.newyorker.com/magazine/2019/09/30/paging-dr-robot [https://perma.cc/B9JR-SQ22].

85. *Nurses Launch New Campaign to Alert Public to Dangers of Medical Technology and More*, NAT'L NURSES UNITED (May 13, 2014), https://www.nationalnursesunited.org/press/nurses-launch-new-campaign-alert-public-dangers-medical-technology-and-more [https://perma.cc/J9XV-G2FC].

86. *Id.*

87. Nat'l Nurses United, *National Nurses United: Nurses v. Computer Care Ad (HD)*, YOUTUBE (May 12, 2014), https://www.youtube.com/watch?v=YthF86QDOXY [https://perma.cc/2RT7-8WKS].

88. *Id.*

89. *Id.*

90. *Id.*

91. *Id.*

92. *See id.* Of course, this can be explained away as a union trying to fight a loss of jobs to machines. But note that the creators of the message did *not* make the video about nurse job losses; the focus of the video is on the harm to patients from robot healthcare. *See infra* Section III.D.

93. H.A. Haenssle et al., *Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma*

triaging diagnoses,[94] and identifying treatment options.[95] For underserved populations, the benefits of algorithms are much greater, as studies suggest that doctors' implicit bias makes them worse than algorithms at diagnosing disease in underrepresented populations.[96]

Yet the desire for human doctors persists. Part of this can be explained away as path dependency, but much of it appears to be a simple preference for human control over the process, a misplaced confidence in human abilities, and a distrust of machines.

3. Employment

We see robophobia in human resources too. Hiring decisions are notoriously plagued by implicit, and sometimes explicit, bias.[97] This is partly a result of the enormous discretion given to individual decision-makers within firms to hire who they see fit—discretion that has been reinforced by the Court in recent years.[98] As a function of the huge number of applicants for a given position, human-resources screeners rely on heuristics like college prestige to screen the first batch of candidates, and these screening mechanisms are enormously biased against diverse

Recognition in Comparison to 58 Dermatologists, 29 ANNALS ONCOLOGY 1836, 1836 (2018) (finding that a deep-learning convolutional neural network significantly outperformed fifty-eight dermatologists in diagnostic classification of lesions).

94. Laura Donnelly, *Forget Your GP, Robots Will 'Soon Be Able to Diagnose More Accurately than Almost Any Doctor*,' TELEGRAPH (Mar. 7, 2017, 10:00 PM), https://www.telegraph.co.uk/technology/2017/03/07/robots-will-soon-able-diagnose-accurately-almost-doctor/[https://perma.cc/82LK-ULEJ].

95. Steve Lohr*, IBM Is Counting on Its Bet on Watson, and Paying Big Money for It*, N.Y. TIMES (Oct. 17, 2016), https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html [https://perma.cc/4TTS-ZWV6]. Watson found treatments when doctors failed in 30 percent of cases. *Id.*

96. *See* Pierson et al., *supra* note 77 and accompanying text.

97. *See, e.g.*, Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable Than Lakisha and Jamal?*, 94 AM. ECON. REV. 991 (2004) (reporting results of an experiment demonstrating implicit bias in hiring decisions consistent with longstanding concerns about discriminatory hiring practices).

98. *See* Wal-Mart v. Dukes, 564 U.S. 338 (2011). Justice Scalia wrote that if a firm does its hiring and firing in a decentralized "largely subjective manner," rather than in a systematic and uniform way, it will not be possible to accuse the firm of systematic bias. *Id*. at 343, 373–74. Because "Wal-Mart has no testing procedure or other companywide evaluation method, [it cannot be] charged with bias." *Id.* at 353.

candidates.[99] As such, algorithms have the prospect of making hiring decisions much less discriminatory.[100]

And yet the use of algorithms in hiring is widely criticized for the prospect of bias.[101] For example, Amazon uses machine-learning tools to screen applicants for possible employment, like many large companies.[102] In 2018, *Reuters* published a story suggesting that Amazon developed an algorithm that continued to generate sexist hiring recommendations—for example, by recommending men over women for engineering roles.[103] The article generated a firestorm of media attention, and Amazon, rather than working to improve the algorithm, decided to continue using human screeners for hiring.[104] The press ran articles with titles like "How Amazon Accidentally Invented a Sexist Hiring Algorithm"[105] and "Amazon's Sexist AI Recruiting Tool: How Did It Go So Wrong?"[106] Despite these alarming headlines, there is no evidence in any of these articles that suggests that the algorithm was actually worse than the human-centered process that it replaced. Yet it appears that Amazon decided it would be

99. Frida Polli, *Using AI to Eliminate Bias from Hiring*, HARV. BUS. REV. (Oct. 29, 2019), https://hbr.org/2019/10/using-ai-to-eliminate-bias-from-hiring [https://perma.cc/X7DF-3WXM] (arguing that while algorithms in hiring have problems, they are fixable and in most cases better than human screeners, who have a long history of discrimination and bias).

100. *Id.*; *see also* Ifeoma Ajunwa, *Automated Employment Discrimination*, 34 HARV. J.L. & TECH. 1, 17 (2021) (surveying the landscape of hiring algorithms that promise to reduce overall bias).

101. *See generally* Ajunwa, *supra* note 100, at 2–26.

102. *See* Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018, 5:04 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [https://perma.cc/965K-F82G].

103. *Id*.

104. *Id.*; *see also* Bo Cowgill & Catherine E. Tucker, Economics, Fairness, and Algorithmic Bias 38, 42–43 (May 11, 2019) (unpublished manuscript), http://www.columbia.edu/~bc2656/papers/JEP-EconAlgoBias-V1.pdf [https://perma.cc/492Y-28HM].

105. *See* Guadalupe Gonzalez, *How Amazon Accidentally Invented a Sexist Hiring Algorithm*, INC. (Oct. 10, 2018), https://www.inc.com/guadalupe-gonzalez/amazon-artificial-intelligence-ai-hiring-tool-hr.html [https://perma.cc/6XQ3-4P87].

106. Julien Lauret, *Amazon's Sexist AI Recruiting Tool: How Did It Go So Wrong?*, BECOMING HUM.: A.I. MAG. (Aug. 16, 2019), https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e [https://perma.cc/7HLT-XY7U]. *But see* Cowgill & Tucker, *supra* note 104, at 44 ("To our knowledge, this essay contains the only publicly expressed skepticism about Reuters' interpretation of Amazon's hiring algorithm.").

better to kill the program, because it might open up the firm to criticism.[107]

Firms are incentivized, from a legal and public-relations standpoint, to continue to rely on biased and deeply flawed human hiring processes. The public seems to have little tolerance for machine errors in hiring, even where those machines promise improvements over the alternative. Unfortunately, this appears to have a material effect on firm behavior.[108]

4. Criminal Justice

Criminal justice is another area where algorithms offer great promise, yet the public reaction has been largely negative. For example, bail determinations—decisions to release or jail people accused of crimes—are riddled with racism, bias, and error.[109] Automated bail systems, in which an algorithm considers risk factors like the nature of the suspected crime and prior records, are increasingly used and hold enormous promise.[110] Notably, they promise *both* to reduce the number of people wrongly detained or wrongly released *and* to reduce the well-documented racial bias of those determinations.[111] That is, while human bail determinations are "plagued by the distortive effects of heuristics, implicit bias, and sheer noise,"[112] algorithms offer a compelling alternative.

Cash bail is a troubling yet common feature of the criminal justice system, requiring defendants to put up money—the cash bail—as collateral. If the defendant fails to appear in court, they lose their money. Cash bail is increasingly seen as ineffective and discriminatory against the least privileged. One recent study found "no evidence that financial collateral has a deterrent effect on failure-to-appear," meaning that the cash-bail system

107. Cowgill & Tucker, *supra* note 104, at 44.

108. Bo Cowgill et al., *The Managerial Effects of Algorithmic Fairness Activism*, 110 AM. ECON. ASS'N PAPERS & PROCEEDINGS 85 (2020).

109. *See, e.g.*, Ian Ayres & Joel Waldfogel, *A Market Test for Race Discrimination in Bail Setting*, 46 STAN. L. REV. 987 (1994) (showing significant racial discrimination in bail bond determinations in Connecticut).

110. *See, e.g.*, Joel Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237 (2018).

111. *Id.* at 241.

112. Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 638 (2020).

is not achieving its goal.[113] Others have suggested that cash bail is unconstitutional.[114] Courts around the country are now looking to alternatives to cash bail, including using risk assessment tools to determine which defendants should be released.[115]

For example, in California, a 2020 ballot initiative proposed to "replace[] cash bail with risk assessments for detained suspects awaiting trials," and supporters of the move said it would eliminate a system in which "the rich can go free" and the poor are "imprisoned solely due to poverty."[116] But the California initiative failed. Opponents of the referendum—which was strongly opposed by the for-profit bail industry—argued that the initiative would use "computer programs to make important justice decisions. These are the same type of algorithms that Big Data companies use to bombard us with ads every day."[117] Another opponent said, "[T]his costly, reckless plan will use racially-biased computer algorithms to decide who gets stuck in jail and who goes free. That's not right."[118] This view—that algorithms are biased, even when they are being developed and implemented to correct current biases—is now commonplace.

Indeed, the scholarly and media reaction to the prospect of algorithms in criminal justice has been overwhelmingly negative. There are hundreds of articles and news reports about the harms of algorithms in criminal justice but comparatively little focus on the potential benefits.[119] In a blockbuster series of reports, ProPublica analyzed data from COMPAS, an algorithmic tool that provides judges with risk scores for use in bail

113. Aurélie Ouss & Megan Stevenson, *Bail, Jail, and Pretrial Misconduct: The Influence of Prosecutors*, at 1 (June 2021) (unpublished manuscript) (on file with author), https://aouss.github.io/NCB.pdf [https://perma.cc/Y78J-CXHS].

114. *See, e.g.*, Sandra G. Mayson, *Detention by Any Other Name*, 69 DUKE L. REV. 1643 (2020).

115. *See* Jenny E. Carroll, *Beyond Bail*, 73 FLA. L. REV. 143 (2021) (summarizing the bail-reform movement of the last twenty years).

116. *California Proposition 25, Replace Cash Bail with Risk Assessments Referendum (2020)*, BALLOTPEDIA, https://ballotpedia.org/California_Proposition_25,_Replace_Cash_Bail_with_Risk_Assessments_Referendum_(2020) [https://perma.cc/Y2HG-N96X].

117. *Id.* (statement of Former Assemblyman Joe Coto).

118. *Id.* (statement of Jeff Clayton).

119. A search of law review and law journal articles regarding algorithmic bias produced over 500 results. Many of these acknowledge the potential upsides of algorithms, but the focus is overwhelmingly on the downsides.

determinations.[120] The conclusions were attention grabbing: COMPAS data appeared to have significant racial bias.[121] The reports made the authors finalists for a Pulitzer Prize "[f]or a rigorous examination that used data journalism and lucid writing to make tangible the abstract world of algorithms and how they shape our lives in realms as disparate as criminal justice, online shopping and social media."[122] But whether the algorithms are in fact examples of "machine bias," as the series suggests—and crucially, whether they are more biased than the humans they displace—is far from clear.[123] Several empiricists have called into question whether the data in the reports support the conclusions ProPublica draws.[124]

None of this is meant to suggest that algorithmic justice is risk free. To the contrary, there are serious and legitimate concerns about the use of biased algorithms in criminal justice.[125] A badly drafted algorithm that is nonreviewable could do enormous damage at a huge scale.[126] The stakes of getting criminal-justice algorithms wrong are enormously high, just as the stakes of a badly designed self-driving car are high. And then there is the scale of the problem: one bad algorithm multiplied over

120. Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/JE68-G6PS].

121. *Id*.

122. *The 2017 Pulitzer Prize Finalist in Explanatory Reporting*, PULITZER, https://www.pulitzer.org/finalists/julia-angwin-jeff-larson-surya-mattu-lauren-kirchner-and-terry-parris-jr-propublica [https://perma.cc/4VE7-8JEV].

123. *See, e.g.*, Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ [https://perma.cc/MC4A-8RJF].

124. *See, e.g.*, *id.*; *see also* WILLIAM DIETERICH ET AL., NORTHPOINTE INC., COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY (2016), https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf [https://perma.cc/LBJ2-EXGK]; *Response to ProPublica: Demonstrating Accuracy Equity and Predictive Parity*, EQUIVANT (Dec. 1, 2018), https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/ [https://perma.cc/6H6F-7LSZ]; Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."*, 80 FED. PROB. 38 (2016); Abe Gong, *Ethics for Powerful Algorithms (1 of 4)*, MEDIUM: ABE GONG (July 12, 2016), https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84#.dhsd2ut3i [https://perma.cc/ZC9F-26NH].

125. *See, e.g.*, Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245 (2016).

126. *See* Mayson, *supra* note 5, at 2280–81.

millions of decisions is going to do more damage than a single bad human decision-maker. Clearly, in both scenarios, we must carefully assess the risks before these machines are put to their intended use.

But to focus overwhelmingly on the risks of these new technologies—rather than assessing those risks as they compare to the promise of these new technologies—is to betray a particular kind of mistrust of algorithms. If we ask if these new systems are perfect, then we may never benefit from these nonhuman deciders, simply because this holds them to a higher standard than the humans they would replace.[127]

Yet the attitude of both courts and many legal scholars has reflected deep skepticism that algorithms are consistent with due process. For example, in *State v. Loomis*,[128] the Wisconsin Supreme Court held that the state constitution allowed judges to consult algorithms for risk scores. However, the court also held that due process demanded that those risk scores "not be considered as the determinative factor in deciding whether the offender can be supervised safely and effectively in the community."[129] The court explained that significant reliance on COMPAS would "raise due process challenges regarding whether a defendant received an individualized sentence."[130]

Scholarship reflects a similar skepticism about machines in criminal justice. Aziz Huq sees this as part of a larger trend in American law toward a right to a human decision, as opposed to a robot-driven decision.[131] He interprets basic elements of constitutional law—especially the right to a jury trial and basic notions of due process, which include both notice and a hearing—as likely incompatible with algorithmic justice.[132] Others have made more normative arguments against machines in criminal

127. *See infra* Section II.A.

128. State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

129. *Id.* at 760.

130. *Id.* at 764.

131. Huq, *supra* note 112, at 622.

132. *Id.* at 625–26 (discussing *Loomis* in this context and noting: "[T]he idea of due process might also be grounds for a mandatory human decision rather than a machine judgment. At its core, the idea of procedural due process is thought to entail 'notice and some kind of hearing.' There is some debate about the timing and the content of a hearing, at least so far as the Constitution's Due Process guarantee is concerned. But it is not hard to see how a question could arise whether due process is supplied by a machine decision. Indeed, it is arguably difficult to make sense of the idea of a "hearing" in the absence of a natural person who is either physically present for verbal arguments, or who reads and evaluates written submissions.")

justice. For example, Kiel Brennen Marquez and Stephen Henderson write that humans, and not robots, should be judges because only humans can be defendants.[133] This idea, they argue, is "intuitive," and the authors note, "We suspect this intuition is widespread."[134] As the authors explain, the point of their article is to "rationalize" the "intuition" that "humans remain 'in the loop' of some decision-making even if it fails to increase—and may well diminish—accuracy and consistency."[135]

To be clear, there are reasons to be concerned about algorithmic bias in criminal justice.[136] But algorithms are tools that can be used for good and for bad. The popular press, the court cases, and the scholarly literature are overly focused on the bad.

5. Discovery & Evidence

Robophobia crops up in civil litigation as well. It has been shown that machines are better than humans—faster, cheaper, more thorough—at many aspects of document review and related discovery tasks, especially over large datasets.[137] Not only are machines more effective than humans at certain kinds of reviews but lawyers are especially bad at them.[138] Lawyers are good at the interpretive task of identifying whether a *particular* document is responsive or not, but they are much worse at accurately plucking the relevant documents from a large stack of irrelevant material.[139] So we might imagine that lawyers would benefit from systems where a machine identifies a potential set

133. Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137 (2019).

134. *Id.* at 140.

135. *Id.* at 139–40.

136. I just want to flag for the reader that criminal justice is also, perhaps counterintuitively, a place where robophilia happens. Machines and algorithms are being deployed around the country *despite* all of this negative coverage. *See infra* Part IV.

137. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 1 (2011).

138. Sam Skolnik, *Lawyers Aren't Taking Full Advantage of AI Tools, Survey Shows*, BLOOMBERG L. (May 14, 2019), https://news.bloomberglaw.com/business-and-practice/lawyers-arent-taking-full-advantage-of-ai-tools-survey-shows [https://perma.cc/89ZB-NMKX] (reporting results of a survey of 487 lawyers finding that lawyers have not well utilized useful new tools).

139. Moore v. Publicis Groupe, 287 F.R.D. 182, 191 (S.D.N.Y. 2012) ("Computer-assisted review appears to be better than the available alternatives, and thus should be used in appropriate cases.").

of documents and lawyers then do the "last yard" of review to determine which documents in the smaller set are, in fact, relevant.

But lawyers generally decline to trust artificial-intelligence tools to conduct document review, despite the evidence that they can work. Surveys of lawyers show a reluctance to rely on technology-assisted review when compared to having lawyers make relevance determinations about every single document.[140] A recent survey of practicing attorneys found that only 31.1 percent of respondents use Technology Assisted Review (TAR) in all or most of their cases.[141] This is so despite the obvious efficiency and accuracy benefits of TAR.[142]

This reluctance is somewhat hard to understand. First, lawyers regularly turn to computers to conduct keyword searches—and, in fact, there is evidence that lawyers tend to be *overconfident* in the responsiveness of these results.[143] Additionally, lawyerly reluctance to use AI might have once been explained by a fear that these determinations would not hold up in court.[144] But, today, "it is now black letter law that where the producing party wants to utilize TAR for document review, courts will permit it."[145] So a fear about judicial acceptance hardly explains attorneys' widespread reluctance to use robots more thoroughly in document review.[146]

---

140. Bob Ambrogi, *Latest ABA Technology Survey Provides Insights on E-Discovery Trends*, CATALYST: E-DISCOVERY SEARCH BLOG (Nov. 10, 2016), https://catalystsecure.com/blog/2016/11/latest-aba-technology-survey-provides-insights-on-e-discovery-trends/ [https://perma.cc/9ZHU-34S2] (noting that "firms are failing to use advanced e-discovery technologies or even any e-discovery technology").

141. Doug Austin, *Announcing the State of the Industry Report 2021,* EDISCOVERY TODAY (Jan. 5, 2021), https://ediscoverytoday.com/2021/01/05/announcing-the-2021-state-of-the-industry-report-ediscovery-trends/ [https://perma.cc/TK4V-PVTK].

142. *Id.*

143. David C. Blair & M. E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMMC'NS ACM 289 (1985).

144. Thomas E. Stevens & Wayne C. Matus, *Gaining a Comparative Advantage in the Process*, NAT'L L.J. (Aug. 25, 2008), https://www.law.com/nationallawjournal/almID/1202423952310/ [https://perma.cc/NP4W-K35L] (describing a "general reluctance by counsel to rely on anything but what they perceive to be the most defensible positions in electronic discovery, even if those solutions do not hold up any sort of honest analysis of cost or quality").

145. Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 127 (S.D.N.Y. 2015).

146. *See* The Sedona Conference, *supra* note 12, at 235–36 ("Some litigators continue to primarily rely upon manual review of information as part of their review process. Principal rationales [include] . . . the perception that there is a lack

There are other explanations for lawyers' reluctance to rely on algorithms, but none of them are convincing. One explanation is that use of an algorithm will invite opposing counsel to demand more transparency into the algorithm than they would demand of traditional human document review. As e-evidence expert Judge Peck noted, "Part of the problem remains requesting parties that seek such extensive involvement in the process and overly complex verification that responding parties are discouraged from using TAR."[147] Yet another explanation is fear over job security. As one lawyer noted, reflecting on a recent conference on AI and legal practice, there was a "palpable" fear of "robots that are coming to take away our jobs and that possibly have even more pernicious goals, up to and including human domination." [148]

6. National Security

Weapons systems are increasingly automated, meaning that many tasks formerly done by humans are now being done by machines. This includes tasks like flying aircrafts, detecting incoming fire, and even deciding how to respond, including firing weapons.[149] In addition to the military advantage these weapons pose, there is a case to be made that they are more ethical than human combatants. As Ronald Arkin notes, lethal autonomous weapons systems may be imperfect, but they promise to be better than human soldiers at reducing casualties and adhering to the laws of war.[150] In addition to never being fatigued or upset, robot weapons need not have a self-preservation instinct.[151] And in the event that the laws of war are broken, robots will be more likely to report the abuse than a human soldier.

---

of scientific validity of search technologies necessary to defend against a court challenge . . . .").

147. Doug Austin, *Learning to Trust TAR as Much as Keyword Search: eDiscovery Best Practices*, EDISCOVERY TODAY (June 28, 2021), https://ediscoverytoday.com/2021/06/28/learning-to-trust-tar-as-much-as-keyword-search-ediscovery-best-practices/ [https://perma.cc/954K-5G68].

148. Robert Ambrogi, *Fear Not, Lawyers, AI Is Not Your Enemy*, ABOVE LAW (Oct. 30, 2017), https://abovethelaw.com/2017/10/fear-not-lawyers-ai-is-not-your-enemy/ [https://perma.cc/W9TD-TK3R].

149. *See, e.g.*, PAUL SCHARRE, ARMY OF NONE 171 (2018).

150. Ronald C. Arkin, *The Case for Ethical Autonomy in Unmanned Systems*, 4 J. MIL. ETHICS 332–341 (2010).

151. *Id.*

Yet there is a large and growing campaign to ban so-called "autonomous weapons"—those weapons that could select and engage targets without human intervention.[152] In 2013, the Human Rights Watch launched the "Campaign to Stop Killer Robots," and the organization has argued repeatedly for the banning of autonomous weapons.[153] The United Nations has convened a working group of governmental experts on autonomous weapons systems,[154] which affirmed the relevance of international law and, in particular, the treaty on Certain Conventional Weapons to autonomous weapons systems.[155] The movement for a global ban on autonomous weapons systems—which increasingly looks like it will succeed[156]—mirrors popular opinion. One study suggests that a majority of American survey respondents oppose autonomous weapons by a two-to-one margin.[157] Thirty countries and 165 NGOs have called for an outright ban on lethal autonomous weapons, citing "ethical concerns, including concerns about operational risk, accountability

152. Bonnie Docherty, *We're Running Out of Time to Stop Killer Robot Weapons,* GUARDIAN (Apr. 11, 2018), https://www.theguardian.com/commentisfree/2018/apr/11/killer-robot-weapons-autonomous-ai-warfare-un [https://perma.cc/9D94-S4DR].

153. *Id.*

154. *Background on LAWS in the CCW*, UNITED NATIONS OFF. FOR DISARMAMENT AFFS., https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/ [https://perma.cc/4AEK-WU6S].

155. United Nations Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to be Excessively Injurious or to have Indiscriminate Effects, Dec. 2, 1983, 1342 U.N.T.S. 137. Importantly, this includes the so-called Martens Clause, stated, for example, in the Additional Protocol I of 1977 to the Geneva Conventions: "In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience." Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), June 8, 1977, 1125 U.N.T.S. 3, 7.

156. *See* Thomas Burri, *International Law and Artificial Intelligence*, 60 GER. Y.B. INT'L L. 91, 98–99 (2017) (arguing that "The Geneva Process Will Result in a Ban on Autonomous Weapons Systems").

157. James Fahey, *Carpenter Presents on Lethal Autonomous Weapons at UN Conference*, UNIV. MASS. AMHERST (June 5, 2014, 6:00 PM), https://polsci.umass.edu/news/carpenter-presents-lethal-autonomous-weapons-un-conference [https://perma.cc/GZ4K-VRE3].

for use, and compliance with the proportionality and distinction requirements of the law of war."[158]

The U.S. government has resisted an outright ban on autonomous weapons, suggesting that they can, in fact, reduce civilian casualties.[159] But the Defense Department's new rules for autonomous weapons systems clearly privilege human judgment over autonomous judgment. The rules require that "[a]utonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise *appropriate levels of human judgment* over the use of force."[160] What is appropriate is unclear; what is clear is that *human* judgment is paramount. As the U.S. government explained in a white paper,

> "appropriate" is a flexible term that reflects the fact that there is not a fixed, one-size-fits-all level of human judgment that should be applied to every context. What is "appropriate" can differ across weapon systems, domains of warfare, types of warfare, operational contexts, and even across different functions in a weapon system.[161]

In addition to the normal weapons-systems review process, the Defense Department's rules also call for the Under Secretary of Defense for Policy, the Chairman of the Joint Chiefs of Staff, and either the Under Secretary of Defense for Acquisition and Sustainment or the Under Secretary of Defense for Research and Engineering to approve the system—an exceptional additional layer of senior-level review.[162]

In other words, international law and domestic regulations demonstrate a clear bias in favor of human deciders over robot

158. KELLY M. SAYLER, CONG. RSCH. SERV., DEFENSE PRIMER: U.S. POLICY ON LETHAL AUTONOMOUS WEAPONS SYSTEMS, IF11150 (2019), https://fas.org/sgp/crs/natsec/IF11150.pdf [https://perma.cc/WWG2-3TDP].

159. U.S., Humanitarian Benefits of Emerging Technologies in the Area of Lethal Autonomous Weapons, U.N. Doc. CCW/GGE.1/2018/WP.4 (Apr. 3, 2018), https://undocs.org/pdf?symbol=en/CCW/GGE.1/2018/WP.4 [https://perma.cc/AJA5-Q4VQ] [hereinafter Humanitarian Benefits].

160. U.S. DEP'T OF DEF., DIRECTIVE 3000.09, AUTONOMY IN WEAPON SYSTEMS 2 (2012), https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf [https://perma.cc/JS97-5329] (emphasis added).

161. U.S., Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, at 2, U.N. Doc. CCW/GGE.2/2018/WP.4 (Aug. 28, 2018), https://undocs.org/en/CCW/GGE.2/2018/WP.4 [https://perma.cc/4WHQ-LG82].

162. Sayler, *supra* note 158.

deciders, despite the real promise of autonomous weapons to reduce civilian casualties.[163] Robot performance is capped at human performance; human performance is a ceiling, not a floor, for robot performance.

***

These are just a few examples where humans hold machines to exceptionally high standards. As noted at the outset, each of these examples shares something with the others but is also distinct. In each example, we see a general skepticism of machines, a wariness, and a demand that they be flawless. Yet in each case there are important differences—in what motivates the concern, in its effects—that will guide our understanding and therefore our response, if any, to the bias.

II. TYPES OF ROBOPHOBIA

As the previous examples suggest, there is a wide range of negative attitudes, feelings, and concerns about algorithms. We can put this antirobot sentiment into different categories, which may be helpful later as we try to think through the possible explanations for it. This Part is again descriptive. There may be good reasons for holding robots to higher process standards—by, say, requiring that they explain themselves more fully than a human judge might—or there may not. For now, what matters is establishing some of the different ways that humans are biased against robot decision-makers across a range of domains. In the next Part, we will examine the explanations for these biases and ask whether any are justified.

A. Elevated Performance Standards

In evaluating where and when to deploy an algorithm, we regularly hold algorithms to higher performance standards than we would a similarly situated human; indeed, the standard is often perfection. The self-driving car examples described above are illustrative. Any algorithmic error, however slight, is cause for news reports, press conferences, and even regulatory changes. Algorithms are held to higher performance standards in a range of other areas as well. Autonomous weapons are held to a higher standard of certainty before being allowed to fire on

163. *See* Humanitarian Benefits, *supra* note 159.

their targets.[164] Patients are less tolerant of mistakes from robot doctors than human doctors.[165] And, in criminal justice, we insist that algorithms both maximize accuracy and all forms of equity—even among mutually exclusive and incompatible values.[166] In many areas, it seems that we expect algorithms not only to outperform the alternative but to be perfect.

B. Elevated Process Standards

One of the ways that scholars and practitioners have reacted to the rise of robots is to ask that they not only achieve near-perfect outcomes but that they explain their decision-making processes clearly and fully.[167] Demands for algorithmic transparency are one example of this. A chorus of commentators argues that algorithms must be transparent and legible—meaning that their reasoning is plain and understandable to reviewers.[168] Others argue that current law already requires this transparency of machine decision-makers. For example, the European Union's privacy regime, the General Data Privacy Regulation, can be understood to require a "right to explanation"—meaning that algorithms must explain to a human how they arrived at a decision.[169]

These calls for algorithmic transparency are welcome, but it is worth noting that they ask for more than is required of humans, who routinely deny visas, decide cases, and decline to lend money. That is, humans frequently make these same decisions without explaining their reasoning—and, unlike machines,

164. SCHARRE, *supra* note 149, at 172.

165. *See supra* Section I.B.2.

166. *See* Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, *in* 8TH INNOVATIONS IN THEORETICAL COMPUTER SCIENCE 43:1 (Christos H. Papadimitriou ed., 2017), https://drops.dagstuhl.de/opus/volltexte/lipics-complete/lipics-vol67-itcs2017-complete.pdf [https://perma.cc/G3R5-GWJ9] (arguing that there is a tension between "competing notions of what it means for a probabilistic classification to be fair to different groups" and documenting how we expect algorithms to satisfy these competing goals simultaneously).

167. *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1089 (2018) (surveying literature calling for algorithms that are legible and transparent).

168. *Id.*; *see also* Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1506 (2013); Kiel Brennan-Marquez, *Plausible Cause: Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1267–68 (2017).

169. *See* Margot Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 1 (2019).

humans can more easily justify their decisions in ways that are duplicitous or intended to hide prejudice.

As another example, human judges regularly issue summary orders—decisions without an explanation.[170] For some, this is a problem[171] and inconsistent with the publicity principle of deliberative democracy.[172] Even so, none of these critics have suggested that courts *must* issue opinions. And when judges do explain themselves, why are their explanations to be believed? They may simply be ex-post reasoning bipartisan hacks. Or they may be engaged in sophisticated efforts that maintain acoustic separation between what is communicated internally and what is communicated externally.[173]

This is not surprising; after all, judges are only human. Humans are not just biased; they are sophisticated post-hoc reasoners and experts at deception.[174] (Deception, it turns out, is closer to the rule than the exception in human communication.[175]) So, to say that an algorithm must always explain itself fully and honestly is to say that it must follow a stricter set of procedures than a similarly situated human.

C. Harsher Judgments

When robots act badly, we judge them more harshly than when humans act badly. In particular, we appear both to find robots more blameworthy and to penalize them more than we would a similarly situated human. A survey of judges showed that they assigned more responsibility to autonomous vehicles than to similar human-operated cars for *exactly the same*

170. *See* Andrew Keane Woods, *The Transparency Tax*, 71 VAND. L. REV. 1, 12–13 (2018).

171. *See* William Baude, *Foreword: The Supreme Court's Shadow Docket*, 9 N.Y.U. J.L. & LIBERTY 1, 3–4 (2015) (describing the Court's use of orders and stays without opinions).

172. David Luban, *The Publicity Principle*, *in* THE THEORY OF INSTITUTIONAL DESIGN 154, 169–72 (Robert E. Goodin ed., 1996) (describing the principle as a component of good governance).

173. *See* Meir Dan Cohen, *Decision Rules and Conduct Rules: On Acoustic Separation in Criminal Law*, 97 HARV. L. REV. 625 (1983).

174. Allison Kornet, *The Truth About Lying*, PSYCH. TODAY (May 1, 1997), https://www.psychologytoday.com/us/articles/199705/the-truth-about-lying [https://perma.cc/HFD8-86EX] (surveying social psychology findings and concluding that dishonesty pervades social interactions).

175. *Id.*

conduct.[176] The same survey found that judges also awarded plaintiffs more damages when the plaintiff was harmed in an accident caused by an autonomous driver than a human driver.[177]

This is consistent with—and perhaps a corollary of—the idea that robots should be held to a higher performance standard than similarly situated humans.[178] Naturally, if we expect machines to perform perfectly, or in any event better than a similarly situated human, then it makes sense to punish them more harshly for their errors—on the theory that they deviated further from their expected performance. That is, if you expect a robot to behave perfectly and it does not, then it is understandable to be upset at any poor performance. On the other hand, if you expect a human to crash their car, you will be less surprised when they do crash and perhaps less likely to punish them severely for conforming with anticipated behavior.

D. Distrust

Suppose you request a ride using your trusted ridesharing app, and the car pulls up. You have never driven with this driver before, and you know little to nothing about the driver's safety record. But you get in without hesitation. Now suppose that the car pulls up without a driver in the driver's seat. Do you get in? If so, do you hesitate? If you would hesitate before getting into the robot-driven car but not the stranger-driven car, you are revealing a distrust of machines. In both scenarios, there is a trusted intermediary—the ridesharing app—that has a strong disincentive against putting passengers in dangerous situations. Still, the evidence suggests that we hesitate to trust the robot.[179]

Lawyers' reluctance to rely on algorithms for discovery is a good example of machine distrust. Sometimes that distrust

176. *See* Jeffrey Rachlinski, *Judging Autonomous Vehicles*, *poster presented at* THE 13TH ANNUAL CONFERENCE ON EMPIRICAL LEGAL STUDIES (CELS), Nov. 9–10, 2018, Univ. of Mich. L. Sch. (on file with author).

177. *Id.*

178. *See supra* Section II.0.

179. Psychologists have found that this is true, though it is more true of older riders than young riders. *See* Hillary Abraham & Chaiwoo Lee, *Autonomous Vehicles and Alternatives to Driving: Trust, Preferences, and Effects of Age*, *paper presented at* THE TRANSPORTATION RESEARCH BOARD (TRB) 96TH ANNUAL MEETING, Jan. 8–12 2017, Washington, D.C., https://agelab.mit.edu/index.php/system/files/2018-12/2017_TRB_Abraham.pdf [https://perma.cc/QD8V-N7XS].

might be based on a fear about someone else's bias—for example, a lawyer reluctant to use a machine because they fear it will be viewed pejoratively by judges. Other times, the lawyer just has less trust in the machine's ability to do the job as well as a human. Doctors are the same way, exhibiting high levels of distrust of algorithms.[180]

There is a large body of literature on the levels of trust in the human-automation interaction, in part because trust is so critical to the relationship's success and in part because humans tend to exhibit so little trust in automated systems.[181] This appears to be especially true for algorithms. Not only do we have less confidence on the front end of robot decisions—before they are made—but we have more doubt on the back end: we second-guess robot decisions in ways that we do not for human decisions. When robot decisions must be confirmed by human review (and a similarly situated human would not be subjected to the same reviewing requirements), we reveal our lack of confidence in robot decision-making.

Relatedly, even when we have initial confidence in automated decisions, that confidence is more fragile than our confidence in human decisions. Berkeley J. Dietvorst and co-authors showed that when experiment subjects trusted an algorithm, they would immediately and almost completely lose faith in the algorithm after seeing it err; the same is not true when humans make errors.[182]

---

180. Keerthi Vedantam, *Venture Cash Is Pouring into AI that Can Diagnose Diseases. Doctors Aren't Sure They Can Trust It.*, DOT.LA (Aug. 7, 2021, 10:48 AM), https://dot.la/medical-ai-venture-2654560192.html [https://perma.cc/X9PN-JHST] ("Despite the sweeping promises of medical imaging AI, doctors remain largely distrustful of the tech.").

181. *See* Kevin Anthony Hoff & Masooda Bashir, *Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust*, 57 HUM. FACTORS 407 (2015) (providing a meta-analysis of the findings of 127 studies and identifying three distinct kinds of human trust in automation); *see also* EMILEE RADER & RICK WASH, TRUSTWORTHY ALGORITHMIC DECISION-MAKING: WORKSHOP REPORT 2 (2017), https://www.rickwash.com/papers/Trustworthy%20Algorithmic%20Decision-Making%202017%20-%20Workshop%20Report%20-%20Final.pdf [https://perma.cc/4XEL-UQAS] (summarizing a workshop that highlighted the importance of trust as a "key factor that helps people decide whether to use systems that engage in algorithmic decision-making").

182. Dietvorst et al., *Overcoming Aversion*, *supra* note 9.

E. Prioritizing Human Decisions

One manifestation of our mistrust of algorithms is the now-common idea that automated systems should maintain a "human in the loop," which privileges human involvement in a particular process over outcomes.[183] In 2017, a professional organization for engineers, the Institute of Electrical and Electronics Engineers (IEEE), unveiled an ambitious set of new ethical guidelines for engineers working on automated and artificial-intelligence systems.[184] One of the guidelines' cornerstone principles was to keep a "human in the loop"—the idea that all automated systems should ensure that humans play a crucial function at some moment during the decision-making or execution process.[185]

Legal scholars have reinforced these calls. Tim Wu writes that artificial intelligence might supplant many aspects of the common law and, therefore, steps should be taken to keep humans involved in judicial decision-making.[186] Wu's work is mostly descriptive but, in predicting a future of hybrid human-machine judging, he makes the normative case for including humans in the loop. Wu is careful to note that there are advantages to machine decision-makers, especially where they might handle "routine procedural matters, like the filing of motions," and leave the hard matters for human judges.[187] But human judges, Wu argues, are normatively more desirable for two reasons: procedural fairness and capability. Because people are robophobic, Wu suggests, "having a major [legal] decision be made by a human may become a basic indicium of fairness."[188] Moreover, in hard cases, Wu argues that human decision-making will be more

183. *See* Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. SCI. 216 (2017) (describing the human-in-the-loop construct and applying it to the transatlantic privacy debate).

184. IEEE GLOB. INITIATIVE ON ETHICS OF AUTONOMOUS & INTELLIGENT SYS., ETHICALLY ALIGNED DESIGN (1st ed. 2019), https://algorithmwatch.org/de/wp-content/uploads/2019/03/IEEE-EAD1e.pdf [https://perma.cc/3TJH-MR9R].

185. *Id.*

186. Tim Wu, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems*, 119 COLUM. L. REV. 2001, 2002–03 (2019) (describing the many advantages of human decisions over algorithmic decisions, including that human decisions are likely to engender greater public support).

187. *Id.* at 2005.

188. *Id.* at 2022.

subtle and intelligent.[189] Others have reached similar conclusions.[190] This is prima facie evidence of a certain kind of bias against machines—that they should not, by design, be allowed to make significant judgments or take actions without human input.

***

Across the board, we treat robots differently than we treat humans, even where the costs of doing so are high. Robophobia is pervasive: it comes in a variety of flavors, and it manifests itself in a number of different ways. What do we know about it? Why does it happen?

III. EXPLAINING ROBOPHOBIA

Our judgment of algorithms is motivated by different intuitions. A passenger's hesitation to get into a self-driving taxi may originate from a different source than an assembly-line worker's hesitation to train a robot. The passenger might be afraid because she knows too little about the driving abilities of the robot; the assembly-line worker might be afraid because she knows all too well the abilities of the robot, which threaten to take her job.

What follows is an effort to unpack some of the different anxieties that drive our algorithmic judgments. This Part is both a descriptive attempt to parse out the distinct motivations and also a normative assessment of those motivations. Many of these explanations for why we fear or distrust robots are just as applicable to human decision-makers, our alternative to robots. When we say that a robot decision-maker is unacceptable because it is inscrutable, we forget that the alternative is a human decision-maker that we have every reason to expect will be just as inscrutable, if not more so. Taken as a whole, what follows are explanations for our judgment of algorithms. However, as we will see, these explanations often fall short as justifications.

189. *Id.* at 2023 ("[S]omething happens when intelligent, experienced, and thoughtful humans are asked to hear reasoned argument and the presentation of proofs to determine how a dispute should be settled.").

190. *See* Brennan-Marquez & Henderson, *supra* note 133. For a counterexample, see Huq, *supra* note 112, at 686 (concluding that the arguments in favor of a right to a human decision-maker are mostly defeated by the technical fact of how modern algorithms operate).

A. Fear of the Unknown

Sometimes we fear robots because we don't know them. When refrigerators were first introduced to the public, there were intense skeptics, despite the obvious public-health advantages over previous food-storage methods.[191] When mechanized tractors first arrived in farmland, advocates of horse-drawn farm equipment launched a massive and popular campaign against the new machines.[192] Similar stories can be told about coffee machines, printing presses, and sound recorders.[193] In short, we are wary of the unknown. And, indeed, with every new technology, there is some combination of too-eager embrace and too-reluctant hesitation.

Reducing aversion to current algorithms may simply be a matter of time and exposure. Many of the examples discussed in this Article, such as artificial intelligence and machine-learning algorithms, are so new that we need more time to understand them. As a recent study of automated recommendation systems explained, "In some cases, it may also be that simply allowing people more experience with recommender systems will increase feelings of understanding over time."[194] This is supported by findings that people working in fields with a longstanding reliance on nonhuman decision-making, such as employees in the financial sector where modeling is an old practice, are less averse to algorithms.[195] So perhaps our fear of machines is merely a temporary condition, one that always trends downward over time. If our fear of algorithms is really a fear of the unknown, then exposure will reduce it.

Yet at times, exposure to robots—even high-performing robots—actually exacerbates our distrust. Dietvorst and

---

191. CALESTOUS JUMA, INNOVATION AND ITS ENEMIES: WHY PEOPLE RESIST NEW TECHNOLOGIES 182–89 (2016).

192. *Id.* at 121–22.

193. *Id.* at 44, 68, 202.

194. Michael Yeomans et al., *Making Sense of Recommendations*, J. BEHAV. DECISION MAKING 1, 10 (2019).

195. *See* Maximilian Germann & Christoph Merkle, *Algorithm Aversion in Financial Investing* (July 2020) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3364850 [https://perma.cc/8CHJ-57UM]. Experimental subjects did not have a strong preference for either humans or algorithms when choosing financial predictions, with the experimenters noting "financial decisions differ from decisions typically studied in the algorithm aversion literature" because they are impersonal and seen as objective. *Id.* at 23.

colleagues showed that people are initially willing to trust algorithms, and they remain so until they see them err; in this case, exposure to the algorithm makes the algorithmic aversion worse, not better.[196] Doctors have known about the advantages of statistical judgment over clinical judgment for years, yet little in the medical field has changed in terms of physician deference to nonhuman deciders.[197] In experimental settings, familiarity with an algorithm can actually *increase* one's likelihood of rejecting an algorithm, principally because more exposure increases the chance that a subject will see the algorithm err.[198] In these cases, it is hard to imagine that what is happening is merely fear or distrust of the unknown. Fear of the unknown cannot explain the bias against the machine.

B. Transparency Concerns

One of the most widely criticized features of algorithmic decision-making is the lack of transparency, which makes algorithms harder to review and challenge.[199] These are understandable concerns. Transparency and reviewability are essential features of due process.[200] But is this criticism convincing? This criticism is primarily leveled in the context of machine-learning algorithms because they are often proprietary and protected as trade secrets.[201] Clearly, the use of a private, nonreviewable criminal justice tool is worrying, but it is hardly the tool's nonhuman nature that is worrying. As Huq points out, "Such secrecy does not plainly distinguish machine from human decisions."[202]

What, then, of the concerns that do not sound in secrecy but instead come from the fact that the algorithm itself might be

196. *See* Dietvorst et al., *Overcoming Aversion*, *supra* note 9.

197. *See* Longoni et al., *supra* note 78.

198. *See* Dietvorst et al., *Overcoming Aversion*, *supra* note 9.

199. *See* Huq, *supra* note 112, at 640. Huq points to two sources: Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018), and PASQUALE, *supra* note 5, at 12–15.

200. *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1297 (2008).

201. *See, e.g.*, Taylor R. Moore, *Trade Secrets and Algorithms as Barriers to Social Justice*, CTR. FOR DEMOCRACY & TECH. (Aug. 2017), https://cdt.org/wp-content/uploads/2017/08/2017-07-31-Trade-Secret-Algorithms-as-Barriers-to-Social-Justice.pdf [https://perma.cc/YK3V-J7R6].

202. Huq, *supra* note 112, at 641.

designed in a way that makes it hard to understand?[203] These concerns suffer two problems. First, it seems that the most common complaints about algorithmic opacity and nonreviewability are compromised or thwarted by technical facts.[204] As a technical matter, there is nothing inherent to machine-learning decisions that makes them impossible to review, and in fact, researchers are making considerable progress interpreting and explaining machine decisions.[205] Indeed, there is a growing literature about designing verifiable and reviewable machine-learning decisions.[206]

Perhaps more importantly, even where algorithmic decisions are inscrutable and there is no novel technology for explaining the decision, that alone is not reason to insist on a human decision-maker. The same criticism can be applied to human decisions, which are often far from transparent as to the author, the audience, and the reasoning.[207] Of course, we would generally prefer legal rulings to be explained and justified in ways that are intelligible and honest. But perfection is not the standard. The standard is human decision-making, and there we regularly tolerate decisions that are explained poorly, deceitfully, or not at all. As Huq's recent survey of this literature suggests, "it cannot be said a priori that [machine-learning decisions] are any more opaque than humans."[208]

C. Loss of Control

One primal fear of robotic decision-making is the fear that robots are not under our control. This goes to both the core of the *Frankenstein* stories and the campaign to ban autonomous robots, among other concerns. We fear that machines have been designed to make decisions that, at some point, may lead to their ability to independently make other types of decisions—ones we

203. *See* Karen Yeung, *Algorithmic Regulation: A Critical Interrogation*, REGUL. & GOVERNANCE 516–17 (2018) (suggesting that algorithms challenge the basic precepts of a liberal society because they are "opaque, inscrutable 'black boxes'").

204. Huq, *supra* note 112, at 640–43.

205. *See* Cary Coglianes & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 49–51 (2019) (summarizing the technical advances in reviewing and verifying algorithms).

206. *Id.*

207. *See* Woods, *supra* note 170.

208. Huq, *supra* note 112, at 640–43.

fear involve harming others or killing their makers. As Huq puts it, "A fearful future looms, one characterized by massive economic dislocation, wherein people have lost control of many central life choices, and basic consumer and political preferences are no longer really one's own."[209]

There is something understandable about this fear. The widespread fear of flying is often driven by a fear of losing control.[210] But one effective treatment for this fear is cognitive therapy that focuses on the fact that getting on the airplane at all was a choice. Passengers typically control their *choice* of travel method, even if they are not in control over its *operation*.[211] Control remains, it is merely shifted.

To be sure, there are examples of automated decision-making systems getting it wrong. One famous example came on September 26, 1983, when an early warning system in a Soviet air-defense bunker near Moscow indicated that an intercontinental ballistic missile was heading from the United States towards the Soviet Union.[212] As the story goes, disaster was averted with intervention by a human operator, Stanislav Petrov, a lieutenant colonel in the Soviet Air Defense Forces, who later said, "I had a funny feeling in my gut."[213] This story is celebrated as a triumph of human intuition over machine error.[214]

Why do we tell these stories? Because we fear handing over control to machines, and these stories confirm that if we release control to the machine entirely, terrible things will happen. Despite these anecdotes lionizing human intuition, the actual data around runaway automated systems are both thin and swamped by the improved decision-making provided by automated systems.[215] We are not assessing the risks rationally but instead trusting our gut—and celebrating when others do too. This is

209. *Id.* at 614–15.

210. *See* Jamie Ducharme, *Why Some People Have a Crippling Fear of Flying — and How They Can Overcome It*, TIME (July 6, 2018, 10:30 AM), https://time.com/5330978/fear-of-flying-aviophobia/ [https://perma.cc/T4EM-VF86].

211. *Id.*

212. *See* Brennen-Marquez & Henderson, *supra* note 133, at 146.

213. Sewell Chan, *Stanislav Petrov, Soviet Officer Who Helped Avert Nuclear War, Is Dead at 77*, N.Y. TIMES (Sept. 18, 2017), https://www.nytimes.com/2017/09/18/world/europe/stanislav-petrov-nuclear-war-dead.html [https://perma.cc/F7BD-SRU8].

214. *See* Brennen-Marquez & Henderson, *supra* note 133, at 146 (describing Petrov as a hero for following his gut).

215. Perhaps the best example of our unfounded fears of automation comes from aviation. *See infra* text accompanying notes 279–280.

consistent with a large body of research that suggests that people's willingness to accept technological risk is governed by factors related not only to the actual risk but also to other characteristics.[216] People are more willing to accept the risks of automation where they feel they can control the machine (which might be true for, say, semi-autonomous weapons systems but not for fully autonomous cars).[217]

D. Job Anxiety

Another explanation for antirobot sentiment is job anxiety—our fear of losing jobs to machines.[218] This is certainly consistent with media coverage of machines in the workplace. Much of this fear stems from a single study out of Oxford that estimated that 47 percent of U.S. jobs are at risk of automation.[219] This study "prompted a myriad of fearful responses in popular media, with articles like 'The AI Revolution Is Coming—And It Will Take Your Job Sooner Than You Think' and 'New Study: Artificial Intelligence Is Coming For Your Job, Millennials.'"[220] Indeed, the fear is so widespread that companies, in an effort to avoid bad press, are reluctant to talk about the use of robotics in

216. *See, e.g.*, Dietvorst et al., *Algorithm Aversion*, *supra* note 9; Lennart Sjöberg, *Factors in Risk Perception*, 20 RISK ANALYSIS 1 (2000); Paul Slovic & Ellen Peters, *Risk Perception and Affect*, 15 CURRENT DIRECTIONS PSYCH. SCI. 322 (2006).

217. *See* PAUL SLOVIC, THE PERCEPTION OF RISK (2000); Dietvorst et al., *Overcoming Aversion*, *supra* note 9; Baruch Fischhoff et al., *How Safe is Safe Enough? A Psychometric Study of Attitudes Towards Technological Risks and Benefits*, 9 POL'Y SCIS. 127 (1978); Harry J. Otway & Detlof von Winterfeldt, *Beyond Acceptable Risk: On the Social Acceptability of Technologies,* 14 POL'Y SCIS. 247 (1982); Paul Slovic, *Perception of Risk*, 236 SCIENCE 280 (1987); Chauncey Starr, *Social Benefit Versus Technological Risk*, 165 SCIENCE 1232 (1969).

218. Indeed, there is strong evidence that people perceive robots as "stealing" their jobs. Armin Granulo, Christoph Fuchs & Stefano Puntoni*, Psychological Reactions to Human Versus Robotic Job Replacement*, 3 NATURE HUM. BEHAV. 1062 (2019).

219. Carl Benedikt Frey & Michael A. Osborne*, The Future of Employment: How Susceptible Are Jobs to Computerisation?*, 114 TECH. FORECASTING & SOC. CHANGE 254, 278 (2017) ("According to our estimate, 47 percent of total US employment is in the high risk category, meaning that associated occupations are potentially automatable over some unspecified number of years, perhaps a decade or two.").

220. Jacky Liang, Ben Ramanauskas & Andrey Kurenkov, *Job Loss Due to AI — How Bad Is It Going to Be?*, SKYNET TODAY (Feb. 4, 2019), https://www.skynet-today.com/editorials/ai-automation-job-loss [https://perma.cc/C9RV-R6YF].

the workplace.[221] The Oxford study was so influential—cited over 4,000 times—yet so often misrepresented that its authors felt the need to clarify what they meant. As they explained in a blog post in 2018, "Our estimates have often been taken to imply an employment apocalypse. Yet that is not what we intended or suggested."[222]

Since the study was released, nearly every major study of the topic has come to the conclusion that computerization will not have such sweeping consequences so soon.[223] The latest wave of AI-powered automation appears unlikely to change a longstanding pattern of technical innovation in the workplace, where machines replace some jobs, augment most jobs, and create new never-before-imagined jobs. A survey of recent studies about the effect of AI on current jobs concluded that the threat of AI to jobs is largely overblown: "Automation will probably displace fewer than 15% of jobs in the near future . . . ."[224] Rather the consensus seems to be that most jobs will be augmented by AI, not replaced.[225]

To be clear, the job anxieties I am addressing here are general worries about the entire economy; the evidence simply does not suggest that machines will bring about mass unemployment. However, as I note in the next Part, even if *overall* job losses will be low, we should be attentive to the distributional consequences of those losses.[226] If self-driving taxis replace human drivers, the overall number of jobs lost may be small by comparison to the entire economy, but those losses will be felt differently by the underprivileged. Moreover, if automation requires workers to adapt and retrain, we can expect that the well-resourced and well-educated will be better situated than others. The idea that

221. Greg Nichols, *Robophobia: 3 Reasons Companies Are Squeamish Talking About Robot Adoption*, ZDNET (May 21, 2019), https://www.zdnet.com/article/robophobia-3-reasons-companies-are-squeamish-talking-about-robot-adoption/ [https://perma.cc/BA2J-G7CJ].

222. Michael Osborne & Carl Benedikt Frey, *Automation and the Future of Work – Understanding the Numbers*, OXFORD MARTIN SCH. (Apr. 13, 2018), https://www.oxfordmartin.ox.ac.uk/blog/automation-and-the-future-of-work-understanding-the-numbers/ [https://perma.cc/P58J-ZLTL].

223. *Id.* ("[O]ne study published by a group of researchers at the University of Mannheim suggests that only 9% of jobs are exposed to automation. And more recently, a study by the OECD suggests that it is actually 14% . . . .").

224. Liang et al., *supra* note 220.

225. *Id.*

226. *See infra* Section IV.A.

the machine age will amplify inequality is a serious concern, even if it is distinct from a general anxiety about mass layoffs.

E. Disgust

Disgust towards robots is another explanation for our mistrust of algorithms. This may seem extreme and even inconsistent with the rise of robots in society, but the more that robots become humanlike, the more they can trigger feelings of disgust. In the 1970s, roboticist Masahiro Mori hypothesized that people would be more willing to accept robots as the machines became more humanlike, but only up to a point, and then human acceptance of nearly-human robots would decline.[227] This decline has been called the "uncanny valley," and it has turned out to be a profound insight about how humans react to nonhuman agents. This means that as robots take the place of humans with increasing frequency—companion robots for the elderly, sex robots for the lonely, doctor robots for the sick—reports of robots' uncanny features will likely increase.

Disgust matters because it can produce judgment errors. Suppose that you find the very best doctor to be physically repulsive. Maybe you dislike their aesthetic appearance for some reason—for example, their race, sex, or something more innocuous like choice of jewelry or clothing. Whatever the reason, despite their qualifications, your disgust is a barrier to accepting their assistance. The same thing happens with robots.

F. Gambling for Perfect Decisions

Why would anyone prefer a human decision-maker to a nonhuman decision-maker if they knew that the algorithm was generally superior? One explanation is that they are gambling for a low-probability-but-high-reward outcome: a perfect decision.[228] A series of studies shows that one motivation for algorithm aversion is that "people choose between decision-making methods on the basis of the perceived likelihood of those methods producing

227. Maya B. Mathur & David B. Reichling, *Navigating a Social World with Robot Partners: A Quantitative Cartography of the Uncanny Valley*, 146 COGNITION 22 (2016) (showing that the "Uncanny Valley" phenomenon is a serious impediment to human-robot social interaction across a range of scenarios).

228. *See* Dietvorst & Bharti, *supra* note 1.

a near-perfect answer."[229] This suggests that while people know an algorithm might be better on average over many decisions, they worry that an algorithm's best decision will not be as good as one made by a human.[230] Put another way, humans may be worse decision-makers on average, but with a human decision, there is a chance of hitting the jackpot and getting a perfect decision.

This relates to the idea of "uniqueness neglect"—the fear that artificial intelligence will not adequately account for the uniqueness of each individual.[231] This is the explanation some social psychologists give for resistance to artificial intelligence in the medical field.[232] The criticisms of algorithmic justice, too, boil down to the claim that machines are not capable of accurately capturing just how unique and distinctive humans are. As John Nay and Katherine J. Strandburg note,

> Critics of automated decision-making raise a number of concerns, but the heart of the argument favoring human adjudicators is a basic skepticism that the "personalization" associated with [machine learning]-based decision tools allows them to generalize as well as human adjudicators to the varied circumstances encountered in real-world cases.[233]

If this is right, then we approach machines with a gambling mindset, ready to trade away the machine's guarantee of a decent result for a chance at a human-driven perfect result. Like other forms of gambling, this is hardly sensible but it is human.

G. Overconfidence in Human Decisions

A related possibility is simply that people prefer humans to anything nonhuman. In law and medicine, some people have strong preferences for a "human touch." What in particular do

229. *Id.*

230. *Id.*

231. *See* Longoni et al., *supra* note 78, at 631. ("[T]he prospect of being cared for by an automated provider evokes a concern that one's unique characteristics, circumstances, and symptoms will be neglected.").

232. *Id.*

233. *See* John Nay & Katherine J. Strandburg, *Generalizability: Machine Learning and Humans-in-the-Loop*, *in* RESEARCH HANDBOOK ON BIG DATA LAW 284, 298 (Roland Vogl ed., 2021).

these patients prefer? The answer is often an intangible quality that cannot be satisfied by a robot, because it is defined as a thing a robot does not have. This might explain the finding that people are especially averse to algorithms when it comes to moral decision-making.[234]

Indeed, for some sorts of robophobia, part of the story is likely overconfidence in our own human abilities.[235] This is particularly true of experts, who are ironically the group of people least likely to trust an algorithm because they are "simply less open to taking any advice."[236] This might explain, for example, why drivers are reluctant to give control over to a robot: we think we are better drivers than we actually are.[237] The same overconfidence in human decision-making abilities might explain a physician's insistence on the "art" of medicine and therefore the rejection of robot doctors.

***

None of these explanations is entirely satisfying. These explanations help to explain why we judge algorithms as we do, but they do not make the costs of our misjudgment any more acceptable.

IV. THE CASE FOR ROBOPHOBIA

But there are good reasons to be wary of algorithms. Before turning to the core normative case for changing how we judge algorithms, we should acknowledge the most compelling reasons we might, despite the costs, decide not to deploy machine decision-makers. Indeed, in deciding where to deploy machines in society, policymakers must identify when to check our biases against machines and when to embrace them.

---

234. *See* Berkeley J. Dietvorst & Daniel M. Bartels, *Consumers Object to Algorithms Making Morally Relevant Tradeoffs Because of Algorithms' Consequentialist Decision Strategies*, J. CONSUMER PSYCH. (2021).

235. *See* Jennifer M. Logg, Julia A. Minson & Don A. Moore, *Algorithmic Appreciation: People Prefer Algorithmic to Human Judgment*, 151 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 90, 91 (2019) (citing the literature describing people's overconfidence in their own judgments, which "raise[s] the question of whether individuals insufficiently trust algorithms (relative to human advisors) or merely overly trust themselves").

236. *Id.* at 99.

237. *See* Amanda N. Stephens & Keis Ohtsuka, *Cognitive Biases in Aggressive Drivers: Does Illusion of Control Drive Us Off the Road?*, 68 PERSONALITY & INDIVIDUAL DIFFERENCES 124 (2014) (showing that both optimism bias and the illusion-of-control bias predict poor driving behaviors).

A. Concerns about Equality

Perhaps the best reason to be wary of machine deciders is that they exacerbate existing distributional problems, even if they make other things better overall. For example, if self-driving cars reduce *overall* fatalities but increase fatalities for a particular subset of the population, we might reasonably decide that this unequal distribution of harms negates the cars' overall benefit. The same could be said of algorithms in the criminal-justice system. Even if such algorithms hold enormous potential to reduce both errors and racial inequities, it is not hard to imagine lawmakers eagerly adopting an algorithm that is "more efficient" or "safer" but has the unintended side effect of amplifying, rather than reducing, racial bias.[238] Inevitably, as new machines are rolled out, there will be benefits and costs, and the analysis of where and when to deploy machines cannot be summed up as a kind of tally of the benefits minus the costs. If the costs are unevenly distributed—and especially if they are particularly bad for groups that have historically been disadvantaged by the criminal justice system—policymakers might reasonably decide that the algorithm is not worth implementing, despite whatever benefits it offers.

In many ways, we are just beginning to understand what role intelligent machines ought to play in society; we are conducting many experiments to see what works well and what does not. Some groups have historically been treated as the subjects of experiments with new technology and not the beneficiaries. Recall the Tuskegee syphilis study, which is just one of many medical experiments that have been conducted at the expense of vulnerable populations.[239] Those experiments had a profound and lasting impact on the trust that Black men place in the American healthcare system.[240] It would be entirely reasonable that people poorly treated by the healthcare system would be wary of future experiments in healthcare, including those involving robots.

---

238. *See* Kleinberg et al., *supra* note 110.

239. *See* Marcella Alsan & Marianne Wanamaker, *Tuskegee and the Health of Black Men*, 133 Q. J. ECON. 407 (2018) (describing the notorious "Tuskegee Study of Untreated Syphilis in the Negro Male" and using a difference-in-differences model to show the study's long-lasting effects on the behavior and health of Black men).

240. *Id.*

Of course, the sentiment might run in the opposite direction—groups that have been historically discriminated against might be *more* willing to use a nonhuman decision-maker if they think it will insulate them from the biases that plague human decision-making. This is an empirical question, and much more work needs to be done in this area.[241]

Ultimately, the distributional consequences of algorithms are worth taking seriously. In a world run by prejudiced human decision-makers, algorithms may be a reason for optimism, as much as they are a reason for skepticism. But that will depend, at least in part, on who develops them and why.

B. The Political Economy of Robots

This raises another good reason to be wary of an efficient or well-tailored algorithm: the political economy in which they are developed. So far, I have described the benefits of better algorithms for individuals and for society. But what about the companies that use and sell them? This is big business.[242] Suppose, for example, that Facebook—a leader in artificial intelligence—developed an algorithm that anyone could deploy on their own devices to enhance the diversity of viewpoints to which they are exposed. Even if it were effective at its task, one could hardly be faulted for distrusting Facebook, which has abused its users' trust before,[243] or for wondering whether the firm had an ulterior motive, such as increasing reliance on the platform.

The core concern here is not that a corporation might figure out how to use algorithms to make money—that is our world—but instead that a powerful and well-resourced company could use algorithms in ways to enhance its dominant position over

241. There is at least some very preliminary evidence that this is the case. *See* Pierson et al., *supra* note 29.

242. GRAND VIEW RSCH., ARTIFICIAL INTELLIGENCE MARKET SIZE, SHARE & TRENDS ANALYSIS REPORT BY SOLUTION, BY TECHNOLOGY (DEEP LEARNING, MACHINE LEARNING, NATURAL LANGUAGE PROCESSING, MACHINE VISION), BY END USE, BY REGION, AND SEGMENT FORECASTS, 2021–2028 (2021), https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market [https://perma.cc/X7FX-B9WD] ("The global artificial intelligence market size was valued at USD 62.35 billion in 2020 and is expected to expand at a compound annual growth rate (CAGR) of 40.2% from 2021 to 2028.").

243. Casey Newton, *Facebook's Trust Problem Isn't About Being Understood*, VERGE (Jan. 31, 2020, 6:00 AM), https://www.theverge.com/interface/2020/1/31/21115104/facebook-mark-zuckerberg-liked-understood-trust [https://perma.cc/B6KN-SN3M].

competitors and users. That is, it is entirely reasonable to resist giving decision-making authority over to an algorithm, even one that improves welfare in the short term, if doing so encourages dependence on the machine and thereby enhances the more powerful to the detriment of the less powerful. This is a related but distinct concern from the distributional problem described above. Even if the algorithm benefits *users* of the algorithm equally, widespread usage might give the *owner* of the algorithm too much power with too little accountability. When the makers and sellers of machines have enormous economic incentive to convince people to embrace those machines, it is sensible to be wary of their widespread adoption.

C. Pro-Machine Bias

Throughout this Article, I have argued that our collective bias against machines is dangerous. But that does not mean that we should have a bias *for* robots. It turns out that sometimes we deliberately prefer robots to humans, which can be a problem. As such, it might make sense to be wary of robots in situations where we know we have a tendency to overrely on them.[244]

Results from a series of experiments show that, at times, people are more willing to follow the advice of an algorithm than the advice of a human.[245] This is consistent with experimental findings in computer science that, in some circumstances, people trust an algorithm more than a person.[246] Despite all of the evidence of mistrust of machines, people also seem to exhibit so-called automation bias—an overconfidence in machine determinations merely because it was determined by a machine.[247] This bias will have greater effect as technology takes over different

244. *See* Germann & Merkle, *supra* note 195 (finding that contrary to algorithm aversion, people prefer algorithms to humans where they think they outperform humans).

245. *See* Logg et al., *supra* note 235.

246. *See* Jaap J. Dijkstra, Wim B.G. Liebrand & Ellen Timminga, *Persuasiveness of Expert Systems*, 17 BEHAV. & INFO. TECH. 155 (1998) (reporting results of an experiment finding that subjects favored advice from "expert systems" over human advisers).

247. *See* Citron, *supra* note 200, at 1271 nn.146–50 (collecting automation bias sources); *see also* Raja Parasuraman & Dietrich H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381 (2010) (explaining that people tend to become complacent when working with an automated system that they routinely trust).

domains of society. As one expert put it, "Automation bias may become increasingly acute in the twenty-first century as our regulatory rules become increasingly intricate."[248]

We show pro-robot biases in other ways too. For example, one study suggests that people are less upset when they find out their job is being taken by a robot and not a human.[249] Some people blame self-driving cars *less* than human drivers for the same crash.[250] Some people cooperate better with robots than they do with humans.[251] Some people worry that law enforcement agents are biased in favor of their algorithmic decision-making aids.[252] And some worry that this bias is "just as likely—if not more likely—to appear in the military context," where operations "occur under greater time pressure than criminal justice decision-making."[253] We even show empathy towards robots in physical pain, albeit less empathy than we show towards other humans.[254]

Pro-robot bias is no better than antirobot bias. If we are inclined *both* to over- and underrely on robots, then we need to correct both problems—the human fear of robots is one piece of the larger puzzle of how robots and humans should coexist. The regulatory challenge vis-à-vis human-robot interactions then is not merely minimizing one problem or the other but rather making a rational assessment of the risks and rewards offered by nonhuman decision-makers. This requires a clear sense of the key variables along which to evaluate decision-makers.

---

248. Linda J. Skitka et al., *Automation Bias and Errors: Are Crews Better than Individuals?*, 10 INT'L J. AVIATION PSYCH. 85, 86 (2000). This may be related to the idea that at times people are biased in favor of quantitative reasoning. Frank Pasquale, *Secret Algorithms Threaten the Rule of Law*, MIT TECH. REV. (June 1, 2017), https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/ [https://perma.cc/37BS-GKRE] ("Judges are all too likely to assume that quantitative methods are superior to ordinary verbal reasoning, and to reduce the task at hand (sentencing) to an application of the quantitative data available about recidivism risk.").

249. Granulo et al., *supra* note 218, at 1062.

250. Edmond Awad et al., *Drivers Are Blamed More than Their Automated Cars When Both Make Mistakes*, 4 NATURE HUM. BEHAV. 134 (2019).

251. Celso M. de Melo, Stacy Marsella & Jonathan Gratch, *Human Cooperation When Acting Through Autonomous Machines*, 116 PROC. NAT'L ACAD. SCIS. 3482 (2019).

252. *See* Pasquale, *supra* note 248.

253. Ashley S. Deeks, *Predicting Enemies*, 104 VA. L. REV. 1529, 1574 (2018).

254. Yutaka Suzuki et al., *Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography*, SCI. REPS. (Nov. 3, 2015), https://www.nature.com/articles/srep15924 [https://perma.cc/CQ7M-B5PJ].

V. THE CASE AGAINST ROBOPHOBIA

We are irrational in our embrace of technology, which is driven more by intuition than reasoned debate. Sensible policy will only come from a thoughtful and deliberate—and perhaps counterintuitive—approach to integrating robots into our society. This is a point about the policymaking process as much as it is about the policies themselves. And at the moment, we are getting it wrong—most especially with the important policy choice of where to transfer control from a human decider to a robot decider.

Specifically, in most domains, we should accept much more risk from algorithms than we currently do. We should assess their performance comparatively—usually by comparing robots to the human decider they would replace—and we should care about rates of improvement. This means we should embrace robot decision-makers whenever they are better than human decision-makers. We should even embrace robot decision-makers when they are less effective than humans, as long as we have a high level of confidence that they will soon become better than humans. Implicit in this framing is a rejection of deontological claims—some would say a "right"—to having humans do certain tasks instead of robots.[255] But, this is not to say that we should prefer robots to humans in general. Indeed, we must be just as vigilant about the risks of irrationally preferring robots over humans, which can be just as harmful.[256]

A. More Than a Preference

It may be tempting to resign ourselves to robophobia, like other biases, as merely a necessary by-product of individual preference: some people *prefer* humans to robots, especially for certain kinds of tasks. Yet we can still calculate the costs of this preference. As we have seen, many people prefer a human doctor to a robot, even when they know the human is less effective.[257] Or they may prefer a human judge to a robot judge, or a human taxi driver to a robot taxi driver. In each scenario, a preference for "the human touch" may come at a cost, usually elevated

255. *See* Huq, *supra* note 112.
256. *See* Logg et al., *supra* note 235, at 100–01.
257. *See* Longoni et al., *supra* note 78.

risk—that is, the risk of being jailed wrongly, treated badly, driven poorly, and so on. At the very least, even if we decide that each individual gets to make this tradeoff as a matter of choice, we should be open about the stakes of that tradeoff.

But also, just as we do with other forms of bias, we should draw a distinction between personal-choice robophobia, choices that affect only the person making the choice, and public-choice robophobia, choices that affect the wider public. In some instances, robophobia is a purely personal choice. You may choose a less-accurate-but-warm human doctor, while I might choose a more-accurate-but-cold robot doctor; the costs of our preferences are mostly internalized (leaving aside insurance pools). If that were the extent of robophobia, it would not be much of a problem. But the reality is that robophobia imposes costs on others, both directly and indirectly.

Indeed, it is hard to imagine a scenario where the preference for robot-deciders does not affect others. Consider an example. Alfred likes human doctors because they have a familiar, warm touch. He is willing to pay more for human care and understands that human doctors have lower success rates than their robot counterparts. Alfred decides that his child should also see a human doctor. He also prefers to drive himself around town because he just does not trust self-driving cars, even though he overestimates his own driving abilities. He gets to work, where he manages a loan portfolio for a bank, and he decides he would rather use his own intuition about the lenders than the algorithm his bank offers. We can see in these examples that Alfred's preferences impose costs on others: his co-insureds, his child, the people he passes in his car, and his bank's stakeholders, among others.

Perhaps these are bad examples because the externalities of an individual choice for a lower-performing human are so obvious. But even in seemingly more difficult examples, an individual's robophobic preferences for human decision-makers carries negative externalities. Consider a more difficult case: the use of robots in one's own trial. Eugene Volokh says that artificial intelligence in the courtroom—especially in the form of brief writers and interpreters—should be held to the same performance standards as humans in those roles.[258] But crucially, Volokh says litigants should be given the choice about algorithms' use:

258. Volokh, s*upra* note 1, at 1140.

an individual litigant could *choose* whether to use a human or AI for their legal representation. This appears to be an example where each litigant's preference for robots or humans as lawyers is internalized. But is it? Suppose that a robot-driven justice process is faster and fairer. Why should society allow people the choice of slower and less fair process? Why should state bar associations allow attorneys to practice law *without* making use of these faster and fairer machines? Why should all of us pay for judges to oversee cases that are slower and less effective merely because one of the litigants has a preference for one sort of representation? You might have an individual preference for a slower and less fair human judge, but it is hard to imagine how that preference will not impose costs on the rest of us.

To be sure, in some domains we can and should preserve individual choice without harming society at large. For example, a patient might opt for a less effective human surgeon and internalize the costs and risks of that choice. We can allow people the autonomy to choose their own medical provider. But that is very different from allowing people to drive on public roads, which imposes a huge risk on other drivers, when the alternative is a safer autonomous vehicle. The costs of robophobia on society are considerably higher in the latter scenario.

B. What Is the Alternative?

As we have seen, we tend to assess algorithmic performance in absolute terms. If a car crashes, then self-driving cars are bad. If a robot doctor errs, then it is unacceptable. We see algorithms err, and our trust evaporates. We often fail to ask the relevant policy question: What is the alternative?

Consider an example. Some people argue against the measles vaccine because it carries some risk of harm.[259] A small portion of children experience flu-like symptoms after vaccination.[260] However, that risk must be weighed against the alternative: the risk of not vaccinating a child against measles. On balance, it is *much* safer to vaccinate a child than to not do so. Choosing not to vaccinate a child is the riskier alternative.

259. *See* Jan Hoffman, *How Anti-Vaccine Sentiment Took Hold in the United States*, N.Y. TIMES (Mar. 26, 2021), https://www.nytimes.com/2019/09/23/health/anti-vaccination-movement-us.html [https://perma.cc/NHN3-UCPA].

260. *Vaccine (Shot) for Measles*, CDC, https://www.cdc.gov/vaccines/parents/diseases/measles.html [https://perma.cc/CLX5-AY4G].

Worse than that, it puts other children in harm's way. Indeed, the rates of measles have gone up in recent years after decades of decline—all because of the so-called anti-vaxxer movement, which is driven by a narrow focus on the risks of vaccination without comparing them to the risks of the alternative of *not* vaccinating children.[261]

In effect, many make the same argument about new machines by asking, Do they have a risk? Instead, we should be asking, How does the risk of using the machine compare to the risk of *not* using the machine? In the context of bail determinations, if we move from human judges to computer algorithms, is there a risk that the algorithm will get it wrong? Yes, absolutely. As one scholar recently put it, "Nowhere is the concern with algorithmic bias more acute than in criminal justice."[262] But the risk of letting a human make the decision is also very high.[263] The relevant question is which is worse? There is convincing evidence that bail-determination algorithms are at least as good as, if not better than, humans.[264] Indeed, evidence shows that the concern about these algorithms has been overstated and that these algorithms can improve decision-making along several variables—for example, by keeping safety levels stable while jailing many fewer people, and by reducing racial biases when determining whom to jail and whom to release on bail.[265]

Consider another example from aviation. Pilots' arguments about fly-by-wire algorithm-driven designs, prior to their widespread adoption, had many of the same flavors of today's anxiety about algorithms. They argued that autopilot programs would introduce new risks that did not exist before[266]—and they were right. Today, there are real risks surrounding pilots who have

261. *See* Hoffman, *supra* note 259.

262. *See*, *e.g.*, Mayson, *supra* note 5, at 2221.

263. *See* David Arnold, Will Dobbie & Crystal S. Yang, *Racial Bias in Bail Decisions*, 133 Q.J. ECON. 1885 (2018) (finding significant racial bias in judges' bail determinations).

264. *See* Mayson, *supra* note 5, at 2225 ("[T]here is every reason to expect that subjective pre- diction entails an equal degree of racial inequality."); Sam Corbett-Davies, Sharad Goel & Sandra Gonzalez-Bailon, *Even Imperfect Algorithms Can Improve the Criminal Justice System*, N.Y. TIMES (Dec. 20, 2017), https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html [https://perma.cc/EY7T-G2LV].

265. *See* Cowgill & Tucker, *supra* note 104, at 35.

266. *See* William Langewiesche, *The Human Factor*, VANITY FAIR, Oct. 2014, at 260–61 (showing that pilots resisted automated flight controls in the 70s for safety reasons).

been lulled into a state of complacency due to automation over-reliance and who do not understand flight technology as well as pilots in past years. Those risks have costs and may be the cause of some of today's largest airplane accidents.[267] But these risks are completely swamped by the gains in safety that the computer-flying revolution has brought. Airplanes now are safer than they have ever been—it is not even close.[268] It would be morally irresponsible to advocate for removing these gains. Yet, in many areas of automated decision-making, that is effectively what we do. We focus on a given technology's risk of harm in a vacuum rather than comparing such risks to the risks of human-based decision-making.

C. Rates of Improvement Matter

Just as there are occasions where it might make sense to hold robots to a higher standard than humans, particularly when we lack information about an algorithm's performance, there are occasions where it makes sense to hold robots to *lower* standards than humans because of their ability to learn. As a class of decision-makers, robots are improving in ability much faster than humans. In just twenty years, robots have learned to translate texts from one language to another, navigate city streets, drive cars, and so on. By contrast, human abilities remain just about where they were twenty years ago.

If comparisons of robot performance to human performance take into account rates of improvement, in many scenarios, it makes sense to embrace robots that currently *underperform* compared to humans, because we can expect them to soon drastically outperform humans. Unlike most human decision-making systems, robots have shown enormous room for improvement.[269] Suppose that self-driving cars are 1.2 times—20 percent—safer than human drivers over all driving conditions. Obviously, keeping these cars off the road is embracing an increased risk of death at the hands of human drivers. What if self-

267. *Id.* (detailing how the combination of human and automated decision-makers led to the 2009 crash of Air France Flight 447, which killed 228 people).

268. *Id.*

269. *See* Jonathan Kay, *How Do You Regulate a Self-Improving Algorithm?*, ATLANTIC (Oct. 25, 2017), https://www.theatlantic.com/technology/archive/2017/10/algorithms-future-of-health-care/543825/ [https://perma.cc/U3EQ-MW7D] (describing how medical algorithms are rapidly advancing in capability).

driving cars were only 80 percent as good as human drivers—20 percent *worse than humans*—but we expected that they could quickly become many multiples better than human drivers with broad deployment? A case could be made for allowing autonomous vehicles, even when their performance is currently below human levels of performance, given the anticipated future benefits. It would be wrong to prohibit such a car from the road because, even though it increases short-term risk, it would considerably lower risk compared to human drivers over the medium and long term.

Not only is robot decision-making improving over time, but it is reviewable. Robots can change—they can be corrected, edited, and educated—in profound ways, whereas humans simply cannot. Put a robot judge on the bench, and if its performance is underwhelming, the robot can be modified. The same cannot be said for a human judge. Give a driver's license or medical license to a human, and it is much harder to monitor their conduct or identify potential risks until after a mistake happens, perhaps at great cost. In contrast, robots are not owed the same kind of privacy nor do we need to worry that their performance degrades under scrutiny.

Of course, if rates of improvement matter, then we must always ask, Over what time period? If a machine is expected to outperform a human in a year, our appetite for mistakes from the machine will be much higher than if we expect the machine to take decades to reach its potential.

D. What Are We Maximizing?

An honest assessment of algorithmic tools requires an honest assessment of our policy goals. When we say that a robot is a "better" decision-maker than a human, what do we mean? In a sense, this is simple: we mean that a robot is a decision-maker that makes fewer errors. But what counts as an error? In self-driving cars, for example, the most commonly discussed metric is safety. But safety is not the only goal. Suppose that self-driving cars were safer than human drivers but they drove at five miles per hour. This would be maddening, and no one would use self-driving cars—no matter their safety record. We want transportation to be both safe and expedient. This is why comparing robot performance to the alternative is so important: it reveals the key variables at stake.

Even seemingly simple concepts like safety have competing and, at times, incompatible definitions. Do we design self-driving cars to minimize fatalities overall or only for their passengers? Surveys show that people generally want autonomous vehicles to aim to reduce overall casualties—including taking steps to protect pedestrians and passengers in other vehicles—but those same people prefer to ride in self-driving cars designed to maximize the safety of the passengers inside.[270]

When comparing two alternatives, there is always a chance that the comparison will flatten and focus too much on a single outcome or variable. Often, the focus is on efficiency or accuracy. But risk of recidivism is not the only variable in bail determinations; discrimination, expediency, and many other values are also essential. Fairness, for example, is key. As Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan note, fairness might be defined several different ways that are "incompatible with each other."[271] A risk score might be defined as "fair" if it identifies White and Black defendants as flight risks at the same rate, if it finds White and Black defendants *not* to be flight risks at the same rate, or if it determines individuals' flight risk independent of their race. These three distinct notions of fairness are mutually exclusive. We can define an algorithm to be "fair" according to one of these definitions but not according to the others. This means, in other words, that algorithmic design forces a policy conversation about what fairness means in bail determinations, what safety means in transportation, and so on. The alternative is to have humans make flight-risk or traffic-safety determinations in an ad hoc, impressionistic manner. If we do not want to be explicit about what policy goal we are maximizing—perhaps because we do not know—then an algorithm is the wrong choice.

***

Our bias against machines is easy to explain but hard to justify. Most of these explanations are driven by intuitions, just like any other kind of judgment error. To be sure, there are some compelling reasons to be wary of algorithms, but none of those

270. *See* Jean-Francois Bonnefon, Azim Shariff & Iyad Rahwan, *The Social Dilemma of Autonomous Vehicles*, 352 SCIENCE 1573 (2016) (showing that people prefer for other cars to be programmed to minimize all casualties, but those same people would prefer to ride in cars programmed to minimize passenger casualties).

271. *See* Kleinberg et al., *supra* note 166, at 43:1.

reasons is a sufficient justification for the kind of society-wide negative reactions to machines we see today.

VI. FIGHTING ROBOPHOBIA

The costs of robophobia are considerable and they are likely to increase as machines become more capable. The greater the difference between human and robot performance, the greater the costs of preferring a human. Unfortunately, the problem of robophobia is itself a barrier to reform. It has been shown in several settings that people do not want government rules mandating robot use.[272] And policymakers in democratic political systems must navigate around—and resist the urge to pander to—people's robophobic intuitions. So, what can be done?

Robophobia is a decision-making bias—a judgment error.[273] Fortunately, we have well-known tools for addressing judgment errors. These include framing effects, exposure, education and training, and, finally, more radical measures like designing situations so that biased decision-makers—human or machine—are kept out of the loop entirely.

A. Switching the Default

One standard debiasing technique changes existing defaults from opt-in to opt-out.[274] The classic example is organ donation.[275] Some people would prefer to donate their organs in the event of an accident, and some people would not. Whether they choose to donate or not appears to depend more on how the question is framed than any personal preference.[276] When the

272. Bonnefon et al., *supra* note 270 (finding that people approve of machines making utilitarian calculations for others but not for themselves and showing strong disapproval of regulations to enforce autonomous car algorithms).

273. This is to distinguish it from identity biases, such as racial or sexual bias. Of course, judgment errors and identity harms can overlap—not hiring doctors of a particular race, for example, would be both a racial bias and a judgment error. For dignitary harms like racial bias, we might prohibit it by law. For judgment errors, we might use softer tools—nudges and education—to debias decision-makers.

274. *See* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 177 (2008).

275. *Id.*

276. Shai Davidai et al., *The Meaning of Default Options for Potential Organ Donors*, 109 PROC. NAT'L ACAD. SCIS. 15201, 15201 (2012) ("This research demonstrates that people's preferences can be dramatically influenced by minor variations in the phrasing of a question . . . .").

default is set to "no organ donation," forcing people to intentionally opt in, people donate organs at drastically lower rates than when the same program is offered with the default set to "organ donation," with the option to opt out.[277] This shows that setting defaults has a powerful effect on people's behavior, and switching from opt-in to opt-out can be a useful tool in designing around judgment errors.

What would this look like in the context of algorithmic decision-making? Currently, our default automatically assumes that humans should do a job unless and until a case has been successfully made for robots to do the work. Humans are the default for many roles—surgeons, judges, taxi drivers, and so on—and we ask whether a robot should instead perform the task. That is, we have an opt-in regime for robot decision-makers in many areas of life.

As an alternative, we could switch the default, with robots assumed to be the right actors for a job unless and until a case can be made for humans to take their place. Suppose instead that we assumed that robots should be surgeons, judges, and taxi drivers unless there was a good reason for them not to be. This may sound fanciful, but it could quickly become a reality with the help of the institutions that design our defaults. Imagine if the Department of Transportation or the local DMV made driver's licenses and taxi medallions automatically available to robot drivers, while human drivers needed to request non-standard licenses. In healthcare, imagine if health insurance providers and HMOs made robotic healthcare the default option where available unless there was a compelling medical reason to use a human. In bail-bond determinations, courts might use algorithms unless there was a compelling due process argument against their use.[278]

To be sure, there very well might be an argument against relying on a machine to perform each of these tasks. If there is a good reason for not using robots in any given setting, let the case be made. The point is not that robots should be doing the jobs of humans but that the dialogue about where and when robots should be deployed is biased. By flipping the default, we harness

277. *Id.* at 15203.

278. And there very well might be, depending on the algorithm's performance and the court's definition of racial fairness. *See* Mayson, *supra* note 5, at 2262 (discussing the different ways algorithms might be optimized to promote equity, none of which would satisfy all critics).

the bias against machines, thereby forcing a conversation about when and where to have robots. Where the merits cash out in favor of using a human and not a robot, it would not be because it *feels* right to use humans but because a case had been made that, on the merits, humans are more effective, fairer, or safer than the default option of a robot.

B. Algorithmic Design

Perhaps we can also design nonhuman systems to address some of the judgment errors described here. For example, machine recommendations are more widely embraced if they are given human characteristics.[279] One study found that "increasing the affective human-likeness of algorithms by providing real examples of algorithms with affective abilities, such as understanding emotion and creating art, can make algorithms seem more effective at performing subjective tasks, which ultimately increases reliance on algorithms for such tasks."[280] Scholars have even suggested that "algorithms pause, as if 'thinking,' before making a recommendation."[281] Making robots more humanlike is an old trick. Another study found that the anthropomorphism of a car predicts trust in the vehicle.[282] That is, participants trusted a self-driving car considerably more when its driving behaviors were anthropomorphized as compared to a self-driving car that was merely trying to drive well but not mimicking human characteristics.[283]

Anthropomorphizing our machines might serve two goals. First, it could encourage people to take as many risks with machines as they do with people. Second, and more importantly, it might also encourage people to think of machines as fallible—to err is human—making them less likely to fall into the trap of automation bias. However, there may be a limit to these anthropomorphic strategies, at least where making machines more

279. Adam Waytz et al., *The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle*, 52 J. EXPERIMENTAL SOC. PSYCH. 113, 116 (2014).

280. Noah Castelo et al., *Task-Dependent Algorithm Aversion*, 56 J. MKTG. RSCH. 809, 811 (2019).

281. Yeomans et al., *supra* note 194, at 412.

282. Waytz et al., *supra* note 279.

283. *Id.*

humanlike triggers the so-called uncanny valley.[284] The most complete study to date shows that people's acceptance of robots increases as the robot becomes more humanlike but only up to a point—and as the robot becomes extremely, even eerily, human, it triggers intense rejection by humans.[285]

Anthropomorphizing algorithms is just one of many design strategies. Another strategy would be for algorithm designers to build in some elements of user control, even if they are minor. It has been shown that people trust algorithms more when they feel they have some control over the algorithm, however slight.[286] Similarly, algorithms could describe their tasks in relatively objective terms, since people's perception of an algorithm's utility is affected by the task the algorithm is assigned and how that task is framed. In short, there are ways we can both design algorithms and frame those design choices to the public that would aid in algorithmic acceptance. Which of these strategies is the most effective will require further study.

C. Education

Perhaps instead of designing robot decisions to track human intuitions, we should decide the best policy and then use education and training to overcome human intuition when it is inconsistent with rational policymaking. Education aimed at a known bias can, at times, counteract it.[287]

In South Korea, a group of researchers developed "Shelly," a tortoise-shelled robot designed to discourage robot abuse in children.[288] When children pet the robotic turtle, it appears

284. Mathur & Reichling, *supra* note 227 (showing that the uncanny valley is a serious impediment to human-robot social interaction across a range of scenarios).

285. *Id.*; *see also supra* Section III.E.

286. Dietvorst et al., *Overcoming Aversion*, *supra* note 9, at 1156.

287. *See* Carey K. Morewedge et al., *Debiasing Decisions: Improved Decision Making with a Single Training Intervention*, 2 POL'Y INSIGHTS FROM BEHAV. & BRAIN SCIS. 129 (2015). Education as a debiasing strategy has generally received insufficient attention from researchers. Baruch Fischoff, *Debiasing*, *in* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES 422 (Daniel Kahneman et al. eds., 1982).

288. Hyunjin Ku et al., *Designing Shelly, a Robot Capable of Assessing and Restraining Children's Robot Abusing Behaviors, in* HRI '18: COMPANION 2018 ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION 161–62, Mar. 5–8, 2018, Chi., Ill., https://dl.acm.org/doi/pdf/10.1145/3173386.3176973 [https://perma.cc/84FW-ARPY].

happy, lighting up and wiggling its arms in delight.[289] But if the robot is hit or kicked, it curls into its shell and stops playing.[290] "At first, we tried to give some feedback that can show that the robot is angry when it gets abused, but we found that those feedbacks can actually foster abuse because children want to see the robot's reaction," noted one of the researchers.[291] The more effective design was to simply have the robot stop playing, which cut robot abuse in half.[292]

Similarly, the general public can also be educated to trust robots. When the media reports on a car accident involving autonomous technology, it would help readers contextualize the crash if the article included a comparison to human-caused accidents over the same time period. When the media reports on a novel technology in healthcare causing a death, it would be helpful to also report the baseline rate of healthcare-related deaths—that is, the rate of healthcare deaths caused by humans in the absence of the technology. Most of us pay too much attention to the news in front of our faces, which often makes the news seem more important than it is in the broader context.[293] This can be mitigated by responsible reporting that provides context for the news.

D. Banning Humans from the Loop

One manifestation of our fear of machines is the now common idea that automated systems must always maintain a "human in the loop."[294] That is, even if robots outperform humans at some tasks, robots can be made even better and safer with human oversight. Either a human can use human-like judgment

289. Katharine Schwab, *Robot Abuse Is Real*, FAST CO. (Mar. 27, 2018), https://www.fastcompany.com/90165541/robot-abuse-is-real [https://perma.cc/MCV7-Q55X].

290. *Id.* ("We concluded that stopping all the interaction for a certain period of time is effective for preventing the robot abuse as children want the robot to keep interacting with them.").

291. *Id.*

292. Ku et al., *supra* note 288, at 162.

293. *See* Amos Tversky & Daniel Kahneman, *Availability: A Heuristic for Judging Frequency and Probability*, 5 COGNITIVE PSYCH. 207 (1973).

294. *See supra* text accompanying notes 183–185. There is some slippage between how this phrase is used in computer science—to explain a machine learning training method—and how the phrase is used among policy advocates. I will focus on the latter, more widespread use of the phrase. Nothing I describe here is a criticism of the idea of human-aided training of machine learning algorithms.

to decide whether to deploy an autonomous system or, in the worst-case scenario of an autonomous system gone rogue, a human can hit an emergency stop button and shut the robot down. This is intuitively appealing, and it speaks to our fear of losing control. But there is considerable evidence in a number of scenarios that keeping humans in the loop eliminates the advantages of having an automated system in the first place and, in some instances, actually makes things worse.

Consider aviation. For a long time, there were essentially two schools of thought in airplane safety. The Airbus approach was to maximize autonomation.[295] The American approach, embodied by Boeing, traditionally emphasized much more human control over airplanes.[296] Airbus planes were traditionally much more automated than their Boeing competitors.[297] But all automated systems in Airbus planes also have a human override, and this specific combination of automated and human systems contributed to the deadly crash of Air France Flight 447 in 2009.[298] There is a risk that automation, which is designed to improve upon human performance, actually "worsens human performance, which begets increasing automation."[299] That is, there is evidence that the introduction of some autonomy actually increases human reliance on the automation, which decreases overall safety.[300]

Developers of autonomous cars worry about the same thing.[301] The National Highway Traffic Safety Administration

295. *See* Huq, *supra* note 112, at 615, 621.

296. Alexander Ibsen, *The Politics of Airplane Production: The Emergence of Two Technological Frames in the Competition Between Boeing and Airbus*, 31 TECH. SOCIETY 342 (2009) (describing two very different regulatory and business approaches to aviation safety).

297. *See* DIGITAL AVIONICS HANDBOOK 224 (Cary R. Spitzer et al. eds., 3rd ed. 2019) (comparing Airbus's approach with Boeing's).

298. Langewiesche, *supra* note 266, at 258 (describing the investigation of the crash and the unique role that ultra-safe automated systems played in the accident).

299. *Id.* at 295.

300. *See* Mica R. Endsley, *Automation and Situational Awareness, in* AUTOMATION AND HUMAN PERFORMANCE: THEORY AND APPLICATIONS 163 (R. Parasuraman & M. Mouloua eds., 1996).

301. *See* Zach Lovering, *Why Direct to Autonomy*, ACUBED (Aug. 2, 2018), https://acubed.airbus.com/blog/vahana/why-direct-to-autonomy/ [https://perma.cc/4NSK-AMCV]; John Markoff, *Google's Next Phase in Driverless Cars: No Steering Wheel or Brake Pedals*, N.Y. TIMES (May 27, 2014), https://www.nytimes.com/2014/05/28/technology/googles-next-phase-in-driverless-cars-no-brakes-or-steering-wheel.html [https://perma.cc/U2AW-TS9Z].

recognizes six levels of automotive autonomy, ranging from 0 (no automation) to 3 (conditional automation) to 5 (full automation).[302] Some people believe that a fully autonomous system is safer than a human driver, but that a semi-autonomous system—where a human driver works with the autonomous system—is actually less safe than a system that is purely human driven.[303] That is, autonomy can increase safety, but the increase in safety is not linear; introducing some forms of autonomy can introduce new risks.[304]

In terms of safety, then, there are scenarios where the safety rating of different levels of autonomy might be listed as follows:

> Full autonomy > no autonomy > partial autonomy

Put another way:

> Robot only > human only > human and robot

The danger that we will misuse partially autonomous systems is highlighted in criminal law. As Megan Stevenson notes, "The policy-relevant question is not 'Is the actuarial tool better at predicting misconduct than the judge' but rather 'Does the judge make better decisions when given access to actuarial predictions?'"[305] Looking at how judges use algorithmic guidance in pretrial release decisions, she found that most judges ignored or overruled the algorithmic guidance.[306] Rather than using the algorithm to enhance their decision-making, "[j]udges may ignore

302. Automated Vehicles for Safety, NAT'L HIGHWAY TRAFFIC SAFETY ADMIN., https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety#topic-road-self-driving [https://perma.cc/6GVK-GFGT]. This is based on the Society of Automotive Engineers taxonomy. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles,* Standard J3016_201806, SAE INT'L (June 15, 2018), https://www.sae.org/standards/content/j3016_201806/ [https://perma.cc/84V8-4V2G].

303. Hod Lipson & Melba Kurman, *Your Robot Car Should Ignore You*, NAUTILUS (May 11, 2017), https://nautil.us/issue/48/chaos/your-robot-car-should-ignore-you [https://perma.cc/5VC3-KL2J] (describing how partial autonomy lulled drivers into dangerous levels of inattentiveness).

304. Kathleen Walch, *Are All Levels of Autonomous Vehicles Equally Safe?*, FORBES (Dec. 8, 2019 1:00 AM), https://www.forbes.com/sites/cognitiveworld/2019/12/08/how-autonomous-vehicles-fit-into-our-ai-enabled-future/#7e3cde7f5df9 [https://perma.cc/3VKA-S5RS].

305. Megan Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 303, 370 (2018).

306. *Id.*

the risk tool in cases where it is correct, or place too much credence on it when it is incorrect," which might eliminate any gains from the algorithm or even make decision-making worse.[307] As others note, this study, read in the context of other examinations of judicial use of algorithms, "suggest(s) a role for limiting judicial discretion."[308]

We might worry about the same thing in the military setting. Imagine if soldiers could decide when and where to deploy an automated weapons system. If the robot is designed to avoid human judgment errors—perhaps firing a weapon out of rage—then giving humans the ability to override robot judgment may undermine those benefits and could even make things worse.

If algorithms can, at times, make better decisions than humans, but *human use* of those algorithms eliminates those gains, what should be done? One answer is to ban semi-autonomous systems altogether; human-robot interaction effects are no longer a problem if humans and robots are not allowed to interact. Another possibility would be to ban humans from some decision-making processes; a purely robotic system would not have the same negative human-robot interaction effects. This might mean fewer automated systems but would only leave those with full autonomy.

If humans misjudge algorithms—by both over- and underrelying on them—can they safely coexist? Take again the example of self-driving cars. If robot-driven cars are safer than human-driven cars but human-driven cars become less safe around robot cars, what should be done? Robots can simultaneously make the problem of road safety better *and* worse. They might shift the distribution of road harms from one set of drivers to another. Or it might be that having some number of robot drivers in a sea of human drivers is actually less safe for all drivers than a system with no robot drivers. The problem is the interaction effect. In response, we might aim to improve robots to work better with humans or improve humans to work better with robots. Alternatively, we might simply decide there are places where human-robot combinations are too risky and instead opt for purely human or purely machine decision-making.

---

307. *Id.* at 334.

308. Cowgill & Tucker, *supra* note 104, at 34 (discussing Stevenson, *supra* note 305, and comparing it to Kleinberg et al., *supra* note 110).

CONCLUSION

One of the most important political decisions of our time is deciding when and where to delegate decision-making authority to machines. Much of the legal scholarship on the topic has focused on the ways in which machines might be biased. Too little legal scholarship has been dedicated to the opposite problem: human misjudgment of machines. The evidence for our deep and widespread judgment errors is overwhelming. This is reflected in our laws and policies, often at enormous cost.

In this Article, I explored relatively standard approaches to what is essentially a judgment error. If our policymaking is biased, the first step is to remove the bias from existing rules and policies. The second step might be to inoculate society against the bias—through education and other debiasing strategies. A third and even stronger step might be to design situations so that the bias is not allowed to operate. For example, if people tend to choose poorer performing human doctors over better performing robot alternatives, a strong regulatory response would be to simply eliminate the choice. Should humans simply be banned from some kinds of jobs? Should robots be required? These are serious questions. If they sound absurd, it is because our conversation about the appropriate role for machines in society is inflected with a fear of and bias against machines.