NEGLIGENCE AND AI'S HUMAN USERS

ANDREW D. SELBST*

ABSTRACT

Negligence law is often asked to adapt to new technologies. So it is with artificial intelligence ("AI"). Though AI often conjures images of autonomous robots, especially autonomous vehicles, most existing AI technologies are not autonomous. Rather, they are decision-assistance tools that aim to improve on the inefficiency, arbitrariness, and bias of human decisions. Decision-assistance tools are frequently used in contexts in which negligence law or negligence analogues operate, including medicine, financial advice, data security, and driving (in partially autonomous vehicles). Users of these tools interact with AI as they would any other form of technological development—by incorporating it into their existing decision-making practices. Accordingly, it is important to understand how the use of these tools affects the duties of care required by negligence law and people's ability to meet them.

This Article takes up that discussion, arguing that AI poses serious challenges for negligence law's ability to continue compensating the injured. By inserting a layer of inscrutable, unintuitive, and statistically derived code in between a human decisionmaker and the consequences of her decisions, AI disrupts our typical understanding of responsibility for choices gone wrong. This Article argues that AI's unique nature introduces four complications into negligence: 1) the inability to predict and account for AI errors; 2) physical or cognitive capacity limitations at the interface where humans interact with AI; 3) the

* Assistant Professor, UCLA School of Law. For extraordinarily helpful insights and comments on earlier drafts, I would like to thank Sareeta Amrute, LaToya Baldwin-Clark, Jack Balkin, Rabia Belt, Kiel Brennan-Marquez, Ryan Calo, Rebecca Crootof, Patrick Davison, Blake Emerson, Kadija Ferryman, Mark Grady, James Grimmelmann, Jill Horwitz, Sonia Katyal, Mark Latonero, Christina Mulligan, Aiha Nguyen, Ted Parson, Sunita Patel, Nicholson Price, Richard Re, Alex Rosenblat, Alex Wang, Rebecca Wexler, and participants at the Yale Information Society Project Ideas Lunch and Fellows Writing Workshop; the NYU Privacy Research Group; and the 2018 Privacy Law Scholars' Conference. Thanks as well to the editors of the *Boston University Law Review* for their excellent and professional work getting this Article ready for publication.

 This work was funded in part by the National Science Foundation (IIS-1633400).

introduction of AI-specific software vulnerabilities into decisions not previously mediated by software; and 4) distributional concerns based on AI's statistical nature and potential for bias. In those contexts where we rely on current negligence law to compensate for injuries, AI's use will likely result in injured plaintiffs regularly losing out, as errors cease being the fault of the operator and become statistical certainties embedded within the technology. With most new technologies, negligence law adapts over time as courts gain familiarity with the technology's proper use. But the unique nature of AI suggests that this may not occur without legislation requiring AI to be built interpretably and transparently, at a minimum, and that other avenues of regulation may be better suited to preventing uncompensated losses by injured parties.

CONTENTS

INTRODUCTION

As with any new technology, once artificial intelligence ("AI") has been adopted widely, there will be injuries, and some will result in lawsuits. Medical AI will recommend improper treatment, robo-advisers will wipe out someone's bank account, and autonomous robots will kill or maim. And just as with any new technology, negligence law will be called on to adapt and respond to the new threat.[1] With most new technologies, we gain familiarity over time, eventually creating a sense of what constitutes reasonable care or a collective intuition on which negligence law can rely as it adapts. But AI may be different. Unlike many technologies before, AI poses challenges for negligence law that may delay the common law's ability to adapt or even prevent adaptation outright.

The large and growing body of scholarship on AI and tort law has mostly set aside discussions of negligence.[2] There is good reason for this. Tort law is most centrally concerned with physical injury, and prior research has focused on robots. Robots are essentially a large, heavy, moving form of embodied AI that can cause severe physical harm if left unchecked.[3] One of the most exciting and doctrinally interesting types of robots in development is the autonomous vehicle, which will likely save countless lives if it becomes commonplace. Prior scholarship has therefore focused heavily on autonomous vehicles, giving rise to two central themes. The first is that by automating the driving task, liability for car accidents will move away from negligence on the driver's part toward product liability for the manufacturer. Because there is no person to *be* negligent, there is no need to analyze negligence, and scholars instead move straight to analyzing product liability's own doctrinal infirmities in the face of AI. The second theme is more policy oriented than doctrinal: a concern that the prospect of tort damages may impede the innovation needed to get autonomous vehicles on the road. Both of these concerns relate specifically to autonomous vehicles and neither calls for an analysis of negligence.

But this is not the full picture of AI. What is missing from this prior research is the recognition that autonomous robots are merely a small subset of AI

---

[1] *See* Mark F. Grady, *Why Are People Negligent? Technology, Nondurable Precautions, and the Medical Malpractice Explosion*, 82 NW. U. L. REV. 293, 293 (1988).

[2] Notable exceptions are Weston Kowert, Note, *The Foreseeability of Human–Artificial Intelligence Interactions*, 96 TEX. L. REV. 181, 183-85 (2017) (analyzing negligence in AI for software developers rather than users) and William D. Smart, Cindy M. Grimm & Woodrow Hartzog, An Education Theory of Fault for Autonomous Systems 4 (Aug. 29, 2017) (unpublished manuscript), http://people.oregonstate.edu/~smartw/papers.php?q=papers&display=detail&tag=werobot2017 [https://perma.cc/85SF-EHMQ] (arguing that "failures in the creation and deployment of unpredictable systems lie in the lack of communication, clarity, and education between the procurer, developer, and users of automated systems").

[3] *See* Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 533-45 (2015).

technologies. Much more common is what can be called "decision-assistance" AI: technology that operates by making a recommendation to a user. Decision-assistance AI is rapidly proliferating to every facet of our lives. Some of the uses are not regulated by tort law—such as employment,[4] lending,[5] retail,[6] policing,[7] and agriculture[8]—but other common uses occur where negligence law (or a negligence analogue) regulates, such as medicine, finance, and data security.[9] Even in the driving context, no fully autonomous vehicle is currently being sold, and partially autonomous vehicles can be seen as decision-assistance technologies.

Though similar in some respects to autonomous robots, AI decision-assistance tools are different enough that they could almost be seen as an entirely different category of technology. Instead of seeking to replicate human capabilities such as driving, they often seek to go beyond human capabilities, recognizing and modeling patterns too complex for humans to process and making decisions in ways humans would not recognize.[10] And instead of operating with the push of a button, human decisionmakers look to them for information and ability enhancement in tasks that they were doing prior to AI assistance. Despite also being based on machine learning techniques, these user-centered technologies differ in fundamental ways that demand a different set of legal analyses.

Recognizing that decision-assistance technologies require users to have an effect situates the relevant tort law conversation in negligence rather than products liability. If a doctor relies on a tool to help her decide to inject a drug or release a patient, we still analyze the case in malpractice despite a tool being involved; we expect the doctor to understand her tools enough to satisfy her duty of care while using them. The same goes for any other user in a context where

[4] Rudina Seseri, *How AI Is Changing the Game for Recruiting*, FORBES (Jan. 29, 2018, 10:34 AM), https://www.forbes.com/sites/valleyvoices/2018/01/29/how-ai-is-changing-the-game-for-recruiting/ [https://perma.cc/9VXK-4YLF].

[5] Breana Patel, *What Role Can Machine Learning and AI Play in Banking and Lending?*, FORBES (Oct. 5, 2018, 9:00 AM), https://www.forbes.com/sites/forbesfinancecouncil/2018/10/05/what-role-can-machine-learning-and-ai-play-in-banking-and-lending/ [https://perma.cc/DZ9M-7QQ-M].

[6] *See* ALEXANDRA MATEESCU & MADELEINE CLARE ELISH, AI IN CONTEXT: THE LABOR OF INTEGRATING NEW TECHNOLOGIES 34-37 (2019).

[7] Matt Burgess, *AI Is Invading UK Policing, but There's Little Proof It's Useful*, WIRED (Sept. 21, 2018), https://www.wired.co.uk/article/police-artificial-intelligence-rusi-report [https://perma.cc/H4GA-87SW].

[8] MATEESCU & ELISH, *supra* note 6, at 18-20 (discussing current disconnect between AI farming technologies and farmers' current resources and methods).

[9] *See infra* Section II.A.

[10] Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1089-99 (2018) [hereinafter Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*].

negligence applies: if a driver cannot operate a car, we do not assume that the manufacturer is to blame. Whether negligence or products liability is the correct framework for compensating injuries is therefore not a technicality but rather a reflection of what we fundamentally envision AI to be. If it is a robot making decisions by "itself," then only products liability matters; if AI is more like a fancy calculator, then we use a negligence analysis.

The normative significance of this depends on the purposes of negligence law. From an accident-prevention perspective,[11] it may not make a significant practical difference whether the governing regime is negligence, products liability, strict liability, or insurance—the torts are instrumental, and society should use whatever tools it can to optimally prevent accidents. But from a corrective justice[12] or civil recourse[13] perspective, it matters whether individuals who are injured or wronged can receive redress, and it matters who pays. A shift to products liability permits actors in negligence-governed contexts to disclaim their duties of care by buying a fancier computer and making the manufacturer liable. If negligence is about responsibility, then we should not so readily accept this blame shifting. This Article starts from the premises that AI today is primarily a tool and that, ideally, negligence law would continue to hold AI's users to a duty of reasonable care even while using the new tool.[14]

[11] *See generally* GUIDO CALABRESI, THE COSTS OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS (1970) (arguing that proper goal of tort law is to achieve efficient balance of cost-saving mechanisms and accident avoidance); WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF TORT LAW (1987) (hypothesizing that common-law tort liability arose because of judges seeking to promote efficient resource allocation); STEVEN SHAVELL, ECONOMIC ANALYSIS OF ACCIDENT LAW (1987); Guido Calabresi & A. Douglas Melamed, *Property Rules, Liability Rules, and Inalienability: One View of the Cathedral*, 85 HARV. L. REV. 1089 (1972).

[12] *See generally* JULES L. COLEMAN, RISKS AND WRONGS (1992) (conceptualizing tort law as corrective justice mechanism); ERNEST J. WEINRIB, THE IDEA OF PRIVATE LAW (1995) (examining tort law as relationship between injured-plaintiff and injurer-defendant); Richard A. Epstein, *A Theory of Strict Liability*, 2 J. LEGAL STUD. 151 (1973) (analyzing tort law and strict liability through individual-responsibility lens rather than moral or economic lens); George P. Fletcher, *Fairness and Utility in Tort Theory*, 85 HARV. L. REV. 537 (1972) (discussing relation between demands of individual and maximizing utility in context of tort liability); Stephen R. Perry, *The Moral Foundations of Tort Law*, 77 IOWA L. REV. 449 (1992) (examining tort liability in context of moral reparations).

[13] John C.P. Goldberg & Benjamin C. Zipursky, *Torts as Wrongs*, 88 TEX. L. REV. 917, 918 (2010).

[14] From a normative perspective, my argument aligns more with fault-based theories than economic theories of tort law because my focus is compensation for individual injuries, not harm prevention. Throughout this Article, however, I try to write in terms that draw from both fault-based and economic theories. I do this because, purposes aside, at the level of abstraction I am most often interested in—examining the operation of negligence as a whole—the mechanics of the argument work equally well in both frames. Ultimately, whether the rationale for negligence is derived from fault or harm reduction, liability only attaches where

This Article therefore takes up the question of whether negligence law can successfully adapt to AI. Irrespective of its ultimate purposes, the operation of negligence law is fundamentally concerned with how people make decisions. A fault-based theorist would say that a tight causal nexus between people's decisions and their outcomes is fundamental to the fair assignment of liability, while an economic theorist would argue that the goals of tort law lie in optimal deterrence or efficient risk allocation. Both would only assign liability where a person is capable of acting in a way that can avoid accidents. No theory of negligence will assign liability where a tortfeasor could not have prevented the harm through greater care.

Decision-assistance AI therefore creates tensions with negligence liability at a fundamental level. The technology is designed to interfere with human decision-making; it replaces or augments human decision processes with inscrutable, unintuitive, statistically derived, and often secret code. AI is sold on the premise that human decision-making is not to be trusted. This is the crux of the challenge. Because AI targets human decision-making directly, negligence law appears particularly unsuited to addressing its potential harms in a way that is not shared by earlier technologies. The concern is that while AI may successfully reduce the overall number of injuries, it will not eliminate them, but it *will* eliminate the ability of the people injured in the new regime to recover in negligence. Accordingly, tort law will likely need assistance in the form of interpretability and transparency requirements for the technology, or we will need to look to other methods to reduce the occurrence of accidents or compensate the victims of AI-caused injuries.

The Article will proceed in three parts. Part I reviews the research on torts and AI and describes the major debates so far. It reveals two general themes: that autonomous vehicles will drive liability regimes away from negligence toward products liability and that the uncertainty of tort damages might interfere with the innovation necessary to get autonomous vehicles on the road. This Part ultimately explains why AI's impact on negligence law has not yet been addressed.

Part II turns to negligence. First, it explains why the AI in autonomous vehicles is a special, narrow case of AI that can be reliably overseen by humans. Then, drawing on examples such as medical AI, robo-advisers, data security AI, and partially autonomous vehicles, it argues that AI creates four new challenges for negligence law: (1) Decision-assistance AI tools often aim to find patterns that are beyond human recognition, often making it difficult to distinguish errors from success and rendering harm from AI errors functionally unforeseeable; (2) The average person's physical and mental abilities are limited in ways that are exposed by interaction with machines, with potentially harmful results; (3) AI introduces operational security concerns into decisions that were not

injury is avoidable through action, and there is an individual component to the determination at trial that is changed by the use of statistically driven AI systems. The difference between the theories goes to whether this is something we should care about.

previously mediated by software; and (4) By substituting statistical reasoning for individualized reasoning and by changing its output based on prospective plaintiffs, AI creates new openings for discriminatory results to enter individual tort cases.

Part III offers observations about how AI's effects interact with certain aspects of the structure and operation of negligence law. First, the practical function of foreseeability in duty, breach, and proximate cause is to limit liability, at least in part because of bounded rationality. We cannot possibly be held responsible for the endlessly rippling effects of our actions because we cannot appreciate and account for them. But AI decision-assistance tools are seen as beneficial precisely because they can exceed the limits of bounded rationality, finding patterns that humans cannot. Stated in those terms, the result that AI errors will be unforeseeable is almost tautological. But this poses a challenge because where the unforeseeable error is the rule, not the exception, negligence law ceases to function. Second, as a common-law regime, negligence would typically adapt to new technologies. Time and experience with the new sociotechnical environments allow us to update standards of reasonable behavior. While that may be possible with AI, there are elements of the AI landscape—such as intense corporate secrecy, the contextual nature of AI, and the speed of AI development—which may prevent legal standards from developing fast enough without outside intervention. Third, the algorithmic bias problem is representative of a larger difficulty of negotiating statistical facts in an area of individual responsibility, for which negligence law has no good answer. This discussion will draw on prior work in algorithmic discrimination, where this is a familiar problem, and demonstrate that any attempt to solve AI problems with individual fault rules may be difficult.

I. TORTS AND THE CREATION OF AI

A large and growing body of scholarship is being written on AI and tort law. Most of this work is about autonomous robots, especially vehicles. This makes sense. Tort law is most centrally concerned with physical injuries, and robots can frequently be large, heavy, moving objects that have the capacity to cause severe physical harm. The scholarship has two central themes that are a direct result of this focus. The first is that due to automation, liability for injuries will move away from negligence toward products liability. The scholarship mostly discusses whether products liability faces new challenges as a result of AI. The second is a concern that the prospect of tort damages may hamper innovation. Both of these concerns relate mostly to autonomous vehicles, focusing on AI's creation rather than its use. This Part briefly reviews the tort and AI literature to date.

A. *Autonomous Vehicles, Products Liability, and Innovation*

Human error is responsible for the vast majority of car accidents.[15] As a result, the ability of autonomous vehicles to separate humans from driving responsibilities is an extremely important achievement. Tort scholars consider this safety enhancement to be the primary benefit of automating driving. There is broad consensus that autonomous vehicles are likely to change the liability calculus, shifting liability to the manufacturers.[16] For some scholars, this is the core of the argument, and for others it is a premise.[17]

One focus of scholarship is the interesting products liability question: how to decide whether certain accidents amount to defects. Courts find manufacturers and sellers of products liable for one of three kinds of product defects: manufacturing defects, design defects, and failures to warn. Manufacturing defects are errors in production—instances where the product differs from the blueprint.[18] The exploding soda bottle is the canonical example.[19] Manufacturing defects lead to strict liability for the manufacturer. Design defects are instead judged by one of two tests: the risk-utility test[20] or the consumer expectations test.[21] The risk-utility test is a cost-benefit analysis that holds a product defective when a "reasonable alternative design" exists, the omission of which "renders the product not reasonably safe."[22] The consumer

[15] NAT'L HIGHWAY TRAFFIC SAFETY ADMIN., U.S. DEP'T OF TRANSP., FEDERAL AUTOMATED VEHICLES POLICY: ACCELERATING THE NEXT REVOLUTION IN ROADWAY SAFETY 5 (2016), http://www.safetyresearch.net/Library/Federal_Automated_Vehicles_Policy.pdf [https://perma.cc/5RYZ-SDQZ] ("94 percent of crashes can be tied to a human choice or error.").

[16] Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 CALIF. L. REV. 1611, 1619 (2017); Dorothy J. Glancy, Robert W. Peterson & Kyle F. Graham, *A Look at the Legal Environment for Driverless Vehicles*, LEGAL RES. DIG., Feb. 2016, at 1, 35-36.

[17] *See* Curtis E.A. Karnow, *The Application of Traditional Tort Theory to Embodied Machine Intelligence*, *in* ROBOT LAW 51, 57-58 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016); Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 VA. L. REV. 127, 134 (2019); Bryant Walker Smith, *Automated Driving and Product Liability*, 2017 MICH. ST. L. REV. 1, 6.

[18] RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2 (AM. LAW INST. 1998).

[19] Escola v. Coca Cola Bottling Co., 150 P.2d 436 (Cal. 1944); RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2 cmt. c, illus. 1.

[20] RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2.

[21] RESTATEMENT (SECOND) OF TORTS § 402A (AM. LAW INST. 1965).

[22] RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(b); Stephen G. Gilles, *The Invisible Hand Formula*, 80 VA. L. REV. 1015, 1047 (1994); David G. Owen, *Design Defects*, 73 MO. L. REV. 291, 315 (2008) ("The Hand defectiveness formula succinctly captures the commonsense idea that a product's design is unacceptably dangerous if it contains a danger that might cost-effectively (and practicably) be removed.").

expectations test defines a defect as a condition that is "dangerous to an extent beyond that which would be contemplated by the ordinary consumer."[23] Both tests aim to address the tradeoff between safety and the cost necessary to find every possible imperfection, and the differences between them may be overstated.[24] Failures to warn employ a similar cost-benefit analysis that asks if the missing warning renders the product unreasonably unsafe.[25]

One of the questions for autonomous vehicles is how to classify a defect.[26] The consequence of classifying an error that leads to a car crash as a manufacturing defect, design defect, or warning defect is stark: A manufacturing defect leads to strict liability and the others receive reasonableness or cost-benefit analyses. A crash also may not be the result of a design defect at all. To prove a design defect, the plaintiff is required to demonstrate that the accident was proximately caused by a decision that the AI made that should have been anticipated and tested for;[27] such a showing seems quite difficult, both conceptually and as a matter of proof.[28] Autonomous vehicles will face unexpected changes: detours from road construction, drivers who break traffic laws or stop very suddenly, or other drivers misapprehending what the automated vehicle itself will do and reacting badly.[29] Each of these will be unique in some way—the timing, the type of stimulus—such that the machine cannot possibly be trained on all of them. Yet the machine will be asked to

[23] RESTATEMENT (SECOND) OF TORTS § 402A cmt. i.

[24] *See generally* MARK A. GEISTFELD, PRODUCTS LIABILITY LAW 69-116 (2012) (analyzing debate surrounding consumer expectations and risk-utility tests).

[25] RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2.

[26] *See* Karnow, *supra* note 17, at 69 (arguing that classification as design defect or manufacturing defect does not make sense for self-learning vehicles that are incomplete off assembly line); Abraham & Rabin, *supra* note 17, at 140-44 (questioning whether idea of "defect" fundamentally makes sense in new regime); Geistfeld, *supra* note 16, at 1633 (arguing that software generally copies with fidelity, so bugs should always be considered design defects); F. Patrick Hubbard, *"Sophisticated Robots": Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 1803, 1854 (2014). Warning defects are not central to the classification discussion because most of the analysis removes the driver—the very person who is supposed to receive the warning.

[27] Note that a failure to avoid the reasonably avoidable crashes is essentially the definition of a design defect for an autonomous vehicle. For this reason, the proximate cause question and defect question are essentially the same. *See* David A. Fischer, *Products Liability—Proximate Cause, Intervening Cause, and Duty*, 52 MO. L. REV. 547, 559-60 (1987).

[28] *See* Kyle Graham, *Of Frightened Horses and Autonomous Vehicles: Tort Law and Its Assimilation of Innovations*, 52 SANTA CLARA L. REV. 1241, 1270 (2012) (arguing that plaintiffs may be prevented from recovering because doing so would require expensive and difficult review of vehicle computer code).

[29] *See generally* Harry Surden & Mary-Anne Williams, *Technological Opacity, Predictability, and Self-Driving Cars*, 38 CARDOZO L. REV. 121, 150-163 (2016) (discussing difficulties that arise because we lack a "theory of mind" about autonomous vehicles).

dynamically handle all of these scenarios. Some scholars have argued that the manufacturer will often lose the cost-benefit argument when, in hindsight, the cost of testing just one more scenario is marginal, and the damage that results is loss of life and limb.[30] But as they have also noted, the test addresses what the programmer could reasonably have known to test for *before* the crash.[31] It would be unreasonable to rely on hindsight to declare that out of the infinitely many possible fact patterns, the one that led to a crash should have been specifically anticipated.[32] To do so would be functionally no different than strict liability for any crash caused by the car, which a court would be unlikely to impose.[33] The reason that calling this a design defect is conceptually more difficult than in a typical product is that the very purpose of autonomous vehicles is to anticipate and respond to unknown scenarios, resulting in no stable sense of what the AI working properly looks like.

Despite this challenge being well understood, it is highly unstable and fact dependent, rendering it unresolvable in the abstract.[34] This has led scholars to propose a number of solutions to augment products liability, including strict

[30] *See* Hubbard, *supra* note 26, at 1854; Gary E. Marchant & Rachel A. Lindor, *The Coming Collision Between Autonomous Vehicles and the Liability System*, 52 SANTA CLARA L. REV. 1321, 1334 (2012).

[31] Hubbard, *supra* note 26, at 1855.

[32] *See id.* at 1854-55 ("[W]ith more than 100 million lines of software code in a modern automobile, it is unclear whether plaintiffs should be able to rely solely on the existence of the error and of a way to fix the error available at the time of trial but not necessarily reasonably available at the time of sale. Arguably, expert testimony of reasonably attainable error elimination at the time of design and sale should also be required."(footnote omitted)); Smart, Grimm & Hartzog, *supra* note 2, at 3.

[33] Professor Mark Geistfeld has argued that if the crash is due to a bug in the code, the manufacturer could be liable under the malfunction doctrine, which applies to "situations in which a product fails to perform its manifestly intended function." Geistfeld, *supra* note 16, at 1634 (quoting RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB., § 3 cmt. b). But it is mathematically impossible to test for every bug in a computer model, s*ee* Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 31 (2017) (discussing the halting problem, which asks "whether all problem statements which have answers also have the property that those answers can be computed algorithmically"), so unless we want to apply strict liability for bugs, it is unclear how the malfunction doctrine should apply to software. Ryan J. Duplechin, *The Emerging Intersection of Products Liability, Cybersecurity, and Autonomous Vehicles*, 85 TENN. L. REV. 803, 825-26 (2018).

[34] *See* Smith, *supra* note 17, at 32.

liability,[35] no-fault insurance,[36] *respondeat superior* applied to autonomous robots,[37] new legislation delineating fault,[38] finding vehicles not defective where aggregate data shows that a car is at least twice as safe as human drivers,[39] and reinvigorating crashworthiness doctrine.[40] A minority of scholars argue that the law will work as it currently stands.[41] The proper response to the uncertainty surrounding liability is the chief debate in the literature on tort law and AI.

A second theme in scholarly work on tort law and AI is innovation. Because autonomous vehicles are seen as a product that will save lives, as is often the case with new technology, there is concern about whether the prospect of uncertain tort liability will hinder innovation. Many articles have called for legal modifications to protect manufacturers;[42] others are more optimistic about the present balance between tort law and innovation, concluding that traditional tort law will adapt adequately to protect the industry.[43] As Professor Bryant Walker Smith has noted, the literature often treats the question of liability as "an obstacle to be removed, the object of consternation rather than contemplation."[44]

---

[35] Sophia H. Duffy & Jamie Patrick Hopkins, *Sit, Stay, Drive: The Future of Autonomous Car Liability*, 16 SMU SCI. & TECH. L. REV. 453, 471-73 (2013); David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 146 (2014).

[36] Kevin Funkhouser, *Paving the Road Ahead: Autonomous Vehicles, Products Liability, and the Need for a New Approach*, 2013 UTAH L. REV. 437, 458-62.

[37] *See generally* SAMIR CHOPRA & LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS 119-91 (2011).

[38] Jeffrey K. Gurney, *Sue My Car Not Me: Products Liability and Accidents Involving Autonomous Vehicles*, 2013 U. ILL. J.L. TECH. & POL'Y 247, 276-77.

[39] Geistfeld, *supra* note 16, at 1653.

[40] Bryan H. Choi, *Crashworthy Code*, 94 WASH. L. REV. 39, 115-16 (2019).

[41] Hubbard, *supra* note 26, at 1865-66; Smith, *supra* note 17, at 2.

[42] Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1, 44-45 (2018) (advocating for creation of incentives for manufacturers to create safer technologies and benefit society in long term); Funkhouser, *supra* note 36, at 458-62 (advocating for no-fault scheme that will ease manufacturer concerns about liability and encourage technological development); Gurney, *supra* note 38, at 277 (advocating for legislation that will provide clarity on potential liability for manufacturers); Marchant & Lindor, *supra* note 30, at 1339-40.

[43] Geistfeld, *supra* note 16, at 1692 (asserting that hardware and software malfunctions will be subject to strict liability, while products liability law and consumer warnings can address other potential suits); Graham, *supra* note 28, at 1270; Hubbard, *supra* note 26, at 1865-66 (arguing that proposals to fundamentally change how liability works with respect to autonomous vehicles—in favor of either plaintiffs or defendants—inappropriately assume that something is wrong with current balance); Smith, *supra* note 17, at 2.

[44] Smith, *supra* note 17, at 2.

B. *Sidelining Users*

Little of the existing scholarship examines how negligence doctrine will treat harms that result from the use of AI systems. The research discussed above is concerned primarily—almost exclusively—with fully automated vehicles. But there is an important difference between partly and fully automated vehicles. Today's "autonomous" vehicles require a human driver—usually called a "safety driver"[45]—to perform a range of driving tasks. The National Highway and Traffic Safety Administration ("NHTSA") has adopted a classification system based on different levels of autonomy.[46] At Level 0, the car is fully manual, though it may include "intermittent warning systems like blind-spot detection."[47] Level 1 incorporates a single automated aspect, such as steering and acceleration, and familiar technologies, such as "parking assist, which only controls steering, or adaptive cruise control (ACC) that only adjusts speed."[48] Level 2 includes systems that combine steering and acceleration.[49] As automobile industry reporter Jonathon Ramsey has explained, "[u]nder all of these level definitions, the driver is still charged with monitoring the environment."[50]

The first level that can be called automated driving in any meaningful sense is Level 3, in which vehicles monitor the entire environment and "can make informed decisions for themselves such as overtaking slower moving vehicles. However, unlike the higher rated autonomous vehicles, human override is required when the machine is unable to execute the task at hand or the system fails."[51] At Levels 4 and 5, no driver input is required.[52] Level 4 vehicles can intervene and self correct if something goes wrong.[53] The only limitation of Level 4 is that it only applies in particular driving contexts, such as highways.[54] That restriction is lifted in Level 5, where a vehicle is expected to be able to do

---

[45] Dana Hull, Mark Bergen & Gabrielle Coppola, *Uber Crash Highlights Odd Job: Autonomous Vehicle Safety Driver*, BLOOMBERG (Mar. 23, 2018, 4:00 AM), https://www.bloomberg.com/news/articles/2018-03-23/uber-crash-highlights-odd-job-autonomous-vehicle-safety-driver.

[46] NAT'L HIGHWAY TRAFFIC SAFETY ADMIN., *supra* note 15, at 9.

[47] Jonathon Ramsey, *The Way We Talk About Autonomy Is a Lie, and That's Dangerous*, DRIVE (Mar. 8, 2017), http://www.thedrive.com/tech/7324/the-way-we-talk-about-autonomy-is-a-lie-and-thats-dangerous [https://perma.cc/7NUB-3QQL].

[48] *Id.*

[49] *Id.*

[50] *Id.*

[51] Jonathan Dyble, *Understanding SAE Automated Driving – Levels 0 to 5 Explained*, GIGABIT (Apr. 23, 2018, 11:42 AM), https://www.technologymagazine.com/ai/understanding-sae-automated-driving-levels-0-5-explained [https://perma.cc/3VLH-XT7U].

[52] *Id.*

[53] *Id.*

[54] *Id.*

everything a human can perform, including, for example, off-roading.[55] There are currently no Level 4 or 5 cars on the market.[56]

While most of the scholarship recognizes that there is a difference between partly and fully autonomous vehicles, the manufacturers are the central characters in the legal analysis, and the users—the drivers—barely register.[57] Because the literature is about the move toward products liability or the concerns about innovation, the focus on the creation of AI makes complete sense. But tort claims arising from the use of AI, as opposed to their creation, will be subject to a negligence analysis rather than a products liability one.

Drivers occasionally appear in the discussions. One cannot discuss warning defects without addressing the drivers to whom the warnings are directed.[58] Drivers also serve as a yardstick to measure how much liability should be imposed on the manufacturers—whether the manufacturer should be wholly or only partially responsible—while assuming that the actual negligence analysis remains unchanged.[59] Just as often, however, the scholarship will eliminate the driver entirely, discussing AI as something deserving of agency or personhood,[60] or proposing doctrinal changes to apply negligence or ascribe reasonableness to a computer.[61]

Some scholars acknowledge the potential for injuries caused by negligent drivers in passing. Professor Gary Marchant and Dr. Rachel Lindor note that if the user ignores the manual's warnings about limiting the vehicle's use in certain weather or the driver fails to operate autonomous mode appropriately, he may be found negligent.[62] They also argue that most of the time the driver is "unlikely

---

[55] *Id.*

[56] Ramsey, *supra* note 47.

[57] *See, e.g.*, Marchant & Lindor, *supra* note 30, at 1326 ("Autonomous vehicles are likely to change the dynamics of who may be held liable. In considering these changes, it is first necessary to distinguish partial autonomous vehicles from completely autonomous vehicles.").

[58] Gurney, *supra* note 38, at 264.

[59] *See, e.g.*, Marchant & Lindor, *supra* note 30, at 1326 ("These partial autonomous systems will shift some, but not all, of the responsibility for accident avoidance from the driver to the vehicle, presumably reducing the risk of accidents (since that is the very purpose of the system)."); *see also* Duffy & Hopkins, *supra* note 35, at 457 ("Driver liability is relatively straightforward and requires little explanation: driver is liable for his own actions in causing an accident, such as negligent or reckless operation of the vehicle."). Even the argument raised by Jeffrey Gurney—who makes four versions of drivers ("Distracted," "Diminished Capabilities," "Disabled," and "Attentive") the centerpiece of his argument—focuses entirely on products liability. Gurney, *supra* note 38, at 257-71.

[60] CHOPRA & WHITE, *supra* note 37, at 153-91; Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231, 1232 (1992).

[61] Abbott, *supra* note 42, at 22-24; Karni Chagal-Feferkorn, *The Reasonable Algorithm*, 2018 U. ILL. J.L. TECH. & POL'Y 111, 115.

[62] Marchant & Lindor, *supra* note 30, at 1327.

to be a factor."[63] Professor Ignacio Cofone refers to negligent supervision as a possibility, analogizing AI to a child.[64] Professor Patrick Hubbard briefly notes that the reasonable use of a sophisticated robot may require special skill.[65] He argues that "in order to satisfy the standard of reasonable care, users of driverless cars would need to use the skills necessary to operate the car reasonably, by, for example, knowing when the driving system was malfunctioning and, to some extent, how to respond to the malfunction."[66] This is the most detailed analysis of negligent driving in an automated vehicle to date.

The one area where negligence for AI use has been discussed in limited fashion is the medical AI context.[67] Medical AI is the other most common form of AI that can result in physical injuries. So far, the scholarship treats the issue as one of malpractice specifically, rather than negligence more generally. In the next Part, drawing on examples from the medical context as well as a few others, I examine challenges that the use of AI generally poses for negligence law.

II. HOW AI CHALLENGES NEGLIGENCE

Outside of the realm of autonomous vehicles, AI today is most commonly seen as a tool to help people make decisions. Most of its uses—in employment, credit, criminal justice—if regulated at all, are not in the purview of traditional tort law. But AI is reaching into every aspect of society, and it should not be surprising that it has also entered several domains that are subject to negligence

[63] *Id.*

[64] Ignacio N. Cofone, *Servers and Waiters: What Matters in the Law of A.I.*, 21 STAN. TECH. L. REV. 167, 191 (2018) ("[T]he driving software has no agency, so its programmers have a more direct relationship with its actions than do parents with those of their children.").

[65] Hubbard, *supra* note 26, at 1861 ("Where the tort system continues to use traditional fault approaches to address the control, use, and service of robots, the application of concepts like reasonable care will change where increasingly sophisticated robots are involved because the legal system measures the level of skill reasonably required by the nature of the activity undertaken.").

[66] *Id.*

[67] *See e.g.*, A. Michael Froomkin, Ian Kerr & Joelle Pineau, *When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning*, 61 ARIZ. L. REV. 33, 61 (2019); Philipp Hacker, Ralf Krestel, Stefan Grundmann & Felix Naumann, *Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges*, ARTIFICIAL INTELLIGENCE & L., Jan. 2020, § 1, § 3.1.2; W. Nicholson Price II, *Medical Malpractice and Black-Box Medicine*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS 295, 300-01 (I. Glenn Cohen, Holly Fernandez Lynch, Effy Vayena & Urs Gasser eds., 2018) [hereinafter Price, *Medical Malpractice and Black-Box Medicine*] ("[A] trained provider should be subject to the exact same standard of negligence irrespective of whether clinical decision-support software is used because any treatment decisions are ultimately his or her own."); Jeffrey M. Senger & Patrick O'Leary, *Big Data and Human Medical Judgment: Regulating Next-Generation Clinical Decision Support*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS, *supra*, at 283, 293-94.

or negligence analogues, including medical malpractice,[68] data security,[69] investment advice,[70] and car accidents in partially autonomous vehicles.[71]

It is therefore important to understand how tort law views *users* of AI, not just its *creators*. Negligence asks whether a person has violated her duty of reasonable care, and if a person could not have reasonably prevented an accident with better decision-making, she will not be held liable. The centrality of decision-making is what makes liability for AI-assisted decisions tricky. While a decision-assistance technology cannot harm people directly, it can significantly interfere with decision processes. AI inserts into decision-making a layer of complex, often inscrutable, computation that substitutes statistics for individualized reasoning and often discovers unintuitive relationships on which to base the decisions.[72] Thus, the troubling question for negligence law is how the insertion of AI changes the decision-making process and whether those changes fundamentally alter the ability of tort law to achieve its compensatory or regulatory goals.

In this Part, I explore the consequences for negligence liability of how users interact with AI. I identify four challenges. The first is epistemic in nature. AI often aims to go beyond human comprehension, and is often likened to an "alien" intelligence.[73] The different way that AI organizes and processes information often makes error detection challenging or impossible in the moment and specific errors unforeseeable. The second challenge concerns limitations on human capacity. The physical and mental abilities of the average person, such as reaction time or persistent attention, are limited in ways that may produce harmful results when the person interacts with machines. The third is about security. AI will introduce operational security concerns into decisions that were not previously mediated by software. Software vulnerabilities are something that negligence doctrine has never addressed well, and AI expands their reach into new contexts. The fourth challenge is distributional. By substituting statistical reasoning for individualized reasoning and by changing its output based on prospective plaintiffs, AI creates openings for algorithmic bias to enter individual cases in a manner that negligence doctrine is not set up to address.

---

[68] *See* Price, *Medical Malpractice and Black-Box Medicine*, *supra* note 67, at 300.

[69] *See* William McGeveran, *The Duty of Data Security*, 103 MINN. L. REV. 1135, 1196 (2019); Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 643 (2014).

[70] *See* Seth E. Lipner & Lisa A. Catalano, *The Tort of Giving Negligent Investment Advice*, 39 U. MEM. L. REV. 663, 668 (2009).

[71] Hull, Bergen & Coppola, *supra* note 45.

[72] Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1089-99.

[73] *See, e.g.*, David Weinberger, *Our Machines Now Have Knowledge We'll Never Understand*, WIRED (Apr. 18, 2017, 8:22 PM), https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/ [https://perma.cc/E5DG-NZ97].

As far as negligence doctrine is concerned, these challenges mostly apply to the breach element. This is because decision-assistance technologies are additions to contexts that already exist: We have preexisting negligence duties in the cases of medical malpractice, financial advice, data security, and driving. What changes is how people's decisions are made in these contexts once AI is introduced and actions ensue. These are breach questions; questions about the scope of the relevant duties, what actions constitute reasonable care, and what consequences are unforeseeable in the new context. The exception is the distributional question, which I argue negligence is ill-suited to address because of a lack of duty to ensure fair outcomes from a distributional perspective.

A. *Unforeseeability of AI Errors*

The goal of negligence law is to determine who should bear responsibility, if anyone, for accidents. Because AI will not prevent all accidents, the promise of AI is to reduce—not eradicate—errors. Thus, when AI is used, there will still be some errors that result in harm. If negligence law works as intended, those harmed will become plaintiffs who can recover in court if the harm was caused by a breach in the user's duty to them. The concept of breaching a duty of care is only coherent, however, if there is some level of care that a person can adhere to that would have prevented the harm. Thus, the ability to determine ahead of time what constitutes reasonable care and a breach thereof is central to negligence liability.

The requirement to take reasonable care applies equally well when the actions taken involve machines as when they do not. Typically, proper use of a machine or a tool is embedded within the idea of a duty of care. The requirement that a person act reasonably does not depend on whether that action is taken with or without the assistance of technology. This general notion is why Hubbard could argue that "in order to satisfy the standard of reasonable care, users of driverless cars would need to use the skills necessary to operate the car reasonably, by, for example, knowing when the driving system was malfunctioning."[74] Users of tools have as much duty as anyone else to act reasonably.

The first challenge AI poses to negligence law is that AI may be a tool where the possibility of using it without error is not responsive to the level of care taken. To understand this claim, a deeper analysis of what AI is and how it works becomes important. In general, we can think of two categories of AI. One type aims to replicate human capabilities, and one aims to exceed them. Autonomous vehicles are an example of AI that typically replicates human capabilities, while decision-assistance systems often try surpass human understanding. As a result, users may often be unable to determine in real time whether the AI is making an error. In those cases, it will often be unclear how a user can satisfy any duty of care in the operation of the AI. No matter which specific standard of care is used

[74] Hubbard, *supra* note 26, at 1861.

in the breach determination, it may be impossible for the AI user to know which side of the line she is on. In many of these applications, expecting such knowledge is unreasonable, as it may be impossible.[75]

This problem can be analogized to a lack of foreseeability—in this case, a claim that specific AI errors are unforeseeable. Foreseeability is a central component of all legal liability. It is a basic principle of tort law that "a defendant is responsible for and only for such harm as he could reasonably have foreseen and prevented."[76] Though the actual doctrine is "a vexing, crisscrossed morass" that is impossible to pin down,[77] it is still conceptually central to the moral

[75] Though products liability is not the focus of this Article, it is worth noting that this distinction is also—and perhaps more obviously—important for product testing. *See, e.g.*, W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419, 440 (2015) [hereinafter Price, *Black-Box Medicine*] ("The third challenge in developing black-box medicine is validation; that is, making sure that the algorithmic models developed by firms are accurate and useful."). It affects the ability of manufacturers to claim that they took reasonable measures to ensure safety, which is an essential component. *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2 cmts. m-n. While this is true of all new technologies to an extent, AI presents some challenges over and above traditional technologies. Whereas with normal machines one can take them apart, test their parts, and examine the mechanical diagrams to understand how the machine should work, AI is rarely decomposable. *See* Zachary C. Lipton, *The Mythos of Model Interpretability*, 2016 PROC. ICML WORKSHOP ON HUM. INTERPRETABILITY MACHINE LEARNING 96, 98-99 (discussing simulatability, decomposability, and algorithmic transparency). AI's results are often otherwise uninterpretable or based on nonintuitive relationships that are difficult for humans to evaluate normatively. *See generally* Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1117-29. Though testing is challenging, there is a lively area of research on interpretability and/or "explainability" within the field of computer science. *See id.* at 1109-17. Practitioners are thinking through risk analyses where explanation is not possible. *See* ANDREW BURT, STUART SHIRRELL, BRENDA LEONG & XIANGNONG (GEORGE) WANG, BEYOND EXPLAINABILITY: A PRACTICAL GUIDE TO MANAGING RISK IN MACHINE LEARNING MODELS (2018), https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf [https://perma.cc/ EVM7-UAK9] (describing explainability as opacity—or lack thereof—of machine learning model). Products liability has encountered products before the makers of which do not completely understand how they work, so this issue may come to a resolution. *See* David G. Owen, *Bending Nature, Bending Law*, 62 FLA. L. REV. 569, 574-80 (2010) [hereinafter Owen, *Bending Nature, Bending Law*]. The most commonly cited instance of this is drugs. *See* Lars Noah, *This Is Your Products Liability Restatement on Drugs*, 74 BROOK. L. REV. 839, 842 (2009); *cf.* Owen, *Bending Nature, Bending Law*, *supra*, at 574 ("More recently, scientists have changed the genetic makeup of food, cloned animals, spliced genes, dispersed cell phones to all corners of the globe, developed new drugs, and have begun to manipulate the atomic construct of everyday products through the marvels of nanotechnology.").

[76] H.L.A. HART & TONY HONORÉ, CAUSATION IN THE LAW 255 (2d ed. 1985).

[77] W. Jonathan Cardi, *Purging Foreseeability: The New Vision of Duty and Judicial Power in the Proposed Restatement (Third) of Torts*, 58 VAND. L. REV. 739, 740 (2005) [hereinafter Cardi, *Purging Foreseeability*].

underpinnings of tort.[78] While the unforeseeable nature of AI errors does not track current notions of foreseeability in negligence doctrine, thinking of this problem as a new type of foreseeability will be useful for understanding its relationship to negligence, as explained below.

1. Two Types of AI

AI systems can be divided into types based on two different high-level goals. One is to find hidden patterns in order to predict relationships that humans cannot predict in an unaided fashion. The other is to replicate human capabilities, but faster, more reliably, and machine-readably. The tasks of autonomous vehicles are an example of the latter. The primary AI in autonomous vehicles is a machine vision system.[79] While it is often supplemented by a broader range of signals than the visual spectrum, potentially including LIDAR, radar, or ultrasonic sensors, it fundamentally seeks to replicate the function of human-vision systems.[80] If a machine vision system is shown a picture of a dog, a bus, or a crosswalk, it will either correctly identify the dog, bus, or crosswalk, or it will not. The result of this approach is the capacity for human oversight. A human can check the machine because "dog," "bus," and "crosswalk" are categories that humans understand and can differentiate from a background image easily.[81] (This is why Google's reCAPTCHA service presents so many pictures of objects on roads; we are collectively training Google's self-driving AI.)[82] The same goes for the act of driving. The machine is attempting to replicate a human activity—driving—and does so by avoiding the same kinds of objects that humans are attempting to avoid but doing it better. If the car hits something, it is clearly an error to anyone watching.

This is not a universal property of machine learning models. Perceptual tasks such as classifying images and spoken language are actually atypical. More often, the primary benefit of an AI system is to learn to do or see things in ways

[78] David G. Owen, *Figuring Foreseeability*, 44 WAKE FOREST L. REV. 1277, 1277-78 (2009) [hereinafter Owen, *Figuring Foreseeability*].

[79] *See generally* Benjamin Ranft & Christoph Stiller, *The Role of Machine Vision for Intelligent Vehicles*, 1 IEEE TRANSACTIONS ON INTELLIGENT VEHICLES 8, 8-9 (2016).

[80] *Id.* at 8.

[81] A well-known example of AI failing to accurately differentiate an image from its background is an AI attempting to differentiate between wolves and huskies. *See* Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1123-24 (citing Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, PROC. 22ND ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 1135, 1142-43 (2016)). Because the AI relied on the background image rather than characteristics of the animals, it produced erroneous results. *See id.*

[82] *See* Michael Lotkowski, *You Are Building a Self Driving AI Without Even Knowing About It*, HACKERNOON (Feb. 25, 2017), https://hackernoon.com/you-are-building-a-self-driving-ai-without-even-knowing-about-it-62fadbfa5fdf [https://perma.cc/VYH6-FFJ5].

humans cannot.[83] A classic example that illustrates AI's strangeness is spam filtering. AI systems learn by example. If a computer is shown many examples of a phenomenon, it can learn the characteristics of different examples that are labeled as corresponding to different outcomes. It would be nearly impossible for a person to try and write out rules for word choice, tone, grammar errors, and other properties that constitute "spam," but by flagging every spam email (and by presuming all others are not spam), we provide labels to a machine so that it can find these patterns that predict likely spam.[84] These rules may not be a perfect definition, and people may not even agree on the total set of rules that would be such a perfect definition, but with enough data, the machine can create a good approximation. But, as Professor Jenna Burrell has pointed out, whereas humans would likely categorize spam in terms of topics—"the phishing scam, the Nigerian 419 email, the Viagra sales pitch"—computers use a "bag of words" approach based on the appearance of certain words with certain frequencies gleaned by seeing millions upon millions of labeled examples of spam.[85]

Even if humans could theoretically write down a long list of rules to define spam, this is not the way we would approach the problem. Consequently, even understanding the automatically generated rules or why they look as they do is difficult. Computer scientists refer to this phenomenon as the "interpretability" problem.[86] Asking for an explanation of how the system works will often invite a reply of "that's what the data says" or a breakdown of which words with which frequencies contribute to the end result.[87] But as Burrell puts it, this is "at best incomplete and at worst false reassurance" because it does not really tell us anything actionable.[88]

With this background, consider AI in three contexts: medicine, finance, and data security. In medicine, AI is increasingly being used to predict things that

[83] *See, e.g.*, ED FELTEN, AI 101: AN OPINIONATED COMPUTER SCIENTIST'S VIEW 14 (2018) (PowerPoint), https://law.duke.edu/sites/default/files/centers/cip/ai-in-admin-state_felten_slides.pdf [https://perma.cc/N7T4-A3VS] ("AI's errors won't be like human errors."); Ed Felten, Professor, Princeton Univ., The 2018 Grafstein Lecture in Communications: Guardians, Job Stealers, Bureaucrats, or Robot Overlords, at 36:32 (Feb. 8, 2018), https://youtu.be/DuQLeZ9Fr4U?t=2177 [https://perma.cc/3G29-MBLL] ("[M]achine mistakes and human mistakes are just very different, and it's indicative of differences in how machines versus people think. So AI errors won't be like human errors."); Weinberger, *supra* note 73.

[84] Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y, Jan.-June 2016, at 1, 7-9.

[85] *Id.* at 9.

[86] *See* Lipton, *supra* note 75, at 39.

[87] Burrell, *supra* note 84, at 9.

[88] *Id.* (footnote omitted).

even well-trained humans (otherwise known as doctors) cannot.[89] Early uses of AI in medicine were aimed at identifying high- and low-risk patients in different contexts.[90] Today, people are developing AI tools to diagnose patients or recommend treatment.[91] Some scholars predict that these tools will become generally more accurate than doctors.[92] Medical diagnostic and treatment tools seek to find and take advantage of patterns that humans would not otherwise recognize. Of course, there is great risk here; a misdiagnosis or mistreatment can be fatal. These risks may not be particularly rare either; with attempts to use

[89] *See* Katie Chockley & Ezekiel Emanuel, *The End of Radiology? Three Threats to the Future Practice of Radiology*, 13 J. AM. C. RADIOLOGY 1415, 1417-19 (2016); W. Nicholson Price II, *Artificial Intelligence in Health Care Applications and Legal Issues*, 14 SCITECH LAW. 10, 10 (2017); Monique Brouillette, *Deep Learning Is a Black Box, but Health Care Won't Mind*, MIT TECH. REV. (Apr. 27, 2017), https://www.technologyreview.com/s/604271/deep-learning-is-a-black-box-but-health-care-wont-mind/.

[90] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm & Noémie Elhadad, *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission*, PROC. 21ST ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 1721, 1721 (2015); I. Glenn Cohen, Ruben Amarasingham, Anand Shah, Bin Xie & Bernard Lo, *The Legal and Ethical Concerns that Arise from Using Complex Predictive Analytics in Health Care*, 33 HEALTH AFF. 1139, 1140 (2014).

[91] Jane R. Bambauer, *Dr. Robot*, 51 U.C. DAVIS L. REV. 383, 387 (2017) (discussing 23andMe and IBM's Watson for Oncology project as examples of AI medicine); Jeffrey De Fauw et al., *Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease*, 24 NATURE MED. 1342, 1342-50 (2018); Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon & A. Aldo Faisal, *The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care*, 24 NATURE MED. 1716, 1716-20 (2018) (discussing use of AI to suggest treatment oprtions for patients diagnosed with sepsis); Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis & Dimitrios I. Fotiadis, *Machine Learning Applications in Cancer Prognosis and Prediction*, 13 COMPUTATIONAL & STRUCTURAL BIOTECHNOLOGY J. 8, 12-16 (2015) (surveying success rate of machine learning applications in cancer treatment); Price, *Black-Box Medicine*, *supra* note 75, at 426; W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421, 425-26 (2017) [hereinafter Price, *Regulating Black-Box Medicine*] ("There are two types of algorithms involved in the process of using relationships in medical data to drive treatment. We might term the first a research algorithm—it is the process by which data are analyzed and relationships are discovered. The second we might call a prediction algorithm—it is the process by which relationships are applied to new data to generate predictions, recommendations, and the like."); Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi & Nadeem Qureshi, *Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?*, PLOS ONE, April 2017, at 1, 1.

[92] *See,e.g.*, Froomkin, Kerr & Pineau, *supra* note 67, at 46; Senger & O'Leary, *supra* note 67, at 291.

IBM's Watson for cancer diagnosis and treatment, it took just fourteen months to move from extreme hype to extreme disappointment.[93]

In finance, a robo-advisor is an automated or semi-automated service that offers advice about investments, insurance, or credit.[94] Most robo-advisors aim to help people without large sums of money automatically build an investment portfolio and rebalance it as needed.[95] Additionally, it is well known that most people who actively trade in the stock market lose money because the stock market is so inherently unpredictable and humans trade emotionally. This seems like a good use case for AI.[96] The model is well tested—machine learning techniques to predict markets have been around since at least the early 2000s[97] and are now used by the majority of hedge funds.[98] But of course, errors are

[93] *Compare* Mallory Locklear, *IBM's Watson Is Really Good at Creating Cancer Treatment Plans*, ENGADGET (June 1, 2017), https://www.engadget.com/2017/06/01/ibm-watson-cancer-treatment-plans/ [https://perma.cc/F4N7-XQ23] ("New data presented this week at the American Society of Clinical Oncology's annual meeting show that IBM's Watson for Oncology suggests cancer treatments that are often in-line with what physicians recommend."), *with* Angela Chen, *IBM's Watson Gave Unsafe Recommendations for Treating Cancer*, VERGE (July 26, 2018, 4:29 PM), https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science [https://perma.cc/D23H-FQKQ] ("[A]ccording to IBM documents dated from last summer, [Watson] has frequently given bad advice, like when it suggested a cancer patient with severe bleeding be given a drug that could cause the bleeding to worsen.").

[94] Tom Baker & Benedict Dellaert, *Regulating Robo Advice Across the Financial Services Industry*, 103 IOWA L. REV. 713, 719-20 (2018).

[95] U.S. SEC. & EXCH. COMM'N, DIV. OF INV. MGMT., GUIDANCE UPDATE: NO. 2017-02 (2017).

[96] Ayn de Jesus, *Robo-Advisors and Artificial Intelligence – Comparing 5 Current Apps*, EMERJ (Nov. 24, 2019), https://emerj.com/ai-application-comparisons/robo-advisors-artificial-intelligence-comparing-5-current-apps/ [https://perma.cc/85W8-BBEK].

[97] *See* Paul D. Yoo, Maria H. Kim & Tony Jan, *Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation*, 2 PROC. INT'L CONF. ON COMPUTATIONAL INTELLIGENCE FOR MODELING CONTROL & AUTOMATION & INT'L CONF. ON INTELLIGENT AGENTS WEB TECHS. & INTERNET COM. 835, 835 (2005) ("As the Internet provides a primary source of event information which has a significant impact on stock markets, the techniques to extract and use information to support decision making have become a critical task."); Vatsal H. Shah, Machine Learning Techniques for Stock Prediction 2 (2007) (unpublished manuscript), https://bigquant.com/community/uploads/default/original/1X/5c6d3b9959a8556a533a58e0ac4568dfc63d6ff4.pdf [https://perma.cc/5FA6-7SKR].

[98] Amy Whyte, *More Hedge Funds Using AI, Machine Learning*, INSTITUTIONAL INV. (July 19, 2018), https://www.institutionalinvestor.com/article/b194hm1kjbvd37/More-Hedge-Funds-Using-AI-Machine-Learning [https://perma.cc/RM7W-ZYJA].

possible and they may lose all of someone's money because of bad data, a bug, or a runaway feedback loop.[99]

Data security offers one more example. AI security tools operate in one of two ways.[100] In the first, AI learns to predict malicious code based on existing examples, improving traditional antivirus software.[101] Previously, antivirus software looked for specific, identifiable pieces of code that functioned as markers for malware. This type of detection can be fooled by small permutations that do not affect the overall structure of the malware. Machine learning allows for a smarter, basic anti-virus software that will recognize families of malware.[102] The second type of AI security tool analyzes typical network traffic patterns, then detects and flags anomalies.[103] Unlike medicine and even stock picking, this is something difficult to imagine humans even attempting without the aid of computers, as there is no preexisting concept that relates to normal network traffic.[104]

Machine learning systems often face a particularly difficult hurdle that other products do not: The models make predictions based on the data that they are given, but that data may not reflect reality well.[105] Data is necessarily reductive; only certain things can be measured, measurements have limited precision, and the very act of deciding how to characterize and order reality changes how we

[99] *See* FIN. INDUS. REGULATORY AUTH., REPORT ON DIGITAL INVESTMENT ADVICE 3 (2016) ("If an algorithm is poorly designed for its task or not correctly coded, it may produce results that deviate systematically from the intended output and that adversely affect many investors.").

[100] Anna L. Buczak & Erhan Guven, *A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection*, 18 IEEE COMM. SURVS. & TUTORIALS 1153, 1153 (2016); Marcus A. Maloof, *Introduction* to MACHINE LEARNING AND DATA MINING FOR COMPUTER SECURITY 1, 1-2 (Marcus A. Maloof ed., 2006).

[101] Lily Hay Newman, *AI Can Help Cybersecurity—If It Can Fight Through the Hype*, WIRED (Apr. 29, 2018, 7:00 AM), https://www.wired.com/story/ai-machine-learning-cybersecurity/ [https://perma.cc/99Y9-7APY].

[102] *Id.*

[103] Martin Giles, *AI for Cybersecurity Is a Hot New Thing—and a Dangerous Gamble*, MIT TECH. REV. (Aug. 11, 2018), https://www.technologyreview.com/s/611860/ai-for-cybersecurity-is-a-hot-new-thing-and-a-dangerous-gamble/.

[104] Trying to imagine a human doing that brings to mind the scene in which Cypher explains to Neo that he sees "blonde, brunette, redhead" in the patterns of the Matrix's clearly indecipherable code. THE MATRIX (Warner Bros. 1999).

[105] *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 683-84 (2016) [hereinafter Barocas & Selbst, *Big Data's Disparate Impact*]; Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 3, 10-12 (Bart Custers, Toon Calders, Bart Schermer & Tal Zarsky eds., 2013) ("[A] particular profile may be entirely correct from a technological perspective, but may still be applied incorrectly.").

perceive it.[106] In the best case, training data makes for a decent approximation of reality, but in the cases where the entire purpose of an AI system is to predict the unobservable, there may be no way to know how far off the approximation is.

Worse yet, in some cases, such as where the category is constructed for the purpose of classification, the very idea of a measurable truth may not even exist.[107] In the data security context, for example, the AI takes normal traffic as a baseline, but the AI is defining normal traffic as it learns; there is no external referent. If it learns "normal" in a compromised system, then the "normal" category is the compromised one. There is therefore no way for an IT professional to oversee the output of the data security model of normal traffic and to judge whether it is compromised. This is fundamentally different than traditional machines, which respond to well-understood and experimentally verifiable physics. It is also different from the machine vision system in a vehicle. The dog, bus, and crosswalk exist or do not, irrespective of what the AI says.[108]

The situation is further complicated where predictions of decision systems will affect the very outcomes they are trying to predict. Take a personalized medical treatment recommendation for example. If it was made in error, the patient will not be aware until an injury occurs. But even then, whether there was an error will not be clear. Once initiated, there is no counterfactual that can undo the treatment. Maybe the treatment was correct and the patient would have been *worse* with a different treatment. There is really no way to know. Generalized medical statistics cannot solve the problem because at issue is the AI's *personalization* of the treatment recommendation; the question is inherently whether a deviation from the general practice was correct. Even a talented physician's catalog of medical knowledge cannot always help. Remember, surpassing human knowledge is a major goal of such an AI system; if second guessing were possible, much of the purpose of these systems would be nullified.

The lack of ground truth is not always a permanent feature of these systems. This is the case only where there is no way to determine "reality" or where the result of the prediction affects a course of action on the ground. Other applications exist. If a machine learning system predicts an aspect of the stock

---

[106] *See generally* GEOFFREY C. BOWKER & SUSAN LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES (1999) (examining how classifications, categories, and standards shape interpretation of information).

[107] *See, e.g.*, Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 105, at 679.

[108] Note that like all categories, humans have created the perceptual linguistic categories of "dog," "bus," and "crosswalk"; they do not exist in the ether. But we agree what they are. To say that humans have created the categories is not to say that there is also no truth value to correctly labeling particular instances of them. *See, e.g.*, BOWKER & STAR, *supra* note 106, at 37-40 (discussing necessity, ubiquity, and materiality of classifications).

market in a given amount of time, there is no ground truth problem. Unless the prediction leads to an action big enough to affect the stock market, the truth will eventually reveal itself.[109] The same is generally true of medical (mis)diagnosis (as opposed to personalized treatment); the truth of the prediction will reveal itself in due time. But in both the stock market and medical diagnosis examples, if the prediction was incorrect, the falsity will be established at the same time the harm is accomplished—in fact, it will be established by the harm itself. Therefore, even though the ground truth problem is eventually resolved, it has little bearing on reasonableness determinations that rely on perceiving AI error in real time.

The result of this epistemic limitation is a lack of a principled basis to contradict AI predictions. Suppose it were otherwise: If a doctor receives a readout that suggests that a patient has a certain rare diagnosis that she missed, how can the doctor determine whether or not to believe the AI and treat the patient accordingly? The doctor could be unsure how to proceed, or could be completely sure that she is correct and turn out to be wrong. There is no reasonable basis on which to make a determination; the choice to use AI in the first place puts the doctor in the position of believing it or not almost as an article of faith.[110]

This implies that the reasonableness of an action in individual cases must be tied to the decision to use AI in the first place.[111] Depending on the state of the science, the decision to use AI may or may not be reasonable. At a certain threshold of error reduction, AI may become a reasonable choice, or the possibility of undetectable errors in individual cases may represent reasons to avoid AI use even if it is safer overall. But the negligence inquiry that is the subject of this Article asks whether people who are harmed when AI is used can still recover damages. By the time the user is put into a position to act reasonably or unreasonably, the choice to use AI has already been made and cannot factor into the analysis. If the use of AI is judged reasonable at the outset, it cannot later be deemed unreasonable to miss errors that are undetectable by humans. Such a standard turns negligence into strict liability.

Now, it will not *always* be the case that people cannot oversee the decisions. There is a strong push within the technical literature to build interpretable

[109] This may be entirely plausible for hedge funds that use machine learning but unlikely for individual robo-advisers.

[110] *See* Price, *Medical Malpractice and Black-Box Medicine*, *supra* note 67, at 300-01.

[111] *See, e.g.*, Froomkin, Kerr & Pineau, *supra* note 67, at 61 ("A physician (or hospital, or insurer) relying on a[machine learning] system will be held to no different a standard than if the physician relied on a human; indeed, from a legal point of view, the decision to rely on [machine learning] will be a human medical judgment like any other.").

machine learning systems or systems capable of ex post explanation.[112] The goal of such systems is to demystify the relationships that the machine learning system uncovers so they can comport better with human understanding.[113] For example, in a well-known study, computer scientists Rich Caruana and colleagues pointed to a model used to predict death from pneumonia, trained on past patients' results in a hospital.[114] Because the model was built to be interpretable, they were able to discover that the model had learned to predict that patients with asthma had reduced risk from pneumonia, a result that makes no sense medically.[115] It was not an error either; it was a real trend present in the data, likely a result of the fact that asthma patients pay more attention than others to breathing problems, self-report pneumonia symptoms earlier, and, once in a hospital, receive emergency treatment.

The authors of this study use it to argue that we need to build systems to be interpretable.[116] An uninterpretable system would have found the same patients to be less risky, but there would be no way to question the result. "Interpretability" and "explainability" in technical systems are terms that stand in for a range of concepts. There are many ways to build interpretable systems, and none of them can get at every meaning of the word.[117] Caruana and colleagues argue for a specific form of interpretability that allows more information about the interaction between input variables. This allowed the link between asthma and pneumonia to appear. Today, the concept of "counterfactual explanations" is in vogue.[118] Counterfactual explanations enable a system to point to the most impactful input variables in order to demonstrate which changes to input variables would most likely result in a different outcome.[119]

The research demonstrates that more interpretable systems can sometimes render the mysterious obvious and take advantage of domain expertise. Thus, interpretability can render some AI errors predictable. In the case of the pneumonia-asthma link, because the system was built to be interpretable, a

[112] *See* Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1110 (stating that methods have emerged "promising to increase interpretability while retaining performance").

[113] *See id.* at 1109-10.

[114] Caruana et al., *supra* note 90, at 1721.

[115] *Id.*

[116] *Id.*

[117] *See generally* Lipton, *supra* note 75 (refining discourse on interpretability, examining underlying motivations, model properties, and techniques thought to confer interpretability, feasibility, and desirability); Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1110-17.

[118] *See* Solon Barocas, Andrew D. Selbst & Manish Raghavan, *The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons*, PROC. CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 80, 80-81 (2020).

[119] *Id.* at 81.

doctor would have enough information that failing to overrule the AI could be considered unreasonable and lead to malpractice liability. But interpretability is not a panacea; it only works for some subset of errors.[120] In Caruana and colleagues' study, pneumonia was intuitively linked to asthma through breathing. But if the model had instead found a correlation between skin cancer and pneumonia, the doctors would be back to scratching their heads.[121] It is not that the doctors would think it is right or wrong; it is that they would not know how to determine that fact—they are left in the same position as if the system were not built interpretably.[122] Similarly, counterfactual explanations will often enable some degree of understanding, but they rely on a plethora of hidden assumptions that render them less demystifying than their proponents argue.[123]

The implications for negligence are as follows: Interpretability and explainability can resolve the foreseeability challenge in some cases. If an interpretable or explainable model happens to demonstrate a correlation that humans can intuitively understand, it can turn the second kind of AI discussed here into the first. The AI would become more like the machine vision system in the autonomous vehicle—one that replicates human knowledge and thus one that we can oversee. But importantly, that is not the general case and can only be used to find errors that accord with human intuition and expertise.[124] Whether the errors that occur will be intuitive is itself unpredictable, however. Thus, while the foreseeability challenge AI poses for negligence is not absolute, it is still a difficulty that will always exist in a subset of cases.

There is another possibility as we become more familiar with AI. If we catalog the cases that AI systems get wrong and use them to better understand the limits of our data, errors that seem mysterious may turn into patterns, and best practices may be reincorporated into the reasonable care standard. Studying the implementation of a sepsis detection AI within the Duke hospital system, Dr. Mark Sendak and colleagues found that medical practitioners were able to

---

[120] Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1123.

[121] *Id.*

[122] *Id.*

[123] *See generally* Barocas, Selbst & Raghavan, *supra* note 118 (demonstrating that "the utility of feature-highlighting explanations relies on a number of easily overlooked assumptions: that the recommended change in feature values clearly maps to real-world actions, that features can be made commensurate by looking only at the distribution of the training data, that features are only relevant to the decision at hand, and that the underlying model is stable over time, monotonic, and limited to binary outcomes"); I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger & Sorelle A. Friedler, *Problems with Shapley-Value-Based Explanations as Feature Importance Measures*, PROC. 37TH INT'L CONF. ON MACHINE LEARNING 8083, 8088-91 (2020).

[124] Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1123.

develop new types of expertise and intuitions related to the AI after some time.[125] If new intuitions develop, new ideas of what is negligent may also. Importantly, there will still be a set of cases that are unpredictable, but existing negligence law allows for the possibility of accidents that are nobody's fault. The goal is not zero risk. Adaptation is the normal result for negligence law in the face of new technologies, and it may occur here. This is discussed more in Section III.A below.

2. A New Kind of Foreseeability Concern

Much of the existing research points to foreseeability as the greatest challenge that AI poses for tort law. A common refrain in discussions of AI is that it is "unpredictable by design."[126] From there, scholars argue that AI systems pose foreseeability problems. As the previous Section suggests, I agree, but it is worth a short digression to be more specific about the point. The epistemic limitation I describe above certainly shares similarities with the foreseeability concerns in the breach or proximate cause elements of negligence. If there is no reason that a defendant can foreseeably connect their action to a plaintiff's harm, then that action cannot be said to be unreasonable. This is true independent of whether the breach standard is one of ordinary care or professional malpractice.[127]

While this concern is certainly related to foreseeability, it does not fit neatly into any of the foreseeability categories in current negligence doctrine. Though foreseeability acts differently in each of the negligence elements of duty, breach, and proximate cause, the doctrine still seeks specific things in each case: Depending on the element, the doctrine asks whether a specific plaintiff, a specific risk, or a specific category of harm is foreseeable.[128] Because much of

[125] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu & Cara O'Brien, *"The Human Body Is a Black Box": Supporting Clinical Decision-Making with Deep Learning*, PROC. CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 99, 106 (2020).

[126] *E.g.*, Calo, *supra* note 3, at 542 ("[T]he mechanisms by which the law sorts fault involve deeply human concepts such as . . . foreseeability . . . which are absent where a system is built to be unpredictable by design."); Jason Millar & Ian Kerr, *Delegation, Relinquishment, and Responsibility: The Prospect of Expert Robots*, *in* ROBOT LAW, *supra* note 17, at 102, 107 (analyzing unpredictability in Watson AI system).

[127] *See, e.g.*, Owen, *Figuring Foreseeability*, *supra* note 78, at 1286 ("[D]ecisions may be considered faulty . . . only if the actor is capable of understanding the *meaning* of those choices—the possible *consequences* of contemplated actions. All decisions, that is, involve choice, choice presumes capacity, and capacity includes foreseeability as a proxy for the actor's will. In short, a person is not meaningfully 'accountable' for causing harm that he or she cannot reasonably foresee and therefore in no sense wills."); Stephen R. Perry, *The Moral Foundations of Tort Law*, 77 IOWA L. REV. 449, 485-86 (1992).

[128] *See* Benjamin C. Zipursky, *Foreseeability in Breach, Duty, and Proximate Cause*, 44 WAKE FOREST L. REV. 1247, 1254-55 (2009) (discussing Third Restatement's treatment of foreseeability in each of these elements of negligence).

the rhetoric around the unpredictability of AI contributes to continued misunderstandings about what AI actually is and does and about how closely it mimics human capabilities and agency, it is important to be specific about how exactly AI is unpredictable and how that affects the analysis.

Let us first examine the foreseeability of categories of harm.[129] Consider AlphaGo, a well-known artificially intelligent Go-playing computer designed by DeepMind. AlphaGo made headlines in 2016 by beating world Go champion Lee Sedol.[130] In Move 37 of the match's second game, AlphaGo made "a move that no human ever would," a move so strange that one expert thought it had to be a mistake.[131] It turns out, though, that it was an excellent move that we could not have predicted or understood in real time because we do not think like computers. While it is true that the machines make unpredictable decisions, there are multiple senses in which a machine can be unpredictable.[132] AlphaGo made an unpredictable move in the game of Go, but ultimately, it was still playing Go. It would have been unpredictable in an entirely different sense if, sensing it was going to lose a game, AlphaGo flipped the board or called in a bomb threat to evacuate the premises.[133]

Predictions of foreseeability issues in existing literature trend more toward AlphaGo's bomb threat. For example, Professor Ryan Calo has offered the following hypothetical:

> Imagine one manufacturer stands out in this driverless future. Not only does its vehicle free occupants from the need to drive while maintaining a sterling safety record, it adaptively reduces its environmental impact. The designers of this hybrid vehicle provide it with an objective function of greater fuel efficiency and the leeway to experiment with system operations, consistent with the rules of the road and passenger expectations. A month or so after deployment, one vehicle determines it performs more efficiently overall if it begins the day with a fully charged battery. Accordingly, the car decides to run the gas engine overnight in the garage—killing everyone in the household.

---

[129] *Cf.* Fischer, *supra* note 27, at 550-51 ("Some negligence cases impose liability only where the type of risk that was foreseeable to the defendant actually occurred. If the defendant's negligence causes harm by fire, he is liable if he could foresee the risk of fire, but not otherwise.").

[130] Cade Metz, *In Two Moves, AlphaGo and Lee Sedol Redefined the Future*, WIRED (Mar. 16, 2016, 7:00 AM), https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/ [https://perma.cc/F3R3-9FPP].

[131] *Id.*

[132] *Cf.* Andrew D. Selbst, *A Mild Defense of Our New Machine Overlords*, 70 VAND. L. REV. EN BANC 87, 89 (2017) ("Even if humans cannot understand machines in the same way we understand each other, that is not to say we cannot understand them at all.").

[133] *Cf.* Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 365 (2016).

> Imagine the designers wind up in court and deny they had any idea this would happen. They understood a driverless car could get into an accident. They understood it might run out of gas and strand the passenger. But they did not in their wildest nightmares imagine it would kill people through carbon monoxide poisoning.[134]

A car killing a household by carbon monoxide poisoning is not—or at least not yet—a realistic risk of automating cars.[135] Cars are limited in their unpredictability; they are programmed only to drive.[136] Similarly, if AI for a medical diagnosis fails, we are still dealing with precisely the category of harm—the injuries that attend misdiagnosis or improper treatment—that one would expect.

If AI becomes more multifunctional and autonomous, category-of-harm foreseeability may become a problem. But while AI is a tool used for a single purpose, it seems no more likely than usual that the category of harm will be anything other than what we would expect. Category foreseeability will be more relevant with something closer to artificial general intelligence ("AGI"), sometimes called "strong AI."[137] Every form of AI currently on the market exists for specified and limited purposes, while AGI is at best many years off and essentially unrelated to existing machine learning technologies.[138] At that point, the foreknowledge that the AI could do anything at all could paradoxically increase the range of what is considered foreseeable.[139] But at least until then, the foreseeable categories of harm should not change simply because AI is used.

Now consider foreseeability of a given risk. For an autonomous vehicle, we might imagine a feedback loop that causes it to accelerate beyond safe speeds and crash. This is a foreseeable driving accident but perhaps not a foreseeable risk. (Of course, anything I can name might actually be foreseeable in some

---

[134] Ryan Calo, *Is the Law Ready for Driverless Cars?*, COMM. ACM, May 2018, at 34, 35.

[135] There is, of course, a general difficulty in naming a specific risk as unforeseeable because to name it in advance, it must be foreseen. Any risk that is truly unforeseeable is inherently also describable as "not realistic." I am not suggesting that the specific harm named is problematic but instead that the very conceptual move to a different category of harm is not realistic for AI that is a single-purpose tool.

[136] *See* Surden & Williams, *supra* note 29, at 128.

[137] *See, e.g.*, John R. Searle, *Is the Brain's Mind a Computer Program?*, SCI. AM., Jan. 1990, at 25, 26 (distinguishing between "strong AI" and "weak AI").

[138] *See, e.g.*, Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 432 (2017) ("[N]othing in the current literature around [machine learning], search, reinforcement learning, or any other aspect of AI points the way toward modeling even the intelligence of a lower mammal in full, let alone human intelligence."); Erik Sofge, *Artificial Intelligence Will Not Obliterate Humanity*, POPULAR SCI., Mar. 19, 2015, at 36, 36.

[139] *See* Owen, *Bending Nature, Bending Law*, *supra* note 75, at 609-10 (describing "paradox of foreseeable unforeseeability").

sense,[140] but let us assume otherwise.) Or, imagine an example raised by Professor Mark Lemley and Bryan Casey, in which a drone was trained to fly to the center of a circle and learned that, as it approached the edge, the fastest way to the center of the circle was to leave it.[141] This occurred because after the drone left the circle, the people training it would turn it off and place it back in the middle of the circle, so it appeared to the system that the edge would magically teleport it back to the center.[142] This magical teleportation was the drone's ground truth.

This kind of unpredictability could easily lead to injuries, but it can only exist in a fully automated system. The whole reason this story is surprising—and even amusing—is that the AI acts in a way that no human would think to. Imagine that instead of moving on its own, the drone was set up with a human in the loop, and the AI's outputs were instructions for the human with the joystick. Then, when the drone tells the joystick operator to run it outside the circle, responsibility would fall on the joystick operator not to do so; the operator would recognize the output makes no sense and ignore it. The human-in-the-loop aspect of the technologies that still rely on negligence law ensure that this type of wildly unexpected AI injury cannot happen, or in fact, the human would be reasonably blamed for it. Recall the caveats stated at the end of Section II.A.1: There will be cases that are obviously wrong, and any concept of reasonable care would still require the human operator to prevent those.

Therefore, neither the category of harm nor the specific risk is unforeseeable with AI decision-assistance technologies. That leaves particular plaintiffs. But a decision-assistance tool applies in a known context, so the people—and thus the plaintiffs—in that context do not differ between the cases with and without AI. A patient and the patient's family are going to be foreseeable victims of medical malpractice, independent of the technology used. Therefore, none of the traditional notions of foreseeability apply. Of course, there will remain the standard foreseeability questions that apply to all injuries, but those are not about AI. If one element of a long *Palsgraf*-like chain of events happens to involve AI, the foreseeability challenge is with the long chain of events, not the AI.[143]

Thus, while the particular chain of decisions that led to AI errors may be impossible to understand, what happened would likely be considered foreseeable under current doctrine. Another way to understand this is that foreseeability does not ask that the *specific* manner of harm be foreseeable, so

---

[140] *See* W. Jonathan Cardi, *Reconstructing Foreseeability*, 46 B.C. L. REV. 921, 951 (2005) [hereinafter Cardi, *Reconstructing Foreseeability*] (describing normative judgment about foreseeability as "indeterminate" and "a point drawn by the decisionmaker on the spectrum of epistemic probability").

[141] Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311, 1313 (2019).

[142] *Id.*

[143] *See* Palsgraf v. Long Island R.R., 162 N.E. 99, 99 (N.Y. 1928).

knowing that AI's use will lead to errors in some cases may be enough to say that those errors are generally foreseeable.

So does this mean that there is no foreseeability problem? I do not believe so, as a practical matter. The prior discussion demonstrates that in many cases, assuming the use of AI is itself reasonable, the choice for an injury is strict liability or no liability with no in-between. One way or another, this is not negligence, as negligence will never demand the impossible as part of what is reasonable. This is why it seems reasonable to treat AI's incomprehensibility as a foreseeability concern. The problems identified here go to the very reasons that foreseeability is so central to the doctrine. The link between actions and responsibility is severed when a person must make a choice without any appreciation of whether that choice will lead to harm in a given case.[144] And from an accident-prevention perspective, a person who cannot predict an outcome cannot be in the best position to prevent it. While AI will not actually challenge traditional notions of foreseeability, unpredictable AI errors can functionally be considered a new type of unforeseeable harm because the other option is just strict liability for *all* AI errors, and that seems unlikely to be the result that courts prefer.

B. *Limitations on Human-Computer Interactions*

Whereas the previous Section concerned humans' inability to foresee AI errors, the use of AI in partnership with human decisions will also encounter limits based on humans' other fundamental cognitive and physical limitations. The field of human-computer interactions ("HCI") is dedicated to studying these sorts of challenges. The best-known example of this is the so-called "handoff problem" with partially autonomous vehicles.[145] In NHSTA's five levels of autonomy, Level 3 cars are also the most inherently dangerous.[146] This is because Level 3 cars are designed to kick control back to the safety driver when the computer runs into trouble, but it turns out that humans are quite bad at continually monitoring a situation without being engaged and then taking over when needed.[147] There is a fundamental limit to humans' ability to reengage

[144] Owen, *Figuring Foreseeability*, *supra* note 78, at 1286.

[145] *See* AM. ASS'N FOR JUSTICE, DRIVEN TO SAFETY: ROBOT CARS AND THE FUTURE OF LIABILITY 14 (2017), http://www.justice.org/sites/default/files/Driven%20to%20Safety%202017%20Online.pdf [https://perma.cc/UNR7-56N3] [hereinafter DRIVEN TO SAFETY] ("Research shows that humans are not well adapted to re-engaging with complex tasks, like driving a vehicle in an emergency situation, once their attention has been allowed to wander.").

[146] *See* Smart, Grimm & Hartzog, *supra* note 2, at 22-23.

[147] *See id.* at 23.

quickly enough to avert an accident. Autonomous vehicle makers believe this strongly enough that some are planning to skip Level 3 automation entirely.[148]

While autonomous vehicle companies can decide to skip Level 3 in anticipation that this problem will resolve itself at "fully autonomous" Levels 4 and 5, other applications do not have this option. Decision-assistance tools are designed for human operation and therefore will never be fully automated. For example, HCI problems also arise in the medical context. Clinical decision support ("CDS") tools have tested well in labs but have mostly "failed when migrating from research to clinical practice," either because doctors do not trust in the system or the system design does not mesh with the way doctors do their jobs.[149] Professors Michael Greenberg and Susan Ridgely have separately written about the phenomenon of "alert fatigue."[150] One application of CDS tools is to create a model of known problems from drug interactions and to alert when such a possibility arises.[151] But as Greenberg and Ridgely write:

---

[148] *See id.*; Alex Davies, *The Very Human Problem Blocking the Path to Self-Driving Cars*, WIRED (Jan. 1, 2017, 7:00 AM), https://www.wired.com/2017/01/human-problem-blocking-path-self-driving-cars/ [https://perma.cc/J5TX-2J98].

[149] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey & James F. Antaki, *Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help*, PROC. ACM CHI CONF. ON HUM. FACTORS COMPUTING SYS. 4477, 4477 (2016); *see also* Srikant Devaraj, Sushil K. Sharma, Dyan J. Fausto, Sara Viernes & Hadi Kharrazi, *Barriers and Facilitators to Clinical Decision Support Systems Adoption: A Systematic Review*, J. BUS. ADMIN. RES., Oct. 2014, at 36, 41-44 (listing barriers to adoption of CDS tools, such as "poor system design" and "prior bad experience"); Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian G.K. Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L. Lewis, Richard M. Wexler & Dominick L. Frosch, *"Many Miles To Go . . .": A Systematic Review of the Implementation of Patient Decision Support Interventions into Routine Clinical Practice*, BMC MED. INFORMATICS & DECISION MAKING, Nov. 2013, at 1, 6 (Supp. II) (reporting that "professional indifference and organizational inertia" inhibit adoption of patient decision-support interventions); Monique W.M. Jaspers, Marian Smeulers, Hester Vermeulen & Linda W. Peute, *Effects of Clinical Decision-Support Systems on Practitioner Performance and Patient Outcomes: A Synthesis of High-Quality Systematic Review Findings*, 18 J. AM. MED. INFORMATICS ASS'N 327, 331-32 (2011) (finding that CDS generally did not result in "benefits on patient outcomes" but "can positively impact healthcare providers' performance with preventative care reminder systems and drug prescription systems"); Kensaku Kawamoto, Caitlin A. Houlihan, E. Andrew Balas & David F. Lobach, *Improving Clinical Practice Using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success*, 330 BRIT. MED. J. 765, 767 (2005) (identifying features associated with CDS's ability to improve patient outcomes including automatic decision support as part of clinician workflow); Jeremy C. Wyatt & Douglas G. Altman, Commentary, *Prognostic Models: Clinically Useful or Quickly Forgotten?*, 311 BRIT. MED. J. 1539, 1539 (1995).

[150] Michael Greenberg & M. Susan Ridgely, *Clinical Decision Support and Malpractice Risk*, 306 JAMA 90, 90 (2011).

[151] *Id.*

> In practice . . . CDS systems often have been overinclusive in the warnings they generate, to a point at which physician "alert fatigue" may in large part undermine the utility the systems offer. . . . One paradoxical result of overly abundant warnings may be to exacerbate malpractice risk for physicians who either ignore or turn off CDS alerts, even as CDS systems create an audit trail to show that those physicians have done so.[152]

How should negligence law treat the driver of the Level 3 car who gets in an accident or the doctor who ignores the wrong alert? There are two possible responses. One is that in order to drive a Level 3 car, reasonable care should be interpreted to require a very high standard of attentiveness. The HCI research does not imply that it *impossible* for humans to remain alert while not driving, just that it requires extraordinary attention and effort.[153] Perhaps the law ought to claim that such attention is a required skill for driving a Level 3 car and that anything short of it is negligent.[154] Similarly for doctors using CDS tools; while ignoring a machine that cries wolf might be typical behavior, perhaps we should expect more of our doctors in that situation. The other possibility is to lean harder on the definition of "reasonable" to suggest that if psychology research says the average person cannot do something, the law should not hold that the reasonable person must.[155] Does "reasonable" care imply an especially high level of care given the facts on the ground, or does it imply a standard of care closer to what an average driver or doctor can plausibly do?[156] Either approach

[152] *Id.*

[153] *See* DRIVEN TO SAFETY, *supra* note 145, at 14.

[154] Such a standard is not totally unthinkable. At common law, common carriers traditionally owed a duty of the "highest degree of care practicable under the circumstances" to their passengers. S. Pac. Co. v. Hogan, 108 P. 240, 241 (Ariz. 1910). But it should be noted that this imposition of the "highest" degree of care is an outlier that is somewhat the result of historical accident, *see* Robert J. Kaczorowski, *The Common-Law Background of Nineteenth-Century Tort Law*, 51 OHIO ST. L.J. 1127, 1157-59 (1990), and even this heightened degree of care requires reasonable care in all circumstances and not "all the care, skill, and diligence of which the human mind can conceive." Nunez v. Prof'l Transit Mgmt. of Tucson, Inc., 271 P.3d 1104, 1105, 1109 (Ariz. 2012) (en banc) (quoting Lunsford v. Tucson Aviation Corp., 240 P.2d 545, 546 (Ariz. 1952)).

[155] Note also that this is a different problem than Hubbard's claim that a reasonable driver must know when software is malfunctioning in order to take over. *See* Hubbard, *supra* note 26, at 1861 ("Similarly, in order to satisfy the standard of reasonable care, users of driverless cars would need to use the skills necessary to operate the car reasonably, by, for example, knowing when the driving system was malfunctioning and, to some extent, how to respond to the malfunction."). That is an epistemic problem and was the subject of the previous Section. Instead, the problem here is that people might be unable to stay engaged enough either to determine whether the car is malfunctioning in a timely manner or to react quickly enough once the determination is made.

[156] Unlike the safety driver, the CDS alert case is not necessarily particular to AI. An annoying non-AI system would be just as quickly ignored. This observation is key to data

can be defended; the former will result in fewer people using the technology, the latter in more uncompensated injuries.

One way to think about this conundrum is as an expansion of the "pocket[s] of strict liability" in negligence law.[157] Tort theorists have observed that negligence law has several pockets of strict liability: places where liability attaches in a negligence regime even though there is no level of care the tortfeasor is capable of that could have prevented the injury. One example is that by applying an objective "reasonable person" standard, the law imposes strict liability on those—such as children, the disabled, or the inept—who may not be able to meet it.[158] Another is vicarious liability, where the employer faces liability but cannot directly affect the overall level of care.[159] One version of this idea, advanced by Professor Mark Grady, seems closest to the HCI problem. Grady argues that a pocket of strict liability comes from the law's requirement of "perfect compliance" with the requirements of reasonable care.[160] Negligence law strictly punishes momentary lapses in attentiveness.[161] While this pocket has always been a feature of negligence law, AI turns the pocket inside out by requiring a higher attention threshold than the average person can keep up. That technology companies are avoiding Level 3 cars suggests that the handoff problem is so troubling because it may become the dominant cause of accidents for Level 3 cars. We accept pockets of strict liability because they are the exception, but AI may make them the rule.

A different, but related, theoretical frame comes from what anthropologist Madeleine Elish has termed "moral crumple zones."[162] Elish argues that because

security as well. *See* Martina Angela Sasse, Sacha Brostoff & Dirk Weirich, *Transforming the 'Weakest Link' — A Human/Computer Interaction Approach to Usable and Effective Security*, BT TECH. J., July 2001, at 122, 123. But the application of AI to new problems, driven by the market power of AI companies and enthusiasm for AI generally, injects software into decision processes that did not have to deal with software before. Thus, the need to deal with more classic HCI problems is at least partly a result of AI. It also introduces other software liability issues discussed in Section II.D.

[157] LANDES & POSNER, *supra* note 11, at 128; *see* SHAVELL, *supra* note 11, at 75 ("Put a little differently, for the inept person archery may be regarded as an ultrahazardous activity; thus it makes sense, in effect, to impose strict liability on the inept person who engages in archery."); Kenneth S. Abraham, *Strict Liability in Negligence*, 61 DEPAUL L. REV. 271, 272 (2012) (noting that negligence law contains elements of strict liability but "defin[es] it out of existence"); Grady, *supra* note 1, at 303 ("In one striking respect, however, the reasonable person is anything but average: he or she never forgets to use a reasonable precaution.").

[158] LANDES & POSNER, *supra* note 11, at 128; SHAVELL, *supra* note 11, at 75.

[159] LANDES & POSNER, *supra* note 11, at 120-21.

[160] Grady, *supra* note 1, at 303.

[161] *Id.*

[162] Madeleine Clare Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAGING SCI. TECH. & SOC'Y 40, 41 (2019) ("I articulate the concept of a *moral crumple zone* to describe how responsibility for an action may be misattributed to a

the machines are—explicitly or implicitly—seen as infallible, there are situations in which humans act as "liability sponges."[163] Even though the human operators may not be able to avert the danger, failure to do so manifests as a decision they can be blamed for—such as not taking the wheel back in time or not paying enough attention to alerts.[164] In March 2018, an Uber-owned vehicle in autopilot mode killed a pedestrian.[165] There was a safety driver behind the wheel who failed to avert the disaster.[166] Elish's framing aptly explains why the safety driver was subsequently charged with criminal negligence.[167] If negligence law requires a higher standard of care than humans can manage, it will place liability on human operators, even where the average person cannot prevent the danger.

The path of tort law in the face of new innovation is not a straight line.[168] While the Uber case seems to point in the direction of moral crumple zones, it is also easy to imagine the reverse—finding that because the average person cannot react in time or stay perpetually alert, failing to do so is reasonable. Ultimately, what AI creates is uncertainty.

C. *AI-Specific Software Vulnerabilities*

As a ubiquitous decision tool, AI introduces software into decisions that were not previously mediated by software. This injects vulnerabilities into decision processes that are not entirely in the control of the decisionmaker. Software crashes. Software can be hacked. Perhaps most importantly, AI creates unique security problems different from other types of software.[169] Rather than simply needing to protect sensitive data for privacy reasons, AI security requires that

human actor who had limited control over the behavior of an automated or autonomous system. Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a highly complex and automated system may become simply a component . . . that bears the brunt of the moral and legal responsibilities when the overall system malfunctions." (footnote omitted)).

[163] *Id.*

[164] *Id.* at 42; *see also* Graham, *supra* note 28, at 1260-66.

[165] *See* Elish, *supra* note 162, at 52.

[166] *See id.*

[167] *See id.* at 53.

[168] *See* Graham, *supra* note 28, at 1269.

[169] *See* Sigal Samuel, *It's Disturbingly Easy to Trick AI into Doing Something Deadly*, VOX (Apr. 8, 2019, 9:10 AM), https://www.vox.com/future-perfect/2019/4/8/18297410/ai-tesla-self-driving-cars-adversarial-machine-learning [https://perma.cc/7BS4-2GUP] (describing adversarial attacks on AI such as using stickers to simulate lines on road to trick machine vision in self-driving cars).

training data be protected to ensure the regular and safe behavior of the AI itself.[170]

An area of research known as "adversarial machine learning" is dedicated to figuring out how to influence a machine learning system's decisions by manipulating the inputs to the model.[171] This research has demonstrated that an attacker can perturb an image slightly enough that a human would not notice the difference while causing an AI to drastically change its interpretation.[172] The canonical example is tricking an AI into classifying an image of a panda as a gibbon with ninety-nine percent certainty.[173] A more recent paper shows that similar attacks can cause a medical-imaging AI to switch its diagnosis of a mole from benign to malignant.[174] Worse yet, if we understand too well how an AI is trained, it is possible to "hack" the real world, rather than the software, to alter the AI's responses.[175] Researchers have demonstrated that adding stickers to a stop sign can cause an AI to see it as a yield sign[176] and a series of white dots on the road can cause a Tesla in semi-autonomous mode to shift lanes.[177] While these AI security concerns are primarily theoretical at the moment, there is good

[170] *See* Ivan Evtimov, David O'Hair, Earlence Fernandes, Ryan Calo & Tadayoshi Kohno, *Is Tricking a Robot Hacking?*, 34 BERKELEY TECH. L.J. 891, 903 (2019).

[171] *See, e.g.*, Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, *Explaining and Harnessing Adversarial Examples*, INT'L CONF. ON LEARNING REPRESENTATIONS, May 9, 2015, at 1, 1; Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik & Ananthram Swami, *Practical Black-Box Attacks Against Machine Learning*, PROC. 2017 ACM ASIA CONF. ON COMPUTER & COMM. SECURITY 506, 506 (2017).

[172] *See* Papernot et al., *supra* note 171, at 1.

[173] *See* Evan Ackerman, *Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms*, IEEE SPECTRUM (Aug. 4, 2017, 6:00 PM), https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms [https://perma.cc/6GGL-NSDL].

[174] *See* Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam & Isaac S. Kohane, *Adversarial Attacks on Medical Machine Learning*, SCI., Mar. 22, 2019, at 1287, 1288 [hereinafter Finlayson et al., *Adversarial Attacks*]; *see also* Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane & Andrew L. Beam, Adversarial Attacks Against Medical Deep Learning Systems 2-4 (Feb. 4, 2019) (unpublished manuscript), https://arxiv.org/pdf/1804.05296.pdf [https://perma.cc/7RSH-ZQKG] (arguing that adversarial attacks are particularly worrisome for medical imaging due to financial incentives and technical vulnerability).

[175] *See* Evtimov et al., *supra* note 170, at 910-12 (discussing examples of changing real-world environment to change behavior of AI systems).

[176] *See* Ackerman, *supra* note 173.

[177] Ariel Bogle, *Hackers Tricked a Tesla, and It's a Sign of Things to Come in the Race to Fool Artificial Intelligence*, ABC NEWS (Apr. 13, 2019, 11:43 PM), https://www.abc.net.au/news/science/2019-04-14/tesla-tencent-study-humans-are-trickable-so-are-computers/10994578 [https://perma.cc/7TX4-ABES].

reason to believe that they will become applicable to real-world contexts soon enough.[178]

Tort law has never fully grappled with software's particular difficulties. Software was originally seen as akin to books, maps, and navigational charts—items for which the information it provides, rather than the tangible product, was considered the important part.[179] Due to software's intangibility, incidents of software crashes usually do not lead to liability, but when software within a larger product fails and leads to physical injuries, such as in plane crashes, courts are more willing to consider the software crash a basis for liability.[180] In a recent article, Professor Bryan Choi examined the history of courts' reluctance to recognize tort liability for software crashes.[181] He reviews three issues: (1) the relegation of software liability to contract law rather than tort law due to the economic loss doctrine which, in certain states, bars tort recovery for purely economic losses on the theory that such losses should be handled by contract; (2) economic protection by courts and Congress of a too-valuable software industry; and (3) the inevitability of software crashes due to software's complexity.[182] Choi's recommendation is to reinvigorate and borrow from "crashworthiness doctrine."[183] The theory of crashworthiness is that cars will inevitably crash—much like software—but when a crash is more impactful than it needed to be, it constitutes a second, independent injury worth holding the automobile manufacturer accountable for.[184] Applying crashworthiness doctrine to software is a promising idea, but only time will tell how courts approach software in the future.

When it comes to computer *security* in particular, reasonable security practice is not primarily enforced by tort lawsuits, but by the Federal Trade Commission ("FTC"). In recent years, the FTC has been trying to enforce data security practices under the "unfairness" prong of its "unfair and deceptive practices" authority.[185] The FTC's theory is that having unreasonable data security practices is inherently unfair and injurious to consumers, and it draws its reasonableness standard from negligence law.[186] The FTC has claimed—and the

[178] *See* Evtimov, O'Hair, Fernandes, Calo & Kobno, *supra* note 170, at 903; Finlayson, et al., *Adversarial Attacks*, *supra* note 174, at 1288-89 (discussing incentives to manipulate input data in context of insurance claims and drug and device approvals).

[179] RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 19 cmt. d.

[180] *See* Michael L. Rustad & Thomas H. Koenig, *The Tort of Negligent Enablement of Cybercrime*, 20 BERKELEY TECH. L.J. 1553, 1578 (2005).

[181] *See generally* Choi, *supra* note 40.

[182] *See id.* at 43-45.

[183] *Id.* at 115-17.

[184] *Id.* at 45-46, 94-95.

[185] *See* Solove & Hartzog, *supra* note 69, at 643.

[186] LabMD, Inc. v. FTC, 894 F.3d 1221, 1231 (11th Cir. 2018) ("The Commission's decision in this case does not explicitly cite the source of the standard of unfairness it used in

Third Circuit agreed in *FTC v. Wyndham Worldwide Corp.*[187]—that its reasonableness standard can be derived from past enforcement actions and a guidebook that lists certain minimum practices, such as using encryption, firewalls, and regular software patches. Although almost all of the cases settle and so little precedent exists, Professors Daniel Solove and Woodrow Hartzog have called the FTC's approach "functionally equivalent to a body of common law."[188] In a recent article, Professor William McGeveran has agreed that, though there are some differences, "the FTC framework is just as clear, and as flexible, as evolving common law jurisprudence."[189] McGeveran surveys fourteen different regulatory frameworks for data security.[190] He notes that reasonableness is the guiding principle throughout and that industry standards define such reasonableness.[191]

Though few states so far have codified specific data-security torts, AI decision-making tools introduce data security into traditional negligence.[192] As a result, negligence law will need to grapple with the question of what duty the AI users have to know or to investigate if they have been compromised, either in the traditional or the AI-specific sense. Consider the doctor who relies on the AI that was hacked to read a malignant mole as benign: Should that doctor be held liable for the AI's security failure? The AI-specific security concerns all involve new research, so, at the moment, it is unclear what a reasonableness standard should or would look like in terms of AI operational security.[193] But

holding that LabMD's failure to implement and maintain a reasonably designed data-security program constituted an unfair act or practice. It is apparent to us, though, that the source is the common law of negligence.").

[187] 799 F.3d 236, 256-57 (3d Cir. 2015) (holding that publication of data security guidebook, previous complaints, and previous consent decrees provided fair notice to defendant).

[188] Solove & Hartzog, *supra* note 69, at 586.

[189] McGeveran, *supra* note 69, at 1150; *see also* Paul N. Otto, Note, *Reasonableness Meets Requirements: Regulating Security and Privacy in Software*, 59 DUKE L.J. 309, 341 (2009) ("Recent approaches to providing protection through laws and regulations have favored the use of broad standards in lieu of specific rules.").

[190] McGeveran, *supra* note 69, at 1139.

[191] *Id.* at 1176-79, 1204.

[192] *See id.* at 1153-54.

[193] Recent research suggests that the availability of adversarial examples is itself a good test to demonstrate where models are insufficiently robust. *See generally* Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran & Aleksander Mądry, *Adversarial Examples Are Not Bugs, They Are Features*, 32 ADVANCES NEURAL INFO. PROCESSING SYS. 1 (2019) (demonstrating that "adversarial examples can be directly attributed to the presence of *non-robust features*: features (derived from patterns in the data distribution) that are highly predictive, yet brittle and (thus) incomprehensible to humans"). If this is the case, then addressing them would seem to be a reasonable expectation, though

one will be necessary to avoid injuries in AI-mediated decisions regulated by negligence law.

D. *Unevenly Distributed Injuries*

Outside of the tort context, the most commonly discussed concerns about algorithms relate to bias and discrimination. Data is not neutral, and decisions made by and with machine learning algorithms may have discriminatory results. This problem has generated government action and a large body of scholarly literature in just a few years.[194] This is such a central problem of algorithmic decision-making that it spawned an interdisciplinary conference and a whole subfield in computer science dedicated to the fairness, accountability, and transparency of these systems.[195] Discriminatory AI models are the result of a wide range of necessarily subjective decisions made throughout the machine learning process—including decisions about how to collect and treat the training data, how the problem is constructed, and how the model itself is trained, among others.[196] The resulting discrimination may be intentional or unintentional,[197]

perhaps one that would more likely lead to a products liability claim than one of negligence on the operator's part.

[194] *See, e.g.*, FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? 7-12 (2016); ANDREW GUTHRIE FERGUSON, THE RISE OF BIG DATA POLICING 199 (2017); Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 105, at 674; Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 522-23 (2018); danah boyd, *Undoing the Neutrality of Big Data*, 67 FLA. L. REV. F. 226, 227 (2016); Matthew Adam Bruckner, *The Promise and Perils of Algorithmic Lenders' Use of Big Data*, 93 CHI.-KENT L. REV. 3, 3-6 (2018); James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 173 (2017); Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM L. REV. 633, 650 (2017); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1053-54 (2019); Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 196-98 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 860-61 (2017); Mary Madden, Michele Gilman, Karen Levy & Alice Marwick, *Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans*, 95 WASH. U. L. REV. 53, 55-56 (2017); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2221 (2019); Frank Pasquale & Danielle Keats Citron, *Promoting Innovation While Preventing Discrimination: Policy Goals for the Scored Society*, 89 WASH. L. REV. 1413, 1415 (2014); Roger W. Reinsch & Sonia Goltz, *Big Data: Can the Attempt To Be More Discriminating Be More Discriminatory Instead?*, 61 ST. LOUIS U. L.J. 35, 37-38 (2016); Tal Z. Zarsky, *Understanding Discrimination in the Scored Society*, 89 WASH. L. REV. 1375, 1385 (2014); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1024-25 (2017) (book review).

[195] ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY, https://facctconference.org [https://perma.cc/Q65E-URC3] (last visited August 15, 2020).

[196] *See generally* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 669-701 (2017) (outlining eight steps involved in creation of machine learning systems).

[197] *See* Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 105, at 677-94.

and, where unintentional, the model may even have been employed with the specific intent of making decision-making fairer by removing human bias.[198]

So far, the legal discussions of algorithmic bias have focused, quite reasonably, on contexts where discrimination is statutorily prohibited, such as employment and credit, as well as contexts regulated by constitutional law, such as policing and criminal justice. But the discrimination present in algorithmic decision-making has implications for negligence as well. This is true not because discrimination law is a species of tort law[199] but because AI is a technology that actually operates differently on different people. If a driver hits someone with a car or a demolition crew damages a home, the race or gender of the plaintiff has no bearing on the facts of the case, at least at the liability stage. With an AI-mediated injury, those traits may well be causally linked to the injury.

To demonstrate, let us consider the example of medical diagnosis. The medical profession knows less about women's bodies and ailments than it does men's.[200] Clinical trials often do not include enough women, so it is often unclear how drugs might affect women's bodies differently.[201] Women's reports

[198] *See* Alex P. Miller, *Want Less-Biased Decisions? Use Algorithms*, HARV. BUS. REV. (July 26, 2018), https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms.

[199] Title VII is commonly referred to as a "statutory tort." *See* Charles A. Sullivan, *Tortifying Employment Discrimination*, 92 B.U. L. REV. 1431, 1432 (2012). Title VII also is sometimes considered a "contemporary extension[] of tort law." Goldberg & Zipursky, *supra* note 13, at 919; *see also* Sandra F. Sperino, *Discrimination Statutes, the Common Law, and Proximate Cause*, 2013 U. ILL. L. REV. 1, 35 ("In [a] general sense, it is appropriate to conceive of federal employment discrimination law as a 'contemporary extension[] of tort law' and placing it within this general category is not problematic." (second alteration in original) (footnote omitted) (quoting Goldberg & Zipursky, *supra* note 13, at 919)). *But see* Sandra F. Sperino, *The Tort Label*, 66 FLA. L. REV. 1051, 1052-54 (2014) (arguing that "tort label," though common, is inaccurate and undermines discrimination law).

[200] *See, e.g.*, PAULA A. JOHNSON, THERESE FITZGERALD, ALINA SALGANICOFF, SUSAN F. WOOD & JILL M. GOLDSTEIN, SEX-SPECIFIC MEDICAL RESEARCH: WHY WOMEN'S HEALTH CAN'T WAIT 3 (2014), https://www.brighamandwomens.org/assets/BWH/womens-health/pdfs/ConnorsReportFINAL.pdf [https://perma.cc/6F2M-RAQ2] ("To ignore [sex] differences challenges the quality and integrity of science and medicine."); Vivian W. Pinn, *Sex and Gender Factors in Medical Studies: Implications for Health and Clinical Practice*, 289 JAMA 397, 397 (2003) ("Although there are arguments that women's health issues have not been studied less than men's health issues, the prevailing lack of information about sex and gender differences or similarities in health and disease has been documented in many publications." (footnote omitted)).

[201] *See, e.g.*, Chiara Melloni, Jeffrey S. Berger, Tracy Y. Wang, Funda Gunes, Amanda Stebbins, Karen S. Pieper, Rowena J. Dolor, Pamela S. Douglas, Daniel B. Mark & L. Kristin Newby, *Representation of Women in Randomized Clinical Trials of Cardiovascular Disease Prevention*, 3 CIRCULATION 135, 135 (2010) ("[T]here remains a concerning gap in the knowledge, understanding, and general awareness of [cardiovascular disease] in women."); Vivek H. Murthy, Harlan M. Krumholz & Cary P. Gross, *Participation in Cancer Clinical Trials: Race-, Sex-, and Age-Based Disparities*, 291 JAMA 2720, 2720 (2004) (concluding

of pain are questioned more frequently by doctors, and women's pain issues are frequently misdiagnosed.[202] Unsurprisingly, these disparities also exist along racial lines[203] and are compounded in intersectional cases.[204]

As discussed above, AI medical diagnostic tools can reduce error rates as compared to doctors.[205] But as in other contexts, medical AI can reproduce or potentially exacerbate human biases.[206] Medical diagnostic tools can use a wide

that women, among other minority groups, were less likely to participate in cancer trials); Amy Westervelt, *The Medical Research Gender Gap: How Excluding Women from Clinical Trials Is Hurting Our Health*, GUARDIAN (Apr. 30, 2015, 3:32 PM), https://www.theguardian.com/lifeandstyle/2015/apr/30/fda-clinical-trials-gender-gap-epa-nih-institute-of-medicine-cardiovascular-disease [https://perma.cc/AUY3-HLZB] (citing example that only one-third of cardiovascular clinical trial subjects are female).

[202] Diane E. Hoffmann & Anita J. Tarzian, *The Girl Who Cried Pain: A Bias Against Women in the Treatment of Pain*, 29 J.L. MED. & ETHICS 13, 13 (2001); *see also* A.C. Pustilnik, *Imaging Brains, Changing Minds: How Pain Neuroimaging Can Inform the Law*, 66 ALA. L. REV. 1099, 1137 (2015) ("Chronic pain claims, like claims of sexual victimization, have long invited doubt and even presumptions of fabrication.").

[203] *See generally* INST. OF MED., UNEQUAL TREATMENT: CONFRONTING RACIAL AND ETHNIC DISPARITIES IN HEALTH CARE (Brian D. Smedley, Adrienne Y. Stith & Alan R. Nelson eds., 2003) (noting that racial and ethnic minorities tend to receive lower quality of healthcare even when insurance status and income are controlled for).

[204] *See* Yolonda Wilson, Amina White, Akilah Jefferson & Marion Danis, *Intersectionality in Clinical Medicine: The Need for a Conceptual Framework*, 19 AM. J. BIOETHICS 8, 10 (2019); P.R. Lockhart, *What Serena Williams's Scary Childbirth Story Says About Medical Treatment of Black Women*, VOX (Jan. 11, 2018, 4:40 PM), https://www.vox.com/identities/2018/1/11/16879984/serena-williams-childbirth-scare-black-women [https://perma.cc/3GBN-7PZD] ("Black women are disproportionately likely to face [pregnancy-related] complications, and they are also more likely to fall victim to America's ongoing maternal mortality crisis . . . .").

[205] *See supra* note 92 and accompanying text.

[206] *See generally* KADIJA FERRYMAN & MIKAELA PITCAN, DATA & SOC'Y, FAIRNESS IN PRECISION MEDICINE (2018), https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf [https://perma.cc/8VJE-AFSY] (discussing biases in computing health data); I. Glenn Cohen & Harry S. Graver, *Cops, Docs, and Code: A Dialogue Between Big Data in Health Care and Predictive Policing*, 51 U.C. DAVIS L. REV. 437 (2017); Sarah E. Malanga, Jonathan D. Loe, Christopher T. Robertson & Kenneth S. Ramos, *Who's Left Out of Big Data?: How Big Data Collection, Analysis, and Use Neglect Populations Most in Need of Medical and Public Health Research and Interventions*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS, *supra* note 67, at 98 (arguing that data must be broadened to increase diversity and reflect population's heterogeneity to ensure marginalized communities secure healthcare advantages and benefits that Big Data can provide). The techniques can also be used to *find* gender-based errors and alert people to them, and such uses should indeed be encouraged. *See* Jennifer Bresnick, *Big Data Shows Gender-Based Medical Error, Patient Safety Patterns*, HEALTH IT ANALYTICS (Sept. 22, 2016), https://healthitanalytics.com/news/big-data-shows-gender-based-medical-error-patient-safety-patterns [https://perma.cc/NQ6M-CN94].

range of inputs. They include data amassed from patient records, newer electronic medical records, patient DNA, insurance claims, medical sensors, and wearables.[207] Indeed, as rapidly as the field is growing, it is almost difficult to imagine what data companies will not try to incorporate, including "diet information [and] social factors," among other information.[208] The end result of vacuuming up as much data as possible, with no theory of why and how it is representative, will be the same as everywhere else it happens: algorithmic biases that are often hidden, invisibly caused, and difficult to correct.[209]

Consider a hypothetical diagnostic tool that is more accurate overall but less reliable for women than men at diagnosing a subset of conditions. Given current trends, it is realistic to assume that the benefits of AI will be distributed unevenly and that men's results will improve more than women's. For the purposes of illustration, let us assume that the AI provides minimal but positive improvement in detecting a given condition in women and a larger improvement in detecting the condition in men. Lastly, suppose that the user does not know the distribution because the manufacturer only tested for overall accuracy or does not provide useful documentation. Given these premises, a doctor will eventually use the tool without knowing about the gender imbalances within the AI, and she will ultimately misdiagnose a woman, leading to the patient's death. How would the use of the AI change the medical malpractice determination?

Without the use of AI, a similarly positioned doctor would have seen a patient and failed to diagnose the condition, resulting in the patient's death. Assuming the condition is one that the reasonable doctor would diagnose correctly, this appears to be classic malpractice, an easy case.[210] The fact that the patient was a woman would not—generally speaking—enter the liability calculus at all. Even though the malpractice would be more likely with respect to women, this is reflected in the fact that more women obtain judgments than men, not that any individual suit takes account of gender.[211]

---

[207] *See, e.g.*, Price, *Regulating Black-Box Medicine*, *supra* note 91, at 427; Chris Rauber, *Lumiata Nabs $6 Million for Personalized Medical Care Software*, S.F. BUS. TIMES (Sept. 11, 2014, 7:04 AM), http://www.bizjournals.com/sanfrancisco/blog/2014/09/lumiata-6-million-funding-personalized-health-data.html [https://perma.cc/Z5UP-DQ28].

[208] Mark van Rijmenam, *Three Innovative Ways How Big Data Will Improve the Healthcare Industry*, DATAFLOQ (Aug. 8, 2013, 8:00 PM), https://datafloq.com/read/three-innovative-ways-big-data-will-improve-health/165 [https://perma.cc/WW3U-ZSPE].

[209] *See* sources cited *supra* note 194.

[210] *See* Michael D. Greenberg, *Medical Malpractice and New Devices: Defining an Elusive Standard of Care*, 19 HEALTH MATRIX 423, 427 (2009); Karyn K. Ablin, Note, *Res Ipsa Loquitur and Expert Opinion Evidence in Medical Malpractice Cases: Strange Bedfellows*, 82 VA. L. REV. 325, 327-28 (1996).

[211] Research by Thomas Koenig and Michael Rustad in the 1990s found that two-thirds of punitive damages awards for medical malpractice and a majority of compensatory damage awards in medical products liability cases go to women. Thomas Koenig & Michael Rustad, *His and Her Tort Reform: Gender Injustice in Disguise*, 70 WASH. L. REV. 1, 61 (1995).

Now consider the AI-assisted case. The same thing occurs; there is a misdiagnosis, and the patient dies. How do we determine whether malpractice occurred? Here, the AI has directly incorporated the medical field's bias. Statistically, many more women will experience medical error than men; in fact, the AI widened the gap by hypothesis. But from the doctor's perspective, her work is identical whether the patient is a man or woman: she takes the scans, feeds them to the AI, and the AI reads out the result. Occasionally there will be errors, but recall from the discussion in Section II.A that except in extreme cases, there may be no clear way for a doctor to question the AI or determine which diagnoses are errors.[212] In this hypothetical, the choice to use the AI will not by itself lead to negligence liability because the AI is—by stipulation—safer than the doctor overall for both women and men. Ultimately, under both scenarios, roughly the same number of women are injured—because the AI's improvement for women was minimal—but in the case of AI use, the injured women suddenly cannot recover in tort.

Scholars have extensively documented ways in which tort law reproduces race and gender hierarchies.[213] Professor Martha Chamallas has argued that more "masculine" physical and pecuniary harms are considered more important than more "feminine" emotional harm[214] and that Black lives and women's activities are devalued, leading to smaller damage awards.[215] Professor Leslie

[212] *See supra* Section II.A (discussing unforeseeability of AI errors).

[213] *See* Naomi R. Cahn, *The Looseness of Legal Language: The Reasonable Woman Standard in Theory and in Practice*, 77 CORNELL L. REV. 1398, 1404-06 (1992) (critiquing "reasonable man" standard); Koenig & Rustad, *supra* note 211, at 58-61. *See generally* MARTHA CHAMALLAS & JENNIFER B. WRIGGINS, THE MEASURE OF INJURY: RACE, GENDER, AND TORT LAW (2010) (discussing how women and minorities have been undercompensated in tort law and that traditional biases have resurfaced in updated forms to perpetuate patterns of disparate recovery based on race and gender).

[214] Martha Chamallas, *The Architecture of Bias: Deep Structures in Tort Law*, 146 U. PA. L. REV. 463, 521-30 (1998) [hereinafter Chamallas, *The Architecture of Bias*] (arguing that social construction of tort categories creates "vicious cycle" where gendered concepts inform definitions); Martha Chamallas & Linda K. Kerber, *Women, Mothers, and the Law of Fright: A History*, 88 MICH. L. REV. 814, 862-64 (1990); *see also* Lucinda M. Finley, *The Hidden Victims of Tort Reform: Women, Children, and the Elderly*, 53 EMORY L.J. 1263, 1266 (2004) ("Noneconomic loss damage caps therefore amount to a form of discrimination against women and contribute to unequal access to justice or fair compensation for women. . . . [W]omen, on average, recover more in noneconomic damages . . . [because] injuries that happen primarily to women are compensated predominantly or almost exclusively through noneconomic loss damages.").

[215] Martha Chamallas, *Civil Rights in Ordinary Tort Cases: Race, Gender, and the Calculation of Economic Loss*, 38 LOY. L.A. L. REV. 1435, 1439 (2005) ("As a practical matter, the use of race and gender-based tables results in significantly lower awards for minority men and women of all races."); Chamallas, *The Architecture of Bias*, *supra* note 214, at 471-80 (citing examples where non-white persons' injuries are statistically "devalued" compared to white persons' injuries). *See generally* Ronen Avraham & Kimberly Yuracko,

Bender has critiqued the "reasonable person" standard as a male standard because it was renamed from "reasonable man" without any of the underlying law changing.[216] Bender has also critiqued how, in certain important cases, foreseeability clearly takes a man's perspective.[217] The AI bias problem here is similar, but distinct. The concerns that Chamallas, Bender, and others raise tend to be the result of analyzing trends across multiple cases. In the AI scenario, however, the problem is that negligence law is ill-equipped to address differing results in a single case where the facts are the same except for the gender or race of the potential plaintiff. I locate this problem in the element of duty because it is the place where public policy considerations most explicitly enter the picture of negligence law.[218] Duty is, depending on the view, owed to the specific plaintiff or to the world,[219] but there is no duty to ensure distributional fairness in individual case outcomes.[220]

This presents a difficult duty issue because it is not obvious what the correct outcome is as a normative matter. Not only does AI import data about potential plaintiffs into the fact pattern, but it imports controversy from discrimination law. If a decision improves diagnoses but improves among some groups more than others, is it morally or ethically wrong? *Should* tort law prohibit these particular kinds of advances? On the one hand, AI can save lives, but on the other hand, it would mostly save men's lives. This problem with tort law and technological advancement is certainly not limited to medical AI; a recent study suggests that vision systems in autonomous vehicles may have an easier time

Torts and Discrimination, 78 OHIO ST. L.J. 661 (2017) (arguing that wage tables create incentives to commit torts against people based on race and gender).

[216] Leslie Bender, *A Lawyer's Primer on Feminist Theory and Tort*, 38 J. LEGAL EDUC. 3, 20-25 (1988).

[217] *See* Leslie Bender & Perette Lawrence, *Is Tort Law Male?: Foreseeability Analysis and Property Managers' Liability for Third Party Rapes of Residents*, 69 CHI.-KENT L. REV. 313, 336 (1993) (concluding that distrust of women's factual experiences causes tort law to conclude women's foreseeability analysis as legally incoherent).

[218] W. Jonathan Cardi, *The Hidden Legacy of* Palsgraf*: Modern Duty Law in Microcosm*, 91 B.U. L. REV. 1873, 1878 (2011) [hereinafter Cardi, *The Hidden Legacy of* Palsgraf]; Mark A. Geistfeld, *Social Value as a Policy-Based Limitation of the Ordinary Duty to Exercise Reasonable Care*, 44 WAKE FOREST L. REV. 899, 903-07 (2009).

[219] Cardi, *The Hidden Legacy of* Palsgraf, *supra* note 218, at 1877-78.

[220] There is no reason there would be such a duty. Moreover, as Professor W. Jonathan Cardi's analysis suggests, in employment discrimination cases—for which there is a statutory mandate to not discriminate—the Supreme Court often performs the equivalent of a duty analysis in mixed-motive cases, finding no duty to not discriminate. W. Jonathan Cardi, *The Role of Negligence Duty Analysis in Employment Discrimination Cases*, 75 OHIO ST. L.J. 1129, 1137 (2014).

detecting light-skinned people, even if they improve safety overall.[221] These questions will lead to a great deal of disagreement over how much overall improvement is permissible to trade off against a discriminatory result and mirrors a debate in discrimination law that sounds in reasonableness and trade-offs about how much cost a decisionmaker such as an employer must bear to ensure nondiscrimination.[222] AI presents the new and difficult normative question of total lives saved as weighed against racial and gender disparities.

III. WHY NEGLIGENCE LAW MAY NOT JUST ADAPT TO AI

The last Part identified four challenges that AI presents when it interacts with activities regulated by negligence. In this Part, I relate these challenges to observations about the structure and operation of negligence law in order to explore whether they are temporary problems that will be addressed with time or whether they are more fundamental challenges. I argue that although negligence law can often adapt to new technologies, the incomprehensibility of AI, extreme corporate secrecy, and AI's replacement of individualized decision-making with statistical reasoning will make it difficult to develop legal standards without outside intervention. I briefly propose interventions that can mitigate some of the challenges, but there is no real way to know yet whether they would work.

A. *Negligence, Bounded Rationality, and AI*

It is a fundamental tenet of negligence law that one cannot be liable for circumstances beyond what the reasonable person can account for.[223] People often lack the information necessary to determine the safest possible course of action, and even if they had it, they could not process all of the information to incorporate it into the decision. This is an instance of the well-known concept of "bounded rationality."[224] First advanced by economist Herbert Simon, the theory modified the concept of the perfectly rational human decisionmaker and eventually ushered in the field of behavioral economics.[225]

---

[221] Benjamin Wilson, Judy Hoffman & Jamie Morgenstern, Predictive Inequity in Object Detection 9 (Feb. 21, 2019) (unpublished manuscript), https://arxiv.org/pdf/1902.11097.pdf [https://perma.cc/8KWR-YTH6].

[222] *See* Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 105, at 728 (arguing that data mining will force "a discussion about what constitutes a tolerable level of disparate impact"); Ernest F. Lidge III, *Financial Costs as a Defense to an Employment Discrimination Claim*, 58 ARK. L. REV. 1, 32-37 (2005); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 191-93 (2017).

[223] *See, e.g.*, Palsgraf v. Long Island R.R., 162 N.E. 99, 99 (N.Y. 1928).

[224] *See* Peter M. Todd & Gerd Gigerenzer, *Bounding Rationality to the World*, 24 J. ECON. PSYCHOL. 143, 144 (2003) (defining "bounded rationality" as combination of external and internal limits on cognitive ability).

[225] *Id.*

Though it is not typically described that way, foreseeability is how negligence law compensates for bounded rationality.[226] Because "almost any outcome is, by some stretch of the imagination, foreseeable," judgments about foreseeability are doing something other than debating whether a risk was literally capable of being foreseen.[227] Rather, as Professor W. Jonathan Cardi explains, "this decision is part fact-finding—determining what the ordinary person would foresee under these circumstances—and part philosophical exercise—deciding what level of epistemic probability should open the door to liability."[228] Phrased another way, these two inquiries correspond directly to the two facets of bounded rationality: (1) What is it reasonable for a person to know?; and (2) How much can we reasonably expect them to be able to process? If perfect information and rationality were possible, foreseeability limitations would be unnecessary.

Decision-assistance AI is most useful when it targets exactly the same concerns; it is usually sold as a remedy to the weaknesses of human decision-making. Humans cannot hold or process nearly as much information as a computer, so AI is a way to take more data into account and make use of it, resulting in "better" decisions. And it works, to an extent. But AI does not actually *solve* bounded rationality; rather, it transforms the problem. The unforeseeable nature of AI errors discussed in Section II.A are a direct result of our inability to process all of the information that an AI system uses.[229] So instead of the traditional doctrinal issues of the foreseeability of the risk, plaintiff, or category of harm, bounded rationality transforms into an inability to completely oversee or understand the AI decision mechanism.

Though the doctrinal entry point is different, the implications for negligence are familiar. Foreseeability is a complex doctrine best explained as a series of policy judgments about the extent to which society should demand the processing of remote possibilities.[230] To say something is unforeseeable is to rule that the reasonable person either could not or need not have taken a given

[226] *See* Shyamkrishna Balganesh, *Foreseeability and Copyright Incentives*, 122 HARV. L. REV. 1569, 1574 (2009) ("Foreseeability connects . . . to the notion of bounded rationality. When certain events or consequences are unlikely to have formed a significant part of an actor's decisions for an action, the law characterizes them as unforeseeable and avoids attributing them to the actor. In economic terms, foreseeability thus enables courts to distinguish between events that are likely to have formed part of an actor's ex ante incentives for action and those that are unlikely to have done so, thereby restricting recovery to the former alone.").

[227] Cardi, *Reconstructing Foreseeability*, *supra* note 140, at 951; *see also* Zipursky, *supra* note 128, at 1256 (arguing that "reasonable foreseeability" is binary and that litigants argue for one side of binary).

[228] Cardi, *Reconstructing Foreseeability*, *supra* note 140, at 940.

[229] *See supra* Section II.A.

[230] Cardi, *Purging Foreseeability*, *supra* note 77, at 762-63; Owen, *Figuring Foreseeability*, *supra* note 78, at 1293.

possibility into account. In the case of AI, the policy judgment is instead about whether there was good reason to question the AI.

Unlike the relatively infrequent cases that give rise to established foreseeability issues, however, the unforeseeable nature of AI errors risks being the exception that swallows the rule. For now, there is no legal requirement for the AI system to be interpretable or explainable. Most AI advocates promote it on the basis that it will be safer as a statistical matter, which does not imply that the AI must be interpretable or explainable. But without interpretable or explainable AI, it is essentially impossible to claim that an AI error should have been foreseen ahead of time. Thus, if AI-error foreseeability is treated like traditional foreseeability, injured plaintiffs would be unable to recover as the rule.

If AI is made interpretable or explainable in some way, then the foreseeability question at least becomes a practical inquiry. Even at that point, however, foreseeable cases will be the exception. Recall the discussion of interpretability and explainability from Section II.A.1. There are many different methods of achieving interpretability or explainability, all of which conflict with each other and work differently in different contexts.[231] In the pneumonia-asthma example from Caruana and colleagues,[232] because asthma was linked to pneumonia through breathing, one might say that injuries due to the AI focusing on asthma would be foreseeable because the link between pneumonia and asthma is intuitive to a doctor.[233] But if the discovered relationship were not as intuitively linked, the injuries would not be foreseeable because a doctor looking at the AI's recommendation could not know if and where the AI went wrong.[234]

We should therefore, at least for the moment, expect that the nonintuitive relationship is the rule rather than an exception. This is because most of the value of machine learning comes from its ability to discover precisely these types of relationships in the data. If we did not think that these nonintuitive relationships were commonly discoverable by AI, there would be less motivation to use AI in the first place, as human experts would be as good or better. Consequently, it is in precisely the contexts where human limitations currently cause the most injuries that demand for AI will be the greatest. Thus, though the injury rates may improve overall with AI, the people who *are* injured—and there may still

---

[231] *See* Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1109-17 (discussing research on interpretability and explainability within field of computer science). Methods include restricting the input variables, paring down models to have fewer internal variables, creating simplified versions of machine learning models, and offering counterfactual inputs as explanations. Each of these methods has strengths and weaknesses.

[232] *See supra* notes 114-20 and accompanying text.

[233] *See* Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1126 (noting that intelligibility problems force default rules, some of which "presume that obviously *correct* relationships will show themselves, so that everything else should be discarded by default").

[234] *See id.*

be many—will be without remedy if negligence treats AI errors as functionally unforeseeable.

This suggests that there will always be some subset of cases that will leave an injured person without recovery. But that alone does not imply that a negligence regime is irreparably broken. After all, the existence of injuries without remedy is equally true today—we just call them accidents. Negligence fails only if such uncompensated cases are common enough that society perceives that AI users should be held to account to prevent more injuries. Thus, if negligence is to be successful, the goal will be to reduce the number of cases that are truly unforeseeable accidents. While this may happen over time with most technology, with AI, if this will be possible at all, it will likely require outside interventions meant to reduce the number of cases that we cannot say anything at all about, such as requiring that machines are built interpretably and transparently, accompanied by documentation.[235]

B. *Updates to Reasonableness with Familiarity and Access*

Society's understanding about what behavior counts as reasonable evolves over time. Negligence law's readjustment in the face of new technology is not only a common occurrence but arguably the primary driving force behind the continued development of negligence law in general.[236] A challenge with technological disruption is not only that the injury rate increases compared to before the technology—it may actually decrease—but also that the events leading to injuries constitute new fact patterns about which we have no intuitions. As we gain familiarity with a new sociotechnical environment, we start to develop new intuitions about what counts as reasonable or unreasonable behavior, as well as what previously unforeseeable events become more regular.[237] Within an industry, new intuitions can be formalized with best practices and customs, and industry custom can become evidence of reasonable care.

What would this look like with respect to the foreseeability of AI errors? One hopeful possibility is that patterns in AI errors will develop over time, and people will learn that the AI gets certain types of cases wrong more often than others. Eventually, if people can better figure out under which circumstances it is appropriate to trust the AI, then we might settle back into a negligence regime because cases will again be differentiable in a fault-based sense. As with the

---

[235] *See id.* at 1134-38.

[236] Froomkin, Kerr & Pineau, *supra* note 67, at 51 ("U.S. tort law recognizes that technology changes what is possible and reasonable, and thus the general standard of care for professions and trades may change too."); Donald G. Gifford, *Technological Triggers to Tort Revolutions: Steam Locomotives, Autonomous Vehicles, and Accident Compensation*, 11 J. TORT L. 71, 142 (2018) ("[W]aves of technological change account for the most significant changes in American tort law . . . ."); Grady, *supra* note 1, at 293.

[237] Mary L. Lyndon, *Tort Law and Technology*, 12 YALE J. ON REG. 137, 141 (1995).

epistemic challenge and foreseeability, the HCI and operational security concerns may also be temporary issues of adaptation, where familiarity will lead to the development of common-law standards of how people should interact with AI. Though the field of HCI has been around for a few decades, the research into human interactions with modern AI systems is still brand new.[238] As mentioned above, the same is true of adversarial machine learning.[239]

Certainly, more familiarity leads to ideas of what proper and improper use looks like. There are, however, three aspects of the current AI landscape that raise doubt that such realignment will occur for AI without intervention: opacity, context-dependence, and speed. The first is a direct result of what was discussed above—if no intervention is made to mandate transparency, interpretability, or explainability, then no one outside the companies will ever know how AI tends to fail, and we will be unable to have regular or concrete enough stories of failure to update our notions of reasonableness.

The famed secrecy of the modern technology industry does not help. Secrecy is more of a concern with AI—and to an extent, software generally—than with older technologies.[240] This is partly driven by business practices; AI companies rely enormously on secrecy to protect their financial interests.[241] There are likely many reasons for this. One is that, as Professor Sonia Katyal has shown, copyright and patent protection for software has tightened and become unpredictable, leading software companies to rely on trade secrecy over other forms of intellectual property.[242] Another is that two practical barriers to trade secrecy claims do not apply to AI. Trade secret claims give way to independent

[238] *See generally* Eric P.S. Baumer, *Toward Human-Centered Algorithm Design*, BIG DATA & SOC'Y, July-Dec. 2017, at 1 (discussing issue of opacity in algorithms). The organizers of a workshop at the major HCI conference explain: "[D]espite the importance of people in the development, deployment, and use of AI systems, Human Computer Interaction (HCI) is often not a core component of these research questions . . . . [More] comprehensive inclusion of HCI's unique perspectives are essential to solving these challenging societal questions. Therefore, through this workshop, we ask the fundamental question: *Where is the human in AI research?*" Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale & Eric P.S. Baumer, *Where is the Human? Bridging the Gap Between AI and HCI*, PROC. 2019 CONF. ON HUM. FACTORS COMPUTING SYS. 1, 3 (2019).

[239] *See supra* notes 171-78 and accompanying text.

[240] *See* Sonia K. Katyal, *The Paradox of Source Code Secrecy*, 104 CORNELL L. REV. 1183, 1188 (2019).

[241] *See, e.g.*, Jeanne C. Fromer, *Machines as the New Oompa-Loompas: Trade Secrecy, the Cloud, Machine Learning, and Automation*, 94 N.Y.U. L. REV. 706, 722-24 (2019); W. Nicholson Price II, *Big Data, Patents, and the Future of Medicine*, 37 CARDOZO L. REV. 1401, 1432-36 (2016) [hereinafter Price, *Big Data*].

[242] Katyal, *supra* note 240, at 1191-1236; Samuel J. LaRoque, Comment, *Reverse Engineering and Trade Secrets in the Post*-Alice *World*, 66 U. KAN. L. REV. 427, 431-35 (2017).

discovery of the secret and reverse engineering of technologies.[243] But while other technologies can be pulled apart, analyzed, stress tested, and reverse engineered, AI's inscrutability makes reverse engineering difficult, if not impossible.[244] Yet another reason might be that trade secret claims are being offered more frequently in contexts where they are less likely to be challenged, such as criminal court.[245] And finally, Congress strengthened trade secret law with the Defend Trade Secrets Act ("DTSA") in 2016.[246] For some combination of these reasons, software companies have moved toward secrecy as a more profitable way to protect their investments.[247] As Professors Robert Brauneis and Ellen P. Goodman have documented, even where algorithms are used in public settings—where presumably the interests in transparency and due process are highest—companies aggressively pursue trade secrecy claims and often require cities to sign non-disclosure agreements.[248]

As a result of the secrecy, we know little of what individual companies have learned about the errors and vulnerabilities in their products. Under these circumstances, it is impossible for the public to come to any conclusions about what kinds of failures are reasonable or not. Even if industry-wide best practices are adopted, there is little indication that knowledge of what happens when those best practices are not followed—and thus which types of errors and injuries are blameworthy—will be made public. If litigants hire experts, an engineer from one company will not know anything about how the model in another company was constructed because the data and the testing is all kept secret, and knowing

[243] Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 475-76 (1974) ("A trade secret law, however, does not offer protection against discovery by fair and honest means, such as by independent invention, accidental disclosure, or by so-called reverse engineering, that is by starting with the known product and working backward to divine the process which aided in its development or manufacture.").

[244] *See* Fromer, *supra* note 241, at 723 ("It is also essentially impossible to reverse engineer these data because they are not discernable from any commercially available software based on machine learning, precisely because they are not contained within the software and because any predictive model built on these data is likely to be too complex to convert back into even a rough approximation of the underlying data."); Michael Mattioli, *Disclosing Big Data*, 99 MINN. L. REV. 535, 553 (2014) ("The recent commentary describing big data's disclosure problem indicates that, unlike software, big data practices cannot be reverse-engineered. That is, an expert cannot decipher just how a set of data was assembled with nothing more to work from than the data itself." (footnote omitted)).

[245] *See* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1396-99 (2018).

[246] 18 U.S.C. § 1836 (2018).

[247] *See generally* Fromer, *supra* note 241; Katyal, *supra* note 240.

[248] Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103, 153 (2018); LaRoque, *supra* note 242, at 435; Wexler, *supra* note 245, at 1350 (noting that after *Alice Corp. Pty. v. CLS Bank Int'l*, 573 U.S. 208 (2014), it became difficult to patent software, causing software firms to turn to trade secrets).

the algorithm or even the source code without the data is not useful.[249] Discovery will not necessarily solve this either; in criminal cases, where defendants probably have their best possible claims for access to the algorithms, judges have deferred to trade secret claims by companies, not by requiring protective orders but by denying defendants access to the "secret" information at all.[250] And even if the code is held to be discoverable, it is often nearly impossible for even experts to trace the inner workings of unfamiliar source code.[251] We need a lot more information about the structure and function of AI systems before we can assume the negligence regime will adapt.

Second, assuming we get the necessary access to information, there is still a risk that patterns may not actually develop in the types of errors that AI produces. Unlike traditional machines, which are stable once they come from the factory, AI is shipped incomplete. It is designed to update its behavior with new data. The operation of AI, if done correctly, is highly context sensitive, and therefore each version coming out of the factory will be trained on local, contextual data and will run differently than others.[252] A more concerning possibility is that, in certain industries, AI tools will be predominantly built in-house, resulting in AI becoming even more dissimilar. Given the complexity of AI and the lack of consensus on how to make systems interpretable, society cannot count on patterns of emerging errors, allowing the law to distinguish errors that it can and cannot fault the user for.

There are efforts underway that may help. Researchers are developing benchmarking and documentation systems that can set standards within the industry.[253] Large membership organizations like the Institute for Electrical and

[249] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 638, 649-50 (2017).

[250] *See* Natalie Ram, *Innovating Criminal Justice*, 112 NW. U. L. REV. 659, 672-73 (2018); Wexler, *supra* note 245, at 1358-59 (discussing case in which death penalty defendant was denied access to source code for forensic software used to convict him).

[251] Kroll et al., *supra* note 249, at 638, 649-50.

[252] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian & Janet Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, PROC. CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 59, 61, 66 (2019).

[253] Matthew Arnold et al., *FactSheets: Increasing Trust in AI Services Through Supplier's Declarations of Conformity*, IBM J. RES. & DEV., July-Sept. 2019, at 1, 1 ("We envision [FactSheets] to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers."); Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III & Kate Crawford, *Datasheets for Datasets*, PROC. 5TH WORKSHOP ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY MACHINE LEARNING 1, 1 (2018) (recommending that datasets "be accompanied with a datasheet documenting its motivation, creation, composition, intended uses, distribution, maintenance, and other information"); Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph & Kasia Chmielinski, *The Dataset Nutrition Label: A*

Electronic Engineering[254] and the Partnership on AI[255] are working to develop best practices for AI use, benchmarking, and documentation. If these standards are adopted and publicly adhered to, the public will at least have some idea of what the software industry thinks best practices are. But while best practices put the industry on a path much likelier to result in consensus, the fact that local and contextual data changes the AI's behavior might remain a problem. Best practices will still not explain what kinds of errors are most likely to occur in specific contexts—e.g., medicine, finance, security, and driving, and how those different errors change with geographic or demographic changes to the data. If the errors look too unique, it will be difficult to build up a common law notion of reasonableness.

Third is the relative speed of development of AI and tort law. It is a well-known maxim that technology outpaces legal development,[256] a gap that is only widening over time.[257] While there might not be anything natural about the so-called "pacing problem"[258]—rather it is an artifact of legal culture that starts

Framework to Drive Higher Data Quality Standards, *in* 12 DATA PROTECTION AND PRIVACY 1, 1 (Dara Hallinan, Ronald Leenes, Serge Gutwirth & Paul De Hert eds., 2020) (proposing "Dataset Nutrition Label" with information about data analogous to that on food and drug labels); Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji & Timnit Gebru, *Model Cards for Model Reporting*, PROC. CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 220, 220 (2019) ("Model cards . . . disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information.").

[254] *See generally* IEEE, ETHICALLY ALIGNED DESIGN: A VISION FOR PRIORITIZING HUMAN WELL-BEING WITH AUTONOMOUS AND INTELLIGENT SYSTEMS (2019) (describing IEEE's standardization efforts for ethical AI). The IEEE describes itself as "the largest technical professional organization dedicated to advancing technology for the benefit of humanity." *Id.* at 13.

[255] *ABOUT ML*, PARTNERSHIP ON AI, https://www.partnershiponai.org/about-ml/ [https://perma.cc/G38J-QJGJ] (last visited August 16, 2020). The Partnership on AI's ("Partnership") stated goals are to "[d]evelop and share best practices . . . in the research, development, testing, and fielding of AI technologies." *About Us*, PARTNERSHIP ON AI, https://www.partnershiponai.org/about/ [https://perma.cc/NP2P-E7R4] (last visited August 16, 2020). By way of disclosure, during my time as a postdoctoral scholar at Data & Society Research Institute, I served as a representative of Data & Society to the Partnership. In my current capacity, I continue to serve as a member of the steering committee for the Partnership's ABOUT ML project. My involvement with the Partnership has not affected this Article except to make me aware of the Partnership's activities.

[256] *See generally* Lyria Bennett Moses, *Recurring Dilemmas: The Law's Race to Keep Up with Technological Change*, 2007 U. ILL. J.L. TECH. & POL'Y 239.

[257] Gary E. Marchant, *The Growing Gap Between Emerging Technologies and the Law*, *in* THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT 19, 19-20 (Gary E. Marchant, Braden R. Allenby & Joseph R. Herkert eds., 2011).

[258] *See* Meg Leta Jones, *Does Technology Drive Law?: The Dilemma of Technological Exceptionalism in Cyberlaw*, 2018 U. ILL. J.L. TECH. & POL'Y 249, 251 (arguing that so-called

with a posture of noninterference—it seems to be a reality of the American legal system. Thus, as technological development intensifies, tort law may not have a chance to reimagine reasonableness before another wave of technological change is upon us.[259]

Secrecy and the pacing problem work together as well. Even with extreme secrecy in the technology industry, some information will get out eventually. Whistleblowers leak, and serious errors may get to the public with the help of the whistleblower protections in the DTSA.[260] Eventually, failures will happen in public, and even if the industry has sat on the research about potential harms for years, as the tobacco industry did, we will eventually learn about them. Secrecy alone may therefore not prevent adjudication and the eventual discovery of a notion of reasonableness. But it might *delay* the discovery significantly enough that the shape of the technology has changed before we settle on new understandings.

If the common law must rely on AI cases converging into patterns, best practices, and canonical stories, anything that could potentially speed up that process will make it more effective. To speed up tort law's responses, we could require or encourage—through tax incentives, safe harbors, or other levers commonly used in policy—disclosure and pooling of knowledge about the common types of AI failures to create a public repository of case studies.[261] One creative example comes from researchers who have recommended copying the idea of "bug bounties" in the computer security industry.[262] Bug bounties work by offering prizes to find errors in code, thereby deputizing members of the

"pacing problem" is "a form of technological determinism wherein technology drives social structures and cultural values").

[259] Compare this to the problem of surveillance and the rise of the internet. We can—and rightly do—argue about various normative concerns, but as privacy and security expert Bruce Schneier put it, "surveillance is the business model of the internet." *"Surveillance Is the Business Model of the Internet,"* OPENDEMOCRACY (July 18, 2017), https://www.opendemocracy.net/en/digitaliberties/surveillance-is-business-model-of-internet/ [https://perma.cc/7D3P-5DHP]; *see also* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM 8-12 (2019). There is a realistic argument that these conversations are happening too late and that, as a practical matter, the business model is too entrenched to reverse course just because people are harmed. That is the danger of waiting to regulate revolutionary technologies.

[260] *See* Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 136-37 (2019).

[261] *See, e.g.*, Price, *Regulating Black-Box Medicine*, *supra* note 91, at 465-72 (arguing for broad information disclosure to enable collaborative governance of medical AI).

[262] *See* Amit Elazari Bar On, *Private Ordering Shaping Cybersecurity Policy: The Case of Bug Bounties*, *in* REWIRED: CYBERSECURITY GOVERNANCE 231, 231-32 (Ryan Ellis & Vivek Mohan eds., 2019) (noting that practice of inviting hackers to perform penetration testing is becoming best practice in cybersecurity and is expanding in evolving market of vulnerabilities).

public to do after-market product testing.[263] Scholars have noted that this idea can be repurposed to apply to algorithmic harms.[264] These can also contribute to a public case study repository.

Another possible intervention is a risk-based analysis for medical malpractice and AI proposed by Professor W. Nicholson Price II.[265] Recall that the goal is to reduce the number of cases in which plaintiffs cannot recover for injuries because we do not know the level of care required. A risk-based analysis operates at a different layer of reasoning than previously discussed and can reduce the number of cases we are concerned with. For minimal-risk recommendations (e.g., extra monitoring or tests), Price proposes that the standard of care should require no special testing of the AI.[266] For riskier recommendations, such as powerful drugs, some sort of process-based validation would be required.[267] Moreover, for certain recommendations that we know to be wrong, such as prescribing thalidomide to pregnant women, the standard of care should never permit that result.[268] This differs from the prior discussion because the reasoning is not about when to use existing knowledge to overrule the AI, but rather about identifying a smaller subset of cases in which we even care about whether the AI is correct. This kind of risk-based analysis could be a threshold question for any negligence analysis of AI use. Practitioners are working on similar approaches in other AI contexts.[269] If the overall goal is to reduce the set of cases where we can have no liability because we cannot attribute error to fault, a risk-based approach will help.

If negligence law is to operate as expected for AI technologies, we must reach a point where we can distinguish blameworthy errors from those that are accidental and reduce the number of non-blameworthy errors to acceptable

[263] *See* Nathan Alexander Sales, *Privatizing Cybersecurity*, 65 UCLA L. REV. 620, 634-36 (2018).

[264] Amit Elazari Bar On, *We Need Bug Bounties for Bad Algorithms*, VICE (May 3, 2018, 10:00 AM), https://www.vice.com/en_us/article/8xkyj3/we-need-bug-bounties-for-bad-algorithms [https://perma.cc/8KQE-UX6S]; Calo, *supra* note 134, at 36; Price, *Big Data*, *supra* note 241, at 1451-52 (arguing that bounties could be implemented for validation purposes); *see also* WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES 164 (2018) (considering that "perhaps the bug hunting could extend to bad privacy design as well as security flaws").

[265] Price, *Medical Malpractice and Black-Box Medicine*, *supra* note 67, at 301; *see also* Greenberg, *supra* note 210, at 424 ("[A]t least some medically induced injuries may better be understood as resulting from complex medical care processes and inherent risk trade-offs . . . .").

[266] Price, *Medical Malpractice and Black-Box Medicine*, *supra* note 67, at 301.

[267] *Id.*

[268] *Id.* at 302.

[269] *See generally* BURT, SHIRRELL, LEONG & WANG, *supra* note 75 (providing template for effective managing of risks associated with machine learning to create better, more accurate, and compliant machine learning models).

limits. This requires that we update our standards of reasonable care, but when AI is developing without well-known failure modes, it will be difficult for the common law to keep up.

C. *Statistical Facts and Individual Responsibility*

So far, I have argued that AI does not interact with negligence law like typical technologies and that there is reason to believe that negligence will not adapt properly. But as discussed in the last Section, it is also possible that with assistance from new legislation, negligence will rebound. The concerns about AI bias are of a different nature entirely. The distributional concern articulated in Part II is similar to the distributional concerns that plague other parts of tort law, as has been extensively documented by scholars.[270] Arguably, the distributional concerns are not in the purview of negligence law at all, concerned as it is with individual responsibility for injury. On this issue, negligence law will not provide an effective answer.

There is a deeper pattern at work, as ever more decisions begin to incorporate AI tools. Specifically, AI employs statistical reasoning in areas of law where we aim to make individualized determinations. In negligence law, the determinations are about whether the individual defendant behaved unreasonably.[271] This is similar to anti-discrimination law. Other statistical-versus-individual challenges appear in the Fourth Amendment's requirement for individualized suspicion[272] or statistical facts in trial evidence.[273] In all of these cases, there is an uncomfortable tension between statistical facts and individual outcomes.

---

[270] *See supra* notes 213-20 and accompanying text (discussing ways in which tort law reproduces race and gender hierarchies).

[271] This is most obviously true under a fault-based approach to negligence. Under an economic approach, the individualized nature of the claim is less obviously important because causal links between breach and injury are not considered necessary, and statistically correct decisions can still optimize for efficient loss allocation. *See generally* William M. Landes & Richard A. Posner, *Causation in Tort Law: An Economic Approach*, 12 J. LEGAL STUD. 109 (1983) (arguing that economic analysis can resolve causation without relying upon causal concepts). Although the cost-benefit analysis defines the reasonable person standard, the individualized determination persists in the requirement that the jury determine whether the individual defendants in the case met their duties. *See* Epstein, *supra* note 12, at 164 ("The concept [of causation] may not be strictly necessary to the development of some theory of tort if the goal of the system is the minimization of the costs of accidents. But its presence reminds us that a system of law which tries to banish it from use may not respond to ordinary views on individual blame and accountability.").

[272] Selbst, *supra* note 222, at 154-57 (discussing how Fourth Amendment's individualization requirement interacts with predictive policing).

[273] *See generally* Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971) (demonstrating mismatch between statistical analysis and particular trial facts).

A brief discussion of the problem as it plays out in anti-discrimination law will demonstrate the parallel. In earlier work, Professor Solon Barocas and I argued that not only would the "necessarily subjective" choices that go into the creation of a machine learning model likely render it biased but that such biases would not be remediable by Title VII, the model for most of American anti-discrimination law.[274] As it is generally understood, Title VII contains two paths to liability for employment discrimination: disparate treatment and disparate impact.[275] Disparate treatment is primarily concerned with intentional discrimination,[276] and while it is possible to intentionally discriminate with machine learning systems, that is not the primary concern.[277] Disparate impact liability is not concerned with intent or motive to discriminate. Rather, it evaluates facially neutral policies with discriminatory effects, asking whether there was a justification for the decision mechanism despite the discriminatory effects.[278]

The doctrine consists of a three-part burden-shifting analysis. First, a plaintiff must demonstrate a disproportionate impact on a protected class.[279] Then, the defendant can respond by demonstrating that the decision mechanism was "job related for the position in question and consistent with business necessity."[280] And finally, in the case of a successful defense, the plaintiff may return with proof of an "alternative employment practice" that the employer "refuse[d]" to use but which was equally effective in the business objective and less discriminatory.[281]

The crux of the doctrine is the business necessity defense. This defense is complex, with many circuits establishing different definitions. While the upshot is that courts give an employer leeway to set criteria for the kind of employee they seek, the more stringent part of the defense is the requirement that the test

[274] Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 105, at 694-714.

[275] *See* 42 U.S.C. § 2000e-2 (2018); *see also* Noah D. Zatz, *Managing the Macaw: Third-Party Harassers, Accommodation, and the Disaggregation of Discriminatory Intent*, 109 COLUM. L. REV. 1357, 1368 (2009) ("Few propositions are less controversial or more embedded in the structure of Title VII analysis than that the statute recognizes only '"disparate treatment" and "disparate impact" theories of employment discrimination.'" (quoting Hazen Paper Co. v. Biggins, 507 U.S. 604, 609 (1993)). *But see* Kim, *supra* note 194, at 867.

[276] Disparate treatment is arguably divisible into two subdoctrines of formal and intentional discrimination, where "formal" refers to the use of protected class identifiers to make choices irrespective of outcome. *See* Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1351 (2010).

[277] *See* Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 105, at 712-14 (discussing masking and problems of proof for finding Title VII liability).

[278] *Id.* at 694.

[279] 42 U.S.C. § 2000e-2(k)(1)(A).

[280] *Id.*

[281] *Id.*

validly predict that trait.[282] The validation question, then, evaluates the machine learning model, asking how well it actually predicts the target variable. Here, courts usually turn to the Equal Employment Opportunity Commission's Uniform Guidelines on Employment Selection Procedures,[283] which prescribe three different validation criteria: "criterion-related, content, and construct validity."[284] Of the three, criterion-related validity is most applicable to machine learning.[285] It "consist[s] of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance."[286] This test so perfectly aligns with AI's predictive aims that, without context, it would not be obvious that the quoted language was the test for validity rather than a description of what the model itself aims to do. Because disparate impact doctrine ties legitimate employment criteria to statistical predictions of future outcomes, properly executed machine learning models will often pass muster.

Importantly, while this result may seem like a specific failure of an older anti-discrimination regime to adapt to widespread machine learning, it is more fundamental than that. Anti-discrimination law sees the problem of discrimination as the result of a choice by a decisionmaker, rather than at least partially the result of people's choices over centuries that have matured into harms derived from social environment.[287] This is what Alan Freeman identified many years ago as the "perpetrator perspective."[288] Freeman contrasted this with the "victim perspective," a view that recognizes the conditions of victims and the harms associated with discriminatory outcomes without needing a perpetrator to have first caused them.[289] Thus, despite occasionally being described as an effects test, disparate impact is still a doctrine focused on finding individual fault.

---

[282] *See* Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 721 (2006) ("Rather than defining the employers' practices as intentional discrimination, the Court allowed employers to use selection methods despite their adverse impact so long as they were demonstrated to be job related.").

[283] 29 C.F.R. § 1607.4(D) (2020).

[284] *Id.* § 1607.5(B).

[285] *See* Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 105, at 708-09 (stating that data mining could be validated by criterion-related validity).

[286] 29 C.F.R. § 1607.5(B).

[287] Selmi, *supra* note 282, at 761.

[288] Alan David Freeman, *Legitimizing Racial Discrimination Through Antidiscrimination Law: A Critical Review of Supreme Court Doctrine*, 62 MINN. L. REV. 1049, 1052-57 (1978).

[289] *Id.* Dean Mario Barnes noted recently that "[i]n the nearly forty years since Professor Freeman published his article, despite the improvements in many areas of race relations—to include the election of the country's first African-American president—the disjuncture Freeman located remains, and, in some ways, has worsened." Mario L. Barnes, *"The More Things Change . . .": New Moves for Legitimizing Racial Discrimination in a "Post-Race" World*, 100 MINN. L. REV. 2043, 2044 (2016).

Viewed through this lens, the business necessity defense is not solely a peculiarity of current doctrine. Rather, business necessity is a type of necessary safety valve that converts disparate impact from a strict liability offense into one that relates to employer fault. If we assume that there are existing inequalities today such that people's current qualifications for certain jobs are shaped at least in part by protected class status, then an employment procedure based on a "perfect" test of merit will still evince a disproportionate impact on protected classes. Without business necessity, each individual employer will be held strictly liable for the inequality that exists in the background of today's society. If Title VII is about policing employer choices, then this result is intolerable, and some version of the safety valve is necessary. Reframing the problem in terms of employer fault also explains why anti-discrimination law is often seen as a species of tort law[290] and why scholars have sometimes likened it to negligence or recklessness.[291]

Now we can see the parallel to the distributional problem described above. Because the operation of negligence law as an ex post remedy is more directly concerned with assigning liability than with rectifying or preventing the harm itself, statistical facts such as disparate impacts will often end up outside its purview. This is certainly true of the cases in which the AI exacerbates disparities but ends up being statistically beneficial for all groups. In the event of a society-wide net benefit but a *negative* result for a given subset of people, a case can be made that the AI user should know not to use the AI for the disadvantaged population. But even this argument is not based on relative disadvantages. Rather, the argument is that the user should have stopped because he should have known that the AI was likely to be harmful for this particular plaintiff. Even in this case, there is no place for discrimination-related reasoning.

It is worth noting that tension between statistical facts and individual liability is also a way to understand the AI error-foreseeability problem. The safety interest will compel the use of a tool that makes us statistically better off. Suppose in a given hospital ten percent of cancer screenings are misdiagnosed, and with AI that number would go down to five percent. But the statistics do not tell the whole story. Because machine errors are different than human errors, this does not result in half of the would-be injured ten percent being saved; rather it

[290] *See* Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 342 (2013) ("When the law grants persons the right to compensation for injury from wrongful conduct, there must be some demonstrated connection, some link, between the injury sustained and the wrong alleged. The requisite relation between prohibited conduct and compensable injury is governed by the principles of causation, a subject most often arising in elaborating the law of torts.").

[291] *See* Stephanie Bornstein, *Reckless Discrimination*, 105 CALIF. L. REV. 1055, 1103-07 (2017); David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899, 931-36 (1993) ("The formulation of a less discriminatory alternative test encourages the consideration of a negligence theory of employment discrimination.").

is likely that the new people who are harmed with the tool are an almost entirely different set of people than would have been harmed without the tool. By using a tool based in statistical reasoning, the hospital prevents many injuries, but from the individual standpoint it also creates an entirely new set of victims that will have no recourse.

Given that the problems are similar to discrimination law, the way to mitigate the harm may be similar too. Looking to interpretability techniques may allow some of the disparities to become clear, and in some cases, prevent the worst. To solve the statistical problem in Title VII, Professor Pauline Kim has argued that Title VII could give employers the burden of demonstrating a "substantively meaningful" relationship between job performance and the variables in the model.[292] Professors James Grimmelmann and Daniel Westreich have argued similarly that employers should have to tell a story about how the model relates to job performance.[293] That is, if employees are chosen because the model shows that good employees in the past have liked the color blue, we must understand why liking that color is predictive of job performance. Grimmelmann and Westreich suggest that it could be because it is a model picking up on a protected class,[294] while Kim is more concerned that it is simply a spurious correlation that will not be stable.[295] In the negligence context, a more granular picture of the causal relationships may still not give rise to liability in the same way, but it will at least shine a light on the problem such that the tool can be fixed in future iterations—and liability could attach for the use of a tool that is not patched. But as discussed earlier, interpretability techniques have their limitations, and it will be difficult to count on this solution.[296] Where injuries are caused by statistical realities, a regime of ex post liability may not be well suited to address the harms.

CONCLUSION

The umbrella term "artificial intelligence" represents a number of technologies. Some can reasonably be seen as autonomous, and in those cases, it makes sense to concern ourselves with their creation and the shift from negligence toward products liability. But the majority of AI technologies on the market today are decision-assistance tools, and it is just as important to pay attention to injuries that result from their use. When injuries result from the use of a tool, we look to negligence law to ask whether the user acted with due care.

This Article has argued that AI technologies can pose various challenges for negligence: unforeseen AI errors, unknown standards of reasonable care in interacting with computers and operational security, and distributional

[292] Kim, *supra* note 194, at 917.

[293] Grimmelmann & Westreich, *supra* note 194, at 174-76.

[294] *Id.* at 173.

[295] Kim, *supra* note 194, at 922.

[296] Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1117.

challenges. Ultimately, because AI inserts a layer of inscrutable, unintuitive, statistically derived, and often proprietary code between the decision and outcome, the nexus between human choices, actions, and outcomes from which negligence law draws its force is tested. While there may be a way to tie some decisions back to their outcomes using explanation and transparency requirements, negligence will need a set of outside interventions to have a real chance at providing redress for harms that result from the use of AI.

Negligence law is hardly the only reason for these transparency and explanation requirements, but without them negligence is unlikely to keep up.[297] So it is worth asking: If negligence no longer works, what options remain? This Article opened by describing the tort literature as overly focused on products liability. But as discussed in Part I, it is not actually clear that products liability provides any better avenue for compensating injuries.[298] AI neither aims for nor can achieve perfect accuracy.[299] As a result, the presence of errors does not imply a defective product required for a finding of products liability. This is once again a matter of the incompatibility of statistical logic with individual case outcomes. If an AI is defective because the error rate is too high, was the injury caused by the defect, or was the error one that a non-defective AI would also make? There is no obvious principled way to answer that. Thus, a move to products liability may not work either. Moreover, in a normative sense, do we really want to simply tell the users and purchasers of complex machinery that they bear no liability for carelessness in its use? Jumping to products liability for legal contexts currently governed by negligence does not appear to be an adequate approach.

Where society decides that AI is too beneficial to set aside, we will likely need a new regulatory paradigm to compensate the victims of AI's use, and it should be one divorced from the need to find fault. This could be strict liability, it could be broad insurance, or it could be ex ante regulation. We could look to existing models. Drugs are the most powerful example of a technology that we use without understanding its inner workings, so perhaps, as Andrew Tutt proposed, we can think about an "FDA for Algorithms."[300] Scholars, advocates, and legislators have proposed Algorithmic Impact Assessments ("AIA"), drawing on the environmental impact assessment model to increase

[297] *See id.* at 1134-38.

[298] *See supra* notes 26-30 and accompanying text (discussing complexities in defect classifications).

[299] *See* PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD 9 (2015) ("If a rule for, say, labeling e-mails as spam is 99 percent accurate, that does not mean it's buggy; it may be the best you can do and good enough to be useful.").

[300] *See generally* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017) (arguing that establishment of regulatory agency would be effective in dealing with challenges posed by complex algorithms).

transparency, explanation, and oversight.[301] Notably, none of these approaches is mutually exclusive. Rather, any of these regulatory models will actually bolster negligence law's ability to catch back up: the FDA model because it will have a centralized body to understand the common fault lines of AI systems, and the AIA model because it will allow the public to have the information needed to do so.

The use of AI decision-assistance tools is rapidly accelerating. Some people will make errors using AI tools, and others will be hurt. Negligence law exists to ensure that people harmed by others' actions have recourse if we consider those actions blameworthy. If we want to ensure that plaintiffs can continue to recover for AI-related injuries, we must either intervene soon to help negligence law adapt or find another way to compensate victims.

[301] *E.g.*, Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. § 2(1) (defining "automated decision system impact assessments"); DILLON REISMAN, JASON SCHULTZ, KATE CRAWFORD & MEREDITH WHITTAKER, ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY 7 (2018) ("A pre-procurement AIA gives an agency the opportunity to engage the public and proactively identify concerns, establish expectations, and draw on expertise and understanding from relevant stakeholders."); Margot E. Kaminski & Gianclaudio Malgieri, *Multi-Layered Explanations from Algorithmic Impact Assessments in the GDPR*, PROC. CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 68, 70-72 (2020) (discussing proposals for algorithmic impact assessments); Selbst, *supra* note 222, at 168-69 (arguing that police should create algorithmic impact statements before adopting predictive policing technology); Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 10, at 1133-35 (arguing that algorithmic impact statements open information to scrutiny); Michael Karlin & Noel Corriveau, *The Government of Canada's Algorithmic Impact Assessment: Take Two*, MEDIUM (Aug. 7, 2018), https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f [https://perma.cc/57D2-KA27] ("[T]he aim of our Algorithmic Impact Assessment is to develop a framework that would help institutions better understand and mitigate the risks associated with Automated Decision-Making Systems by providing the appropriate governance, oversight and reporting, and audit requirements." (citing Michael Karlin, *A Canadian Algorithmic Impact Assessment*, MEDIUM (Mar. 8, 2018), https://medium.com/@supergovernance/a-canadian-algorithmic-impact-assessment-128a2b2e7f85 [https://perma.cc/RET8-6T6T])).