WHEN AIS OUTPERFORM DOCTORS: CONFRONTING THE CHALLENGES OF A TORT-INDUCED OVER-RELIANCE ON MACHINE LEARNING

A. Michael Froomkin,* Ian Kerr** & Joelle Pineau***

Someday, perhaps soon, diagnostics generated by machine learning (ML) will have demonstrably better success rates than those generated by human doctors. What will the dominance of ML diagnostics mean for medical malpractice law, for the future of medical service provision, for the demand for certain kinds of doctors, and—in the long run—for the quality of medical diagnostics itself?

This Article argues that once ML diagnosticians, such as those based on neural networks, are shown to be superior, existing medical malpractice law will require superior ML-generated medical diagnostics as the standard of care in clinical settings. Further, unless implemented carefully, a physician's duty to use ML systems in medical diagnostics could, paradoxically, undermine the very safety standard that malpractice law set out to achieve. Although at first doctor + machine may be more effective than either alone because humans and ML systems might make very different kinds of mistakes, in time, as ML systems improve,

---

* Laurie Silvers & Mitchell Rubenstein Distinguished Professor of Law, University of Miami. Member, University of Miami Center for Computational Science; Fellow, Yale ISP. Thanks to Peter Asaro, Jack Balkin, Caroline Bradley, Ryan Calo, Kate Crawford, Brad DeLong, Ed Felten, Colleen Flood, Jonathan Frankle, David Froomkin, Sue Gluck, Bob Glushko, James Grimmelmann, Woody Hartzog, Margot Kaminski, Gregory Keating, Daniel Kluttz, Mark Lemley, Amanda Levendowski, Christopher Millard, Deirdre Mulligan, Helen Nissenbaum, Paul Ohm, Frank Pasquale, Laurel Riek, Cynthia Rudin, Robin Schard, Andrew Selbst, Latanya Sweeney, participants in a University of Miami School of Law faculty seminar, participants in a Yale ISP Ideas Lunch, and participants in We Robot 2018 for advice and information. Thanks to the Social Sciences and Humanities Research Council and the Canada Research Chairs program for their generous support.

** Canada Research Chair in Ethics, Law & Technology, University of Ottawa, Faculty of Law, with cross appointments to the Faculty of Medicine, Department of Philosophy, and School of Information Studies.

*** William Dawson Scholar and Associate Professor, School of Computer Science, McGill University.

Electronic copy available at: https://ssrn.com/abstract=3114347

effective ML could create overwhelming legal and ethical pressure to delegate the diagnostic process to the machine. Ultimately, a similar dynamic might extend to treatment also. If we reach the point where the bulk of clinical outcomes collected in databases are ML-generated diagnoses, this may result in future decisions that are not easily audited or understood by human doctors. Given the well-documented fact that treatment strategies are often not as effective when deployed in clinical practice compared to preliminary evaluation, the lack of transparency introduced by the ML algorithms could lead to a decrease in quality of care. This Article describes salient technical aspects of this scenario particularly as it relates to diagnosis and canvasses various possible technical and legal solutions that would allow us to avoid these unintended consequences of medical malpractice law. Ultimately, we suggest there is a strong case for altering existing medical liability rules to avoid a machine-only diagnostic regime. We argue that the appropriate revision to the standard of care requires maintaining meaningful participation in the loop by physicians the loop.

TABLE OF CONTENTS

Electronic copy available at: https://ssrn.com/abstract=3114347

INTRODUCTION

Someday, perhaps sooner,[1] perhaps later,[2] machines will have demonstrably better success rates at medical diagnosis than human physicians—at least in particular medical specialties.[3]

We can reasonably expect that machine-learning-based diagnostic competence, which we will sometimes call "AI" for short, will only increase. It is thus appropriate to consider what the dominance of machine-based diagnostics might mean for medical malpractice law, the future of medical service provision, the demand for certain kinds of physicians, and—in the long run—for the quality of medical diagnostics itself.

In this Article, we interrogate the legal implications of superior machine-generated diagnosticians, particularly those based on neural networks, currently a leading type of machine learning used in prediction.[4] We argue that existing medical malpractice law will eventually require superior ML-generated medical diagnosis as the standard of care in clinical settings. We further argue that—unless implemented carefully—a physician's duty to use ML in medical diagnostics could, paradoxically, undermine the very safety standard that malpractice law set out to achieve. Once computerized diagnosticians demonstrate better success rates than their human trainers, effective machine learning will create legal (and ethical) pressure to delegate much, if not all, of the diagnostic process to the machine. If we reach the point where the bulk of clinical outcomes collected in databases are ML-generated diagnoses, this may result in future decision scenarios that are

1. *See infra* text accompanying notes 12–22.
2. *See infra* text accompanying notes 30–35.
3. *See infra* text accompanying notes 24–26.
4. Machine learning (ML) is the discipline of automated pattern recognition and making predictions based on patterns that are detected. Neural networks are one of several types of ML. "Deep Learning," another term of use, refers to neural networks with many layers. "AI" is a more general term applied to automated techniques that produce outputs which appear to mimic human reason or behavior. Thus, deep-learning systems are a subset of neural networks, which are a subset of ML, which is itself a subset of AI. IBM's Watson, which we also discuss, is perhaps the best-known example of a neural-network-based medical diagnostic system. *See infra* text accompanying notes 38–45.

Electronic copy available at: https://ssrn.com/abstract=3114347

difficult to validate and verify. Many ML systems currently are not easily audited or understood by human physicians, and if this remains true, it will be harder to detect sub-par performance, jeopardizing the system's efficacy, accuracy, and reliability. Once ML systems displace doctors in a specialty, the demand for such doctors will shrink as will training opportunities for human experts. Because we will continue to need humans to generate much of the training data for future ML systems, this reduction in human competence may create roadblocks to the continuing improvement of ML systems especially once new diagnostic sensors are available. We maintain that such unintended consequences of medical malpractice law must be avoided and canvass various possible technical and legal solutions.

Our story has four acts.

1) We begin with the effect of existing law on the use of ML diagnostic technology, be it neural networks or some other form of AI. We argue that once a machine is demonstrably superior to human diagnosticians, malpractice law will require the use of the superior technology in certain sectors of medical diagnostics. Medical service providers who do not use ML systems will be said to fall below the appropriate standard of care in cases where things go wrong, and hospitals that use human physicians rather than ML systems will be subject to claims in negligence—as will the treating physicians themselves.

2) Next, we consider the consequences that these novel legal requirements might have on the overall demand for physicians of certain types and the potentially diminished role that they might play in medical practice. We suggest that the advent of superior ML diagnosticians will reduce the demand for human physicians,[5] much like the enhanced safety and efficacy of self-driving trucks will increase the demand for robot drivers and decrease the demand for human drivers.[6] These consequences, flowing from the requirements imposed by medical malpractice law, give rise to various narratives. To the extent that patient outcomes are now better and perhaps even cheaper—depending on automated-system service-provider pricing—these newly imposed legal requirements offer a desirable neoliberal result: better living through technology. Of course, the possible outcomes also comport just as well with the classic account of deskilling: over-reliance on these machines could render obsolete the human cultivation of medical skills and know-how developed over centuries.[7] Indeed, robotic surgery—

5. It will likely increase demand for certain types of medical technicians. A similar economic logic applies to robot surgeons and other medical specialties as they get robotized.

6. Olivia Solon, *Self-Driving Trucks: What's the Future for America's 3.5 Million Truckers?*, GUARDIAN (June 17, 2016, 7:00 AM), https://www.theguardian.com/technology/2016/jun/17/self-driving-trucks-impact-on-drivers-jobs-us ("Driverless trucks will be safer and cheaper than their human-controlled counterparts . . . ."); Scott Santens, *Self -Driving Trucks Are Going to Hit Us Like a Human-Driven Truck*, MEDIUM (May 14, 2015), https://medium.com/basic-income/self-driving-trucks-are-going-to-hit-us-like-a-human-driven-truck-b8507d9c5961.

7. *See, e.g.*, HARRY BRAVERMAN, LABOR AND MONOPOLY CAPITAL: THE DEGRADATION OF WORK IN THE TWENTIETH CENTURY 118–19 (Monthly Review Press 1998)

Electronic copy available at: https://ssrn.com/abstract=3114347

which can perform some tasks more quickly and more accurately than humans[8]—is already being accused of causing a loss of surgical skill among medical trainees.[9] That law has mandated the use of a new technology that produces improved health outcomes might also make this tale a happy outlier to more familiar stories of the law's interaction with technology—those in which law is disrupted by the technical change and in which self-interested parties may seek to hold off the change.[10]

3) Regardless of which narrative best describes our second act, we believe there is a third act that must also be considered: the development of a diagnostic monoculture and other dangers associated with an over-reliance on ML. By "diagnostic monoculture" we mean a scenario in which the medical and legal systems standardize on a mechanized approach to diagnosis in a given sub-specialty. Diagnostic monoculture exemplifies a more general problem that arises when society comes to rely, to its detriment, on a dominant mode of thinking to the exclusion of other possible solutions. In this case, a diagnostic monoculture that leads to less input from human physicians could make quality control of diagnostic databases much more difficult. The problem becomes far more serious once reliance on ML goes beyond diagnosis to treatment. The reduction in new data from physicians—that is to say the creation of a loop in which outcomes added to the database are solely or overwhelmingly the result of ML-informed treatment decisions—creates scenarios in which we cannot rule out the risk that sub-optimal conclusions are reached. If a set of symptoms is consistently producing an erroneous ML diagnostic, and physicians act on that erroneous diagnostic, where will ML get the data to suggest a different diagnosis which leads to better treatment? If the answer is "nowhere" then we have a problem. Worse, it is not

---

(1974); THE DEGRADATION OF WORK?: SKILL, DESKILLING, AND THE LABOUR PROCESS 11–12 (Stephen Wood ed., 1982); Stanley Aronowitz & William DiFazio, *High Technology and Work Tomorrow*, 544 ANNALS AM. ACAD. POL. & SOC. SCI. 52 (1996), doi: 10.1177/0002716296544001005 (arguing that technology tends to destroy high-skill jobs and replace them with low-skill jobs); *but see* Paul Attewell, *The Deskilling Controversy*, 14 WORK & OCCUPATIONS 323, 323 (1987), doi: 10.1177/0730888487014003001 (offering theoretical and empirical critique of deskilling thesis).

8. *See* Hannah Devlin, *The Robots Helping NHS Surgeons Perform Better, Faster – and For Longer*, GUARDIAN (July 4, 2018, 6:00 AM), http://www.theguardian.com/society/2018/jul/04/robots-nhs-surgeons-keyhole-surgery-versius.

9. *See* Matthew Beane, *Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail*, 64 ADMIN. SCI. Q. 87, 87–88 (2018), doi: 10.1177/0001839217751692.

10. For example, the Digital Millennium Copyright Act (DMCA) is sometimes accused of propping up outdated or anticompetitive business models in the face of easy content-sharing. *See, e.g.*, Ryan J. Shernaman, *The Digital Millennium Copyright Act: The Protector of Anti-Competitive Business Models*, 80 UMKC L. REV. 545, 545–46 (2011). Likewise, DMCA-type legislation has also been shown to undermine privacy. Ian Kerr, *If Left to Their Own Devices . . . How DRM and Anti-Circumvention Laws Can be Used to Hack Privacy*, *in* IN THE PUBLIC INTEREST: THE FUTURE OF CANADIAN COPYRIGHT LAW (Michael Geist ed., 2005), https://ssrn.com/abstract=902448.

Electronic copy available at: https://ssrn.com/abstract=3114347

even clear that either the ML system or an outside observer necessarily would know that the results were sub-optimal. From a human perspective, the challenges associated with understanding and auditing an ML system's predictive diagnostic process will become significant. Those challenges become greater if the output of the ML diagnostic system is then fed into a second ML treatment system. In that case, absent personalized medicine, for any given set of symptoms one might get consistent treatment decisions leading to less variegated treatment-to-outcome data. The lack of variety in treatment could further mask any issues caused by sub-optimal diagnoses and could lead to bad decision-making and, potentially, tragic medical outcomes. To guard against this possibility, we will need a mechanism. And until we know how to automate that too we may need a substantial corps of medical researchers on tap to help audit and monitor the machines to spot anomalies.

4) The approach taken in our fourth act is speculative and involves exploring different possible future scenarios and potential solutions. Our starting point imagines a future in which the reliability of the diagnostic ML is high enough that the human physician seems unnecessary or even—to the extent she may overrule valid diagnoses—unhelpful insofar as her inputs tend to reduce the probability of a successful outcome. We consider technological fixes in response to an ML monoculture and whether better liability rules might avoid or at least postpone the problem. One complicating factor that we must consider is that law is not the only driver here: even without the malpractice push, if the price is right, economics could incentivize a very similar evolution. In either case, it is essential to examine several potential means of avoiding the risks associated with an ML diagnostic monoculture and an over-reliance on ML.

If we are correct that tort law will provide the wrong incentives, the question is what one can or should do about it. Countries with national health systems featuring strong centralized control might find an administrative method of overcoming the problems we describe. But in the United States, in which both medical service provision and insurance remain relatively decentralized, the tort system—malpractice law—serves as an important source of incentives and thus de facto regulation of medical service provision. A possible legal strategy would be to change existing medical malpractice rules and thus reduce the incentives that drive medicine to reduce its reliance on people. We propose meaningful human participation in diagnostics as an essential requirement of the standard of care. This will blunt the legal aspect of the push toward replacing physicians with ML.

Furthermore, as probabilistically superior AIs come to work alongside humans, we must find ways to combat malpractice law's tendency to stay the human's hand in individual cases: if a physician overrides the machine, the physician (and his or her employer) will be taking a terrible malpractice risk if it remains the case that the machine has a significantly better probability of success on its own then does the physician. We thus will also need to formulate new rules that balance the social interest of having human judgment in the loop with the individual patient's interest in getting the best outcome. However, this requires that we consider thorny ethical and legal issues. Unless we are very confident in our technical solutions, we argue, there is a strong case for altering existing medical liability rules to maintain focus—when it comes to determining the

Electronic copy available at: https://ssrn.com/abstract=3114347

appropriate role of humans and machines in medical diagnostics—on both ethics[11] and cost rather than defensive medicine. A revision of the standard of care to avoid allowing a machine-only diagnostic regime would require meaningful participation by people in the loop. As such, it risks being expensive because the machine will cost money and the rule we propose will negate potential cost savings from reducing the number of physicians in reliance on the new technology. However, we suggest that our proposal could be a first step in preventing law from overriding these other important considerations, preserving many long-term beneficial outcomes that would otherwise be at risk due to pressure from the legal system and from cost-cutting.

I. ONCE A MACHINE LEARNING SYSTEM IS DEMONSTRABLY SUPERIOR, MALPRACTICE LAW WILL REQUIRE THAT MEDICAL SERVICE PROVIDERS USE IT

It seems inevitable that—at least for some medical specialties—ML diagnosticians someday will have demonstrably better success rates than human physicians. A number of ongoing initiatives suggest that ML will have, or perhaps already has,[12] great diagnostic power for a variety of diseases and conditions ranging from oncology to drug discovery. Google's neural net diagnoses skin cancer as effectively as do experienced dermatologists.[13] Google has tested an AI-based system that successfully identified eye diseases in retinal fundus photographs[14] and one that reaches or exceeds that of experts on a variety of sight-

11. *See* Jason Millar & Ian Kerr, *Delegation, Relinquishment, and Responsibility: The Prospect of Expert Robots in* ROBOT LAW 102, 115 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016).

12. *See* Ian Steadman, *IBM's Watson Is Better at Diagnosing Cancer than Human Doctors*, WIRED (Feb. 11, 2013), http://www.wired.co.uk/article/ibm-watson-medical-doctor.

13. *See* Andre Esteva et al. *Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks*, 542 NATURE 115, 118 (2017), doi: 10.1038/nature21056. *But see* Zachary C. Lipton & Jacob Steinhardt, *Troubling Trends in Machine Learning Scholarship*, ARXIV:1807.03341 (July 27, 2018), http://arxiv.org/abs/1807.03341 ("The comparison to dermatologists conceals the fact that classifiers and dermatologists perform fundamentally different tasks. Real dermatologists encounter a wide variety of circumstances and must perform their jobs despite unpredictable changes. The machine classifier, however, only achieves low error on [static] test data.").

14. Varun Gulshan et al., *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs*, 316 J. AM. MED. ASSOC. 2402, 2402 (2016), doi:10.1001/jama.2016.17216; *see also* Ariel Bleicher, *Teenage Whiz Kid Invents an AI System to Diagnose Her Grandfather's Eye Disease*, IEEE SPECTRUM (Aug. 3, 2017, 5:00 PM), https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/teenage-whiz-kid-invents-an-ai-system-to-diagnose-her-grandfathers-eye-disease (describing creation of "Eyeagnosis, a smartphone app plus 3D-printed lens that seeks to change the diagnostic procedure from a 2-hour exam requiring a multi-thousand-dollar retinal imager to a quick photo snap with a phone").

Electronic copy available at: https://ssrn.com/abstract=3114347

threatening retinal diseases.[15] Other programs already beat humans: an AI beat humans at predicting heart attacks—without even considering the effects of diabetes or lifestyle.[16] A different AI beat humans at diagnosing brain tumors and predicting hematoma expansion.[17] So too with predicting certain heart diseases: "Machine-learning significantly improves accuracy of cardiovascular risk prediction, increasing the number of patients identified who could benefit from preventive treatment, while avoiding unnecessary treatment of others."[18] Researchers at MIT and Harvard are using ML for Alzheimer detection.[19] Meanwhile, "Chinese researchers have developed an artificial intelligence system which can diagnose cancerous prostate samples as accurately as any pathologist."[20] A deep-learning system using convolutional neural networks, trained with 100,000 images, found 95% of melanomas in a study, while human dermatologists only found 86.6% of them.[21] Similarly,

> Watson for Drug Discovery rank ordered all of the nearly 1,500 genes within the human genome and proposed predictions regarding which genes might be associated with ALS. . . . [E]ight of the top 10 ranked genes proved to be linked to the disease. More significantly, the study found five never before linked genes associated with ALS.[22]

---

15. Jeffrey De Fauw et al., *Clinically Applicable Deep Learning For Diagnosis and Referral In Retinal Disease*, 24 NATURE MED. 1342, 1348 (2018), doi: 10.1038/s41591-018-0107-6 (noting that training data was only 14,884 scans).

16. Lulu Chang, *Machine Learning Algorithms Surpass Doctors at Predicting Heart Attacks*, DIGITAL TRENDS (Apr. 17, 2017, 6:21 AM), http://www.digitaltrends.com/health-fitness/ai-algorithm-heart-attack/.

17. For the brain tumors the BioMind AI system "made correct diagnoses in 87 percent of 225 cases in about 15 minutes, while a team of 15 senior doctors only achieved 66-percent accuracy;" it correctly predicted the hematomas 83% of the time while the humans only managed a 63% accuracy rate. Xinhua, *China Focus: AI beats human doctors in neuroimaging recognition contest*, XINHUANET (June 6, 2018), http://www.xinhuanet.com/english/2018-06/30/c_137292451.htm.

18. Stephen F. Weng et al., *Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?*, 12 PLOS ONE e0174944, Apr. 4, 2017, at 2, https://doi.org/10.1371/journal.pone.0174944.

19. *See Predicting Change in the Alzheimer's Brain*, MIT CSAIL (Oct. 6, 2015), http://www.csail.mit.edu/predicting_change_in_the_alzheimers_brain; Adrian V. Dalca et al., *Predictive Modeling of Anatomy with Genetic and Clinical Data*, MIT (2015), http://www.mit.edu/~adalca/files/papers/miccai2015_predictiveModelling_precr.pdf.

20. Science Business Reporting, *Artificial Intelligence Can Diagnose Prostate Cancer as Well as a Pathologist*, SCIENCE|BUSINESS (Mar. 19, 2018), https://sciencebusiness.net/healthy-measures/news/artificial-intelligence-can-diagnose-prostate-cancer-well-pathologist.

21. H. A. Haenssle et al., *Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists*, 29 ANNALS ONCOLOGY 1836, 1839 (2018), doi: 10.1093/annonc/mdy166. The 58 human dermatologists came from 17 countries; just over half were "expert"-level, but 29% had less than two years' experience. *Id.* at 1838–39.

22. *Barrow Identifies New Genes Responsible for ALS using IBM Watson Health*, CISION (Dec. 14, 2016, 12:30 PM), http://www.prnewswire.com/news-

Electronic copy available at: https://ssrn.com/abstract=3114347

Diagnostic medicine seems a particularly good fit for what today's AIs can do best—pattern recognition—as well as being an area with real room for improvement. Five percent of U.S. adults who seek outpatient care each year experience a diagnostic error, leading to 6%–17% of adverse events in hospitals.[23]

Radiology seems to be a specialty particularly suited to replacement by ML.[24] One study reports that an AI correctly detected 92.4% of breast-cancer tumors compared to the 73.2% detected correctly by human doctors.[25] Indeed University of Toronto Professor Geoffrey Hinton argues that radiologists are about to be obsolete:

> I think that if you work as a radiologist you are like Wile E. Coyote in the cartoon . . . . You're already over the edge of the cliff, but you haven't yet looked down. There's no ground underneath. . . . It's just completely obvious that in five years deep learning is going to do better than radiologists.[26]

Some radiologists vehemently disagree, because "radiologists do not just look at pictures."[27] Hyperbole notwithstanding, many ML experts share Hinton's vision

---

releases/barrow-identifies-new-genes-responsible-for-als-using-ibm-watson-health-300378211.html.

23. *See* Nicolas P. Terry, *Appification, AI, and Healthcare's New Iron Triangle*, 21 J. HEALTH CARE POL'Y 117, 174 (2018) (citing INSTITUTE OF MEDICINE, IMPROVING DIAGNOSIS IN HEALTH CARE) (2015)), doi: 10.2139/ssrn.3020784.

24. *See* Katie Chockley & Ezekiel Emanuel, *The End of Radiology? Three Threats to the Future Practice of Radiology*, 13 J. AM. COLL. RADIOL. 1415, 1417–19. (2016), doi: 10.1016/j.jacr.2016.07.010.

25. Yun Liu et al., *Detecting Cancer Metastases on Gigapixel Pathology Images*, ARXIV:1703.02442 [CS] (Mar. 8, 2017), http://arxiv.org/abs/1703.02442 (stating "[a]t 8 false positives per image, we detect 92.4% of the tumors, relative to 82.7% by the previous best automated approach. For comparison, a human pathologist attempting exhaustive search achieved 73.2% sensitivity"). Currently, however, the ML system's false-positive rate remains greater than that of humans. *See* Dayong Wang et al, *Deep Learning for Identifying Metastatic Breast Cancer*, ARXIV (June 18, 2016), https://arxiv.org/pdf/1606.05718.pdf.

26. Siddhartha Mukherjee, *A.I. Versus M.D.*, NEW YORKER (Apr. 3, 2017), http://www.newyorker.com/magazine/2017/04/03/ai-versus-md.

27. "[W]ithout radiologists, a hospital simply cannot function." Hugh Harvey, *Why AI Will Not Replace Radiologists*, TOWARDS DATA SCI. (Mar. 11, 2018), https://towardsdatascience.com/why-ai-will-not-replace-radiologists-c7736f2c7d80; *see also* Will Knight, *Google X-Ray Project Shows Ai Won't Replace Doctors Any Time Soon*, MIT TECH. REVIEW (Mar. 27, 2018), https://www.technologyreview.com/s/610552/google-x-ray-project-shows-ai-wont-replace-doctors-any-time-soon/. On the other hand, ML is making inroads into the radiological treatment process also. *See* Ian Sample, *'It's Going to Create a Revolution': How AI is Transforming the NHS*, GUARDIAN (July 4, 2018), http://www.theguardian.com/technology/2018/jul/04/its-going-create-revolution-how-ai-transforming-nhs (describing use of IBM's "InnerEye" system to markup scans automatically for prostate-cancer patient, saving time and—it is hoped but not yet proved—increasing quality of treatment).

Electronic copy available at: https://ssrn.com/abstract=3114347

regarding the inevitable demise of human medical diagnosis for conditions where we have large amounts of high-quality data.[28]

IBM promoted Watson as using oncological data to diagnose cancers that humans have difficulty identifying.[29] Skeptics point to issues with current trials and suggest that ML superiority remains purely speculative,[30] and that IBM's advertising over-promises what Watson can do.[31] Oren Etzioni, CEO of the Allen Institute for AI, went as far as to say that "IBM Watson is the Donald Trump of the AI industry—outlandish claims that aren't backed by credible data."[32] Indeed, IBM Watson's "Oncology Expert Advisor" suffered a high-profile setback when the University of Texas's cancer center canceled a flagship collaboration because the project foundered on incompatibilities with the hospital records system as well as alleged violations of hospital procurement regulations.[33] In the end, the "project appeared to fall apart because of cost overruns related to incompatible IT platforms and the extraordinarily complex work involved in structuring and preparing massive amounts of data to be ingested by Watson's machine learning systems."[34] Even a state-of-the-art AI was no match for "the idiosyncrasies of medical records: the acronyms, human errors, shorthand phrases, and different styles of writing."[35]

---

28. *See, e.g.*, Chockley & Emanuel, *supra* note 24.

29. *See* Steve Lohr, *IBM Is Counting on Its Bet on Watson, and Paying Big Money for It*, N.Y. TIMES (Oct. 17, 2016), http://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html.

30. *See, e.g.*, Casey Ross & Ike Swetlitz, *IBM Pitched Watson as a Revolution in Cancer Care. It's Nowhere Close*, STAT (Sept. 5, 2017), https://www.statnews.com/2017/09/05/watson-ibm-cancer/.

31. A particularly egregious example is IBM, *Watson at Work*, YOUTUBE (Feb. 10, 2016), https://www.youtube.com/watch?v=7zKLEyLTqNU, in which "Watson" has a dialog with basketball scouts on the court—although reportedly, the Toronto Raptors are in fact using a version of Watson to help them rank scouted players based on various numerical metrics. *See* IBM, *Seeing Things the Other Teams Can't is the Key to Victory*, https://www.ibm.com/watson/stories/ca-en/basketball-with-watson.html (last visited Jan. 16, 2018).

32. Jennings Brown, *Why Everyone Is Hating on IBM Watson—Including the People Who Helped Make It*, GIZMODO (Aug. 10, 2017) (quoting Oren Etzioni), http://gizmodo.com/why-everyone-is-hating-on-watson-including-the-people-w-1797510888.

33. *See* Matthew Herper, *MD Anderson Benches IBM Watson in Setback for Artificial Intelligence in Medicine*, FORBES (Feb. 19, 2017), https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine.

34. John Battelle, *A Trio of Tech Takedowns*, NEWCO SHIFT (July 17, 2017), https://shift.newco.co/a-trio-of-tech-takedowns-b931c0df5ef6; *see also* Herper, *supra* note 33.

35. Casey Ross & Ike Swetlitz, *IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show*, STAT (July 25, 2018), https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/; *see also* B L Holman et al., *Medical impact of unedited preliminary radiology reports.*, 191 RADIOLOGY 519, 520 (1994), doi: 10.1148/radiology.191.2.8153332 (reporting

Electronic copy available at: https://ssrn.com/abstract=3114347

Watson's Sloan Kettering system apparently erred badly, engineers said, exhibiting "multiple examples of unsafe and incorrect treatment recommendations" due to faulty synthetic (hypothetical rather than real patient) training data.[36] IBM defends its Watson program by pointing to other successes, particularly in Watson for Genomics.[37]

There is no question that Watson has enjoyed a friendly press and significant hype.[38] It is also the case that not everything IBM currently markets as "Watson" is true ML. For example, "Watson for Oncology" has been touted as giving "the same recommendations as professional oncologists in 99 percent of the cases" in a test at the University of North Carolina.[39] But the program is really a decision-support tool enhanced with preprogrammed suggestions based on what a committee of doctors at Sloan Kettering said they would do when presented with various symptoms and scenarios.[40] And it is also likely that some diagnostic problems can be solved with simpler ordinary non-ML models that predict as well or almost as well as ML while enabling much greater transparency as to the reasons for a diagnosis.[41]

---

that 5.4% of unedited radiology reports examined had significant errors, and that even after editing 1.8% would have caused either unnecessary testing or actual danger to patients); Hugh Harvey, *Synoptic Reporting Makes Better Radiologists, and Algorithms*, MEDIUM (Mar. 25, 2018), https://towardsdatascience.com/synoptic-reporting-makes-better-radiologists-and-algorithms-9755f3da511a (discussing ease with which errors creep into medical records, especially those generated from free text and natural-language parsing).

36. Ross & Swetlitz, *supra* note 35.

37. *See* John E. Kelly III, *Watson Health: Setting the Record Straight*, WATSON HEALTH PERSP. (Aug. 11, 2018), https://www.ibm.com/blogs/watson-health/setting-the-record-straight/.

38. *See* Mary Chris Jaklevic, *MD Anderson Cancer Center's IBM Watson Project Fails, and so Did the Journalism Related to It*, HEALTHNEWSREVIEW.ORG (Feb. 23, 2017), https://www.healthnewsreview.org/2017/02/md-anderson-cancer-centers-ibm-watson-project-fails-journalism-related/. Internal IBM documents reveal that doctors were livid about Watson's performance: "This product is a piece of shit," a doctor at Florida's Jupiter Hospital said to IBM. "We bought it for marketing and with hopes that you would achieve the vision. We can't use it for most cases." Ross & Swetlitz, *supra* note 35.

39. Ben Dickson, *How Artificial Intelligence Is Revolutionizing Healthcare*, NEXT WEB (Apr. 13, 2017), https://thenextweb.com/artificial-intelligence/2017/04/13/artificial-intelligence-revolutionizing-healthcare/.

40. "That training does not teach Watson to base its recommendations on the outcomes of these patients, whether they lived, or died or survived longer than similar patients. Rather, Watson makes its recommendations based on the treatment preferences of Memorial Sloan Kettering physicians." Ross & Swetlitz, *supra* note 30.

41. *See* Cynthia Rudin & Berk Ustun, *Optimized Scoring System: Towards Trust in Machine Learning for Healthcare and Criminal Justice*, 48 INTERFACES 449 (2018). Rudin and Ustun argue that if models that are given a choice between a black-box ML model and a non-ML model "that is so simple it can fit on an index card" we might prefer the simpler, more transparent model even if it can only "predict almost equally well." *Id.* at 450. That may be a persuasive argument in the criminal-justice context, which is constrained by Due Process concerns among others; it is less obvious in the medical system

Electronic copy available at: https://ssrn.com/abstract=3114347

However, we should not allow the real ML wheat to be obscured by the marketing chaff. ML systems are being used for everything from dress designing to cooking, roadside assistance, business messaging, education, and movie direction.[42] In April 2018, the Food and Drug Administration (FDA) approved IDx-DR for sale, making it the first-ever AI-based software approved to detect diabetic retinopathy.[43] Notably, "IDx-DR is the first device authorized for marketing that provides a screening decision without the need for a clinician to also interpret the image or results, which makes it usable by healthcare providers who may not normally be involved in eye care."[44] Meanwhile, researchers are using ML systems, including Watson, to find tumors in radiological data,[45] making these the paradigmatic examples of the genre.

A. Machine Learning

1. ML Algorithms Today

At their core, ML systems are simply algorithms designed to draw on data to answer questions.[46] Depending on the design of the algorithm, and the type and amount of data available, an ML system can answer very simple questions, such as predicting the expected weight gain for a patient receiving a given medication or more complex questions, such as analyzing brain scans and delineating the location of a tumor.[47]

The basic components of an ML system include:

- **Input**: The training examples fed into the algorithm. The examples are described by a set of features—e.g., doctors' notes, clinical results, time-series recordings, images, etc.—that the machine will observe.[48]
- **ML Algorithm**: The computer program that will digest the data and make a prediction—e.g., linear regression, neural networks, decision trees. We include in this component both the computer's representation of the knowledge extracted and the optimization routine used to train the representation.[49]

---

where medical ethics, patients, and the tort system all create very great pressures to choose the technology that is demonstrably best. *Id.* at 449–50.

42. Will Knight, *IBM's Watson Is Everywhere—But What Is It?*, MIT TECH. REV. (Oct. 27, 2016), https://www.technologyreview.com/s/602744/ibms-watson-is-everywhere-but-what-is-it/.

43. Press Release, FDA, FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems (Apr. 11, 2018), https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm.

44. *Id.*

45. *See* Chockley & Emanuel, *supra* note 24.

46. *See* CHRISTOPHER BISHOP, PATTERN RECOGNITION AND MACHINE LEARNING 1 (2006).

47. *See* Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, 55 COMM. ACM 78, 78 (2012), doi: 10.1145/2347736.2347755.

48. *See* BISHOP, *supra* note 46, at 2.

49. *Id.* at 5.

Electronic copy available at: https://ssrn.com/abstract=3114347

- **Output**: The information that is produced by the algorithm for given examples—e.g., predicted weight gain, tumor location, primary health outcome, recommended treatment strategy, prescribed medication dosage.[50]
- **Evaluation**: The criteria by which we measure the algorithm's performance—e.g., classification accuracy, prediction error, false-positive rate.[51]

In this Article, we distinguish between ML systems that make *predictions* and ML systems that make *interventions*. Most of the components may be very similar in both cases, so the distinction is primarily in terms of the output. *Prediction-type* ML systems produce outputs designed to inform medical personnel and enhance their knowledge, situational awareness, and understanding, which they can incorporate in their own decision-making about treatment strategy. *Intervention-type* ML systems produce outputs that are actionable and can be applied directly, such as a request for a clinical test, a prescription, or in some cases a direct intervention. Examples of interventions include the case of a neuro-stimulation device using ML to decide the timing and intensity of electrical stimulation applied to a patient with epilepsy in hopes of reducing the incidence of seizures,[52] or an artificial pancreas using ML to adapt the dosage of an implanted insulin pump on a diabetic patient.[53]

While from a technical perspective Prediction-type ML and Intervention-type ML can be built using analogous technology and data, the distinction between them is potentially important in the context of discussing medical malpractice law because of the different degrees of human intervention that occur before the ML output is applied to a patient. It might seem obvious that a human's liability for relying on ML will be greater in the Intervention-ML scenario than in the mere Prediction-ML scenario. After all, if ML is only being used for prediction, there clearly is a human in the loop making the treatment decision rather than—dare we say—mechanically following the dictates of the Intervention-ML. However, in our view the liability distinction between the two is less sharp than it may seem: if the downstream human's reliance on the Prediction-ML was the source of the patient's bad outcome, but this reliance was reasonable given the Prediction-ML's track record or its being part of the standard of care, then the liability of the human under the Prediction system may be no greater than under the Intervention system.

---

50. *See id.* at 2.

51. *Id.* at 32.

52. *See* Ali Hossam Shoeb, *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*, MIT (2009), https://dspace.mit.edu/handle/1721.1/54669.

53. *See* Melanie Katrin Bothe et al., *The Use of Reinforcement Learning to Meet the Challenges of an Artificial Pancreas*, 10 EXPERT REV. MED. DEVICES 661, 661 (2013), doi: 10.1586/17434440.2013.827515; Eric Wicklund, *Can Watson Help mHealth Predict Health Emergencies?*, MHEALTH INTELLIGENCE (Jan. 22, 2016), https://mhealthintelligence.com/news/can-watson-help-mhealth-predict-health-emergencies; *FDA approves clinical testing of AI-powered bionic pancreas for diabetes*, SCI. SERV. (May 29, 2018), https://www.dr-hempel-network.com/digital-health-technolgy/beta-bionic-ai-powered-bionic-pancreas-for-diabetes/.

Electronic copy available at: https://ssrn.com/abstract=3114347

Neural networks are but one type of ML algorithms designed to answer questions using data. Earlier methods, including linear regression, decision trees, and simple probabilistic models, have been used for years to make predictions.[54] Currently, researchers are making particularly rapid progress in training neural networks, especially those with many layers ("deep learning"), to recognize increasingly complex patterns in data.[55] Neural networks are now the method of choice to analyze high-dimensional data, including images of all types, sound, and natural-language text.[56] Their power resides in their ability to extract patterns from large data sets with relatively little prior knowledge about useful features or variables.[57]

A critical element of deep learning is that it trains synthetic neurons in multiple layers, both of which extract information at different levels of abstraction.[58] One can think of each neuron as a simple unit of computation (typically performing a linear equation, followed by a non-linear transform).[59] Groups of neurons are assembled into layers; each neuron in a layer is in communication with the ones in the layer above it; each successive layer tends to learn to recognize more general features of the network's input.[60] The neurons in the very first layer observe the Input (raw) data. The neurons in the final layer are responsible for producing the Output.[61]

A common denominator of all ML algorithms, including neural networks, is that they require training. Training methods vary, but they all depend on access to a sufficient—and usually quite large[62]—body of accurate training data. For tumor detection, the data set might be a set of input images, along with the annotations from expert radiologists about the target output—e.g., simple tumor/no-tumor classification, or a detailed tumor-contour segmentation.[63] The fact that the images come with a human-annotated label is crucial.[64] The ML

---

54. *See* TOM MITCHELL, MACHINE LEARNING 15 (1997).

55. *See generally* Yoshua Bengio, Aaron Courville & Pascal Vincent, *Representation Learning: A Review and New Perspectives*, 35 IEEE TRANSACTIONS ON PATTERN ANALYSIS & MACHINE INTELLIGENCE 1798 (2013), doi: 10.1109/TPAMI.2013.50.

56. *See* IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 19 (2016).

57. *Id.*

58. A more formal description appears in David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams, *Learning Representations by Back-Propagating Errors*, 323 NATURE 533, 533 (1986), doi: 10.1038/323533a0.

59. *See* GOODFELLOW, BENGIO & COURVILLE, *supra* note 56, at 165.

60. *See id.*

61. *See id.*

62. *See* Prakash Jay, *Transfer Learning Using Keras Towards Data Science*, MEDIUM (Apr. 15, 2017), https://medium.com/towards-data-science/transfer-learning-using-keras-d804b2e04ef8 (noting that with "small" datasets of under 40,000 examples "it is difficult to achieve decent accuracy" for computer vision problems).

63. *See* Perelman School of Medicine, *Multimodal Brain Tumor Segmentation Challenge 2017*, SBIA, http://braintumorsegmentation.org/ (last visited Jan 29, 2018).

64. Some ML algorithms are trained by *unsupervised learning*, i.e., by recognizing patterns using test data that has not been labeled, classified, or categorized by

Electronic copy available at: https://ssrn.com/abstract=3114347

algorithm relies on having that pairing between Input and Output in the data, and the process of "training" the ML system corresponds to the computer learning how to set its own representation so as to reliably select a good output for any new input it might observe.[65] A key component of the training procedure is to assess the expertise level of the ML algorithm throughout training. This is typically done by keeping a portion of the data—e.g., 10%—aside as a "validation set," against which the results of the training will be evaluated using the specified Evaluation criteria.[66]

Another significant feature for our purposes is that neural network systems are rarely static. Even after the successful processing of the initial training data, there are many reasons why one would want to give a deep learning ML additional data to digest.[67] The most obvious is that additional data offers the possibility of better predictions.[68] This is true when the new data is simply a greater quantity of the same type of data—e.g., more x-rays graded by experts—and when assuming the data comes from the same distribution—i.e., collected in the same way, annotated in the same way, from the same type of patients. However, it is not inevitably the case that more data is always better; in particular, data collected from a different hospital, potentially with slight variations in procedure, may confuse the ML system. It is important to be vigilant about the quality of the data used to train the system and, in particular, to ensure that the data used for training is collected under the same conditions as the ML system will be used in practice.[69] If the inputs from which the ML is to make its decision change in some way over time, the deep-learning system will need to be retrained with new representative data. Changes in data distribution are not uncommon and might be due to quality degradation caused by aging equipment[70] or quality improvements resulting from the invention of better and more accurate data-acquisition equipment—e.g., the invention of better-quality imaging machines. Without representative examples of the new information, the AI will not be able to make the best predictions from them[71] and indeed could, in theory, go badly wrong.[72]

Due to the very large number of variables, large neural networks are often thought to have a black-box quality. In reality, it is possible to track very precisely

---

humans. Current state-of-the-art for these techniques still lags behind supervised learning, which uses data tagged by humans, so we do not dwell on these approaches here.

65. *See* BISHOP, *supra* note 46, at 2.
66. *Id.* at 11.
67. *See generally id.*
68. *Id.* at 6.
69. *Id.* at 9–10.
70. *See* ETHEM ALPAYDIN, INTRODUCTION TO MACHINE LEARNING 275 (2014).
71. *See id.* at 286.
72. *See* MASASHI SUGIYAMA & MOTOAKI KAWANABE, MACHINE LEARNING IN NON-STATIONARY ENVIRONMENTS 3–19 (2012).

Electronic copy available at: https://ssrn.com/abstract=3114347

the computation at each neuron and each layer.[73] However, it is often difficult to extract a simple explanation for the decision at the end layer (output), because it depends on the combination of many small decisions by each neuron.[74] This highlights an important distinction: most ML algorithms have high traceability (they run on a computer, and can be re-run several times to generate the same results), but poor explainability (they cannot extract a compact narrative explaining the logic behind their reasoning).[75] In contrast, humans tend to have poor traceability (difficult to track, at the neural level, reasons for our decisions), but high interpretability (we can easily construct narratives to explain our behaviors[76]). Neural networks, in particular, do not typically extract causal relationships between inputs and outputs; therefore, it is important to interpret any relationship between input and output as a predictive one, no matter how intuitive such relationships might look on the surface.[77]

2. Our Assumptions About Tomorrow

For the purposes of this Article, we make two predictive assumptions: one about AI's capabilities and one about its limits. Regarding AI's abilities, we assume that at some future date—which may come soon—an ML will be shown to be measurably superior to humans in some specialized aspect of diagnostic medicine. We make this assumption because current trends point strongly in that direction given ML's advances in tumor-detection[78] as well as other areas.[79] For our purposes—and those of the legal system—a new diagnostic technique, such as an ML system, is superior if its diagnostic accuracy is greater to a statistically significant degree. For simplicity, we assume here that the ML system either makes fewer false positives (Type-I errors) and no more false negatives[80] (Type-II errors), or that it makes fewer false negatives and no more false positives, or that

---

73. *See* Dave Gershgorn, *MIT Researchers Can Now Track AI's Decisions Back to Single Neurons*, QUARTZ (July 11, 2017), https://qz.com/1022156/mit-researchers-can-now-track-artificial-intelligences-decisions-back-to-single-neurons/.

74. *See* Bengio, Courville & Vincent, *supra* note 55, at 1803.

75. *See* Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter & Lalana Kagal, *Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning*, ARXIV:1806.00069 (2018), https://arxiv.org/abs/1806.00069.

76. However, that those narratives are in fact accurate ought not to be assumed. *See* Zachary C. Lipton, *The Mythos of Model Interpretability*, ARXIV:1606.03490 (Mar. 6, 2017), http://arxiv.org/abs/1606.03490 (noting "black box nature of human brain").

77. *See* Cary Conglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1173 (2017); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION 87 (2016); FRANK PASQUALE, THE BLACK BOX SOCIETY (2015).

78. Chockley & Emanuel, *supra* note 24.

79. For example, ML has made significant progress advancing computer vision, speech recognition, and machine translation. *See supra* text accompanying notes 12–26.

80. Unsurprisingly, false negatives are the errors most likely to create malpractice claims in radiology. *See* Antonio Pinto & Luca Brunses, *Spectrum of Diagnostic Errors in Radiology*, 2 WORLD J. RADIOL. 377, 377 (2010), doi: 10.4329/wjr.v2.i10.377.

Electronic copy available at: https://ssrn.com/abstract=3114347

the ML system improves on humans to a statistically significant extent in both types of error.[81]

It is also likely that even if an ML system has a better success rate than the average human doctor, ML and humans combined might be even better.[82] There are some reasons to suspect that today the combination might beat either one alone, as is the case in "centaur chess."[83] We also know that, at present, neural networks can make confident but erroneous identifications that no human would make.[84] Keeping a human around protects against those obvious errors and might protect against other kinds of errors as well.

Indeed, if machine + human is demonstrably better than machine alone, then the combination should become the standard of care through the ordinary operation of the legal system without the need for external intervention unless the combination is seen as prohibitively expensive.[85] At least until ML gets very good, there are scenarios in which the human doctor's role evolves more than evaporates. If ML makes prediction and correlation cheaper, that arguably increases the value of other inputs.

However, even in this scenario machine + human remains the standard of care only so long as AI technology does not improve to where the ML system alone is as good at some activity as machine + human. At that point, we posit, the ML system alone becomes, or suffices to meet, the standard of care for that

---

81. It is also possible that malpractice law might determine that an ML system that made substantially fewer false-negative diagnoses but also a small number of increased false positives was legally superior either on its own or in conjunction with a human diagnostician, but we need not consider that distracting case to make our argument.

82. For context, see *infra* text accompanying notes 324–31.

83. "The best chess players in the world are human-machine teams"—so long as teams are not time-limited for moves. PAUL SCHARRE, CTR. FOR A NEW AM. CENTURY, AUTONOMOUS WEAPONS AND OPERATIONAL RISK 39 (2016).

84. *See* Anh Nguyen, Jason Yosinski & Jeff Clune, *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, 4 IEEE COMP. VISION & PATTERN RECOGNITION 427 (2015) (discussing "a project that used neural networks to predict the probability of death for patients with pneumonia, so that low-risk patients could be treated as outpatients. The results were generally more accurate than those that came from handcrafted models that applied known rules to the data. But the neural network clearly indicated that asthmatic pneumonia patients are at low risk of dying and thus should be treated as outpatients. This contradicts what caregivers know, as well as common sense. It turns out that the finding was caused by the fact that asthmatic patients with pneumonia are immediately put into intensive care units, resulting in excellent survival rates"); *see also* David Weinberger, *Alien Knowledge*, WIRED: BACKCHANNEL (Apr. 18, 2017), https://backchannel.com/our-machines-now-have-knowledge-well-never-understand-857a479dcc0e.

85. For an interesting description of a user-centered design intended to overcome physician reluctance to consult an AI, see Cliff Kuang, *An Ingenious Approach to Designing AI that Doctors Trust*, CO.DESIGN (Jan. 17, 2018), https://www.fastcodesign.com/90157144/an-ingenious-approach-to-designing-ai-that-doctors-trust (describing work of Prof. John Zimmerman on decision support for cardiac surgeons).

Electronic copy available at: https://ssrn.com/abstract=3114347

activity—e.g., diagnosis—and the problems discussed below all reappear, making a policy intervention necessary. Perhaps at that point humans will need to switch to other activities such as "the application of ethics, and for emotional support"—and indeed, if ML allows us to diagnose and treat more diseases, the demand for those activities could increase.[86]

Conversely, for simplicity, we assume that the diagnostic specialty in which the AI excels is one that ordinarily takes place away from the point of care, or if it is at the point of care forms only a part of the care-provider's diagnostic responsibilities. This second assumption allows us to assume that there will still be a physician present at the point of care, e.g., an oncologist who ordinarily would be informed by consulting with a radiologist but instead turns to an ML system.[87] In so doing we can avoid engaging, at least for now, with long-standing medical-ethics debates about the appropriateness of fully robotic care.[88]

As set out in the next Section, once ML diagnostics are statistically superior to humans, it will only be a short while before legal systems, including in the United States, treat machine diagnosis as the "standard of care." That designation will mean that any physician or hospital failing to use machine diagnosis without a good excuse will be running a substantial risk of malpractice liability if the patient is incorrectly diagnosed.[89] In a fairly short time, every insurance company and every hospital will require the use of ML, at least as an assistant to physicians, because failure to do so will be actionable in the event of a bad outcome. There are some variables that might alter how quickly this will happen: notably cost and whether courts continue to make distinctions between types of practices and types of practice situations, e.g., teaching hospitals versus rural hospitals versus sole practitioners. But these are primarily questions of speed and detail rather than of trend. In fairly short order, it seems highly plausible that ML systems will be prescribed not by doctors but by tort law for certain forms of diagnosis and that medical service providers will comply. And, if an ML system

---

86. Ajay Agawal, Joshua Gans & Avi Goldfarb, *The Simple Economics of Machine Intelligence*, HARV. BUS. REV. (Nov. 17, 2016), https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence.

87. One very substantial difference between consulting with a human oncologist and "consulting" with a computerized system is that there is no opportunity for any discussion or give and take. An AI gives a report but can neither explain it nor alter it in light of reasoned argument. This could be a real loss to the quality of care, although it is possible that increasing reliance on electronic health records as means of communication between specialists has already eroded those conversations and relationships.

88. The so-called Standard View of biomedical ethics holds "that the *practice of medicine and nursing* are ineluctably human." KENNETH W. GOODMAN, ETHICS, MEDICINE, AND INFORMATION TECHNOLOGY 26 (2015) (citing R.A. Miller, *Why the Standard View is Standard: People, Not Machines, Understand Patient's' Problems*, 15 J. MED. & PHIL. 581, 581 (1990)).

89. *See* Patricia Kuszler, *Telemedicine and Integrated Health Care Delivery: Compounding Malpractice Liability*, 25 AM. J.L. & MED. 297, 316–17 (1999); Kori M. Klustaitis, *Dr. Watson Will See You Now: How the Use of IBM's Newest Supercomputer Is Changing the Field of Medical Diagnostics and Potential Implications for Medical Malpractice*, 5 BIOTECHNOLOGY & PHARMACEUTICAL L. REV. 88, 101–02 (2011–2012).

Electronic copy available at: https://ssrn.com/abstract=3114347

proves statistically superior for treatment, then a similar argument will also apply. In which case, hospitals and other medical service providers will carry out AI-recommended treatment plans unless there is a very clear reason to do otherwise.

B. How Tort Law Incorporates Technical Change

Medical malpractice law is a species of negligence law, which itself is a type of tort, a civil wrong.[90] Physicians can commit malpractice by failing to get informed consent (an issue not especially relevant here), or by breaching their duty to provide the appropriate standard of care in a manner that causes injury to the patient.[91] Defining the relevant standard of care is thus a central issue in many malpractice cases.[92]

The standard of care for a doctor is, at the most general level, that of a reasonably competent physician,[93] i.e., one who uses a reasonable degree of care and skill.[94] While there can of course be evidentiary issues as to what a physician actually did, in cases that involve whether a physician should have used a particular, relatively new technology there can also be complicated questions as to whether the use of the new technology—or the failure to use the new technology—is itself negligence.[95] Using new technology also may invite claims that perhaps the people who used it were not yet sufficiently familiar with it and thus used it improperly.[96]

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.[97] Indeed, where once "custom"—what most people in the trade or profession do and have generally done—was the starting point for measuring the appropriate standard of care, U.S. courts today are somewhat suspicious of custom-based arguments on the theory that these arguments provide too little incentive to modernize and may favor entrenched modes of service provision at the expense of the victim.[98]

This modernizing tendency traces back at least as far as the oft-cited *T.J. Hooper* case, where Judge Learned Hand ruled that it was negligent for a tugboat sailing the Atlantic in 1928 to fail to have a working radio on board to hear storm-weather warnings.[99] The trial court had found that if the *T.J. Hooper* had carried a

---

90. "Professional negligence is commonly called malpractice." VICTOR SCHWARTZ ET AL., PROSSER, WADE, AND SCHWARTZ'S TORTS: CASES AND MATERIALS 183 (13th ed. 2015) (citing Restatement (Second) of Torts § 299A (1979)).
91. *See id.* at 188, 201–02.
92. *See id.* at 190 (citing Boyce v. Brown, 77 P.2d 455 (Ariz. 1938)).
93. *See id.* at 188 (citing Lucas v. Hamm, 364 P.2d 685 (Cal. 1961)).
94. *See id.* at 193.
95. *See id.* at 193–94; Morrison v. MacNamara, 407 A.2d 555 (D.C. 1979).
96. *See* SCHWARTZ ET AL., *supra* note 90, at 198.
97. *See id.* at 189–91 (citing Boyce v. Brown, 77 P.2d 455 (Ariz. 1938)).
98. *See id.* at 161–64 (citing Trimarco v. Klein, 436 N.E.2d 502 (N.Y. 1982)).
99. The T.J. Hooper, 60 F.2d 737 (2d Cir. 1932) (L. Hand, J.).

Electronic copy available at: https://ssrn.com/abstract=3114347

radio, it likely would not have foundered.[100] On appeal, Judge Hand first noted that there was no general and established custom of carrying a radio among coastwise carriers, and he admitted that courts sometimes treated the absence of such a custom as a full defense.[101] But he also noted that a suitable radio was not expensive[102] and that custom should not be definitive:

> [A] whole calling may have unduly lagged in the adoption of new and available devices. It never may set its own tests, however persuasive be its usages. Courts must in the end say what is required; there are precautions so imperative that even their universal disregard will not excuse their omission. But here there was no custom at all as to receiving sets; some had them, some did not; the most that can be urged is that they had not yet become general. Certainly in such a case we need not pause; when some have thought a device necessary, at least we may say that they were right, and the others too slack.[103]

The rule in *T.J. Hooper* has, to some extent, been subsumed into more general negligence rules about how to balance the cost of prevention against expected benefits or risks. These modern rules are often traced to another Learned Hand opinion, in the even more celebrated *Carroll Towing* case.[104] This case gave rise to the so-called Hand Formula in which the test for negligence is whether the cost of a precaution would be more or less than the expected value of the gain in safety.[105] Then-Professor Richard Posner identified *Carroll Towing* as "one of the few attempts to give content to the deceptively simple concept of ordinary care."[106]

Since *The T.J. Hooper* and *Carroll Towing*, U.S. courts have not been shy about demanding additional precautions where the likely benefits seemed to outweigh the costs even when an industry resisted them[107]—except in the case of medicine, where until recently the courts have been more cautious.

C. Medical Variations: Custom and Localities

To succeed in a medical malpractice case, the plaintiff must show that his or her injury, more likely than not, resulted from the treating physician's departure from "the generally recognized and accepted practices and procedures that would

---

100. The T.J. Hooper, 53 F.2d 107, 111 (S.D.N.Y. 1931), *aff'd* 60 F.2d 737 (2d Cir. 1932).

101. The T.J. Hooper, 60 F.2d at 740.

102. *Id*. at 739.

103. *Id.* at 740 (citations omitted).

104. *See* United States v. Carroll Towing Co., 159 F.2d 169 (2d Cir. 1947) (L. Hand, J.).

105. For a discussion of the origins of the Hand Formula as expressed in *Carroll Towing*, see Stephen G. Gilles, United States v. Carroll Towing Co.*: The Hand Formula's Home Port, in* TORTS STORIES 11 (Robert L. Rabin & Stephen D. Sugerman eds., 2003).

106. Richard Posner, *A Theory of Negligence*, 1 J. LEGAL STUD. 29, 32 (1972).

107. *See, e.g.*, Bimberg v. Northern Pacific Ry., 14 N.W.2d 410, 413 (Minn. 1944) ("Local usage and general custom, either singly or in combination, will not justify or excuse negligence.").

Electronic copy available at: https://ssrn.com/abstract=3114347

be followed by the average, competent physician in the defendant's field of medicine under the same or similar circumstances."[108] What constitutes average competence in a given field of medicine is a question of fact, for which parties commonly offer expert testimony.[109]

In contrast, who makes up the set of comparable physicians is primarily an issue of law.[110] For many years, physicians, almost alone among professionals and tradespeople, enjoyed two special protections from professional-negligence liability, both relating to who counted as comparable: a heightened ability to plead custom as a defense,[111] and the "locality rule."[112] The effect of these two rules was to insulate physicians from liability so long as they provided treatment no worse than was common in their community.[113] Because physicians were reluctant to testify against their colleagues until fairly late in the twentieth century, these rules worked to greatly limit malpractice claims.[114]

1. The Waning of the Locality Rule

The locality rule reflected a judicial belief that it would be unfair to apply a single standard of care to all physicians.[115] Physicians vary as to their training

108. Hoard v. Roper Hosp., Inc., 694 S.E.2d 1, 4 (S.C. 2010); *see also* Pike v. Honsinger, 49 N.E. 760 (N.Y. 1898). The basic elements of the tort of negligence are duty, breach, causation, and injury.

109. "In most charges of negligence against professional persons, expert testimony is required to establish what the reasonable practice is in the community. The conduct of the defendant professional is adjudged by this standard. Without such expert testimony a plaintiff cannot prove negligence." Getchell v. Mansfield, 489 P.2d. 953, 955 (Or. 1971).

110. *See*, *e.g.*, Brune v. Belikoff, 235 N.E.2d 793 (Mass. 1968) (upholding decision by trial court that, as a matter of law, relevant comparatives for specialist doctor were national not local).

111. Tim Cramm, Arthur J. Hartz & Michael D. Green, *Ascertaining Customary Care in Malpractice Cases: Asking Those Who Know*, 37 WAKE FOREST L. REV. 699, 699–700 (2002) ("Medical malpractice law has long modified the ordinary tort duty of reasonable care. Health care professionals must exercise the same care that other professionals customarily exercise. Thus, the duty applied to medical professionals is a purely factual one, unlike the normative 'reasonable care' standard invoked for non-professionals."). *But see* Steven Hetcher, *Creating Safe Social Norms in a Dangerous World*, 73 S. CAL. L. REV. 1 (1999) (critiquing reliance on custom as a measure of negligence).

112. *See infra* Subsection I.C.1.

113. *See* Theodore Silver, *One Hundred Years of Harmful Error: The Historical Jurisprudence of Medical Malpractice*, 1992 WIS. L. REV. 1193, 1234 n.133 (1992) (collecting cases).

114. Dean William Prosser, for example, referred in the 1955 edition of his Torts treatise to "[t]he well known reluctance of doctors to testify against one another, which has been mentioned now and then in the decisions." WILLIAM PROSSER, TORTS § 31, at 134 (2d ed. 1955).

115. *See* Small v. Howard, 128 Mass. 131, 132 (1880) (holding that defendant small-town surgeon "was bound to possess that skill only which physicians and surgeons of ordinary ability and skill, practising in similar localities, with opportunities for no larger experience, ordinarily possess; and he was not bound to possess that high degree of art and

Electronic copy available at: https://ssrn.com/abstract=3114347

and specialization, and also in their practice settings. A general practitioner should not be expected to have the same skill as a specialist, at least in matters touching on that specialty.[116] A small, rural practice does not have access to the same equipment as a large, urban teaching hospital;[117] many courts also seemed influenced by the idea that it would be unfair to expect the prototypical rural practitioner to be as up-to-date as someone affiliated with a major hospital.[118] Precisely what the comparatives were varied slightly: other physicians with similar training in the same or a similar community, or perhaps other physicians with similar training in similar communities in the state.[119]

Today the standard of care for physicians is increasingly national, reflecting the relative standardization of medical training. Physicians continue to be held to a varying standard depending on their training and type of practice, but the standard applied to members of a given specialty is more or less uniform nationally.[120] The standard of care is that established by the "relevant community," which is now understood to be the national group of practitioners in that specialty.[121] To whatever the extent the locality rule lives on, it applies primarily to general practitioners.[122]

2. Custom in Medical Malpractice Meets Technological Change

U.S. courts have, at least until recently, tended to accept evidence of customary practices as persuasive defenses against claims of medical

---

skill possessed by eminent surgeons in large cities, and making a specialty of the practice of surgery").

116. *See* James O. Pearson, Jr., Annotation, *Modern Status Of "Locality Rule" in Malpractice Action Against Physician Who Is Not a Specialist*, 99 A.L.R.3d 1133 (1980).

117. "[C]ity doctors are likely to be more advanced than their rural counterparts, teaching hospitals are more likely to employ the latest techniques and technologies than are nonteaching hospitals." Mark F. Grady, *Better Medicine Causes More Lawsuits, and New Administrative Courts Will Not Solve the Problem*, 86 NW. U. L. REV. 1068, 1073 (1992) (reviewing PAUL C. WEILER, MEDICAL MALPRACTICE ON TRIAL (1991)).

118. "The rule, in its early form, was demonstrably calculated to protect the rural and small town practitioner, who was presumed to be less adequately informed and equipped than his big city brother." Jon R. Waltz, *The Rise and Gradual Fall of the Locality Rule in Medical Malpractice Litigation*, 18 DEPAUL L. REV. 408, 410 (1969).

119. *See* Scott A. Behrens, Note, *Call in Houdini: The Time Has Come to Be Released from the Geographic Straitjacket Known as the Locality Rule*, 56 DRAKE L. REV. 753, 754–64 (2008) (tracing origins and evolution of the locality rule); Pearson, *supra* note 116.

120. Jordan v. Bogner, 844 P.2d 664, 666 (Colo. 1993); *see also* Gerald L. Michaud & Mark B. Hutton, *Medical Tort Law: The Emergence of a Specialty Standard of Care*, 16 TULSA L.J. 720, 730 (1981); Waltz, *supra* note 118, at 418.

121. *See Jordan*, 844 P.2d at 666.

122. *See* STUART M. SPEISER ET AL., 4 AMERICAN LAW OF TORTS § 15:19 (March 2018 Update) (surveying varying application of locality rule to non-specialist doctors). *But see* Waltz, *supra* note 118, at 420 (concluding that there will soon be a national standard for general practitioners, albeit one lower than for specialists).

Electronic copy available at: https://ssrn.com/abstract=3114347

negligence.[123] The rule has been strongly criticized for deterring medical innovation.[124] If the standard of care is defined by custom, then any physician who innovates takes on the risk of deviating from custom. If the innovative practice or device causes harm, that creates an exposure to malpractice liability for "unreasonable" behavior even if, on average, the innovation is beneficial.[125]

In part due to such criticism, and perhaps also due to the erosion of the view that physicians should be above criticism,[126] the privileged position of physicians that allowed them to plead custom in malpractice cases has greatly diminished:

> Gradually, quietly and relentlessly, state courts are withdrawing this legal privilege. Already, a dozen states have expressly rejected deference to medical customs and another nine, although not directly addressing the role of custom, have rephrased their standard of care in terms of the reasonable physician, rather than compliance with medical custom.
>
> Even more important than the raw numbers is the trend revealed by the decisions. The slow but steady judicial abandonment of deference to medical custom began in earnest in the 1970s, continued in the 1980s, and retained its vitality through the 1990s. Showing no signs of exhaustion, this movement could eventually become the majority position.
>
> Furthermore, many of the states that theoretically continue to defer to custom actually apply the custom-based standard of care in a way that operates very much like a reasonable physician standard.[127]

In other words, in more and more states,[128] the physician's duty under malpractice is being normalized and brought into alignment with the ordinary tort duty of care, permitting courts to hold that even widespread medical practices can

---

123. How and why that came to be is itself controversial. *See* Silver, *supra* note 113 (arguing that the move away from ordinary negligence rules for the medical profession was a mistake).

124. *See generally* Gideon Parchomovsky & Alex Stein, *Torts and Innovation*, 107 MICH. L. REV. 285 (2008).

125. *See id.*

126. Public deference to the judgment of medical professionals has gradually declined since World War II. *See generally* Philip G. Peters, Jr., *The Quiet Demise of Deference to Custom: Malpractice Law at the Millennium*, 57 WASH. & LEE L. REV. 163 (2000).

127. *Id.* at 164; *see also* Behrens, Note, *supra* note 119, at 770–72 (concluding "[t]he movement of nearly all jurisdictions has been to incorporate a national standard of care, and those that have not had the right case arise have continued to loosely apply the similar locality rule").

128. By 2009, "almost half of the states [had] adopted an objective 'reasonable care' standard" instead of one "based on what the majority of medical practitioners actually do." Michael D. Greenberg, *Medical Malpractice and New Devices: Defining an Elusive Standard of Care*, 19 HEALTH MATRIX 423, 428–29 (2009).

Electronic copy available at: https://ssrn.com/abstract=3114347

be negligent[129]—particularly if the innovations that the physician has not adopted are "precautions so imperative that even . . . universal disregard will not excuse their omission."[130] Indeed, as a general matter, the standard of care is not only national but also subject to reasonably rapid change when confronted with a breakthrough technology.[131] Thus, for example, courts routinely required doctors to use x-rays to diagnose fractures "very quickly after the technology was introduced."[132] It was not long until the failure to take a diagnostic x-ray was "so clearly negligent as to constitute *res ipsa loquitur* . . . an obvious failure to follow accepted medical practice."[133]

---

129. Peters, *supra* note 126, at section II.B. (citing cases). Interestingly, studies show that as states switch from a custom-based measure of the standard of care to a national standard based on reasonableness, the rate of adoption of innovations converged to the national mean. This suggests that "this change in behavior was motivated by the change in tort law's test of reasonable care, not by any independent medical evaluation of whether compliance with the local or national custom was in the best interests of the patient." Mark Geistfeld, *Does Tort Law Stifle Innovative Medical Treatments?*, JOTWELL (June 2, 2015) (reviewing Anna B. Laakmann, *When Should Physicians Be Liable for Innovation?*, 36 CARDOZO L. REV. 913 (2015)), http://torts.jotwell.com/does-tort-law-stifle-innovative-medical-treatments/).

130. The T.J. Hooper, 60 F.2d 737, 739 (2d Cir. 1932) (L. Hand, J.).

131. *See* Patricia Kuszler, *Telemedicine and Integrated Health Care Delivery: Compounding Malpractice Liability*, 25 AM. J. L. & MED. 297, 316–17 (1999). On the physician's duty to keep informed of new treatment methods, see Jolene S. Fernandes, *Perfecting Pregnancy via Preimplantation Genetic Screening: The Quest for an Elusive Standard of Care*, 4 U.C IRVINE L. REV. 1295, 1308–12 (2014); Alan Weintraub, *Physician's Duty to Stay Abreast of Current Medical Developments*, 31 MED. TRIAL TECH. Q. 329 (1985); Carter L. Williams, Note, *Evidence-Based Medicine in the Law Beyond Clinical Practice Guidelines: What Effect Will EBM Have on the Standard of Care?*, 61 WASH. & LEE L. REV. 479, 508–12 (2004). Consider also Harbeson v. Parke-Davis, Inc., 656 P.2d 483 (Wash. 1983) (holding that physician's failure to conduct literature search on side effects of Dilantin justified liability for wrongful birth).

It should also be noted that some advances in the standard of care are not due to the workings of the tort system and instead arise from statute or regulation. For example, mass screening for Phenylketonuria (PKU) quickly became a national standard after Robert Guthrie discovered a cheap and easy PKU test that could reliably identify asymptomatic infants—this being the time when the potential for treatment is greatest. Mass screening soon followed, but mostly due to government prodding via state laws requiring testing. *See* Dianne B. Paul, *The History of Newborn Phenylketonuria Screening in the U.S.*, *in* PROMOTING SAFE AND EFFECTIVE GENETIC TESTING IN THE UNITED STATES at app. 5 (Neil A. Holtzman & Michael S. Watson eds., 1997), https://biotech.law.lsu.edu/research/fed/tfgt/appendix5.htm.

132. William J. Curran, *The Unwanted Suitor: Law and the Use of Health Care Technology*, *in* THE MACHINE AT THE BEDSIDE: STRATEGIES FOR USING TECHNOLOGY IN PATIENT CARE 119, 123 (Stanley Joel Reiser & Michael Anabar eds., 1984).

133. *Id*. As early as 1928, the court in Lippold v. Kidd, 269 P. 210, 213 (Or. 1928), accepted that failure to take an x-ray of an injured eye could establish a prima facie case for medical negligence. This was a sea change, as less than 20 years earlier a Washington court had held that, in light of testimony that x-rays were used only "as a matter of extreme care," failure to use an x-ray could not be grounds for a directed verdict. Wells

Electronic copy available at: https://ssrn.com/abstract=3114347

More recently, the automated external defibrillator became the standard of care for first responders in 1988 when the Advanced Cardiac Life Support (ACLS), a working group of the American Heart Association, endorsed it.[134] The first articles about clinical use of those defibrillators had appeared in medical journals only a decade earlier, but the national consensus crystalized quickly after studies published in the late 1980s demonstrated their value in improving patient survival.[135]

Indeed, some doctors have criticized the legal system for anointing some technologies as the standard of care too quickly—before the proof is in that they are helpful—and sticking to that judgment even after studies suggest the technology does not live up to its promise. For example, initial studies suggested that Electronic Fetal Monitoring, now used in the large majority of births in the United States, would lead to a 50% reduction in intrapartum deaths, mental retardation, and cerebral palsy.[136] Later studies undermined those optimistic predictions, but the malpractice verdicts continued.[137]

Even custom, when it reigned, did not always prove an iron-clad defense. In 1974, when medical custom was still king, the Supreme Court of Washington held that even though the national standard of care of ophthalmologists did not require routine glaucoma tests, in light of the low cost of the test,

> reasonable prudence required the timely giving of the pressure test to this plaintiff. The precaution of giving this test to detect the incidence of glaucoma to patients under 40 years of age is so imperative that irrespective of its disregard by the standards of the opthalmology [sic] profession, it is the duty of the courts to say what is required to protect patients under 40 from the damaging results of glaucoma.[138]

Although the decision was much criticized at the time, and "no other courts . . . followed the *Helling* case directly,"[139] it was a harbinger of things to come.

---

v. Ferry-Baker Lumber Co., 107 P. 869, 870 (Wash. 1910). By 1961, the requirement of x-rays in cases of injury was clearly established. *See*, *e.g.*, Gonzales v. Peterson, 359 P.2d 307, 310 (Wash. 1961).

134. *See* Richard O. Cummins, *From Concept to Standard-of-Care? Review of the Clinical Experience with Automated External Defibrillators*, 18 ANNALS EMERGENCY MED. 1269, 1270 (1989), doi: 10.1016/S0196-0644(89)80257-4.

135. *See id.* at 1269–70.

136. "EFM is used in approximately 85% of annual births and is the most common obstetrical procedure in the United States during labor." Michael Brook & Kary Irle, *Litigating Intraoperative Neuromonitoring (Iom)*, 45 U. BALT. L. REV. 443, 465 (2016).

137. *Id.*; *see also* Margaret Lent, *The Medical and Legal Risks of the Electronic Fetal Monitor*, 51 STAN. L. REV. 807 (1999) (arguing that auscultation should replace EFM as standard of care); Thomas P. Sartwelle, *Electronic Fetal Monitoring: A Bridge Too Far*, 33 J. LEGAL MED. 313 (2012), doi: 10.1080/01947648.2012.714321 (blaming greedy trial lawyers and "junk science" for rise and persistence of EFM as standard of care).

138. Helling v. Carey, 519 P.2d 981, 983 (Wash. 1974) (citing T.J. Hooper, 60 F.2d 737, 740 (2d Cir. 1932)).

139. Curran, *supra* note 132, at 125.

Electronic copy available at: https://ssrn.com/abstract=3114347

As we have seen with the x-ray and the automated external defibrillator, the standard of care indeed can change quickly. The rise of evidence-based medicine (EBM), which encourages physicians to apply current scientific evidence even before it becomes a custom,[140] arguably encourages this trend. If EBM becomes the meta-standard, then physicians may become liable for not considering the latest evidence, potentially causing fast tracking of malpractice liability.[141]

D. Nature of Machine Learning Removes Common Obstacles to the Adoption of New Medical Technology

Much of the writing and thinking about the interaction between medical negligence rules and technical change concerns clinical techniques or devices that are not unambiguously good for the patients to whom the new technology may be applied. Most of these technologies create new risks as well as benefits;[142] frequently they require new training without which physicians may fear they could fail to reap the benefits of the new technology or even misuse it in a harmful way.[143] Frequently there is concern that not all the long-term risks of the new techniques or devices will necessarily be evident at the time that the physician must decide whether to use the familiar procedure or the new one.[144] Each of these properties creates the specter of tort liability if something goes wrong, creating disincentives that may balance out or even overcome the purported advantages: a bad outcome following a new surgical procedure creates the risk that the patient may claim improper training; a new implantable device creates risks of unforeseen long-term complications or even failure; a new invasive diagnostic procedure may have side effects; some advanced diagnostic equipment may be too expensive to have in every hospital, much less in every physician's office.

ML systems are different from these common examples in many important respects. From the point of view of malpractice risk management, AI diagnostics should be much easier to implement than other recent medical advances that have required expensive equipment be present on-site. ML can be trained to work with any diagnostic materials that can be reduced to standardized data, notably including radiographic images. As the ML is fundamentally a computer program, the analysis need not be done on-site but can instead live anywhere else or even in the cloud.[145] Any medical facility capable of capturing

---

140. Proponents define EBM as "the integration of best research evidence with our clinical expertise and our patient's unique values and circumstances." *See* SHARON E. STRAUS ET AL., EVIDENCE BASED MEDICINE: HOW TO PRACTICE AND TEACH IT 1 (2011).

141. *See* E Monico, C Moore & A Calise, *The Impact of Evidence-Based Medicine and Evolving Technology on the Standard of Care in Emergency Medicine*, 3:2 INTERNET J. OF L, HEALTHCARE AND ETHICS 1 (2004).

142. *See* Greenberg, *supra* note 128, at 436.

143. *See id.* at 435–36.

144. *See id.* at 430, 445–46.

145. Any remote location and especially cloud-based services raise issues of security and privacy outside the scope of this Article. *See, e.g.*, Sebastian Zimmeck, *The Information Privacy Law of Web Applications and Cloud Computing*, 29 SANTA CLARA COMPUTER & HIGH TECH. L.J. 451, 469–82 (2013) (surveying risks of information disclosure); Warwick Ashford, *Cloud Computing Presents a Top Security Challenge*,

Electronic copy available at: https://ssrn.com/abstract=3114347

clinical information, digitizing it, and transmitting it, could presumably access an ML-based computer located anywhere else, so long as the cost was affordable.

In short, the data collection needs to be done at the point of care, where the patient is—the data input and the processing can be done anywhere. Rather than being equipment or a technique, ML systems present as a service. Unless the pricing is extortionate, this will not only increase the rate at which medical service providers adopt ML systems, but also increase the speed with which hospitals, and even local physicians, feel legal pressure to use ML.[146]

However, there is one way in which ML may not be different from other medical innovations: it will not be immune to all malpractice claims. Even if we can prove that an ML system, on average, is a better diagnostician than the average physician, that will not mean it is incapable of actionable error. For example, a patient misdiagnosed by an ML might claim that, even if the ML's overall average is better than most or all humans, a significant part of the ML's success occurs in cases where humans would have failed, and that a significant part of the ML's errors fall on a group of patients who might have fared better with a human doctor.[147] The misdiagnosed patient could claim that he or she fell into the group who would have fared better with an average—or a particular—human physician. Simply put, humans and ML systems might make very different kinds of mistakes. And these differences might affect the manner in which liability is assessed.

Currently, we tend to train ML systems from databases that reflect the best judgments of panels of practicing physicians.[148] One could, in theory, train on actual real-world outcomes if the medical system commonly annotated diagnostic data files with outcome data at regular intervals. At present, however, it is not common to find, say, a database containing radiological images linked with data about whether and which tumors manifested in the patients over a set period of time. Given the hypothesis on which this Article is based—that an ML system has managed to do substantially better on average than do human physicians—we

---

COMPUTERWEEKLY (Dec. 10, 2008, 4:43 PM), http://www.computerweekly.com/Articles/2008/12/10/233839/cloud-computing-presents-a-top-security-challenge.htm; J. Aikat et al., *Rethinking Security in the Era of Cloud Computing*, IEEE SECURITY PRIVACY, May–June 2017, at § 3.1, doi: 10.1109/MSP.2017.80 (summarizing top cloud-security threat types).

146. Watson as a service also raises some complex issues of what standards of liability would apply to Watson's errors. *See* Jessica S. Allain, Comment, *From Jeopardy! to Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems*, 73 LA. L. REV. 1049 (2013). It also raises potentially difficult problems of proof, as one would need a perfect snapshot of the entire medical database on which the ML could have relied at the moment of treatment to prove that had the ML been consulted it would have made a better decision than the human. Unfortunately, these issues are beyond the scope of this Article.

147. *See* Millar & Kerr, *supra* note 11.

148. *See* Fei Jiang, Yong Jian, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen & Yongjun Wang, *Artificial Intelligence in Healthcare: Past, Present and Future*, 2 STROKE & VASCULAR NEUROLOGY 230 (2017), doi:10.1136/svn-2017-000101.

Electronic copy available at: https://ssrn.com/abstract=3114347

would not expect in the short term[149] that the ML system's errors would tend to be in cases that humans would, on average, have diagnosed correctly. Nevertheless, because that tendency is only a matter of *probability*, the *possibility* cannot be excluded as a provable or mathematical certainty in general or indeed in any given case. Worse, as described in Subsection I.A.1, the current state of the art for neural networks, with its lack of interpretability, creates some circumstances in which there is no practical way for humans to examine the reasoning for any given decision.[150] Furthermore, the lack of causal connections of the sort humans typically use to understand reasoning makes it difficult to pinpoint a specific source of error in the ML-based prediction system. Any given diagnosis is the result of correlations based on the entire medical database available at the moment of diagnosis. As a result, given current technology,[151] a physician or hospital relying on a neural network cannot back up any particular decision with evidence of a reasoned decision-making process beyond pointing to the program's overall batting average and perhaps (if the system is programmed to provide it) to an evidence profile that shows how it weighed different classes of information[152] or perhaps to some number indicating the neural network's degree of confidence in its diagnosis.[153] Thus, for example, if a hospital is relying on ML for its diagnosis, both parties in a resulting malpractice action will be free to provide *ex post* rationalizations based on expert testimony by humans, but while defendants relying on the ML system will have a chance to argue that the ML system made the right call on the merits, the defendants may have the disadvantage of not being able to explain how the actual decision came to be.

A neural network can learn from its successes and its mistakes—that is the key to how it is trained initially. So long as its decisions are being reviewed by human physicians on an ongoing basis we would hope that its success rate continues to improve as its training data incorporates new information based on the physicians' input.[154] Likewise, such systems will improve as the quality and quantity of data increases.[155] Most commonly this would happen in batch mode, not real time: scientists train models first and deploy them into the wild in a static form.[156] They might then release updated versions later that take into account new

149. We return to the issue of relative long-term accuracy in Part III.

150. *See supra* text accompanying note 74.

151. For a discussion of ongoing efforts to provide explanation see *infra* text accompanying notes 324–28.

152. For an example of this in the Jeopardy game-show context, see David Ferrucci et al, *Building Watson: An Overview of the DeepQA Project*, AI MAG., Fall 2010, http://www.aaai.org/Magazine/Watson/watson.php, in which weights are given to "location," "passage support," "popularity," "source reliability," and "taxonomic" categories for the answer to the question "Chile shares its longest land border with this country."

153. *See* G. Papadopoulos, P.J. Edwards & A.F. Murray, *Confidence Estimation Methods for Neural Networks: A Practical Comparison*, 12 IEEE TRANSACTIONS ON NEURAL NETWORKS 1278 (2001), doi: 10.1109/72.963764.

154. *But see supra* text accompanying note 70.

155. *See* Jay, *supra* note 62.

156. *See id.*

Electronic copy available at: https://ssrn.com/abstract=3114347

data. Working in batch mode allows for testing between releases and makes it easier to avoid error that can occur if the neural network is learning in real time.[157] Furthermore, we would expect that, prior to the adoption of ML diagnosticians, researchers would have studied ML's outcomes carefully to see if any patterns of error emerge. Perhaps doctors using ML diagnoses could be warned not to rely on them for any identifiable sub-classes of cases where humans were still superior. However, it is worth noting that the search for such patterns of error likely would require a careful review process external to the ML system because the ML itself is unlikely to be able to make these distinctions unless the sub-classes to consider can be defined for it in advance. Worse, while doctors should be able to identify some false positives (Type I errors) fairly quickly—e.g., if they operate but find no tumor[158]—false negatives (Type II errors) may take longer to manifest; this may pose real risk to patients if they are misdiagnosed as a result of reliance on the ML system because early detection and intervention are the key to cancer survival rates.[159] Ideally, rigorous external review would keep the number of meritorious malpractice claims based on a robust ML system's diagnoses low and should keep the number of successful claims low as well, but the technical obstacles to achieving this ideal may be substantial.

E. Malpractice Law Will Require Machine Learning Systems When They Are Demonstrably Better

It is important to recall two basic rules of malpractice law: bad outcomes do not necessarily mean there was malpractice, and physicians are not expected to be perfect.[160] Sadly, there are some cases that cannot be cured with even the best medical care in the world. A physician (or hospital, or insurer) relying on an ML 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 ML will be a human medical judgment like any other. As noted above, the law requires only that physicians exhibit the ordinary skill and judgment of a reasonably competent, similarly situated physician.[161] Thus, a physician, hospital, or insurer relying on an ML diagnosis will, at least initially, be held to no higher standard than that of the ordinary physician. Once ML itself becomes the standard of care, ML will raise

157. For an example of the dangers of continual real-time learning, see James Vincent, *Twitter Taught Microsoft's AI Chatbot to be a Racist Asshole in Less than a Day*, THE VERGE (Mar 24, 2016), https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

158. However, it should be noted that some oncological treatment regimens do not involve initial surgery, for example those relying instead on chemotherapy. Error may be harder to detect in such cases since the absence of a subsequent cancer might falsely be attributed to successful treatment.

159. *See Early Detection Facts and Figures*, CANARY FOUND., http://www.canaryfoundation.org/wp-content/uploads/EarlyDetectionFactSheet.pdf (last visited Feb. 22, 2019).

160. "In most situations the best medical treatment in the world cannot provide an absolute guarantee of success; medicine is not an exact science in that sense." McBride v. United States, 462 F.2d 72, 75 (9th Cir. 1972).

161. *See supra* Subsection I.C.1.

Electronic copy available at: https://ssrn.com/abstract=3114347

the bar. But even though a higher level of accuracy will now be the standard, the malpractice exposure of ML-users will actually shrink because by relying on ML they will be complying with the professional standard;[162] at that point, reliance on human diagnosticians will become the risky legal strategy both for failing to use an increasingly common technology of which they should have been aware and because (by hypothesis) the risk of error is in fact greater.

In states that have changed the standard of care to align with general tort principles, one would expect the legal pressure to adopt ML to be very strong once the evidence was clear that an ML system was better than human physicians, for it would be unreasonable to fail to adopt ML unless the cost was very high.[163] In the decreasing number of states that still allow custom to act as a defense, medical malpractice law's definition of the standard of care can act as a brake on innovation. In those states, the legal push to use ML will not be as great until ML is in common use nationally in the relevant specialty; at that point, ML usage itself becomes customary, and we would expect the law to provide a strong push toward compliance with the relevant general norm for any late adopters.[164]

There are more than 15,000 medical malpractice claims filed against healthcare providers in the United States every year.[165] Although a 2015 report by the National Academy of Sciences called diagnostic error the "blind spot" in modern medicine,[166] diagnostic error is increasingly recognized as a major

---

162. One small caveat ought to be noted here: were an ML system to provide a clearly ludicrous diagnosis, one that any reasonable physician ought to have noticed was wrong or inapposite, then—even after it becomes the standard of care—relying on ML in those circumstances could easily be characterized as negligence, and plausibly as gross negligence. This entails a need for continued comprehensive human training, even if the role of human physicians, like pilots, becomes secondary to the role played by machines. *See, e.g.*, Madeline Elish & Tim Hwang, *Praise the Machine! Punish the Human!: The Contradictory History of Accountability in Automated Aviation*, DATA & SOCIETY (Feb. 24, 2015), https://www.datasociety.net/pubs/ia/Elish-Hwang_Accountability AutomatedAviation.pdf.

163. We address this issue in Part II below.

164. There is one persistent exception to this trend: the "two schools of thought" doctrine. Under this doctrine doctors have a powerful defense against a malpractice claim based on failure to adhere to the standard of care if the defendant can show that the treatment provided is supported by a minority of professionals in the field due to disagreement as to which is the optimal treatment. *See generally* Douglas Brown, *Panacea or Pandora' Box: The Two Schools of Medical Thought Doctrine after* Jones v. Chidester, 44 J. URBAN & CONTEMP. LAW 223 (1993). Note that this defense would not generally apply if the minority consisted of doctors unwilling to modernize in the face of a demonstrably better new technique or technology. Use of the defense would be limited to situations where evidence as to which "school" is better is disputed in the medical literature or among experts.

165. *Most Common Causes of Medical Malpractice Claims*, OHIO TIGER, https://ohiotiger.com/common-causes-medical-malpractice-claims/ (last updated April 15, 2016).

166. NATIONAL ACADEMY OF SCIENCE, IMPROVING DIAGNOSIS IN HEALTH CARE 1 (Erin P. Balogh, Bryan T. Miller & John R. Ball eds., 2015).

Electronic copy available at: https://ssrn.com/abstract=3114347

problem: estimates of the prevalence of diagnostic error range from 5% to 20% of physician-patient encounters;[167] "cognitive factors," particularly "premature closure" (being satisfied with an initial conclusion) are a major cause, perhaps even the primary cause, of these errors.[168] Doctors, hospitals, insurers, and any other participants in the healthcare system with exposure to malpractice liability should be particularly attracted to any new technology that promises a substantial reduction in diagnostic error, a major source of malpractice claims.[169]

From the point of view of the tort-law theorist, at least of the law-and-economics persuasion, the idea that fear of malpractice liability would push medical care providers toward using a technology with a lower error rate is a happy story as tort law seems poised to do exactly what theorists would want it to do: it incentivizes a profession to adopt a new technology that likely will save lives.[170] Indeed, even if tort law was neutral or a possible brake, as in the case of custom-dependent states before the national trend develops,[171] once ML's success rate is demonstrably superior to human physicians we would expect that both medical ethics and cost considerations would drive medical care providers to choose to consult an ML system and to rely on its judgments unless they could articulate good reasons not to. Thus, if ML's track record is significantly better than most humans', then arguably ethics would counsel (most[172]) humans to rely on the ML even if they believed they had a superior diagnosis.[173] In time, perhaps even in a short time, a provably superior ML becomes the standard of care for diagnosis in a specialty in many jurisdictions, and certainly throughout the United States.

We turn now to the economic drivers toward ML—and to some speculation about ML's economic consequences. Our happiness may prove temporary.

---

167. Paul A. Bergl et al., *Diagnostic Error in the Critically Ill: Defining the Problem and Exploring Next Steps to Advance Intensive Care Unit Safety*, 15 ANNALS OF THE AM. THORACIC SOC'Y 903, 903 (2018), doi: 10.1513/AnnalsATS.201801-068PS.

168. *See* Mark L. Graber et al., *Diagnostic Error in Internal Medicine*, 165 ARCHIVES OF INTERNAL MED. 1493, 1493, 1498 (2005), doi: 10.1001/archinte.165.13.1493.

169. Misdiagnosis is the most common cause of malpractice claims in outpatient settings; surgical errors are the most common cause of malpractice claims in hospital settings. *Most Common Causes of Medical Malpractice*, *supra* note 165.

170. *See* GUIDO CALABRESI, THE COSTS OF ACCIDENTS 26 (1970) ("I take it as axiomatic that the principal function of accident law is to reduce the sum of the costs of accidents and the costs of avoiding accidents.").

171. *See* Parchomovsky & Stein, *supra* note 124, at 303–08.

172. Presumably Dr. House would demur.

173. *See* Millar & Kerr, *supra* note 11. The argument in the text presupposes that the human physician at least accepts that Watson's diagnosis is plausible. If the human physician believes Watson's diagnosis is erroneous, then he or she will have a duty to step in. *See supra* note 162; *see also infra* text accompanying notes 196–97 (discussing how errors can happen).

Electronic copy available at: https://ssrn.com/abstract=3114347

II. MACHINE LEARNING AND THE DEMAND FOR SPECIALIST PHYSICIANS

A. Machine Learning and the Market for Diagnostic Physicians

Physicians are expensive to train, and expensive to keep on staff.[174] Given the necessity of acquiring training data, formatting it, and establishing compatible data-exchange regimes with hospitals and other medical care providers,[175] we presume that ML diagnostics will follow the path of many other digital technologies and exhibit high fixed costs but relatively low marginal costs.[176] The fixed costs will be the presumably high cost of first priming the system with training data, then arranging for compatible data input from the treating physician's office. The costs of processing individual requests we presume to be low by comparison, although this is, at best, only informed speculation on our part. Magnetic resonance imaging (MRI) may, however, be instructive: early MRI machines cost around $2 million plus $1 million for installation.[177] Modern state-of-the-art devices can cost up to $3 million.[178] Yet failure to use one would in many cases be malpractice. As the high capital cost of an MRI machine can be shared by the many patients who will use it during the machine's lifetime, the per-patient cost is low enough to make an MRI the standard of care, and therefore the standard diagnostic tool, for many different diseases and sets of symptoms.[179]

---

174. In 2017 the median U.S. wage for an internist, one of the lowest-paid medical specialties, was $198,370, while the median anesthesiologist received $265,990. U.S. Bureau of Labor Statistics, *Physicians and Surgeons*, https://www.bls.gov/ooh/healthcare/print/physicians-and-surgeons.htm (last modified June 11, 2018). In 2015, the median salary for a radiologist was about $400,000. R. C. Semelka et al., *Radiologist Income, Receipts, and Academic Performance: An Analysis of Many Nations*, 57 ACTA RADIOLOGICA 1497, 1500 (2016), doi: 10.1177/0284185116633914. Doctors also impose substantial overheads, plus require offices and support staff.

175. For more on the importance of acquiring training data, see *infra* text accompanying notes 308–13.

176. A typical example is an online multiplayer game, or any other service subject to a network effect. These typically involve a fixed setup cost, but the marginal cost of adding additional users is relatively low. *See, e.g.*, Pachinco, *MMORPG's and How They Turn a Profit*, Post on Gaming Discussions Forum, NEOGAF (Nov. 12, 2008), https://www.neogaf.com/threads/mmorpgs-and-how-they-turn-a-profit.341822/ (discussing fixed and variable costs of World of Warcraft). The extreme example is digital publishing, for "[o]nce a work is created, the marginal cost of making an unlimited number of digital copies and distributing them worldwide is zero." Raymond Shih Ray Ku, *The Creative Destruction of Copyright: Napster and the New Economics of Digital Technology*, 69 U. CHI. L. REV. 263, 300 (2002). We assume that because an ML system is fundamentally (expensive) software it tends toward the software-publishing side of the spectrum.

177. Ben L. Holmes, *Current Strategies for the Development of Medical Devices* in TECHNOLOGY AND HEALTH CARE IN AN ERA OF LIMITS 219, 220 (INSTITUTE OF MEDICINE STAFF 1992).

178. Lacie Glover, *Why Your MRI or CT Scan Costs an Arm and a Leg*, FISCAL TIMES (July 21, 2014), http://www.thefiscaltimes.com/Articles/2014/07/21/Why-Your-MRI-or-CT-Scan-Costs-Arm-and-Leg.

179. Klustaitis, *supra* note 89, at § III.2.

Electronic copy available at: https://ssrn.com/abstract=3114347

At present, the smart bet seems to be that ML systems will not be as expensive as a human physician: "Once a model has been 'trained,' it can be deployed on a relatively modest budget."[180] In any plausible cost scenario, however, the medical services provider's financial problem is that unless ML replaces all or part of some other cost—the human doctor being the natural target—ML is just one more cost, whether small, medium, or large. And as is well known, the medical sector is under pressure to cut costs.[181]

Whatever the pricing scenario, the more that an ML system becomes the diagnostician of choice, the less there should be demand for similar human diagnosticians.[182] Instead, all that will be necessary is for someone to collect the patient's data and feed it to the system. Recall our second simplifying assumption above, that Prediction-ML is replacing a consulting specialist, not the point-of-care physician.[183] The legal issues created by purely automated medicine of the Treatment-ML variety are both more remote in time and more complex than those discussed here.[184] If it becomes the case that all that ML requires is the input of data, in many cases those data could be collected by less-trained technicians, just as today nurses or trained medical technicians, not physicians, take blood samples, Electrocardiograms (EKGs), MRIs, and CT-scans. Or, in time, other specially trained AIs may do the intake interview as well.[185]

If someday we remove human doctors entirely from the treatment protocol and have patients treated only by machines, the tort-law frame could change from medical malpractice to products liability. That day, however, is likely much farther away than the scenario we focus on here: one in which point-of-care

---

180. Andrew Beam & Isaac S. Kohane, *Translating Artificial Intelligence into Clinical Care*, 316 J. AM. MED. ASSOC. 2368, 2369 (2016).

181. U.S. health-care spending is projected to outpace growth in the United States' Gross Domestic Product (GDP) by about one percentage point in the coming decade, leading to estimates that health-care spending may account for 19.7% of GDP by 2026, up from 17.9% in 2016. Gigi A. Cuckler et al., *National Health Expenditure Projections, 2017–26: Despite Uncertainty, Fundamentals Primarily Drive Spending Growth*, 37 HEALTH AFFAIRS 482, 482 (2018). "Rising health care costs pose a direct threat to workers' take-home pay, the federal budget, and state government finances." *Cutting Health Care Costs*, CTR. FOR AM. PROGRESS (Aug. 2, 2012) (quoting Citigroup, Inc Vice President Peter Orszag, a former director of the Office of Management and Budget), https://www.americanprogress.org/issues/healthcare/news/2012/08/02/11970/cutting-health-care-costs/.

182. For a general argument that "the number of workers-intellectual as well as manual-is reduced by quantum measures in computer-mediated labor" see Aronowitz & DiFazio, *supra* note 7, at 53.

183. *See supra* text accompanying notes 87–88.

184. For a taste of the issues see Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1 (2018).

185. For an account of an early attempt to train an AI to do patient interviews in China see *Baidu Announces Melody, a New AI-Powered Conversational Bot for Doctors and Patients*, MKT. WIRED (Oct. 11, 2016), http://www.marketwired.com/press-release/baidu-announces-melody-a-new-ai-powered-conversational-bot-for-doctors-and-patients-nasdaq-bidu-2165197.htm.

Electronic copy available at: https://ssrn.com/abstract=3114347

doctors use an AI first as a decision-support tool and then as a substitute for consulting experts in back-office specialties, such as radiology or pathology.

Even in our more imminent scenario, patients injured by AIs on which doctors rely may have product-liability claims against the AI's supplier as well as malpractice claims against its user.[186] Possible patient claims against the people responsible for providing the AI raise complicated questions including whether one would characterize what was provided as a good or a service,[187] and whether a buggy AI would be characterized as suffering from a product defect or a design defect.[188] That characterization could have legal consequences because product-defect claims tend to be strict liability,[189] while the nature and evidentiary requirements of design defect claims are currently contested terrain.[190] Although historically U.S. courts have been reluctant to allow strict-liability claims against doctors using medical technology,[191] the specter of product-liability claims should incentivize AI suppliers to take care to provide high-quality diagnostic services because they will wish to avoid lengthy or expensive lawsuits. At the same time, it is important to acknowledge the plausible counterargument that a characterization of ML-gone-wrong as *any kind* of defect could be misguided because ML is premised on the idea that the software will transcend its initial programming.[192] When an ML learns to make decisions that are unpredictable or unintended, it may not be because anything has gone wrong or because the product (or service) is defective in its performance or design.[193] Emergent behavior is often the very reason to deploy an ML; its departure from human decision-making is often a feature, not a bug.[194]

Either way, medical service providers and insurers will, at first, treat ML diagnosis simply as another tool that is available to physicians. Thus at first, hospitals will feel required to keep the same number of physicians around to double-check what the ML does. This will be costly because the hospitals and insurers will have to pay both the physicians and whoever provides the diagnostic service. In addition, as big-data-based diagnosis takes off, hospitals may be

---

186. *See, e.g.*, Ian Kerr, Jason Millar & Noel Corriveau, *Robots and Artificial Intelligence in Healthcare*, *in* CANADIAN HEALTH LAW AND POLICY 257 (5th ed. 2017).

187. *See* Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 390 (2016) (describing this as a "thorny issue").

188. *See generally* DAVID G. OWENS, PRODUCTS LIABILITY LAW (3d ed. 2015).

189. *See* RESTATEMENT (SECOND) OF TORTS § 402A (Am. Law Inst. 1965); RESTATEMENT (THIRD) TORTS: PRODUCTS LIABILITY §§ 1–2 (Am. Law Inst. 1988).

190. For a spirited salvo in this debate, which also describes the issues, see generally George W. Conk, *Is There a Design Defect in the* Restatement (Third) of Torts: Product Liability*?*, 109 YALE L.J. 1087 (2000).

191. *See, e.g.*, Nicolas P. Terry, *When the "Machine That Goes 'Ping'" Causes Harm: Default Torts Rules and Technologically-Mediated Health Care Injuries*, 46 ST. LOUIS U. L.J. 37, 53–58 (2002).

192. *See generally* Kerr, Millar & Corriveau, *supra* note 186.

193. *See generally* Millar & Kerr, *supra* note 11.

194. *See id.*

Electronic copy available at: https://ssrn.com/abstract=3114347

expected to collect increasing amounts of data to supply the AI with the information it needs to continue to learn to improve its diagnoses.[195] Thus, hospitals will find themselves paying for more recording equipment, for more nurses and technicians to apply the recording equipment, for the same number of physicians, and for the AI. Again, in the short run, bills go up.

However, once confidence in the AI increases, insurers will inevitably seek cost savings by decreasing the use of physicians to do diagnosis. These savings are likely to be small in comparison to what might be achieved from having machines do treatment as well as diagnosis, but one could see these small savings as the vanguard of a possible future in which the push to replace doctors with machines is more widespread. We suspect that the real action will occur once ML capably encroaches on areas of medical treatment—including not only the development of treatment plans but also their delivery.

Initially, rather than remove humans entirely from the diagnostic loop, hospitals and insurers likely will seek to have a physician review ML diagnoses. Because the cost savings are predicated on reducing the number of physicians, the inevitable result of this "human in the loop" policy is that each remaining physician will be tasked with reviewing a larger number of cases per day than they previously handled. At some point, perhaps quite soon, the load on the physicians will rise to the point where one might question their ability to do more than a basic reality check.[196] Even that check undoubtedly will have some value, because at present MLs can become confused—such as when the Jeopardy-playing Watson suggested Toronto is a U.S. city.[197]

However, we question how often a physician presented with a large volume of cases would be able to detect relatively subtle errors. As the load

---

195. How this plays out will depend on the regulatory and competitive environment. Data collection might be mandated, or it might be the subject of negotiation between at least the AI vendor(s) and the hospitals. Hospitals (or even patients?) might, for example, expect to be paid for their valuable data.

196. *Cf.* Juan Mateos-Garcia, *To Err Is Algorithm: Algorithmic Fallibility and Economic Organisation*, NESTA (May 10, 2017), http://www.nesta.org.uk/blog/err-algorithm-algorithmic-fallibility-and-economic-organisation. Mateos-Garcia argues that

> supervisors need to check each decision individually. This means that as the number of decisions increases, most of the organisation's labour bill will be spent on supervision, with potentially spiralling costs as the supervision process gets bigger and more complicated. . . . . When considered together, the decline in algorithmic accuracy and the increase in labour costs . . . are likely to limit the number of algorithmic decisions an organisation can make economically.

Id.

197. Despite surface appearances, it is not. For an explanation of the error, see Steve Hamm, *Watson on Jeopardy! Day Two: The Confusion over an Airport Clue*, SMARTER PLANET BLOG (Feb. 15, 2011), http://asmarterplanet.com/blog/2011/02/watson-on-jeopardy-day-two-the-confusion-over-an-airport-clue.html. For other entertaining examples of ML errors, see Janelle Shane, *When Algorithms Surprise Us*, AIWEIRDNESS.COM (Apr. 13, 2018), http://aiweirdness.com/post/172894792687/when-algorithms-surprise-us.

Electronic copy available at: https://ssrn.com/abstract=3114347

increases, the carefulness of the review must inevitably decrease; meanwhile, it seems probable that the human's malpractice liability would remain the same, making the physician a moral and possibly financial "crumple zone."[198] Ultimately, either the physicians will rebel, or the cost of their insurance will wipe out at least a chunk of the savings, or MLs will become so reliable that insurance companies and hospitals force physicians out of the loop. In this scenario, bills go down unless ML providers react to the removal of the human doctors by charging even higher monopoly prices—something that presumably would be prohibited by the Sherman Act.[199]

Indeed, the removal of humans from the practice of radiology has already begun. Krista Jones wrote of her son's decision to become a radiology technician:

> After seeing what this radiation treatment was able to do for me, my son applied to a university program in radiology technology to explore a career path in medical radiation. He met countless radiology technicians throughout my years of treatment and was excited to start his training off in a specialized program. However, during his application process, the program was cancelled: He was told it was because there were no longer enough jobs in the radiology industry to warrant the program's continuation.[200]

Whatever the current demand for radiologists, future doctors, and even radiology technicians, they are being exposed to strong signals that radiology is a field with no future: "They should stop training radiologists now," asserts University of Toronto Professor Geoffrey Hinton.[201] Hinton's view is extreme: radiologists do more than view films; for example, interventional radiologists oversee radiation and other treatment for patients.[202] Nevertheless, Hinton's

198. *See* Madeline Elish, *Moral Crumple Zones: Cautionary Tales in Human Robot Interaction* (Columbia Univ. & Data Soc'y, We Robot 2016 Working Paper), http://robots.law.miami.edu/2016/wp-content/uploads/2015/07/ELISH_WEROBOT_cautionary-tales_03212016.pdf. ML can be used to choose which cases are most uncertain and present those only to reduce the volume. But there remains the risk that the ML system gets it wrong, i.e., misses some important cases that need to be reviewed, and we are back to the problem of humans having too many cases to review.

199. *See* Sherman Antitrust Act, 15 U.S.C. §§ 1–7 (making it a felony to "monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among the several States, or with foreign nations").

200. Krista Jones, *I Was Worried About Artificial Intelligence-Until It Saved My Life*, QUARTZ (Aug. 20, 2017), https://qz.com/1056817/i-was-worried-about-artificial-intelligence-until-it-saved-my-life/.

201. Quoted in Mukherjee, *supra* note 26.

202. *See* supra note 27; *see also* David Kessel, *What is Interventional Radiology?*, BRITISH SOCIETY OF INTERVENTIONAL RADIOLOGY, https://www.bsir.org/patients/what-is-interventional-radiology/ (last visited Jan. 6, 2019).

Electronic copy available at: https://ssrn.com/abstract=3114347

overstatement represents the perceptual zeitgeist of many incoming students, who will increasingly turn their focus toward other specialties.[203]

That said, the future in which a patient in the United States consults an AI directly for treatment without seeing even a primary care physician seems highly implausible if not far, far away—not only from a scientific point of view but also from a legal perspective. Direct-to-patient services of this type face legal and regulatory obstacles of their own, not the least of which is unauthorized-practice-of-medicine claims in many states.[204] Doctorless diagnosis, on the other hand, may not be so far away. In 2015, the Federal Trade Commission settled claims against marketers of "MelApp" and "Mole Detective" for "deceptively claiming their mobile apps could detect symptoms of melanoma, even in its early stages."[205] But only three years later, the FDA approved an AI-based program that can detect diabetic retinopathy.[206] Although the system is marketed to health-care professionals, it requires no input from a trained doctor,[207] so it seems fair to speculate that at-home use may not be so far away. Meanwhile, HealthTap's "Dr. AI" offers online services and mobile-phone apps that use an algorithm to respond

---

203. *See* Bo Gong, et al., *Influence of Artificial Intelligence on Canadian Medical Students' Preference for Radiology Specialty: A National Survey Study*, ACAD. RADIOLOGY (forthcoming 2019), https://www.academicradiology.org/article/S1076-6332(18)30471-9/pdf, doi: 10.1016/j.acra.2018.10.007 (reporting that survey of students at 17 Canadian medical schools showed fear of displacement by AI "discouraged many medical students from considering the radiology specialty" and suggesting that "[t]he radiology community should educate medical students about the potential impact of AI, to ensure radiology is perceived as a viable long-term career choice"). *But see* Kush Purohit, *Growing Interest in Radiology Despite AI Fears*, ACAD. RADIOLOGY (forthcoming 2019), https://www.sciencedirect.com/science/article/pii/S1076633219300406, doi: 10.1016/j.acra.2018.11.024 (noting that 2017–2018 U.S. National Residency Match Program saw highest percentage of applicants to diagnostic radiology programs since 2010).

204. However, they might have promise for countries with less-developed economies or large, dispersed, rural populations. *See, e.g.*, *Your Face Could Reveal if You Have a Rare Disease*, WIRED UK, http://www.wired.co.uk/article/fdna-rare-disease-facial-recognition-algorithms (last visited Jun 11, 2017) (describing use of phones to detect rare diseases).

205. FEDERAL TRADE COMMISSION, FTC CRACKS DOWN ON MARKETERS OF "MELANOMA DETECTION" APPS, https://www.ftc.gov/news-events/press-releases/2015/02/ftc-cracks-down-marketers-melanoma-detection-apps (last visited Sep 30, 2017).

206. *See* FDA, *supra* note 43.

207. The IDx-DR analyzes "images of the eye taken with a retinal camera called the Topcon NW400." *Id*. According to its manufacturer, the Topcon NW400 "can be operated by someone who isn't a physician and it only takes nonmedical personnel a few hours to learn to take a picture." *See* Marcia Frellick, *AI Speeds Diabetic Retinopathy Diagnosis Without Specialist*, MEDSCAPE (Aug. 28, 2018), https://www.medscape.com/viewarticle/901297#vp_2. Indeed, the manufacturer of the camera provides online education for technicians wishing to learn to operate it. *See Online Diagnostic Instrument Training*, EYE ON EDUCATION, https://eye.opted.org/1988 (last visited Jan. 6, 2019).

Electronic copy available at: https://ssrn.com/abstract=3114347

to medical queries by steering users to a library of informative articles or to a physician who can answer by text or video chat—or ultimately make a referral to a doctor or to the emergency room.[208]

B. Machine Learning and the Deskilling Debate

Medical observers have repeatedly warned that new technology causes the loss of old skills.[209] It is too early to say whether ML will cause the loss of diagnostic skills[210] and "reduced interest in and decreased ability to perform holistic evaluations of patients, with loss of valuable and irreducible aspects of the human experience such as psychological, relational, social, and organizational issues"[211] or whether we should better "hypothesize that the use of [ML], especially their ability to identify and rank differential diagnoses, might actually improve diagnostic acumen."[212] We may never know; if ML actually eliminates all or most of the demand for the diagnostic services of physicians in a given specialty, inevitably there will be some kind of loss of human know-how, however one characterizes it. The reduction in demand for physicians in a specialty will have knock-on effects in medical schools, as students, and especially interns and residents, steer away from the subject. Soon, hiring committees will decide to use scarce resources elsewhere. The knowledge is not lost—it lives on in the few remaining specialists and researchers and in a database—but it is no longer being added to in the same manner because humans contribute few, if any, new diagnoses paired with outcomes to the ML system's database. Instead, new data about outcomes come primarily from situations where ML itself provided the diagnosis. One can only speculate about the extent to which the future of human medical knowledge will be compromised after a generation or two of diagnostic or treatment decisions generated exclusively by machines.

ML may also have other deskilling effects beyond the elimination of a specialty. We will still need physicians to act upon ML's conclusions and to do the surgery—at least until we have good robot surgeons, which seems to involve a much more complex set of challenges.[213] On the other hand, we may not need physicians to interview the patient. An ML system could do the job, or perhaps—initially—a nurse practitioner (or even a nurse) might do the interview, if guided

---

208. *See What We Make*, HEALTHTAP, https://www.healthtap.com/what_we_make/story (last visited Jan. 6, 2019).

209. *See* Robert Lehr Goodman, *Commentary: Health Care Technology and Medical Education: Putting Physical Diagnosis in Its Proper Place*, 85 ACAD. MEDICINE 945, 946 (2010), doi: 10.1097/ACM.0b013e3181dbb55b (lamenting that "the exam skills of even today's most seasoned examiner pale in comparison with those of earlier eras"); *see also* GOODMAN, *supra* note 88, at 56 (noting that "[e]very generation enjoys the services of at least a few pessimists who despair of the current state of affairs" in medicine).

210. For a warning, see Federico Cabitza et al., *Unintended Consequences of Machine Learning in Medicine*, 318 JAMA 517, 517 (2017), doi: 10.1001/jama.2017.7797 ("A major issue related to incorporation of ML-DSS in medicine could be overreliance on the capabilities of automation.").

211. *Id.*

212. GOODMAN, *supra* note 88, at 58.

213. *See, e.g.*, Kerr, Millar & Corriveau, *supra* note 186, at 257.

Electronic copy available at: https://ssrn.com/abstract=3114347

by questionnaires, updated on the fly, provided by an expert system; tomorrow the questionnaire may be informed by a full AI interacting with information from real-time sensors.[214] The more that AI medicine provides occasions for substituting less expensive personnel for physicians and other highly paid medical service providers,[215] the more we can expect simple economic pressure to push toward the same ends we ascribed to malpractice liability above. A further push likely will come from the need to force the data collected to be as standardized as possible, in order to become quality fodder for future AI training and testing.

Anticipating some version of this future, an opinion column in the *Journal of the American Medical Association* recently suggested that in order to maintain their relevance, perhaps radiologists and pathologists should rebrand themselves as "Information Specialists" "whose responsibility will not be so much to extract information from images and histology but to manage the information extracted by artificial intelligence in the clinical context of the patient."[216] Even so, the article suggested that there would be enormous economies of scale, allowing the specialists to export their work: "A single information specialist, with the help of artificial intelligence, could potentially manage screening for an entire town in Africa."[217] Indeed, this more or less is the business model of the startup Alexapath.[218]

Extrapolating the future of AI-based diagnostic medicine is not easy. Current trials offer hope that ML systems will find cures for new diseases without human help, particularly at the molecular level.[219] In a world of partial successes, we would expect ML to be able to identify which treatments work best.[220]

---

214. *See supra* note 185. For a discussion of the difference between sensor data and electronic health-record data, and the greater utility and ease of analysis of the sensor data, see Iyad Batal, *Temporal Data Mining for Healthcare Data*, in HEALTHCARE DATA ANALYTICS 379, 380 (Chandan K. Reddy & Charu C. Aggarwal, eds. 2015). Many U.S. states have regulated limits on the role of so-called physician extenders which might block this scenario. *See* Amanda Swanson & Fazal Khan, *The Legal Challenge of Incorporating Artificial Intelligence into Medical Practice*, J. HEALTH & LIFE SCI. L. 90, 116 (2012).

215. We can dream about replacing hospital administrators, but they likely will be the last to go.

216. Saurabh Jha & Eric J. Topol, Viewpoint, *Adapting to Artificial Intelligence*, 316 J. AM. MED. ASSOC. 2353, 2354 (2016).

217. *Id.*

218. *See* Jessica Leber, *Kill Time in Traffic by Diagnosing Cancer*, NEO.LIFE (Sept. 28, 2017), https://medium.com/neodotlife/lou-auguste-and-alexapath-46f7b5f724ca.

219. For a suggestive example of AI being used to find a drug to cure a new disease, see Jordana Divon, *Toronto startup has a faster way to discover effective medicines*, GLOBE & MAIL (July 27, 2015), https://www.theglobeandmail.com/report-on-business/small-business/startups/toronto-startup-has-a-faster-way-to-discover-effectivemedicines/article25660419/?arc404=true (describing use of AI to find potential treatment for Ebola).

220. This includes noting correlations that have escaped humans. *See* Andrew H. Beck et al., *Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival*, 3 SCI. TRANSLATIONAL MED. 108 (2011), doi: 10.1126/scitranslmed.3002564 (describing use of "C-Path (Computational

Electronic copy available at: https://ssrn.com/abstract=3114347

Researchers are also working on using ML to customize treatments for patients based on their genetics or on the similarity of their symptoms to earlier success stories.[221]

III. DANGERS OF OVER-RELIANCE ON MACHINE LEARNING IN MEDICINE

Part III is the most speculative, in part because it imagines events farthest in the future. ML works by using as inputs what is, in effect, big data of medicine: symptoms, test results, diagnoses, and outcomes from a substantial number of patients.[222] In the case of ML and radiology, the "outcomes" are the opinions of a panel of physicians who, for example, score images as being of tumors or not tumors.[223] In other cases, and perhaps for future iterations of ML too, the inputs might be based on real-life outcomes.[224] In still other cases, the inputs could be "synthetic" training data created to train the system, if only as a way of initiating the system before graduating to what could be a smaller quantity of genuine patient data.[225] In each of these cases, the training process is path dependent, and the quality of answers depends on how the system is trained.[226] Inevitably, the quality of an AI's outputs is subject to the quality of the data—GIGO (garbage in, garbage out) remains as true as ever.[227] Indeed, there are reasons to worry that "medical datasets currently available for use by AI researchers are notoriously biased" in part because their data is drawn from a population that is "extremely

---

Pathologist)" to identify stromal morphologic structures, a "previously unrecognized prognostic determinant for breast cancer"); *see also* David L. Rimm, *C-Path: A Watson-Like Visit to the Pathology Lab*, 3 SCI. TRANSLATIONAL MED. 1, 2 (2011), doi: 10.1126/scitranslmed.3003252 (noting importance and limits of study).

221. Not that this prospect is itself without unique legal issues. See W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419 (2015), for a survey.

222. *See* Jay, *supra* note 62.

223. *See* Marc Kohli, Luciano M. Prevedello, Ross W. Filice & J. Raymond Geis, *Implementing Machine Learning in Radiology Practice and Research*, 208 AM. J. ROENTGENOLOGY 754, 758 (2017), doi: 10.2214/AJR.16.17224.

224. *See generally* Tom J Pollard & Leo Anthony Celi, *Enabling Machine Learning in Critical Care*, 17 ICU MANAG. PRACT. 198 (2017).

225. Note, however, that choosing to use synthetic data is not without its risks. In particular, the data may have built-in biases or fail to reflect important but perhaps unnoticed aspects of genuine data. *Cf. supra* text accompanying note 35 (noting that poor synthetic data was blamed for failure of Watson's Sloan Kettering experiment).

226. *See* Syed Shariyar Murtaza et al, *How to Effectively Train IBM Watson: Classroom Experience*, 49th HAWAII INT'L CONF. ON SYSTEM SCIS. (2016), https://www.computer.org/csdl/proceedings/hicss/2016/5670/00/5670b663.pdf.

227. *See, e.g.*, Hugh Harvey, *Separating the Art of Medicine from Artificial Intelligence*, TOWARDS DATA SCIENCE (Dec. 21, 2017), https://towardsdatascience.com/separating-the-art-of-medicine-from-artificial-intelligence-6582f86ea244 (summarizing literature showing that "[n]ot only are radiologists really quite bad at writing accurate reports on chest X-rays, they also write entirely different reports to each other given the same chest X-rays"); Vincent, *supra* note 157.

Electronic copy available at: https://ssrn.com/abstract=3114347

male and extremely white,"[228] which may increase the attractiveness of proprietary data sets.[229] Nonetheless, as we have seen in Part II, there may come a point where the reliability of AI is so high that human physicians seem unnecessary or even—to the extent they may overrule valid diagnoses—unhelpful in that their inputs tend to reduce the probability of a successful outcome.

As we have seen, some believe that it is already foreseeable that ML will so displace the diagnostic functions of radiology as to make it a much less attractive specialization in the near future.[230] But what happens once we take the human physicians out of the equation? Now the outcome data being input into the ML system are no longer produced by human decisions or AI-plus-human decisions, but only from outcomes based on ML-generated diagnoses.

This could happen in either of two ways, depending on whether we rely on ML solely for diagnosis or use it also for identifying the course of treatment dictated by the diagnosis.

A. Scenario One: Machine Learning Takes Over Diagnosis Only

First, and earlier in time, assume the ML takes over the diagnostic function from people but human doctors continue to choose the appropriate treatment. We expect the ML system will be trained from an initial batch of data. However, when the ML needs new training data—for example as new and improved sensors or imaging equipment come online—if humans with the necessary diagnostic training are no longer available because they have been displaced by machines for too long, we face a problem. Where previously we could create new training data by consulting expert physicians, now we face the problem that those expert physicians no longer exist, or perhaps are in very short supply, since the clinical demand for their services has evaporated.

Relying on ML trained on old training data has problems.[231] In this scenario, there is a danger that the diagnostic decisions in a closed universe of ML systems might take a wrong path: one not as good as the one that would have been taken if human physicians continued to provide training data. On the other hand, trying for an evidence-based approach in which we examine treatment outcomes based on human treatment decisions and then associate those outcomes with the diagnostic materials introduces substantial problems of its own. One is that it is a lot of work. Another is that it can take a long time, because all of the outcomes we are interested in may take years to manifest.

The training-data problem is potentially very serious, but it is also complex and subtle, and its severity will be dependent on variables, some of which are difficult to predict. The nature of the problem likely will depend on both the future course of ML development and whether the sensor technology producing

228. Dav Gershgorn, *If AI is Going to be the World's Doctor, It Needs Better Textbooks*, QUARTZ (Sept. 6, 2018), https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/.

229. *See infra* text accompanying notes 308–19.

230. *See supra* text accompanying note 200.

231. *See supra* text accompanying notes 70–72.

Electronic copy available at: https://ssrn.com/abstract=3114347

the raw data that must be graded to produce training data is collected by means that are invasive or dangerous.

1. Will Machine Learning Continue to Require Huge Data Sets?

At present, most ML systems require substantial quantities of training data.[232] If ML technology were to improve to a point where smaller training sets sufficed, then it would take fewer doctor-hours to produce training data.[233] Unless and until we achieve that innovation, very large data sets will remain the prerequisite (and also barrier to entry) to fielding an ML system. Even if the equipment being used on patients does not change, we may need new data if the conditions under which the system is used change. For example, if a tumor-detection system is called on to diagnose smaller tumors, earlier in their growth, then we need to have data that reflects this type of tumor. Whenever our understanding of the condition being measured changes, we will need to redo the training data in light of this new knowledge; that may require regrading training data used for the first version of the ML and combining that with new data.

The situation changes further when the sensor technology being used on patients changes. Imagine, for example, that someone invents a higher-resolution scanner that takes sharper images than its predecessor. Human beings who could recognize tumors on the old photos might have little or no difficulty recognizing the same tumors on the new, sharper images; ideally, humans might also be able to see new things they had not been able to discern or become able to better distinguish previously ambiguous results. Unfortunately, ML systems do not work like that.[234] To an ML system, the new, higher-resolution image is a completely new thing, and anyone wanting to field an ML system that can use the new equipment will first need a whole new corpus of higher-resolution training data based solely on the new, higher-resolution images.

2. Can Old ML Train New ML?

At present, there is one shortcut that might make producing new training data easier, but it will be possible only in some cases—and radiology is not one of them. There are methods by which one ML system can train another, but only if there is a way of linking the data the old system used to the data input in the new one.[235] Thus, for example, if the ML system were being trained to identify skin cancers from photographs, it should be possible to take two photos of the suspected area: one with the old, lower-resolution camera, and one with the new,

---

232. *See* Jay, *supra* note 62.

233. For one attempt to achieve this see Zhe Li et al., *Thoracic Disease Identification and Localization with Limited Supervision*, ARXIV:1711.06373 [CS, STAT] (2017), http://arxiv.org/abs/1711.06373.

234. See the discussion on datapoints that are "independent and identically distributed" in BISHOP *supra* note 46, at 26.

235. *See* Xingchao Peng, Judy Hoffman, Stella X. Yu & Kate Saenko, *Fine-to-Coarse Knowledge Transfer for LowRres Image Classification*, IEEE CONFERENCE ON IMAGE PROCESSING, 2016, at 3683, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7533047.

Electronic copy available at: https://ssrn.com/abstract=3114347

higher-resolution camera. The lower-resolution photos could be input to the old ML, and its diagnosis could be used to tag the higher-resolution photo of the same area that then would become part of the training data for the new ML system.

Using one ML to train the other in this manner can be effective.[236] Subfields of AI, such as pretraining and transfer learning, are concerned with this problem.[237] The downside is that because a substantial amount of training data will be required, a large group of patients will have to be subjected to two parallel diagnostic procedures: the old and the new. If the procedure is a photo of a person's skin, that is largely a management problem. If it is an MRI, it is also a substantial, but necessary, expense. But if the procedure is invasive or harmful, like an x-ray[238] or especially a CT-scan,[239] then it requires exposing a very large number of patients to additional risk and also ensuring that the patients who consent to undergo the risk have conditions that are representative of the population as a whole.

While unsupervised learning, that is training on raw data that has not been labeled, classified, or categorized, has often been touted as the future of ML, it is unlikely to be the solution here. Unsupervised learning is most useful in combination with supervised learning; for example, where unsupervised data is used for pretraining, outlier removal, and low-dimensional data projection. But some amount of labeled data is still necessary to express the core concept that the machine should learn. [240]

B. Scenario Two: Machine Learning Takes Over Diagnosis and Treatment

The situation looks even more concerning if ML systems also take on the job of choosing and applying the course of treatment. Now, we face a closed-loop system, one in which the outcomes themselves owe their origins to ML-generated choices. In such a scenario, the very distribution of observed cases and outcomes is

---

236. *See id.*

237. *See* Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436 (2015), doi: 10.1038/nature14539; Sinno Jialin Pan & Qiang Yang, *A Survey on Transfer Learning*, 22 IEEE TRANSACTIONS ON KNOWLEDGE & DATA ENGINEERING 1345, 1345–59 (2010), doi: 10.1109/TKDE.2009.191.

238. Current estimates of the risk of cancer from diagnostic x-rays are small. A 2004 paper estimated that 0.6% of cancers in the UK were due to diagnostic x-rays, although in other countries that used them more frequently the toll might climb as high as 3%. Amy Berrington de González & Sarah Darby, *Risk of Cancer from Diagnostic X-rays: Estimates for the UK and 14 Other Countries*, 363 LANCET 345, 345 (2004).

239. A 2009 study found that "the patients who are most frequently imaged have cumulative risks significantly greater than the typical patient. The top percentile . . . have estimated LARs of cancer incidence in excess of 2.7% (above the baseline 42% cancer rate), equating to 6% or more of their total expected cancer incidence." Aaron Sodickson et al., *Recurrent CT, Cumulative Radiation Exposure, and Associated Radiation-Induced Cancer Risks from CT of Adults*, 251 RADIOLOGY 175, 180 (2009), doi: 10.1148/radiol.2511081296.

240. *See* Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent & Samy Bengio, *Why Does Unsupervised Pre-training Help Deep Learning?* 11 J. MACHINE LEARNING RES. 625 (2010); *see also supra* note 64.

Electronic copy available at: https://ssrn.com/abstract=3114347

a result of the ML system's decision strategy. If the ML system does not consider the right optimization function, things may derail.[241] When clinicians are in the decision loop, they have the ability to adjust the optimization criteria—e.g., balance symptom reduction with side-effects—and incorporate additional variables into that criteria—e.g., multiple types of side-effects—to refine the decision strategy.[242] An ML system optimizes a fixed-performance criteria,[243] but it does not have the same normative ability to self-correct and gradually incorporate new dimensions to its value system.

Before going any further, it may be useful to emphasize the relative modesty of our claim regarding this scenario. We are not claiming that closed-loop retraining *must* result in the degradation of an AI's predictive abilities. And we are certainly not echoing Juan Mateos-Garcia's claim that "'entropic forces' that degrade algorithm accuracy will win out in the end: no matter how much more data you collect, it is just impossible to make perfect predictions about a complex, dynamic reality"[244]—not least because this claim is addressed primarily to systems where humans have an incentive to game against the AI,[245] a condition that we trust does not apply to diagnostic medicine. Rather, our concern is whether in the closed-loop scenario we can be confident that over time the AI's diagnoses will remain of the high quality that originally led the medical and legal systems to prefer the AI to human diagnosticians. And even if we have some confidence that degradation is unlikely, as we explain below, there is the larger risk that improvement will not continue; indeed, especially if we rely on ML to plan and deliver treatments upon diagnosis, there is some real risk of the ML system reinforcing its original decisions when some other path might be better.[246] If, as we believe, both legal and medical ethics should require that we have this confidence before we rely solely on AI diagnosticians, then we may have a problem.

Statistical systems require feedback.[247] "The ideal technique for testing the obtained model is to use an external validation dataset that is collected independently of the training dataset on which the model was built."[248] Indeed, this testing and improvement is a continual process.[249] Ideally, one would check and

241. *See* GOODFELLOW, BENGIO & COURVILLE, *supra* note 56, at 102.

242. *See* Albert Freitas, Altamiro Costa-Pereira, Pavel Brazdil. *Cost-Sensitive Decision Trees Applied to Medical Data*, PROC. 9TH ANN. INT'L CONF. ON DATA WAREHOUSING & KNOWLEDGE DISCOVERY, 2017, at 303.

243. *See* GOODFELLOW, BENGIO & COURVILLE, *supra* note 56, at 102.

244. Juan Mateos-Garcia, *To Err Is Algorithm: Algorithmic Fallibility and Economic Organisation*, NESTA (May 10, 2017), https://www.nesta.org.uk/blog/err-algorithm-algorithmic-fallibility-and-economic-organisation.

245. *See id.*

246. *See infra* text accompanying notes 247–70.

247. O'NEIL, *supra* note 77, at 6.

248. Sanjoy Dey et al., *Predictive Models for Integrating Clinical and Genomic Data*, in Reddy & Aggarwal, *supra* note 214, at 433, 450.

249. *See* Martin Zinkevich, *Rules of Machine Learning: Best Practices for ML Engineering*, MARTIN ZINKEVICH, http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf (last visited Feb. 22, 2019).

Electronic copy available at: https://ssrn.com/abstract=3114347

retrain the AI on new data, making for a workflow of collect data, train a model, get new data, retrain, repeat.[250] Retraining does not necessarily require a human in the loop. But for more complex, real-life problems, retraining may require human input to check data quality and to generate labels for the new data.[251] And here is where the problem lies: if the AI always recommends a particular drug regime for a given type of cancer, we will never get any new data on the efficacy of radiation. As a result, we will never learn whether radiation could end up being better in some circumstances. In essence, the AI's initial diagnosis decisions will decide the training examples available downstream.[252] Of course, similar problems bedevil cancer treatments run by humans: ethics and humanity prevent the use of control groups of patients with deadly diseases.

How much humans need to be involved in ML retraining varies with the type of problem being solved. Physical processes that can be observed and measured objectively, like object grasping or motor learning in robotics, lend themselves to automated retraining,[253] essentially via trial and error using reinforcement learning methods.[254] However, we do not wish to subject patients to random error as an ML system learns by doing. Automated retraining works best for problems where the preferred objective can be described precisely (mathematically), such as winning or losing in the game of Go.[255] Indeed, DeepMind's latest Go-playing AI, AlphaGo Zero, learned using no external training data at all: "With each iteration of self-play, the system learns to become a stronger player."[256] "It can do this efficiently because all the other uncertainties are known. . . . There is complete information. . . . There is a way to measure success. In short, the behavior of the game of Go is predictable, real world systems however are not."[257] In contrast to playing Go, retraining on diagnostic technique will require human input and supervision until such a time as we can sufficiently describe the conditions we are testing for in advance.[258]

---

250. *See id.*

251. *See id.*

252. *See supra* text accompanying notes 241–51.

253. Retraining with no humans in the loop is sometimes called "self-supervised learning." *See, e.g.*, Dave Gershgorn, *Google's Robots Are Learning How to Pick Things Up*, POPULAR SCI. (Mar. 8, 2016), http://www.popsci.com/googles-robots-are-learning-hand-eye-coordination-with-artificial-intelligence.

254. *See generally* RICHARD S. SUTTON AND ANDREW G. BARTO, REINFORCEMENT LEARNING: AN INTRODUCTION (2017).

255. Google's AlphaGo Zero is the perfect example of a system that can train itself in a closed loop. "The network learns by comparing itself not from external training data but from synthetic data that is generated from a previous version of the neural network." Carlos E. Perez, *Why AlphaGo Zero is a Quantum Leap Forward in Deep Learing*, MEDIUM (Oct. 22, 2017), https://medium.com/intuitionmachine/the-strange-loop-in-alphago-zeros-self-play-6e3274fcdd9f. AlphaGo Zero can do this, however, only because the rules of Go can be described mathematically. *Id.*

256. *Id.*

257. *Id.*

258. For a description of the technique of "sparse representations"—in which an AI is trained with general criteria that require less and more general training data, then left

Electronic copy available at: https://ssrn.com/abstract=3114347

One might reasonably ask why, once the AI is up and running and routinely outperforming human doctors, it cannot simply learn from its mistakes. One part of the answer is that the machine has no natural notion of a "mistake," and it must be taught the concept from a human.[259] Another part of the answer is that, at least in the case of tumor detection, we may only learn of the machine's mistakes several years after the fact, if then.[260] Even assuming that medical systems are engineered to gather the feedback years later, that still leaves the possibility of an AI running on the wrong path for some significant period of time. Indeed, AI applications with long delays between prediction and real-world validation are among those at the greatest risk of "concept drift," a known source of error.[261] Another risk is that learning from new training data can overwrite the learning from older data, which may not lead to an improvement in performance,[262] although this danger ought to be able to be mitigated by careful validation against the original training data.

Worse, in some cases, especially if the initial training data has systematic errors, is that automated feedback, and even human-assisted feedback, can amplify the errors rather than correct them.[263] Thus, for example, if a crime database is biased because officers have tended to stop minorities or to patrol disproportionately in minority neighborhoods, a predictive system based on that data will continue to steer police in those directions, and the arrests they make will be seen as confirmation of the initial bias.[264]

---

to train itself, then "fine-tuned" by humans (which includes checking to see if the results make any sense at all)—see Dinggang Shen et al., *Deep Learning in Medical Image Analysis*, 19 ANN. REV. BIOMEDICAL ENGINEERING 221 (2017), doi: 10.1146/annurev-bioeng-071516-044442.

259. See the notion of "Evaluation" in Domingos, *supra* note 47, at 1–2.

260. "[C]ertain tumours … grow slowly and it's even possible that a person can die from a completely different cause with a tumour firmly lodged inside them." Dance Baltina, *How Fast do Tumours Grow?*, NOTES ONCOLOGIST (Feb. 26, 2018), https://notesofoncologist.com/2018/02/26/how-fast-do-tumours-grow/.

261. Adam Gelencser, *Concept Change in Machine Learning*, WARWICK INST. FOR SCI. CITIES, https://www.wisc.warwick.ac.uk/files/6814/7922/2663/AdamG.pdf (citations omitted) (last visited Feb. 24, 2019).

> In the real world, concepts and data distributions are often not stable but change with time. This problem, known as "concept drift," complicates the task of learning a model from data and requires special approaches, different from commonly used techniques, which treat arriving instances as equally important contributors to the target concept. Among the most popular and effective approaches to handle concept drift is ensemble learning, where a set of models built over different time periods is maintained and the best model is selected or the predictions of models are combined. A. Tsymbal, M. Pechenizkiy, P. Cunningham & S. Puuronen, *Handling Local Concept Drift with Dynamic Integration of Classifiers: Domain of Antibiotic Resistance in Nosocomial Infections*, 19TH IEEE SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS, 2006, at 679.

262. *See* CARLOS E. PEREZ, THE DEEP LEARNING AI PLAYBOOK 110 (2017).

263. *See* BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 146–60 (2007) (describing "ratchet effect").

264. *See* O'NEIL, *supra* note 77, at 87; PASQUALE, *supra* note 77, at 40–41.

Electronic copy available at: https://ssrn.com/abstract=3114347

For these and other reasons, some computer experts, such as Cathy O'Neill, have suggested that AI-based predictions should only be relied on if someone is continuously checking predictions against reality.[265] O'Neill thinks AIs are too prone to error for us to rely on them when making important decisions unless a human remains in the loop.[266] Of course, humans are known to suffer from the same problems, which is what causes bias in the data to begin with. Having a human in the loop may help mitigate problems of bias, but it is not in itself any guarantee.

Some types of updating cause new difficulties. Typically, including new sensor data in a training set means we can no longer use the old data. And of course, that new sensor data needs to be associated with "correct" diagnoses for which, at present, we rely on human experts. Plus, a diagnostic ML with revised training data based on data derived from improved technology will need to demonstrate anew that it is at least as good as its predecessor. That requires validation data, also at present created by humans. However, as noted above, producing that new data becomes even more difficult if treatment decisions, as well as diagnoses, have become the province of machines.

Conversely, imagine a period in which new types of data are not coming on stream, but the ML system is making poor diagnoses. What does it do then? If the same set of symptoms is producing the same diagnosis in all cases, where will the ML get the data to suggest which different diagnosis would be better? If the answer is "nowhere" then we have a problem.[267] Again, the problem is likely even more serious if ML takes over treatment as well as diagnosis.

Or, even worse, imagine that the data on which the AI relies has been modified in some way, turning it into a "BadNet."[268] How long would it take before doctors first suspected, then were able to confirm, the existence of a problem? As a leading report on robotics and AI recently warned,

> The whole field of formal modelling, verification measurement and performance evaluation of [Robotics and AI (RAI)] systems is still very much in its infancy: it is critical that one should be able to prove, test, measure and validate the reliability, performance, safety and ethical compliance–both logically and statistically/

265. *See* O'NEIL, *supra* note 77, at 208–10.

266. *See id.*

267. Admittedly, this problem is not unique to ML: clinicians and practitioners have had to update their priors about symptoms, diagnosis, and treatments as new studies quantify the race, gender, socioeconomic, and other inequities reflected in medical research and treatment. *See* Editorial, *Clinical Trials Have Far Too Little Racial and Ethnic Diversity*, SCI. AM. (Sept. 1, 2018), https://www.scientificamerican.com/article/clinical-trials-have-far-too-little-racial-and-ethnic-diversity/. But at least clinicians and practitioners are able to critique each other. In the case of ML dominance there is a risk that it could take even longer to identify problems unless there is a corps of experts able to monitor and identify poor outcomes.

268. For chilling scenarios, see Tianyu Gu et al., *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain*, ARXIV:1708.06733 [CS] (Aug. 22, 2017), http://arxiv.org/abs/1708.06733.

Electronic copy available at: https://ssrn.com/abstract=3114347

> probabilistically–of such RAI systems before they are deployed. It should be noted that the verification of systems that adapt, plan and learn will involve the development of new modelling and verification approaches; moreover, such modelling and verification is a prerequisite for informed certification and regulation of RAI systems, which in turn is a factor in public acceptance of RAI.[269]

Even with better validation protocols than currently exist, human observers may have real difficulty observing that a problem exists: as systems become more complex, "human operators may have greater uncertainty regarding the conditions under which the system will fail" due to an inability to confidently verify the behavior of the system under all possible operating conditions.[270]

A further complexity arises if tort law were to respond to the removal of humans from the decision loop by shifting the frame from malpractice to product liability. Even if, as discussed above, the unexpected or unforeseen performance of some ML is not easily understood as a product defect in the usual sense, treating the ML system as the product invites potential plaintiffs to investigate if there might have been errors in the design of the ML system, in its "production", or in its use.[271] These concepts emerge from the law's encounter with the assembly line.[272] They map imperfectly at best to a creation process in which the "product" is an algorithm perhaps unknown to its creators, produced by collecting (and perhaps creating) a mass of data that was used to train the system. Tort law commonly imposes strict liability for production errors, but the extent of liability for design errors is controversial, sometimes less, and can be subject to very high hurdles of proof.[273] Thus, whether the data collection and creation process is considered "design" or something else could have very great consequences for the potential liability of the creators (and users) of an ML system that makes a harmful error.[274] Because we see these complex issues as unlikely to arise until the still somewhat-far-off day when ML systems do treatment as well as diagnosis, we leave these interesting and important questions for another day.

Whether the law will treat ML systems as products or services likely applies to deep-learning systems in general, and it might be unfair to expect that future proponents of AI-based healthcare solve it on their own. Either way, there are two extremely important problems that accompany the delegation of medical diagnostics and treatment to ML: the extent to which legal as well as economic

---

269. *Joint written evidence submitted by AAAI and UKCRC (ROB0021)*, DATA.PARLIAMENT.UK, http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/robotics-and-artificial-intelligence/written/32533.html (cited with approval in SCIENCE AND TECHNOLOGY COMMITTEE, ROBOTICS AND ARTIFICIAL INTELLIGENCE, 2016–17, HC 145, at 16 (UK)) (last visited Feb. 22, 2019).

270. Scharre, *supra* note 83, at 17.

271. *See supra* text accompanying note 189.

272. *See generally* LAWRENCE M. FRIEDMAN, A HISTORY OF AMERICAN LAW (1973).

273. *See generally* Conk, *supra* note 190.

274. *See supra* text accompanying notes 188–90.

Electronic copy available at: https://ssrn.com/abstract=3114347

pressure will drive actors to prefer the AI over humans, and the risk to life that might be caused by an over-dependence on AI-produced training data in the future.

In our next Part, we canvass possible solutions to the risk of over-reliance on AI diagnosticians.

IV. SORTING POTENTIAL SOLUTIONS

One of the simplest potential solutions, at least conceptually, is to impose legal rules or other governance mechanisms that ensure we have an adequate cadre of human-physician diagnosticians. Of course, the goal is not merely to impose a quota of warm bodies. It is to retain and retrain scientists and physicians who will continue experimentation with better solutions,[275] and who will maintain a meaningful and complementary role, working with ML to create new training data, adjust the performance criteria, and certify the decisions of the ML system. This aim is clearly in tension with the trends suggested in Parts I and II above and would certainly be costly.[276] Nevertheless, we return to this idea after first canvassing a variety of other potential technical, economic, and legal solutions. However, one change we do not address is switching the United States to a single-payer health system. In a single-payer system, such as the one in Canada, it might be possible to make rules centrally that address the downsides of the success of ML diagnostic medicine.[277] If the United States is to move to single-payer or some other form of nationalized health system, it will be for reasons of social policy larger than the encroachments of ML systems on diagnosis and treatment.

A. Desiderata

The perfect, or at least good, solution to avoiding a scenario in which both legal rules and economic choices result in vastly reduced, if not outright collapse of, human participation in the improvement of various diagnostic and treatment specialties (thus eliminating the expertise needed to monitor the performance of ML systems and to create new training data when needed), would have the following properties:

- It would be consistent with *primum non nocere*, in that it would not involve any rule change with negative side-effects on other areas of law, ethics, or technology.
- It would at best create incentives to give patients the best medical treatment affordable. At the very least it would impose no impediment to an evolving standard of care and would never incentivize the definition of a legal standard

275. We would also need to ensure that there is a mechanism which allows ML systems to respond quickly to scientific and medical findings by overriding whatever the ML systems had previously been doing.

276. *See supra* text accompanying note 174 ("Physicians are expensive to train, and expensive to keep on staff.").

277. In a government-run, or even private, single-payer medical system, an administrative order or a national payment rule would presumably suffice to induce compliance with rules relating to when ML would be allowed to replace doctors, or how ML should be used more generally.

Electronic copy available at: https://ssrn.com/abstract=3114347

of care worse than what could reasonably be provided given the overall state of the art.

- It would not create incentives that would tend to reduce the progress of medical research nor tend to leave us less well-able to react to medical emergencies, such as new diseases and epidemics.
- It would be resistant to, or ideally invulnerable to, the dangers of monoculture and over-reliance on ML as identified above in Part III.
- It would at best allow capture of any cost savings enabled by new technology. At the very least it would incentivize cost savings consistent with the ethical and legal obligations to give patients at least the standard of care, given the overall state of the art.
- It would have a bottom line that is consistent with the "Standard View" of biomedical ethics; namely, "that the practice of medicine and nursing are ineluctably human."[278]

Spoiler alert: we do not have a perfect solution that meets all these criteria. In what follows we discuss various imperfect solutions and warn against particularly bad ones. Even our best solution has negative characteristics.

One challenge that seems to emerge from what follows results from the interaction between economic and legal incentives. A change to legal rules that fails to adequately deal with the effects of the economic incentives likely will not achieve much because economic imperatives could still dominate: even if malpractice law does not *require* reliance on ML, insurers and others may choose to demand it, to the extent that law permits, if ML cuts costs. So, to be viable, it would appear that a solution must overcome both sets of incentives.

In spite of this one-two punch, it is important to state as a framing principle that we should not allow the entanglement of law and economics to become an impermeable barrier. If pressure from law and cost does indeed lead us down a path of over-reliance on mechanized medicine, and this truly does create a risk of either bad outcomes or a reduction in the creation of better outcomes, then in accord with our bottom-line desiderata stated above, we must be sure not to relinquish the human element in medicine. This especially includes access to and human control over the creation of medical knowledge. This point distinguishes our approach to economic considerations regarding ML from how one might approach other crucial diagnostics tools, such as functional magnetic resonance imaging (fMRI). One could decide to bite the bullet on costs with either technology purely on the basis of the medical benefits that they provide, but the potential long-run consequences of ML—especially with regards to our ability to understand, control, and access future medical knowledge—remind us that, in this case, we need to look beyond short-run economic benefits: both Kantian- and

278. *Supra* note 88.

Electronic copy available at: https://ssrn.com/abstract=3114347

utilitarian-based ethics may support the need for a human-centered approach to medicine.[279]

B. Should We Trust the Private Sector to Solve the Problem?

A common answer, at least in the United States, to problems that have both an economic and legal component is that we should let the market decide. Proponents of market solutions tend to argue that the market should be seen as the default and that any claim for government intervention must be justified by the existence of a (significant) market failure.[280] These arguments ring somewhat hollow in the context of the U.S. health-care market, an arena in which the market is notoriously dysfunctional due to issues on both the demand side (patients are not able to shop well due to pricing and quality opacity[281] plus bounded rationality,[282] and even more so when the patient is ill or unconscious) and the supply side (local monopolies,[283] distortions caused by our payment and insurance rules).[284]

ML systems that displace doctors will add an additional market imperfection to the pile. We have suggested that, left to operate in the market such as it currently is, there is a danger that effective ML diagnostic systems will create conditions in which doctors no longer get the training and experience that they need to become expert enough to create high-quality training data.[285] It is as if physicians today, by learning on the job, are creating a positive externality: acquiring the skill and judgment needed to create great training data.[286] The

---

279. *See generally* Jharna Mandal, Dinoop Korol Ponnambath & Subhash Chandra Parija1, *Utilitarian and Deontological Ethics in Medicine*, 6 TROPICAL PARASITOLOGY 5 (2016), doi: 10.4103/2229-5070.175024.

280. *See* GREGORY MANKIW, PRINCIPLES OF ECONOMICS 12 (8th ed. 2016) ("There are two broad rationales for a government to intervene in the economy and change the allocation of resources that people would choose on their own: to promote efficiency or to promote equality. . . . Economists use the term market failure to refer to a situation in which the market on its own fails to produce an efficient allocation of resources.").

281. Stephen R. Latham, *Richard Epstein on Healthcare*, 19 QUINNIPIAC L. REV. 727, 733–34 (2000) ("Healthcare markets suffer from a number of imperfections that virtually assure that a series of completed voluntary transactions will not maximize social utility. For example: demand for health services is irregular and unpredictable; uncertainty as to the quality and efficacy of proposed medical treatments plagues both the demand and the supply side, and is not resolved by post-treatment observation of outcomes; the pervasive use of insurance creates risks of moral hazard; and even without insurance there are ineradicable agency problems in the patient-doctor relationship."); *see also* Abigail R. Moncrieff, *The Individual Mandate as Healthcare Regulation: What the Obama Administration Should Have Said in* NFIB v. Sebelius, 39 AM. J.L. & MED 539, 544–47 (2013).

282. *See, e.g.*, Daniel Young, *Curing What Ails Us: How the Lessons of Behavioral Economics Can Improve Health Care Markets*, 30 YALE L. & POL'Y REV. 461, 468 (2012).

283. *See, e.g.*, Clark C. Havighurst & Barak D. Richman, *The Provider Monopoly Problem in Health Care*, 89 OR. L. REV. 847, 848 (2011).

284. *See, e.g.*, Moncrieff, *supra* note 281, at 562.

285. *See supra* Part II.

286. We are indebted to Andres Sawicki for this analogy.

Electronic copy available at: https://ssrn.com/abstract=3114347

introduction of the ML system removes the opportunity for gaining this experience and in time removes the replacement supply of doctors in the effected specialty, thus removing the occasion for the positive externality's production (or, if you prefer, the ML system is causing a negative externality of its own).[287]

Of course, the classic answer to an externality problem is to internalize it. And it could be argued that in the case of ML systems the internalization comes built-in: the firms that want to market next-generation ML systems will have all the incentive needed to ensure that they have a stable of well-qualified physicians able to create training data whenever it is required.

We are not prepared to say this could never happen; it is theoretically possible. However, we are quite skeptical that it would actually happen for a number of reasons. In order for the market to overcome the effects we have described one must believe all of the following things strongly enough to base public-health policy on them (in order of decreasing plausibility):

- ML-system providers will have large enough income streams to keep a significant number of doctors on staff full- or part-time. Firms will do so despite the fact that the pace of technical change is notoriously unpredictable and it might be years between generations of sensors that would necessitate a new set of training data.[288]
- ML-system providers will find a way to train their doctors other than having them diagnose patients in a world where both patients and healthcare providers prefer the machine. Firms could, for example, ask their staff doctors to shadow machines and compare their diagnoses to the ML systems' diagnoses.
- Persons attracted to the practice of medicine will find this work, which does have long-run benefits to society, sufficiently interesting and fulfilling to choose it over medicine with more immediate and tangible benefits to patients.[289]
- What is more, those persons will be doctors of comparable quality and, in time, experience to the doctors currently relied on for training data.[290] (Recall

---

287. Medicine is not known for dealing well with externalities. Consider, for example, how the over-use of antibiotics has contributed to the evolution of antibiotic-resistant bacteria.

288. For example, the average life-span of an MRI scanner exceeds 11 years. *See Average MRI Scanner Nearing Adolescence*, DIAGNOSTIC IMAGING (Feb. 5, 2014), http://www.diagnosticimaging.com/mri/average-mri-scanner-nearing-adolescence (stating average age in 2013 was 11.4 years).

289. For evidence that medical students are already avoiding radiology due to the fear of displacement by AI, see Bo Gong et al., *supra* note 203.

290. For one suggestive account of how this goes wrong, see Beane, *supra* note 9 at 1. Beane's ethnographic study found that the introduction of robotic surgery gravely harmed the training of new surgeons:

> [R]obotic surgery greatly limited trainees' role in the work, making approved methods ineffective. Learning surgery in this context required

Electronic copy available at: https://ssrn.com/abstract=3114347

the GIGO principle—unless the training data are of high quality, the ML system's diagnoses cannot be.)

To trust in the market solution, one needs to believe all these things and to believe them strongly enough to gamble public health on them.

C. Possible Technical and Economic Changes

We could attempt to engineer the national health system to enjoy as much of the benefit of ML's enhanced diagnostic abilities as possible without falling into the trap of monoculture or an over-reliance on ML. Depending on their nature, technical changes can be required by law, by the imposition of agreed standards, or self-imposed in response to ethical or market concerns.

1. Create a Control Group?

A potential technical solution would be to divide the population into two groups. One group would receive ML-informed care, while the other group, the control, would not. This is likely a non-starter if one is convinced that ML is better than physicians, because the control group would then be getting substandard care. The ethical and legal difficulties are complex.[291]

Beyond ethical questions are the practical concerns: running a very large control group would be highly impractical. Not only would it be difficult to decide how big the control group needed to be, but it would be equally challenging to decide how long the experiment needed to run before we reach conclusive results.[292] For most ML systems, there is at present no obvious point beyond which we can safely say that if the problems we have identified have yet to manifest we are likely in the clear forever.[293] Conversely, there is no extant standard by which we can decide the ML is so good that the problems we highlighted above are no longer a concern.[294]

Yet, without a control group, relying on human physicians to spot and correct an ML system's errors or especially failures to improve is perilous because the human doctors may not have anything to compare to in order to help them notice. If competing firms have equal access to the entire database, or have access to separate databases that are roughly equal in size and quality, competition might supply the needed monitoring. Unfortunately, for reasons discussed below, access

---

> what I call ''shadow learning'': an interconnected set of norm- and policy-challenging practices enacted extensively, opportunistically, and in relative isolation that allowed only a minority of robotic surgical trainees to come to competence.

Id.

291. *See generally* CHARLOTTE LEVY, THE HUMAN BODY AND THE LAW: LEGAL & ETHICAL CONSIDERATIONS IN HUMAN EXPERIMENTATION (1975).

292. *See* Kenneth Jung, Nigam H. Shah, *Implications of Non-Stationarity on Predictive Modeling Using EHRs*, 58 J. BIOMED. INFORMATICS 168, 174 (2015), doi: 10.1016/j.jbi.2015.10.006.

293. *See id.*

294. *See id.*

Electronic copy available at: https://ssrn.com/abstract=3114347

to data may prove to be a substantial barrier to entry unless the law changes in some way.[295]

2. Require a "Red Team" and a "Blue Team"?

A slightly less bad variant on the control-group solution might be to divide the population into two or more groups, each of which would be separate for database purposes, and have the different groups' data be used by different ML systems. Thus, in effect, we have Dr. Abdul Watson, Dr. Betty Watson, and Dr. Chia Watson and so on, each using a different population's data to shape their advice. Every so often—how often? and how?—they would have a virtual medical conference in which they exchange their "best ideas" (or would that be their most telling data?) and in effect upgrade each other's diagnostic suggestions. This seems a poor solution because in the usual case an ML system's accuracy is positively correlated with the size of the database.[296] It follows that splitting the database into shards creates a risk of sub-optimal care for everyone. Furthermore, different systems may offer different trade-offs—e.g., more/less Type I vs Type II error; more explainability vs more accuracy—so cannot be compared directly.

3. Alternate AIs?

A third, and perhaps better although somewhat unlikely, technical solution might be to allow each ML to have the same full database[297] but require that their programming or training differ in some meaningful way—if this difference can be defined, measured, and (most importantly) maintained, all without subjecting one group to inferior treatment. Using multiple models can add accuracy; were one model best, ethics and law might force us to use it uniquely.[298]

If this condition holds over time, the diagnostic problem becomes akin to the hurricane-forecasting problem currently faced by meteorologists. There are several competing models, some with different algorithms, others with different coverage, and "[t]he best forecasts are made by combining the forecasts from three

---

295. *See infra* text accompanying notes 308–19.

296. While this is generally true it also depends on factors such as data quality and sometimes also data composition, such as the ratio between negatives to positives in the data set. *See, e.g.*, Rafal Kurczab & Andrej J. Bojarski, *The Influence of the Negative-Positive Ratio and Screening Database Size on the Performance of Machine Learning-Based Virtual Screening*, 12 PLOS ONE, Apr. 6, 2017, https://doi.org/10.1371/journal.pone.0175410.

297. A valuable byproduct of a national ML system is that we would not only have more and thus better data for ML systems to chew on, but we would also have valuable public-health data. Identifying environmental issues, e.g., cancer clusters, will be much easier if all patients' diagnostic info is going into a national database in a standard format.

298. This follows from the argument in Part I, that if an ML system is better than humans, it will become the required standard of care. Logically, the same should apply if one is choosing between competing ML systems: if there are consistent differences between the different ML systems, then unless there are great cost differences, we would expect the best one to become the standard of care.

Electronic copy available at: https://ssrn.com/abstract=3114347

or more models into a 'consensus' forecast."[299] One group of researchers recently demonstrated that a consensus of multiple models plays Atari video games better than any of the models alone.[300] Because Atari video games are like Go in that identifying the "success" criteria is automatic and requires no human input,[301] the applications to medical diagnostics remain, at best, for the future. Nonetheless, the use of ensemble learning has often been shown to surpass a single learner.[302]

Achieving this scenario would require us to overcome a number of legal and economic complexities. First, we would probably need to have multiple competing providers of AI diagnostic services, for it is hard to see what would incentivize a single firm to provide multiple possibly conflicting diagnostic suggestions. Second, we would need to evolve a standard of care that addressed whether it would suffice to consult (purchase) just one AI model or whether multiple AI opinions would be required. Third, we would need to evolve a method of combining, or sorting among, the competing diagnoses if AIs disagreed that would not expose the person making the decision to unreasonable liability.

Having multiple competing providers of AI diagnostic services that each use a different algorithm should prevent diagnostic monoculture. But any plan that intends to rely on multiple providers must address economic and legal obstacles to creating and sustaining multiple providers.

The economic obstacle arises from the nature of the industry, a special case of the winner-take-all phenomenon often observed in markets relying on new technology.[303] We noted above that the economics of deep-learning neural networks involved high fixed costs, including the cost of gathering and formatting the training data, the cost of designing and tuning the relevant algorithms, and perhaps (although here predictions vary) the cost of the equipment hosting the

---

299. Jeff Maters, *Hurricane and Tropical Cyclones*, WEATHER UNDERGROUND, https://www.wunderground.com/hurricane/models.asp (last visited Feb. 20, 2017). We are indebted to Jonathan Frankle for pointing us to weather models as an analogy.

300. Matteo Hessel et al., *Rainbow: Combining Improvements in Deep Reinforcement Learning*, ARXIV:1710.02298 [CS] (Oct. 6, 2017), http://arxiv.org/abs/1710.02298.

301. *See supra* text accompanying notes 255–56.

302. BISHOP, *supra* note 46, at 653; Saso Džeroski & Bernard Ženko, *Is Combining Classifiers Better than Selecting the Best One?*, 54 MACHINE LEARNING 255, 267 (2004), doi: 10.1023/B:MACH.0000015881.36452.6e.

303. For discussions of the general phenomenon of winner-take-all in high-technology industries see, for example, Mark A. Lemley & David McGowan, *Legal Implications of Network Economic Effects*, 86 CAL. L. REV. 479 (1998); Ronald Cass, *Antitrust And High-Tech: Regulatory Risks for Innovation And Competition*, FEDERALIST SOCIETY (June 28, 2013), https://fedsoc.org/commentary/publications/antitrust-and-high-tech-regulatory-risks-for-innovation-and-competition; Thomas A. Piraino, Jr., *A Proposed Antitrust Approach To High Technology Competition*, 44 WM. & MARY L. REV. 65, 87 (2002); Cass R. Sunstein, Robert H. Frank, Sherwin Rosen & Kevin M. Murphy, *The Wages Of Stardom: Law And The Winner-Take-All Society: A Debate*, 6 U. CHI. L. SCH. ROUNDTABLE 1 (1999).

Electronic copy available at: https://ssrn.com/abstract=3114347

AI.[304] Indeed, a widely quoted analysts' report recently cast doubt on the profit potential of IBM's Watson despite its being "one of the more mature and broad cognitive computing platforms today" precisely because users face a high cost of data gathering and curation.[305] However, in contrast, the marginal cost of diagnosing a patient is comparatively small.[306] This account of high fixed costs and low marginal costs resembles the economic profile of a so-called natural monopoly in most respects,[307] save one: other than the contingent question of whether there is sufficient demand to support the capital costs of running multiple competing AIs, there is nothing that is an absolute barrier to entry.

For the multiple-competing-provider scheme to work, all providers need access to sufficient training data,[308] and ideally, they all would have access to all of it because large data sets tend to increase accuracy.[309] Some firms may, however, be able to interpose a legal obstacle to their rivals' access to training data. Training data is not inherently rivalrous. Training an AI is not like siting a water turbine on a river, where there can be only one at any point.[310] But early indications are that would-be providers of AI health-related services see their access to data as a

304. *See supra* text accompanying notes 176–80.

305. James Kisner et al., *Creating Shareholder Value with AI? Not so Elementary, My Dear Watson*, EQUITY RESEARCH AMERICAS: JEFFERIES FRANCHISE NOTE (July 12, 2017), https://javatar.bluematrix.com/pdf/fO5xWjc (rating IBM "underperform" due to doubts about Watson).

306. *See* AJAY AGRAWAL, JOSHUA GANS & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE 7–20 (2018).

307. OECD GLOSSARY OF INDUSTRIAL ORGANIZATION ECONOMICS AND COMPETITION LAW 62 (R. S. Khemani & D. M. Shapiro eds., 1993) ("Generally speaking, natural monopolies are characterized by steeply declining long-run average and marginal-cost curves such that there is room for only one firm to fully exploit available economies of scale and supply the market.").

308. "Deep learning requires very large quantities of data in order to build up a statistical picture." Alex Hern, *Why Data is the New Coal*, GUARDIAN (Sept. 27, 2016) (quoting Imperial College Professor Murray Shanahan), https://www.theguardian.com/technology/2016/sep/27/data-efficiency-deep-learning.

309. To this end, the U.S. Department of Energy and the National Cancer Institute are partnering in a "three-year pilot project called the Joint Design of Advanced Computing Solutions for Cancer," designed to assemble and integrate large amounts of data about how tumors respond to treatment. Argonne National Laboratory, *Cancer's Big Data Problem*, COMMS. ACM (Oct. 21, 2016), http://cacm.acm.org/careers/208869-cancers-big-data-problem/fulltext.

310. *See* James Bradford Delong & A. Michael Froomkin, *Speculative Microeconomics for Tomorrow's Economy*, *in* INTERNET PUBLISHING AND BEYOND: THE ECONOMICS OF DIGITAL INFORMATION AND INTELLECTUAL PROPERTY 6 (Brian Kahin & Hal Varian Eds., 2000) (discussing economic consequences of non-rivalrous nature of data). However, patenting an ML system would, create at least a temporary monopoly. For a discussion of how to draft patent specifications for an ML system see Vincent Spinella-Mamo, *Patenting Algorithms: IP Case Law and Claiming Strategies*, IPFOLIO BLOG, http://blog.ipfolio.com/patenting-algorithms-ip-case-law-and-claiming-strategies (last visited Feb. 24, 2019).

Electronic copy available at: https://ssrn.com/abstract=3114347

strategic asset to which they wish to have exclusive access.[311] If our strategy for avoiding monoculture relies on having multiple equally competent providers, then as Amanda Levendowski has argued in the context of avoiding training bias, the legal system may need to remove existing regulatory obstacles to data sharing. Levendowski suggests that using training data be will often be a fair use.[312] But if trade secret and proprietary first-mover advantages are among the main obstacles to access,[313] then even a copyright workaround may not be enough; in time we may need to impose some sort of compulsory-licensing scheme on holders of the data. Compulsory-license schemes require the owner of an intellectual-property right to share it on reasonable terms.[314] U.S. law does not tend to give compulsory licenses, but they do exist as antitrust remedies[315] and in relatively unusual provisions of existing law relating to patents in essential foods[316] and atomic energy,[317] and for copyrights in certain music.[318] Then again, foreign companies

311. An example is Google's DeepMind's deal to get access to data provided by the UK's National Health Service. The terms of the deal caused a panel of external reviewers to warn that DeepMind could "exert excessive monopoly power" by using technological means to deny competitors effective access to the data. *See* Natasha Lomas, *UK Report Warns DeepMind Health Could Gain 'Excessive Monopoly Power'*, TECHCRUNCH (Jun 15, 2018), https://techcrunch.com/2018/06/15/uk-report-warns-deepmind-health-could-gain-excessive-monopoly-power/. DeepMind later handed the patient data to Google despite "explicit reassurances made by DeepMind's founders that there was a firewall sitting between its health experiments and its ad tech parent, Google." Natasha Lomas, *Google Gobbling DeepMind's Health App Might Be the Trust Shock We Need*, TECHCRUNCH (Nov. 14, 2018), https://techcrunch.com/2018/11/14/google-gobbling-deepminds-health-app-might-be-the-trust-shock-we-need/.

312. Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 619–30 (2018).

313. For a daunting list of obstacles, see Richard Blunk & Eric Armstrong, *Technology Legal Interoperability: Initial Steps Towards an Analytical Framework*, PRIVACY & DATA SEC. LAW RES. CTR. (BLOOMBERG BNA), http://privacylaw.bna.com/pvrc/7057/split_display.adp?fedfid=121122251&vname=pvlrnotallissues&jd=0000015da36bd172abdfeb7fbdf90002&split=0 (last visited Sept. 22, 2017).

314. Strictly speaking, in the United States the government sets the price of the license, so while the price will be lower than what the holder of the IP would have charged, it will not inevitably be reasonable; operationally compulsory licensing is much more efficient once the government determines the need for the license because the price negotiations cannot be contentious beyond a point. *See* Srividhya Ragavan, Brendan Murphy & Raj Davé, *Frand v. Compulsory Licensing: The Lesser of the Two Evils*, 14 DUKE L. & TECH. REV. 83, 116 (2015).

315. *See* United States v. Besser Mfg. Co., 343 U.S. 444, 447 (1952) (imposing compulsory licensing on a "fair" basis).

316. 7 U.S.C. § 2404 (2012) (empowering Secretary of Agriculture to "declare a protected variety open to use on a basis of equitable remuneration to the owner, not less than a reasonable royalty, when the Secretary determines that such declaration is necessary in order to insure an adequate supply of fiber, food, or feed in this country and that the owner is unwilling or unable to supply the public needs for the variety at a price which may reasonably be deemed fair").

317. 42 U.S.C. § 2183 (2012).

318. 17 U.S.C. § 115 (2012).

Electronic copy available at: https://ssrn.com/abstract=3114347

based in countries that have national policies designed to encourage access to training data as part of a pro-AI industrial policy may fill the gap without the need for radical changes in U.S. law.[319]

4. Encourage Transparency?

A big part of what makes the monoculture story troubling is how difficult it could be to detect a problem if it occurred. As we noted above, decision-making by deep-learning-based AI is notoriously opaque.[320] For example, IBM Watson, as currently engineered, does not clearly explain its decision-making processes in terms that are understandable to most humans.[321] It is possible to formally trace (in the computer's memory) how Watson made its decisions, but it takes time and effort to understand the result of that trace.[322] The same problem is present in other ML systems.[323]

Although researchers are increasingly aware of the need for "explainable AI," we are still far from something the average doctor could use in real time to help decide what weight to put on a diagnosis.[324] To the extent, for example, that the explanation consists of a set of weights of various bits of evidence without much in the way of context as to how the neural network chose those weights,[325]

---

319. As Chinese AI expert and investor Kai-Fu Lee says, "[t]he U.S. and Canada have the best AI researchers in the world, but China has hundreds of people who are good, and way more data." Will Knight, *China's AI Awakening*, MIT TECH. REV. (Oct. 10, 2017) (quoting Mr. Lee), https://www.technologyreview.com/s/609038/chinas-ai-awakening/; *see also* Dame Wendy Hall & Jérôme Pesenti, *Growing the Artificial Intelligence Industry in the UK*, UK DEP'T FOR DIGITAL, CULTURE, MEDIA & SPORT AND UK DEP'T FOR BUS., ENERGY & INDUS. STRATEGY (Oct. 15, 2017), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf (making multiple recommendations to facilitate UK-based AI access to training data).

320. *See supra* note 74–76; *see also* Finale Doshi-Velez et al., *Accountability of AI Under the Law: The Role of Explanation*, ARXIV:1711.01134 [CS, STAT] (Nov. 3, 2017), http://arxiv.org/abs/1711.01134 (discussing technical requirements for AI systems that could provide kinds of explanations that are currently required of humans in light of EU GDPR); Aaron M. Bornstein, *Is Artificial Intelligence Permanently Inscrutable?*, LEARNING NAUTILUS (Sept. 1, 2016), http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable (last visited Sep 7, 2016); Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (describing "Deep Patient" and AI that can "anticipate the onset of psychiatric disorders like schizophrenia surprisingly well" using methods opaque to its designers).

321. *See* Hamm, *supra* note 197.

322. *See id.* (describing how Watson erroneously concluded Toronto was in the United States). Similar attempts have been made to reconstruct AlphaGo's move #37 in game #2 of the first match against Lee Sedol. Cade Metz, *In Two Moves, AlphaGo and Lee Sedol Redefined the Future*, WIRED (Mar. 16, 2016), https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.

323. *See supra* text accompanying notes 73–75.

324. *See supra* text accompanying notes 41 and 75.

325. *See supra* text accompanying note 74.

Electronic copy available at: https://ssrn.com/abstract=3114347

we are a long way from the user-friendly, easy-to-use summary a doctor would need. Moving in that direction, we now have neural networks that can provide a confidence number with the decision.[326] Humans can then use that information to prioritize checking the results with lower confidence.[327] However, this presumes that the confidence estimate is sufficiently well informed, i.e., that the machine "knows what it knows." So far ML can only guarantee this in some limited settings.[328]

Researchers today are actively working on the explainability problem,[329] and thus there is reason to hope that it will get better. The more that an ML system can provide an explanation for its diagnoses, the more scope there will be for people to evaluate it meaningfully and, one presumes, spot mistakes or add value.[330] It follows that the "centaur" model[331] is most likely to endure if AI becomes less opaque, because there will still be something meaningful for people to do. However, as noted above, should there come a point where the AI is so good that humans are not adding value, all the arguments we make here come rushing back into play.

---

326. *See generally* Robert Tibshirani, *A Comparison of Some Error Estimates for Neural Network Models*, 8 NEURAL COMPUTATION 152 (1996), doi: 10.1162/neco.1996.8.1.152.

327. *See* Richard Dybowski & Stephen J. Roberts, *Confidence Intervals and Prediction Intervals for Feedforward Neural Networks*, *in* CLINICAL APPLICATIONS OF ARTIFICIAL NEURAL NETWORKS 298, 298–326 (Richard Dybowski & Vanya Grant eds., 2001).

328. *See* Zachary C. Lipton, *The Mythos of Model Interpretability*, ARXIV:1606.03490 [CS, STAT] (June 10, 2016), http://arxiv.org/abs/1606.03490,

329. Examples include Dong Huk Park et al., *Attentive Explanations: Justifying Decisions and Pointing to the Evidence*, arXiv:1612.04757v2 (July 25, 2017), https://arxiv.org/abs/1612.04757 (using neural-network-based, natural-language processing, and generation techniques to cooperatively explain the behavior of other neural networks); Leilani Gilpin, *Reasonableness Monitors*, TWENTY-THIRD AAAI/SIGAI DOCTORAL CONSORTIUM, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17361/16430 (last visited Jan. 19, 2019) (using deductive reasoning to create a "reasonableness monitor" that detects when cyberphysical systems violate rules encoded in formal logic); Tao Lei, Regina Barzilay & Tommi Jaakkola, *Rationalizing Neural Predictions*, arXiv:1606.04155v2 (Nov. 2, 2016), https://arxiv.org/abs/1606.04155 (exploring how to determine the minimum fragment of the input to a neural network necessary for the decision it reached, thus offering some clarity about the network's rationale). We are grateful to Jonathan Frankle for pointing us to these examples. *See also* Rudin & Ustun, *supra* note 41, at 1 (arguing that "[t]here is new technology to build transparent machine learning models that are often as accurate as black box machine learning model"). For a cautionary view, however, see Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018) (warning that "explanation" may make ML systems less inscrutable but will not necessarily make it easier to understand whether their conclusions are justified).

330. *See* Editorial, *Towards Trustable Machine Learning*, NATURE BIOMEDICAL ENGINEERING 2, 709–10 (Oct. 10, 2018), doi: 10.1038/s41551-018-0315-x.

331. *See supra* note 83 and accompanying text (discussing centaur chess).

Electronic copy available at: https://ssrn.com/abstract=3114347

5. Tax ML to Change Incentives?

If the medical industry seeks to substitute ML for the work of a medical specialty, such as radiology, we would expect that in the short term radiologists' salaries might drop, blunting the economic pressure to eliminate them. But, as we have argued above, in the longer run, demand could shrink to near zero; meanwhile, those medical students whose choice of specialty is influenced by salary will avoid that specialty.

One way to discourage over-reliance on ML, therefore, is to change the economic calculus using tax law. If we can maintain a role for doctors in a manner that is more attractive financially, that will remove the economic incentive to undermine human participation in diagnostic decisions and the planning and delivery of treatment. The malpractice-law incentive to choose ML would remain, but as we discuss below, there are some possible legal solutions that do not address the economics, and thus a tax solution might be combined with a legal solution.

To the extent that we see the growth of ML as imposing a negative externality on the medical system as a whole (or undermining an existing positive externality), a classic remedy would be a Pigouvian tax (or subsidy).[332] A Pigouvian tax on a negative externality (or subsidy on a positive one) is designed to reflect the true social cost (or value) of the activity.[333] Thus, in theory, one could either tax the use of ML, subsidize the employment of human physicians, or both—perhaps even having the ML tax provide the funds for the subsidies. The idea of a robot tax is a popular one, having been endorsed by none less than science and tech celebrities such as Bill Gates,[334] Elon Musk,[335] and Stephen Hawking.[336] The idea of a tax has also been criticized as impractical, given we do not have agreed definitions of what constitutes a robot,[337] a critique that applies with nearly equal force to AI and ML. The EU Parliament flirted with the idea of a

---

332. "Most economists believe that the government should impose Pigouvian taxes on firms that produce negative externalities like pollution." Jonathan S. Masur & Eric A. Posner, *Toward a Pigouvian State*, 164 U. PA. L. REV. 93, 138 (2015).

333. *Id.*

334. *See* Kevin J. Delaney, *The Robot That Takes Your Job Should Pay Taxes, Says Bill Gates*, QUARTZ (Feb. 17, 2017), https://qz.com/911968/bill-gates-the-robot-that-takes-your-job-should-pay-taxes/.

335. *See* Catherine Clifford, *Elon Musk: Robots Will Take Your Jobs, Government Will Have to Pay Your Wage*, CNBC (Nov. 4, 2016), https://www.cnbc.com/2016/11/04/elon-musk-robots-will-take-your-jobs-government-will-have-to-pay-your-wage.html.

336. *See* Doug Bolton, *Stephen Hawking Says Robots Could Make Us All Rich and Free – But We're More Likely to End up Poor and Unemployed*, INDEPENDENT (Oct. 9, 2015), https://www.independent.co.uk/life-style/gadgets-and-tech/stephen-hawking-says-robots-could-make-us-all-rich-and-free-but-were-more-likely-to-end-up-poor-and-a6688431.html.

337. *See, e.g.*, Robert J. Kovacev, *The Challenges of Administering a Robot Tax*, LAW360 (Sept. 25, 2017), https://www.law360.com/articles/967115/the-challenges-of-administering-a-robot-tax.

Electronic copy available at: https://ssrn.com/abstract=3114347

robot tax but ultimately rejected it.[338] The biggest problem, not considered by any of the proposals mentioned here is that, in our view, the ultimate aim of the tax is not to create *en masse* disincentives for the development of effective medical ML but, rather, to incentivize the successful development of (centaur-type) ML that leaves a meaningful role for human doctors and, most importantly, avoids monoculture by ensuring human access to future medical knowledge and know-how.

How to devise a tax strategy that achieves these ends might prove an insurmountable challenge. In any event, a tax on ML would ultimately be a loss for patients, who would see costs rise; a subsidy from general revenues would not hurt patients as directly.[339] But to the extent that the tax discouraged medical service providers from using ML, patients would suffer from being deprived of a diagnosis that (ex-ante) has a higher probability of being correct.

6. Tax ML to Support an Expert Corps of Radiologists?

Rather than trying to change incentives, which involves nearly impossible measurement issues, a more interesting scenario would be to set the ML tax at a level sufficient to support a corps of expert radiologists who would be charged with keeping tabs on the ML systems' accuracy, creating new training data as needed, conducting research to improve detection and analysis of scan data, and responding to medical emergencies.

Because there will be few if any relevant market signals, one should not underestimate the difficulty of fixing the right size of such a corps, determining its budget, and recruiting and training highly competent persons to join it. Nevertheless, the idea of a reserve corps of specialists at the National Institute of Health, or perhaps spread out among teaching hospitals, does have some allure. Because it would be much smaller than the current number of radiologists, supporting a group of experts would presumably be less expensive than attempting to preserve the entire profession, even at reduced salaries.

An important challenge in setting up such a corps is in designing the appropriate training curriculum for these experts. The ideal profile would be people with both medical training and advanced ML training.[340] This is a challenging program of study.[341] The shift in curriculum, requiring medical students to incorporate training in probability, statistics, and algorithms, may prove hard to sell for some of the more conservative medical faculties.

---

338. *European Parliament Calls for Robot Law, Rejects Robot Tax*, REUTERS (Feb. 16, 2017), https://www.reuters.com/article/us-europe-robots-lawmaking/european-parliament-calls-for-robot-law-rejects-robot-tax-idUSKBN15V2KM.

339. Patients may suffer indirect harm to the extent that the subsidy from general revenues requires additional taxes that either fall on them or on others who increase prices or reduce wages as a result.

340. *See* Patricia Balthazar, *Training Medical Students and Residents for the AI Future*, DATA SCI. INST. AM. C. OF RADIOLOGY (Nov. 17, 2018), https://www.acrdsi.org/Blog/Medical-schools-must-prepare-trainees.

341. *See id.*

Electronic copy available at: https://ssrn.com/abstract=3114347

D. Possible Changes to Legal Rules

1. Revive the Locality Rule?

In Section I.C we showed how the demise of the locality rule eliminated the ability of physicians to assert a defense of custom, local or otherwise. This, we argued, makes malpractice an engine that will drive the progression toward AI monoculture or at least toward a potentially dangerous over-reliance on ML. Would a return to the locality rule stop this trend and thus prevent malpractice law from creating the incentives that would tend to make ML displace too many doctors?

The answer is that it would not. Even if the revival of the locality rule was able to delay or blunt malpractice law's impetus to switch to ML, it seems unlikely that a (politically improbable) revival of the locality rule would do much to prevent the problems we have identified above: so long as ML seems to offer significant accuracy increases and cost savings, the push to adopt it and in time reduce the use of human doctors will remain strong. As a result, the hospitals, insurers, and private medical practices that choose not to use ML will in time find themselves painted as outliers and laggards even when compared to other hospitals and physicians who are similarly situated geographically or by type of practice.[342]

Furthermore, unless the revival of the locality rule was narrowly cabined to AI-based medical technology, it could have vast and unpredictable side-effects as it infected first malpractice claims generally, and then perhaps other areas of the law of professional negligence. As law and economics scholars have shown, the locality rule imposes substantial costs on society because it disincentivizes innovation, which means that patients will lose the advantages they would have gained from the adoption of new medical technology.[343] Intuitively, the long-term costs in lost advances would seem very likely to exceed the value of any temporary gains.

2. Create a Broad "ML Exception" to Malpractice Law?

Perhaps, therefore, instead of looking for a broad-brush solution, we should just create a judicial or legislative "ML Exception" to malpractice law, by which we would agree that failing to use an ML system in diagnosis is not malpractice.

Unfortunately, this broad ML Exception suffers from most of the same problems as the idea that we might revive the locality rule: it fails to take account of economic incentives to deploy ML, which exist independently from the push provided by malpractice law.[344] Also, like the locality-rule revival, the broad ML Exception also seems likely to impose greater social costs than benefits, for to the extent that it removes an incentive to use ML even carefully, it degrades the quality of patient care.

---

342. *See supra* Subsection I.C.1.

343. *See generally* Parchomovsky & Stein, *supra* note 124.

344. *See supra* Section I.A. The incentives could, however, be overcome by taxes. *See supra* Subsection IV.C.5.

Electronic copy available at: https://ssrn.com/abstract=3114347

3. Create a Narrow "ML Exception" to Malpractice Law?

If a broad ML Exception is too much, how about a more narrowly tailored one, such as a rule that a human doctor's overruling of an ML system is not malpractice unless grossly negligent, but that failing to do so when needed would be actionable error. In other words, the standard of care would still require *consulting* the ML, but it would not be *per se* error to deviate from its diagnostic conclusions. Indeed, we might go further and say the ML's diagnosis was not admissible evidence, although this is probably only a short-term fix at best: over time one would expect that juries would come to understand that ML was the norm and expect to hear about its diagnosis.[345]

This narrower exception would not relieve medical providers from liability for failing to use ML once it became the standard of care but would provide a safe harbor from liability for overruling an ML system unless the human's decision was indefensible. We suggested above that under current liability rules, especially in the increasing number of states that have abandoned the locality rule, even human doctors who believe with some justice that their diagnoses are better than the computer's will face moral risks and obstacles in displacing the AI's suggestion.[346] If nothing else, we suggested, the fact that ML has a better success rate will mean that the physician will run a very great malpractice risk in supplanting its judgment, and that insurers will be loath to permit such decisions as a result. The second form of the ML Exception removes, or at least greatly reduces, this risk. In so doing, it departs from the pattern in other contexts, such as piloting, where we believe machines outpace humans.[347]

The second part of the exception, in which human doctors are liable for failing to overrule an ML system when they should have, is not on its face a change from current law. Under current law, an ML system, being a machine, has no identity nor agency for legal purposes, and hence its decisions will in all cases be ascribed to the human(s) or corporation(s) responsible for acting on its diagnoses.[348] On the other hand, once ML has a better batting average than the average human, it will, as we've said repeatedly, be a courageous human who overrules it in any but the most obvious cases.[349] Under current law, cases in which the computer's decision was arguably plausible but courageously overruled anyway will invite litigation if the outcome goes badly, but cases where the doctor

345. As the use of AI becomes increasingly routine and enters into popular culture, we would expect that jurors will expect to hear about what the system recommended, much like the "CSI effect," see Caroline Kensey, *CSI: From the Television to the Courtroom*, 11 VA. SPORTS & ENT. L.J. 313, 318–31 (2012), is said to shape juror demands for scientific evidence today. *Id.* at 320–21.

346. *Cf.* Millar & Kerr, *supra* note 11.

347. "A court may . . . infer negligence on the part of the pilot from evidence that suggests that the pilot switched from automatic pilot to manual in a crisis situation." James E. Cooling & Paul V. Herbers, *Considerations in Autopilot Litigation*, 48 J. AIR L. & COMM. 693, 710 (1983).

348. *See* Neil M. Richards & William D. Smart, *How Should the Law Think About Robots*, *in* ROBOT LAW 4 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2017).

349. *See supra* Section I.E.

Electronic copy available at: https://ssrn.com/abstract=3114347

should have overridden the computer but did not will be much harder for plaintiffs to prove if and when ML alone becomes the standard of care.[350]

Thus, the second part of the exception can be characterized as no more than a savings clause: a way to emphasize that while liability for overruling ML is changing, liability for not using ML and for not overruling it remains in place. Alternately, one can see the second clause as a means to emphasize the importance of keeping a human in the loop: liability will lie not only for failing to use ML when one should but also for failing to overrule it when one should.

Although undoubtedly preferable to any of the rules canvassed so far, the social-welfare consequences of this narrower ML Exception are hard to predict with any certainty. Even if we assume, somewhat heroically, that on average humans will overrule ML approximately as often as we would want them to, that leaves open the door for errors in both directions, i.e., overruling the ML system when it was right, and failing to overrule the ML system when it was wrong. The patients in the first group, who would have had the benefit of the ML system's correct diagnosis, will be made worse off compared to the treatment they would have received if the narrow ML Exception did not exist. In contrast, the patients in the second group, who would have suffered from the machine's error in any case, are no worse off than they would have been.

How we measure the cost of the errors to the first group is inevitably difficult, but without any defensible idea of how big that group would be—something we could only establish empirically—it is even more impossible to say. Unfortunately, we can say with some confidence that humans will feel freer to overrule ML systems under this rule than under the current default rule because under the current rule an overruling decision would run a greater risk of being found to depart from the (machine) standard of care. Arguably, this means that the number of patients harmed by a doctor's ignoring ML's correct diagnosis ought to grow above the baseline.

Furthermore, if this narrow exception suffices to incentivize medical service providers and malpractice insurers to keep a human doctor fully in the loop, then we also will lose all or part of any cost savings from having ML replace humans, with the size of the loss depending on both the relative costs and the extent to which human doctors can work more efficiently when paired with ML—i.e., diagnose more quickly and/or more accurately.

Against these costs, one should put the speculative, but potentially large, gains caused by creating a data set of human decisions and resulting outcomes that can be used to provide ongoing training data for ML systems. If—and we stress that this may be a big "if"—humans end up deciding enough cases differently from ML to provide enough examples for training purposes, this may suffice to head off what would otherwise be the monoculture of training data that we warned about in Part III.

---

350. *See supra* text following note 162.

Electronic copy available at: https://ssrn.com/abstract=3114347

One other caveat should be noted: for the human-generated training data to have real value, it needs to include a significant number of cases in which the human's decision was better than ML's, something which likely will turn on how great ML's success rate is. As this point may be obscure, a short elucidation is in order. We assume ML is on average more accurate than people. But neither is 100% accurate. The less accurate the humans are, the less accurate ML needs to be in order to be noticeably better than humans. The less accurate a better-than-humans ML is, the more scope will remain for potential cases in which, were a human to overrule the ML system, they might improve the patient outcome. (Of course, there is also the possibility that they might both be wrong in different ways, but we can collapse that scenario by defining "right" as "better than the other diagnosis.") Conversely, the more accurate ML is overall, the less frequently we would expect to see a human decision to override the ML diagnosis lead to a better outcome.

4. Define the Standard of Care to Require a Human Doctor Plus ML?

Rather than create a malpractice exception for human–ML interactions, we could instead fix the legal standard of care (either legislatively or judicially) to require ML plus meaningful review by a human doctor. At present—while human diagnosticians remain on average superior to ML—any doctor who uses ML as a decisional aid is in effect subject to this standard of care. We suggested above that once ML is provably superior to the average human the standard of care would change, setting off a chain of events ending in the lack of meaningful human participation in certain diagnostic functions—a state we fear could be deleterious in the long term.[351] Freezing the standard of care to require meaningful human participation would head off those consequences. Indisputably, "meaningful" is a somewhat vague term, and it invites some fact-based debate as to what level of review by a human doctor would suffice. In the abstract, however, it is very hard to define the appropriate level of review with any precision; litigation in courts may actually be a good way of developing the factual records needed to put more detail into this standard.

Both the broad and narrow ML Exceptions to malpractice take large swaths of human liability out of the equation; in so doing they leave the choice of using a person or an AI to other factors, namely ethics[352] and cost.[353] In contrast, setting the standard of care to require both ML and humans invokes law to override those ethical and economic concerns, but it does so at the possible price of forgoing a larger number of beneficial outcomes that will not happen because the AI plus physician is too expensive.[354] The risk here is that some people may not be able to afford the care that they otherwise might have had.

On the other hand, freezing the standard of care makes it more likely than does the narrow ML Exception that the rate of human overrides of ML will tend

351. *See supra* Part III.
352. *Compare* Millar & Kerr, *supra* note 11, with sources cited *supra* note 88.
353. *See supra* Section II.A.
354. *See supra* text accompanying notes 195–97.

Electronic copy available at: https://ssrn.com/abstract=3114347

toward the optimal level, where "optimal" refers to individual-patient outcomes without considering systemic effects on training data. Under the narrow exception, humans are protected from liability for overruling ML in the absence of gross negligence, and this opens the door to excessive overrides. In contrast, setting the standard of care leaves current standards for reviewing a doctor's conduct in place.[355] Plaintiffs who wish to argue that a physician should have deferred to the ML will not be able to argue a *per se* violation of the standard of care, but doctors challenged for overriding ML will have to make the ordinary fact-based showing that their decisions were appropriate.

Even if the above is correct, and this proposal comes closest to incentivizing an optimal rate of human overrides of ML diagnoses, we cannot be confident that it will necessarily provide a sufficient supply of human-generated, accurate training data. How much data people will create depends on a number of variables that can only be estimated once ML is up and running full speed. The two chief variables are ML's failure rate and what fraction of those failures are detected and corrected by the human reviewers. (Recall that when humans wrongly override a correct diagnosis, this does not produce useful training data for ML; it might, however, provide useful training data for medical students.) We cannot know at this early stage whether the correct corrections will suffice, but this option probably gives as much hope as any, and more than most; the only one that comes close is the narrow ML Exception, and that is because its incentive effects are likely to be similar.

CONCLUSION: THE LEAST-WORST SOLUTION WILL BE EXPENSIVE

We have argued that if and when AI can outperform human doctors both malpractice law and, if pricing warrants it, economic imperatives will push providers to substitute machines for human doctors. This is not as wonderful as it may sound to technophiles because it creates a subtle risk of a closed loop as well as the obvious (short-run) opportunity for better patient care.

The risk is a result of AI's great promise. If, as we assumed for the purposes of this Article, some future ML system becomes significantly better at some types of diagnosis, such as reading x-rays and other radiological studies, then medical skills may suffer; if and when ML takes over treatment, some specialties may all but disappear. The problem we are concerned with is not directly the employment prospect of present or future radiologists. The problem is that the over-reliance on AI, and the resulting loss of medical knowledge, can create a closed loop in which future training and validation data sets are the result of decisions by the AI itself. At that point, we may lose the ability to discover new, better treatments, in the case where the ML system settles for a sub-optimal

355. Recall that the issue in a medical malpractice case is whether the claimed injury resulted from the treating physician's departure from "the generally recognized and accepted practices and procedures that would be followed by the average, competent physician in the defendant's field of medicine under the same or similar circumstances." *Supra* text accompanying note 108.

Electronic copy available at: https://ssrn.com/abstract=3114347

solution or the ML chooses a solution that optimizes a narrow performance criterion.

We can head off this scenario in a number of ways. The simplest legal change would be to require that a human be fully and meaningfully in the loop in all cases. Preventing an ML alone from becoming the standard of care, and thus defining the standard as an ML plus a physician meaningfully involved in reviewing the diagnostic decision, could alleviate the problem. We may also need to tinker with malpractice rules to prevent humans from being too unwilling to overrule an AI for fear of liability.

Admittedly, keeping physicians fully in the loop is likely to prove expensive compared to an AI-only world. Further, even if it may be a long-term fix, we should not expect it to be permanent. We will need to continue to revisit the level at which machines and humans integrate and exchange information and make decisions. Perhaps worst of all, our solution has more than enough of a whiff of the Luddite to make any robot or AI enthusiast uncomfortable. Nevertheless, we see no better answer at present; the remaining challenges will focus on the proper alignment of humans and machines to integrate and exchange information, and to make and carry out medical decisions. Figuring out how best to deal with the alignment questions will be a key consideration in the modernization of medical-school curricula so that the next generation of medical professionals are adequately trained to work with ML.

Modern auto-pilots are capable of making complex decisions while flying jets, decisions which may be too complex for human pilots to follow; in some cases human intervention prevents accidents, but in others it causes accidents that the autopilot might have prevented.[356] Yet we still require human pilots to be in the cockpit for the entire flight in case of emergency and despite the arguable duplication of expense.[357] Meanwhile, whether over-reliance on autopilots is dangerous, in part due to deskilling of pilots, is a live debate.[358] Now it's medicine's turn.

356. Gary Brown, *Out of the Loop*, 30 TEMP. INT'L & COMP. L.J. 43, 48–49 (2016).

357. *See* 14 C.F.R. § 91.3(a) (2018) ("The pilot in command of an aircraft is directly responsible for, and is the final authority as to, the operation of that aircraft."); *cf.* Brouse v. United States, 83 F.Supp 373, 374 (N.D. Ohio 1949) ("The obligation of those in charge of a plane under robot control to keep a proper and constant lookout is unavoidable.").

358. *See* Carolyn Presutti, *FAA Study Issues Recommendations to Correct Pilot Overreliance on Automation*, VOICE AM. (Nov. 22, 2013), https://www.voanews.com/a/faa-study-issues-recommendations-to-correct-pilot-overrelaince-on-automation/1795995.html (noting FAA's concern that "pilots are not as skilled at manually flying a plane in emergencies or when transitioning back from automation to manual").

Electronic copy available at: https://ssrn.com/abstract=3114347