2019

Artificial Intelligence in the Health care Space: How We Can Trust What We Cannot Know

Robin Feldman
UC Hastings College of the Law, feldmanr@uchastings.edu

Ehrik Aldana
UC Hastings College of the Law, aldanaehrik@uchastings.edu

Kara Stein
UC Hastings College of the Law, steinkara@uchastings.edu

ARTIFICIAL INTELLIGENCE IN THE HEALTH CARE SPACE: HOW WE CAN TRUST WHAT WE CANNOT KNOW

Robin C. Feldman*, Ehrik Aldana** & Kara Stein***

As AI moves rapidly into the health care field, it promises to revolutionize and transform our approach to medical treatment. The black-box nature of AI, however, produces a shiver of discomfort for many people. How can we trust our health, let alone our very lives, to decisions whose pathways are unknown and impenetrable?

As challenging as these questions may be, they are not insurmountable. And, in fact, the health care field provides the perfect ground for finding our way through these challenges. How can that be? Why would we suggest that a circumstance in which we are putting our lives on the line is the perfect place to learn to trust AI? The answer is quite simple. Health care always has been a place where individuals must put their faith in that which they do not fully understand.

Consider the black box nature of medicine itself. Although there is much we understand about the way in which a drug or a medical treatment works, there is much that we do not. In modern society, however, most people have little difficulty trusting their life to incomprehensible treatments.

This article suggests that the pathways we use to place our trust in medicine provide useful models for learning to trust AI. As we stand on the brink of the AI revolution, our challenge is to create the structures and expertise that give all of society confidence in decision-making and information integrity.

* Arthur J. Goldberg Distinguished Professor of Law and Director of the Center for Innovation, University of California Hastings College of the Law.

** Research Fellow, Center for Innovation, University of California Hastings College of the Law.

*** Senior Research Fellow, Center for Innovation, University of California Hastings College of the Law; former Commissioner of the Securities and Exchange Commission.

INTRODUCTION

Artificial Intelligence (AI) is moving rapidly into the health care field. Personalized medicine, faster and more accurate diagnostics, and accessible health apps boost access to quality medical care for millions. In a similar vein, data from doctor visits, clinical treatments, and wearable biometric monitors are collected and fed back into ever-learning and ever-improving AI systems.

As AI advances, it promises to revolutionize and transform our approach to medical treatment. As any cancer specialist will attest, however, transformation may be for the good, or it may not.[1] Such is the case with AI. On the one hand, AI may have the ability to revolutionize society's discovery of disease treatment—as well as enhance our ability to rapidly deliver that treatment in a manner tailored to an individual's needs. On the other hand, the black box nature of AI produces a shiver of discomfort for many people. How can we trust our health, let alone our very lives, to decisions whose pathways are unknown and impenetrable?

The black box nature of artificial intelligence raises concern whenever such technology is in play. For example, suppose an autonomous car makes a decision that leads to injury or death. If we do not understand the pathways that led to the choices made, we may be reluctant to trust the decision. Moreover, if an algorithm is used by a court to determine whether a defendant should receive bail, without the factors and analysis transparently available to the public, we may be reluctant to trust that decision as well. Of course, when a human driver makes a choice that leads to injury or death, we may not fully understand the decision pathway either. It would be a stretch to say that any reconstruction of an event could accurately dissect the mental pathways. After all, human memory is frail, and humans are remarkably able to remember events in a way that casts them in the most favorable light. Nevertheless, our legal system is grounded in the concept of open deliberation, and the notion that one cannot even try to unravel the reason for a decision creates discomfort. More important, although the notion may be somewhat misguided, individuals may be more likely to trust those who are similar to them. And nothing seems more different from a human being than an algorithm.

As challenging as these questions may be, they are not insurmountable. We suggest that the health care field provides the perfect ground for finding our way through these challenges. How can that be? Why would we suggest that a

1. *See* Robin Feldman, *Cultural Property and Human Cells,* 21 INT'L J. CULTURAL PROP. 243, 248 (2014) (explaining that tumors can operate in a systems approach; if treatments cut off one approach to the tumor's growth, the tumor may develop work-arounds which can be more dangerous and damaging than the original pathway).

circumstance in which we are putting our lives on the line is the perfect place to learn to trust AI? The answer is quite simple. Receiving health care treatment always has been, and always will be, one of the moments in life when individuals must put their faith in that which they cannot fully understand.

Consider the black box nature of medicine itself. Although there is much we understand about the way in which a drug or a medical treatment works, there is still much that we do not. In modern society, however, most people have little difficulty trusting their lives to often incomprehensible treatments. Such trust is all the more important given evidence that confidence in medical treatment affects treatment outcome. In other words, those who believe their medicine will work stand a better chance of being healed.[2]

Trust is vital to developing and adopting health care AI systems, especially for health (medicine you trust works better) and AI (the more adoption, the better it becomes). The "black box" mentality we use to conceptualize AI reduces trust, as well as stalling the development and adoption of potentially life-saving treatments discovered or powered by AI. However, just because we can't completely understand something doesn't mean we shouldn't trust it. Medicine is a prime example of this—the original "black box." Despite the challenges, medicine has overcome the "black box" problem with the help of policy and regulatory bodies. This Article suggests that the pathways we use to place our trust in medicine provide useful models for learning to trust AI. The question isn't whether we know everything about how a particular drug might work or how an AI reaches its decision; the question is whether there are rules, systems, and expertise in place that give us confidence. As we stand on the brink of the AI revolution, our challenge is to create the architecture that will give all of society confidence in AI decision-making. And of course, society must ensure that such confidence is deserved—that we can trust the integrity of the information being used by AI and the reliability AI decisions.

I. TRUST & INTERPRETABILITY

A. *The Importance of Trust in Health Care and AI*

In recent years, we have seen that AI systems in a variety of fields are able to match, and in some cases exceed, the ability of humans to perform specific tasks. Widely known for defeating humanity's best in games like Chess,[3] Jeopardy,[4] and Go;[5] AI systems now are expanding into more practical and

2. Johanna Birkhäuer et al., *Trust in the Health Care Professional and Health Outcome: A Meta-analysis*, PLOS ONE 9 (Feb. 7, 2017), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5295692.

3. IBM, DEEP BLUE — OVERVIEW, IBM100, https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue.

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

professional endeavors such as driving,[6] law,[7] labor hiring and management,[8] furniture assembly,[9] and even investment advice.[10]

In the health care space, AI also is making waves, and is involved in drug discovery, diagnostics, surgery, patient information management, and psychological treatment. For example, automated medical-image software systems are beginning to arrive at expert-level diagnostic accuracy,[11] impacting medical specialties such as radiology, ophthalmology, dermatology, and pathology.[12] AI also has been deployed to discover drugs to treat particular

---

5. *Artificial Intelligence: Google's AlphaGo Beats Go Master Lee Se-dol,* BBC NEWS (Mar. 12, 2016), https://www.bbc.com/news/technology-35785875.

6. BRANDON SCHOETTLE & MICHAEL SIVAK, A PRELIMINARY ANALYSIS OF REAL-WORLD CRASHES INVOLVING SELF-DRIVING VEHICLES, UNIVERSITY OF MICHIGAN TRANSPORTATION RESEARCH INSTITUTE No. UMTRI-2015-34, 15 (2015), http://umich.edu/~umtriswt/PDF/UMTRI-2015-34.pdf (showing crash rates between self-driving and human-driven vehicles are within the margin of error of one another, suggesting "[the investigators] currently cannot rule out, with a reasonable level of confidence, the possibility that the actual [crash] rates for self-driving vehicles are lower than for conventional vehicles.").

7. LAWGEEX, COMPARING THE PERFORMANCE OF ARTIFICIAL INTELLIGENCE TO HUMAN LAWYERS IN THE REVIEW OF STANDARD BUSINESS CONTRACTS 2 (2018) (comparing the contract analysis abilities of a legal AI platform LawGeex against human lawyers. In the study, human lawyers achieved an eighty-five percent average accuracy rate, while AI achieved ninety-five percent accuracy. Moreover, the AI completed the task in only twenty-six seconds, while the humans took an average ninety-two minutes).

8. *See* Sean Captain, *This AI Factory Boss Tells Robots & Humans How to Work Together,* FAST COMPANY (Aug. 7, 2017), www.fastcompany.com/3067414/robo-foremen-could-direct-human-and-robot-factory-workers-alike (describing a "Boss AI" project Siemens is working on in which jobs are assigned to human workers and robotic workers based on the worker's skill and the job requirements); Don Nicastro, *5 Things to Consider When Using AI for Hiring,* CMS WIRE (Nov. 8, 2018), https://www.cmswire.com/digital-workplace/5-things-to-consider-when-using-ai-for-hiring (explaining that nearly all Fortune 500 companies use automation to support the hiring process and citing a 2018 LinkedIn report that seventy-six percent feel AI's impact on recruiting will be at least somewhat significant).

9. Francisco Suárez-Ruiz et al., *Can Robots Assemble an IKEA Chair?* 3 SCI. ROBOTICS 2 (2018).

10. *See* Swapna Malekar, *Ethics of Using AI in the Financial/Banking Industry,* DATA DRIVEN INVESTOR (Sept. 16, 2018), https://medium.com/datadriveninvestor/ethics-of-using-ai-in-the-financial-banking-industry-fa93203f6f25 (describing Royal Bank of Canada's experiments with using personal, social, commercial and financial customer data to provide personalized recommendations to end users).

11. *See* Geert Litjens et al., *A Survey on Deep Learning in Medical Image Analysis,* 42 MED. IMAGE ANAL. 60, 68-69 (2017), *available at* https://arxiv.org/pdf/1702.05747.pdf (reviewing over 300 research contributions to medical image analyses, finding that "[e]specially CNNs [convolutional neural networks] pretrained on natural images have shown surprisingly strong results, challenging the accuracy of human experts in some tasks."). *See also* ULTROMICS, http://www.ultromics.com/technology (last visited May 14, 2019) (describing their AI diagnostics system for heart disease).

12. *See* Kun-Hsing Yu et al., *Artificial Intelligence in Healthcare,* 2 NATURE BIOMEDICAL ENGINEERING, 719, 722-725 (2018); *see also* Huiying Liang et al., *Evaluation and Accurate Diagnoses of Pediatric Diseases Using Artificial Intelligence,* 25 NATURE MED. 433, 433 (2019), https://www.nature.com/articles/s41591-018-0335-9 (Chinese AI system consistently outperformed humans in pediatric diagnoses); *see also* Azad Shademan

diseases.[13] These drugs potentially could be tailored to avoid negative side effects, such as a new chemical scaffold for the opioid receptor.[14]

Perhaps the most exciting and intriguing arenas for AI in the health care field are devices for monitoring health and for delivering treatment. This brave new world includes wearable and implantable devices, or what one of the authors calls "implantable nurses,"[15] along with AI practitioners who can deliver care to a patient.[16] A report from the Association of American Medical Colleges anticipates a national shortage of about 46,000 to 90,000 physicians by 2025.[17] These shortfall predictions highlight the health care system's need to find innovative ways to efficiently and safely deliver medical care. In that context, health care AI systems are filled with promise. For example, a chatbot nurse in the future could perform an initial diagnosis or engage in triage. It could do this by asking a patient's symptoms, examining data from wearable devices, and looking at easily accessible and crowdsourced health records of other patients from around the world.

One can expect that AI's continued use and proper development will result in improved health and life outcomes for a truly immense number of future patients. However, despite these potential benefits, there exist significant barriers to adoption of the health care AI systems. For instance, a 2018 survey of health care decisionmakers found that respondents saw lack of patient and clinician trust in AI as a significant barrier to adoption.[18] This lack of trust in

---

et al., *Supervised Autonomous Robotic Soft Tissue Surgery*, 8 SCI. TRANSLATIONAL MED. 337, 342 (2016) (detailing a 2016 study where an autonomous robotic system during an *in vivo* intestinal procedure showed, in a laboratory setting, better suturing quality than human surgeons).

13. *See* Evan N. Feinberg, *AI for Drug Discovery in Two Stories*, MEDIUM (Mar. 14, 2018), https://medium.com/@pandelab/ai-for-drug-discovery-in-two-stories-49d7b1f019f3.

14. *See id.*

15. *See Federal Trade Commission, Hearings on Emerging Competition, Innovation, and Market Structure Questions Around Algorithms, Artificial Intelligence, and Predictive Analytics* (2018) (statement of Robin Feldman, Professor of Law, University of California Hastings Law), *available at* https://www.ftc.gov/news-events/audio-video/video/ftc-hearing-7-nov-14-session-2-emerging-competition-innovation-market.

16. *See* Miguel Hueso et al., *Progress in the Development and Challenges for the Use of Artifical Kidneys and Wearable Dialysis Devices*, 5 KIDNEY DISEASES 3, 4 (2018), *available at* https://www.karger.com/Article/FullText/492932 (discussing the potential for novel wearable dialysis devices with contributions from AI); *see also* Erin Brodwin, *I Spent 2 Weeks Texting a Bot About My Anxiety—and Found It to Be Surprisingly Helpful*, BUS. INSIDER (Jan. 30, 2018), https://www.businessinsider.com/therapy-chatbot-depression-app-what-its-like-woebot-2018-1; Daniel Kraft, *12 Innovations that Will Revolutionize the Future of Medicine*, NAT'L GEOGRAPHIC (January 2019), https://www.nationalgeographic.com/magazine/2019/01/12-innovations-technology-revolutionize-future-medicine (discussing smart contact lenses with biosensors).

17. IHS INC., THE COMPLEXITIES OF PHYSICIAN SUPPLY AND DEMAND: PROJECTIONS FROM 2013 TO 2015, 28 (2015), *available at* https://www.aamc.org/download/426248/data/thecomplexitiesofphysiciansupplyanddemandprojectionsfrom2013to2.pdf.

18. INTEL CORP., *Overcoming Barriers in AI Adoption in Healthcare*, https://newsroom.intel.com/news-releases/u-s-healthcare-leaders-expect-widespread-adoption-artificial-intelligence-2023/ (last visited 2018).

AI is exhibited in other fields as well. For example, another study demonstrates that only eight percent of people would trust a machine offering mortgage advice, compared to forty-one percent trusting mortgage advice from a human mortgage broker (with nine percent of participants claiming they would trust a horoscope).[19] It is worth noting that banking customers seemingly would have more trust in a horoscope than in machine learning.

Trust plays a vitally important role in various aspects of the health care system, particularly those grounded in the patient-provider relationship.[20] Without established trust between the patient and provider, a patient has little to no incentive to seek care and advice, share sensitive information, or follow the treatment plans and preventative recommendations of a provider.[21] Mistrust also can pose serious health consequences for the public, especially if individuals choose not to get flu shots or vaccinate their children due to lack of trust in medical providers.[22] As such, creating, preserving, and enhancing trust is understood to be one of the fundamental goals of medical ethics, as well as health care law and public policy.[23]

Trust is all the more important when considering how confidence in medical treatment may measurably affect treatment outcomes. Trust is widely believed to be essential to therapeutic outcomes and an effective course of treatment.[24] Commentators speculate that trust is a key factor in the mind-body interactions that underlie placebo effects and unexplained variations in outcomes from conventional therapies.[25] Meta-analysis has supported these hypotheses, indicating a correlation between trust and healthcare outcomes, such as beneficial health behaviors, fewer symptoms of illness, higher quality of life, and more satisfaction with treatment.[26] In other words, those who

19. HSBC, *Trust in Technology* 4 (2017), https://www.hsbc.com/-/files/hsbc/media/media-release/2017/170609-updated-trust-in-technology-final-report.pdf?download=1 (discussing independent survey of over 12,000 respondents in eleven countries on technology perceptions in habits).

20. Mark A. Hall et al., *Trust in Physicians and Medical Institutions: What Is It, Can It Be Measured, and Does It Matter?*, 21 MILLBANK Q. 4, 613 (2001).

21. David H. Thom et al., *Measuring Patients' Trust In Physicians When Assessing Quality Of Care*, 23 HEALTH AFF. 4 (study finding sixty-two percent of patients with high levels of trust always take their medications prescribed by providers, while only fourteen percent of those with low levels of trust do); Dhruv Kullar, *Do You Trust the Medical Profession?*, N.Y. TIMES (Jan. 23, 2018), https://www.nytimes.com/2018/01/23/upshot/do-you-trust-the-medical-profession.html ("[A] study found that trust is one of the best predictors of whether patients follow a doctor's advice about things like exercise, smoking cessation and condom use.").

22. *See* Kullar, *supra* note 21.

23. *See* David Mechanic, *The Functions and Limitations of Trust in the Provision of Medical Care*, 23 J. HEALTH. POL. L. 4: 661-86 (1996); David Mechanic & Mark Schlesinger, *The Impact of Managed Care on Patients' Trust in Medical Care and Their Physicians*, 275 JAMA 21: 1693-1697 (1998).

24. *See* Hall, *supra* note 20, at 614.

25. *Id.*

26. *See* Birkhäuer et al., *supra* note 2.

believe their medicine and treatment will work stand a better chance of being healed.

Finally, because clinical trials and enrollment rely on the trusting participation of patients, trust is a necessary component of innovation and research in the health care space. Patients play a vital role in medical innovation. They are the ones who participate in clinical trials that allow doctors and scientists to experiment and develop new treatments. If patients do not trust their providers in particular,[27] or their health care in general, they will be unlikely to participate in relevant studies of new treatments and technologies.

The fact that lack of trust may stifle participation in clinical trials is especially relevant to health care AI systems, which require diverse and large amounts of data (from people) in order to optimize outcomes.[28] The more that people use AI, the more data can be fed back into AI systems to iterate and improve them. Without sufficient and representative amounts of training data, current AI systems cannot function effectively—in some cases affecting certain populations disproportionately.[29]

Finally, some types of AI systems can only provide maximum benefit if participation is widespread or even universal.[30] Consider autonomous cars. One of the greatest impediments to successful utilization of autonomous cars, and one of the greatest dangers on the road, is the fact that human drivers are puzzlingly irrational.[31] Imagine a world in which a majority of cars are

---

27. *See, e.g.*, Doris T. Penman, *Informed Consent for Investigational Chemotherapy: Patients' and Physicians' Perceptions*, 2 J. CLINICAL ONCOLOGY 7: 849-55 (1984) (finding that cancer patients considering experimental chemotherapy indicate that trust in their physician was a primary reason for participating in a clinical trial); *see also* Giselle Corbie-Smith et al., *Attitudes and Beliefs of African Americans Toward Participation in Medical Research*, 14 J. GEN. INTERNAL MED. 9: 537-46 (1999); Angeliki Kerasidou, *Trust Me, I'm a Researcher!: The Role of Trust in Biomedical Research*, MED. HEALTH CARE PHIL. 1: 43-50, 43 (2017); Mark Yarborough and Richard R. Sharp, *Restoring and Preserving Trust in Biomedical Research*, 77 ACAD. MED. 8, 9 (2002).

28. *See* Yu et al., *supra* note 12, at 719-20.

29. Danton S. Char, Nigam H. Shah, & David Magnus, *Implementing Machine Learning in Health Care – Addressing Ethical Challenges*, 378 NEW ENG. J. MED. 981, 982 (2018) ("An algorithm designed to predict outcomes from genetic findings will be biased if there have been few (or no) genetic studies in certain populations. For example, attempts to use data from the Framingham Heart Study to predict the risk of cardiovascular events in nonwhite populations have led to biased results, with both overestimations and underestimations of risk.").

30. Robin C. Feldman, *Artificial Intelligence: The Importance of Trust & Distrust*, 21 GREEN BAG 2D 201, 205-07 (2018) (discussing networked AI systems and explaining that "some of the power of AI systems depends not just on whether humans can be coaxed into using them at all but also whether the use is widespread, even ubiquitous"); *see also* Kristen Hall-Geisler, *All New Cars Could Have V2V Tech by 2023*, TECH CRUNCH (Feb. 2, 2017), techcrunch.com/2017/02/02/all-new-cars-could-have-v2v-tech-by-2023.

31. *See* Matt Richtel & Conor Dougherty, *Google's Driverless Cars Run into Problems: Cars with Drivers*, N.Y. TIMES (Sept. 1, 2015), https://www.nytimes.com/2015/09/02/

autonomous and networked together, while a few cars are driven by irrational humans. To put it simply, those humans are likely to gum up the works.

In the health care context, imagine a system in which AI bots perform certain tiny, delicate functions in a complex operation requiring different surgical specialties. If some of those doctors decline to participate so that the AI system is working partially with (1) human doctors who will work with an AI surgical device; and (2) human doctors who will work only with manual devices—analogous to humans who will drive in a driverless car and those who wish to drive on their own—the AI cannot operate to its highest potential. For these circumstances and many others, trust will be essential in ensuring the full benefits. To begin that process, one must understand why the public might not trust an AI health care system.

B. *AI's "Black Box" Barrier*

A primary concern with the use of health care AI systems (and AI systems generally) is their potential "black box" nature.[32] AI systems are often labeled black boxes by both the media[33] and academics[34] because while their inputs and outputs are visible, the internal process of getting from the input to the output remains opaque. People can describe the degree of accuracy of an AI system for its given purpose, but given the current state of the field, they cannot explain or recreate the system's reasoning for its decision. The degree to which a human observer can intrinsically understand the cause of a decision is described by the machine learning community as an AI system's "interpretability" or "explainability."[35]

---

technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html?_r=0.

32. *See* Intel Corporation, *supra* note 18.

33. *See, e.g.*, Jeff Larson, Julia Angwin, & Terry Parris Jr., Breaking the Black Box: How Machines Learn to Be Racist, PROPUBLICA (Oct. 19, 2016), https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist?word=Clinton; Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH REVIEW (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai.

34. *See, e.g.*, Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J. L. TECH 889, 891 (2018), https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf.

35. Richard Tomsett et al, "Interpretable to Whom? A Role Based Model for Interpretable Machine Learning Systems," 2018 ICML Workshop on Human Interpretability in Machine Learning (2018) at 9. For a discussion of the discourse surrounding the definition of "interpretability" in the context of AI/machine learning, see Zachary C. Lipton, "The Mythos of Model Interpretability," 2016 ICML Workshop on Human Interpretability in Machine Learning (2016).

Today, discussions about AI generally are about deep learning.[36] Although a full discussion of deep learning is beyond the scope of this article, the term typically refers to using historical data to train a computer model to make predictions about future data and to direct computer choices based on that data. Using a *very* loose analogy to the human brain, we call these computer models neural networks. Neural networks operate by applying a series of mathematical algorithms or transformations, with each transformation referred to as a layer of the neural network. Thus, deep learning simply means a system that has many iterations. These iterations contribute to the difficulty that the public, policymakers, and even developers have in explaining the reasoning behind a deep learning AI system.

Amongst the public, this lack of interpretability in AI is often seen as a significant barrier to trust.[37] In a 2017 survey of CEOs, seventy-six percent of respondents said potential for biases and a lack of transparency were impeding AI adoption in their businesses.[38] Indicators such as these point to how a lack of interpretability may significantly deter trust in, and ultimately the adoption of, health care AI systems. Government agencies and the public will want to know more than just the computer's outcome; they will want to know how the computer reached that outcome.[39]

In an age in which medical symptoms and treatment information are readily available online, society cannot expect patients to blindly trust what they don't understand. This is especially true for cases in which mistakes or misclassifications in machine-learning models may have a high cost.[40] Trust issues involve more health care actors than just patients. Doctors are also less likely to trust what they cannot understand. They also will not be able to convey the necessary degree of understanding to their patient, possibly eroding an additional dimension of trust between patients and their providers.

The black box nature of AI can also make it more difficult to figure out technical problems or vulnerabilities. This also makes it more difficult to come

---

36. The description of AI and deep learning in this paragraph is adapted from Feldman, *supra* note 30, at 202-03 and FEDERAL TRADE COMMISSION HEARINGS (statement of Professor Robin Feldman), *supra* note 15.

37. *See, e.g.*, Cliff Kuang, *Can A.I. be Taught to Explain Itself?*, N.Y. TIMES MAG. (Nov. 21, 2017), https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html; Ian Sample, *Computer Says No: Why Making AIs Fair, Accountable, and Transparent Is Crucial*, GUARDIAN (Nov. 5, 2017, 7:00 AM), https://perma.cc/5H25-AQC7. *See also* Zachary C. Lipton, The Doctor Just Won't Accept That! (Dec. 7, 2017) (unpublished submission, presented at Interpretable ML Symposium (NIPS 2017)), https://arxiv.org/abs/1711.08037.

38. Anand Rao and Chris Curran, *AI Is Coming. Is Your Business Ready?* (Sept. 26, 2017), http://usblogs.pwc.com/emerging-technology/artificial-intelligence-is-your-business-ready.

39. *See* Kuang, *supra* note 37. See *also* Feldman, *supra* note 30, at 206-07.

40. *See* Muhammad Aurangzeb Ahmad et al., *Interpretable Machine Learning in Healthcare*, 19 IEEE INTELLIGENT INFORMATICS BULL. 1, 1 (2018), http://www.comp.hkbu.edu.hk/~cib/2018/Aug/article1/iib_vol19no1_article1.pdf.

up with possible technical solutions. To maximize proper functioning of systems and determine potential areas of failure, those who interact with an AI system must be able to understand it. In a well-known example, a research team created a relatively simple pneumonia risk assessor using machine learning classifying tools.[41] Curiously, the resulting computer model suggested that patients who have asthma have a lower risk of in-hospital death when admitted for pneumonia. That determination, however, was not accurate: patients with asthma actually have a higher risk for complications from pneumonia that can lead to in-hospital death.[42] The model "implicitly" accounted for the fact that pneumonia patients with asthma generally receive significantly more attention by providers, thereby increasing their chance of survival. Lacking this data, the model merely inferred correlations from the data it had. This error was detected by a human reviewer who used common sense reasoning. However, the computer model, as a black box, was not able to explain (in a manner understood by humans) why it made the choices that it did. In the context of more complex medical situations, this type of problem could be particularly worrisome because errors are less easily detectable through human review and intuition.

Of course, the easiest way to solve the issues caused by the black box nature of AI would be to eliminate the black box nature of AI. If researchers could find a simple way to make the box crystalline, or even create a clearer view into the box, in a way that most humans could easily understand, the problem would be solved. That, however, is more difficult than it appears, and even machine learning and computer science experts are in a debate over whether we might be able to achieve interpretability.[43] On the technical level, AI systems are opaque largely because they constantly modify their own parameters and rules.[44] In other words, an AI system works not simply by following the rules that are inputted, but by creating and testing its own hypotheses, and then adjusting and readjusting the hypotheses along the way

41. *See* Rich Caruana et al., Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission 1721, 1721 (2015) (unpublished submission, 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), https://scinapse.io/papers/1996796871.

42. *Id.*

43. *See* The Artifical Intelligence Channel, *The Great AI Debate - NIPS 2017 – Yann LeCun*, YOUTUBE (Dec. 9, 2017), https://www.youtube.com/watch?v=93Xv8vJ2acI (discussing how interpretability is necessary for machine learning).

44. *See* Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y, Jan.–June 2016, at 1, 10, https://journals.sagepub.com/doi/abs/10.1177/2053951715622512, *cited in* NUFFIELD COUNCIL ON BIOETHICS, BIOETHICS BRIEFING NOTE: ARTIFICIAL INTELLIGENCE (AI) IN HEALTHCARE IN RESEARCH (2019), http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf.

until it reaches a conclusion. These are the so-called "layers" of the deep learning.[45]

Cognizant of AI's black box problem, researchers have begun trying to remedy the issue. For example, in 2016, the Defense Advanced Research Projects Agency (DARPA) launched the Explainable Artificial Intelligence (XAI) Initiative. The Initiative hopes to translate decisions made by machine learning systems into something accessible to human understanding.[46] Although these are important steps for AI systems, researchers currently do not know when or if the field might succeed in achieving widespread interpretability to a degree that adequately satisfies stakeholders such as patients in the health care context—or even to a degree that experts can understand. Solving the tension that exists regarding whether deep learning AI systems can eventually be explained or if they are truly too complex to be knowable, however, is not necessary to resolve at this point. In either case, we need to construct a path forward, so that the development of potentially life-saving technologies does not stall unnecessarily. Such a path forward should accommodate for both realities.

Moreover, even if it becomes technologically feasible to open the black box so that people can better peer in, there may be societal and legal constraints. For example, private sector stakeholders are likely to worry that detailed explanations about the inner workings of a proprietary machine learning system could undermine the intellectual property interests they hold in the technology.[47] As one of the authors has suggested, "a company's first instinct is unlikely to encompass throwing open the doors to its technology, particularly if competitors are peering into the open doorway."[48] The same may hold true for post-hoc examination and explanation of an AI system's decision pathway. If the explanations of decisions are too thorough, it would potentially be possible to reverse-engineer the product. In fact, protection of intellectual property has always been the major argument against those advocating for "opening the black box."[49]

For this reason, one of the authors has suggested that intellectual property protection for AI inventions should follow the pathway of data protection for

---

45. *See supra* text accompanying notes 36-39 (briefly explaining deep learning and neural nets).

46. *See* Sara Castellanos & Steven Norton, *Inside Darpa's Push to Make Artificial Intelligence Explain Itself*, WALL ST. J. (Aug 10, 2017), https://blogs.wsj.com/cio/2017/08/10/inside-darpas-push-to-make-artificial-intelligence-explain-itself; David Gunning, *Explainable Artificial Intelligence (XAI)*, DARPA.MIL (last visited 2019), https://www.darpa.mil/program/explainable-artificial-intelligence.

47. For discussion of the rush to patent AI, see Tom Simonite, *Despite Pledging Openness, Companies Rush to Patent AI Tech*, WIRED (July 31, 2018, 7:00 AM), https://www.wired.com/story/despite-pledging-openness-companies-rush-to-patent-ai-tech.

48. Feldman, *supra* note 30, at 206.

49. *See* Sarah Wachter, Brent Mittelstadt, & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 85-86.

new pharmaceuticals. For example, in exchange for sharing clinical trial data on safety and efficacy with government regulators and allowing generic competitors, branded drugs receive four or five years of intellectual property protection before a generic competitor can file for approval using that data. Similarly, in the AI field, innovators could receive a short period of time for protection—four to five years—in exchange for allowing safety and efficacy data to be used by both the government and competitors.[50] No such change in the law, however, is currently on the horizon.

As society waits for technological and legal innovation to solve these challenges, AI need not necessarily sit idle. Scientific innovation marches forward at its own pace, regardless of whether law and society are ready for it. And as counter-intuitive as it may sound, the health care system may provide the ideal pathway for thinking through a framework that establishes greater trust in AI, despite the lack of clarity in AI processes.

II. PATHWAYS TOWARD TRUST WITHOUT CLARITY

A. *Is Medicine Already a Black Box?*

As noted above, it is presently unclear how or whether scientists technically will be able to overcome the black box problem. This suggests two possible scenarios for the future: (1) AI systems that use deep learning techniques will never be interpretable to experts or the general public; or (2) AI systems will eventually be interpretable to experts, who likely will experience difficulty explaining the algorithmic decision-making processes to the general public. In order to accommodate the innovation, development, and adoption of AI systems while optimizing for safety and efficiency, a robust pathway forward should account for both of these scenarios.

There are a myriad of examples in which users do not know how a product works, and yet they continue to trust and use that product. One does not need to be a computer programmer or an automotive engineer in order to trust that technology will work as promised. For example, consumers don't need to know how their cell phones or cars work, but they do need to trust that (1) their equipment will perform certain tasks when requested; and (2) their devices aren't performing other tasks without their knowledge, such as spying on them, that could in some way harm them.

As described in the first scenario detailed above, it may seem implausible that consumers trustingly adopt a product whose inner workings the developer doesn't even understand. Nevertheless, this is already the case in the high-stakes health care realm. Consider pharmaceuticals. In pharmacology, scientific

50. *See* Robin C. Feldman & Nick Thieme, *Competition at the Dawn of Artificial Intelligence*, *in* THE EFFECTS OF DIGITALIZATION, GLOBALIZATION AND NATIONALISM ON COMPETITION (forthcoming Edward Elgar Publishing 2019), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3218559.

researchers aim to understand the cause behind a drug's effects based on the drug's chemical properties and how those properties physiologically interact with living organisms.[51] Put another way, researchers aim to find a drug's mechanism of action. For example, the mechanism of action for selective serotonin reuptake inhibitors (SSRIs)—commonly used to treat depression and one of the most prescribed therapeutic treatments in medicine—is well known. SSRIs inhibit the reuptake of serotonin, which increases the level of serotonin in the brain and improves a person's mood.[52]

At the same time, there are many common drugs that do not have a known mechanism of action and yet are regularly prescribed by medical providers and used by patients.[53] Take for example the widely prescribed drug compound *acetaminophen*, more commonly recognized by its brand name Tylenol. If someone has a headache or fever, there would be little hesitation in going to a pharmacy or grocery store to pick up a bottle, even though consumers and researchers don't know how the drug works.[54] Of course, we all grew up with acetaminophen, and it seems deeply familiar and trustworthy to us. But someone had to start using it sometime. Moreover, there are numerous drugs today whose mechanisms of action are unknown, including the muscle relaxant metaxalone, the diabetes-related drug metformin, and the cough suppressant guifenesin. Government regulatory bodies determine *whether* many drugs and treatments are safe and effective, but the answer of *how* the drug works is not a necessary condition.[55] Thus, no one knows how they really work, yet doctors are quite comfortable prescribing them, and patients are quite comfortable taking them.

So how has society reached this level of trust around drugs whose mechanism of action is unknown? At the moment, we convince patients of a drug's safety and efficacy through rigorous testing procedures. These mechanisms of trust are further mediated by expert regulatory bodies—the FDA in the United States—and by doctors—whose focus on patient care and

---

51. *See Subjects: Pharmacology*, NATURE, https://www.nature.com/subjects/pharmacology (last visited Apr. 24, 2019).

52. *See* Stephen M. Stahl, *Mechanism of Action of Serotonin Selective Reuptake Inhibitors: Serotonin Receptors and Pathways Mediate Therapeutic Effects and Side Effects*, 51 J. AFFECTIVE DISORDERS 3, 215 (1998), https://www.ncbi.nlm.nih.gov/pubmed/10333979.

53. *See* Carolyn Y. Johnson, *One Big Myth About Medicine: We Know How Drugs Work*, WASH. POST (July 23, 2015), https://www.washingtonpost.com/news/wonk/wp/2015/07/23/one-big-myth-about-medicine-we-know-how-drugs-work ("Knowing why a drug works has historically trailed the treatment, sometimes by decades.").

54. *See* Carmen Drahl, *How Does Acetaminophen Work? Researchers Still Aren't Sure*, 92 CHEMICAL & ENGINEERING NEWS 29, 31 (2014), https://cen.acs.org/articles/92/i29/Does-Acetaminophen-Work-Researchers-Still.html.

55. *See* Russell Katz, *FDA: Evidentiary Standards for Drug Development and Approval*, 1 THE AM. SOC'Y FOR EXPERIMENTAL NEUROTHERAPEUTICS 307, 316 (2004) ("Theories about mechanism of action of a drug or disease mechanisms play important parts in drug development and approval, but they are entirely subsidiary to the fundamental questions that must be answered in the course of drug approval; namely, is a drug effective, and is it safe in use."); *Mechanism Matters*, 16 NATURE MED. 347 (2010).

patient bonding help bridge the understanding gap and establish trust. Of course, one could ask which came first: Did we develop regulatory systems and expert mediators because they were the only ways to engender trust, or did we trust black box medical treatments because regulatory systems and expert mediators exist? Mimicking these mechanisms of trust, however, does not require answering that question. The fact that this avenue exists—however it may have developed—makes it easier to travel down the same road again.

We should also note at this point that the existence of trust, without clarity, in the health care system may not be perfectly parallel in other arenas in which AI may be used. For much of history, medicine operated on the model that "the doctor knows best," and it is only in more recent memory that the field has evolved so that patients play a more active role in their health care. The same cannot necessarily be said of all areas in which AI may operate.

Although the costs and benefits of regulatory and expert mediation for AI systems may not be exactly commensurate with those involving the FDA and doctors, there is much that is similar. For example, the FDA must monitor and respond to concerns that emerge long after a drug has been approved. The FDA also must continually monitor whether pharmaceutical manufacturing facilities meet proper safety regulations.[56] That being said, AI systems evolve far more rapidly and constantly than pharmaceuticals, shifting according to the AI's learning and experience. Moreover, the field itself is evolving at light speed. The foundation for all of modern neural networks emerged only five years ago.[57] One might compare this dynamic to a drug manufacturer constantly adjusting a drug's formula as it was simultaneously undergoing testing.

Despite these distinctions, at least in the health care field, clarity (or even the possibility of clarity) is not absolutely fundamental in order for stakeholders to trust in the effectiveness of the medical treatment. Health care has proven that it can overcome both scenarios described at the beginning of this section—that scientists may never be able to fully understand AI decision pathways and that, even if they do, experts may not be able to explain them to the public. These scenarios also suggest a framework for developing trust in AI health care systems, one that could eventually expand to AI systems in general.

---

56. *See* FDA, CURRENT GOOD MANUFACTURING PRACTICE REGULATIONS (2018), https://www.fda.gov/Drugs/DevelopmentApprovalProcess/Manufacturing/ucm090016.htm.

57. *See* Feldman, *supra* note 30, at 203 (explaining that the difference between two of the latest versions of Google's AlphaGo would be analogous to the difference between the first rudimentary touchscreen phone twenty five years ago and the new iPad Pro, except that the AI advancement took place over two years, rather than twenty five). For the work that provided the basis of modern neural nets, see Ian J. Goodfellow et al., *Generative Adversarial Nets*, NEURAL INFO PROCESSING SYS. CONF. (2014), papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

B. *Pathways Forward: Using Existing Structures in the Health Care to Enhance Trust in AI*

As a starting point, one might ask what features exist that allow societal trust in health care to thrive despite lack of interpretability, and how we might apply similar features in the context of health care AI systems. Although trust in health care exists along many different dimensions, a few aspects stand out that could help in establishing regulatory pathways or an architecture of trust in health care AI systems. This section will look at the following elements of trust: (1) provider competence, (2) patient interest, and (3) information integrity. As a starting point, however, the authors would like to point out the importance of linguistic framing. Language such as "black box" or even "artificial intelligence" itself can evoke frightening images of technology run amok. In a similar vein, we use the frightening terms such as "dark pools" to describe alternative trading venues for trading stocks and other investment contracts. Perhaps society would be better served by less evocative language. This is really about computer software that can help doctors be better doctors. These computers are diagnostic tools—albeit highly sophisticated and dynamic ones. Our desire to romanticize may actually be helping create barriers to trust as well.

Provider competence. In medicine, patients widely trust and expect that their health care providers have a high degree of competence.[58] Given that most patients may not be able to directly assess the competence of their providers,[59] trust must be established through other mechanisms. For example, the medical profession is highly selective. People who want to become doctors must have outstanding academic achievement, take and pass difficult entrance examinations, and demonstrate interpersonal competence. Medical training itself is highly rigorous. Even after training, providers are expected by the profession to maintain high licensing standards and are subjected to regular certification testing throughout their careers in order to maintain the ability to practice. In light of these mechanisms, doctors have been extraordinarily successful, compared to other professions, in establishing a deep level of trust in their abilities and competence.

Just as it often is difficult for patients to determine the actual competence of a physician, the actual competence of a health care AI system will be difficult for patients to measure as well. Although we will discuss in a moment the potential for creating pathways similar to those that have allowed patients to develop trust in their doctors, there may be a simpler and easier route to take right off the bat. Specifically, can society count on or utilize the trust that already exists in physician decision making? In other words, will physicians themselves be able to sprinkle the fairy dust necessary to establish a patient's comfort with an AI system?

---

58. *See* Mechanic, *supra* note 23, at 664; Hall et al., supra note 20, at 621-22.

59. Hall et al., *supra* note 20, at 62.

Throughout medicine's history, doctors have been presented with new technologies and therapies to enhance and augment their ability to treat their patients. Patients expect their doctors to be competent in the use of new scientific discoveries and technologies. For example, we rarely question a health care provider's use of an MRI, especially given that a provider has access to specialists in the relevant area of technology (e.g., radiology). Perhaps the same could be true of AI technologies prescribed or recommended by one's physician. Of course, society would want to ensure that members of the physician's team have the proper expertise to evaluate the value of the technology and safely implement it. Similar to the certifications required for airplane pilots or medical professionals, one would want to ensure that those who prescribe and operate AI systems are worthy of the trust that will be conveyed. Nevertheless, the pathway of "physician as conveyor of trust" may be useful for establishing comfort first with health care AI systems, and later with the use of AI in other arenas.

In cases in which the AI system itself is effectively acting as a provider (e.g., making autonomous and independent decisions about diagnosis), society would do well to find some other proxy or means for establishing and conveying trust. Proxy organizations, either public or private, play the role of conveying trust and ensuring that such trust is deserved. For example, numerous standard-setting bodies exist covering technological issues from electricity to telecommunications. If public, the proxy organization could follow the model of a regulatory body such as the FDA. If private, there could be a body as such the Financial Accounting Standards Board (FASB) that sets the standards for the accounting industry, or the Financial Industry Regulatory Authority (FINRA), that sets the standards for broker-dealers. In effect, there could be a proxy organization that is an industry standards-setting body for AI developers or machine learning researchers developing products in the health care space.

Standards or certifications to disclose certain features of algorithms or qualifications for operators can help reassure patients and other stakeholders that the AI system itself is competent in performing its intended functions (and avoiding unintended negative consequences like harmful biases). Requiring standards for representative data sets may, for example, be an approach to addressing bias in data. This could circumvent the need to audit an AI system for explainability.

Standardized codes of ethics and conduct for developers (similar to a Hippocratic Oath for developers) could provide strong signals of ethical competence in regard to patients and their health. With either a public or private entity, however, society would need to grapple with the potential for industry capture—a common concern with both public and private regulatory bodies—and the risk that lost trust is difficult to regain.

Protecting the patient's interest. Ethical standards in medicine have existed since ancient times (e.g., the Hippocratic Oath), and they remain a foundational element in establishing trust in the medical profession today. In a

recent survey of public attitudes toward medicine, more than two-thirds of the public (sixty-nine percent) rated the honesty and ethical standards of physicians as "very high" or "high."[60] A key component of this trust in medicine is the belief that medical professionals will put the bests interest of the patient first.[61] A potential threat to this trust is the idea that medical providers would prioritize their own financial interest over the interests of patients. For example, physicians' financial relationships with pharmaceutical companies might bring into question the intent, judgment, and effectiveness of a prescribed drug treatment.

In the U.S., there have been various policy mechanisms put into place to ensure that a patient's interests are protected from perverse incentives and outside financial interests, thus reinforcing trust in the medical provider. One of the oldest and most well-known examples of such a mechanism is the federal Anti-Kickback Statute (AKS).[62] The AKS prohibits remuneration—broadly defined as anything of value including direct payment, excessive compensation for consulting services, and expensive hotel stays and dinners[63]—that would incentivize medical providers to recommend products or services which are paid for or covered by federal health care programs (e.g., Medicare, Medicaid). In addition to criminal penalties, the Office of the Inspector General for the Department of Health and Human Services can pursue additional civil penalties. Such kickbacks can lead to negative patient outcomes including overutilization, unfair competition, excessive treatment costs, and reduced agency in treatment plans for patients. By eliminating opportunities for corrupt decision making on the part of medical providers, the AKS limits these potential negative patient outcomes and helps reinforce trust by patients in their medical professionals.

Another more recent example of a policy that promotes trust is the Physician Payments Sunshine Act,[64] enacted by Congress in 2010 as part of the Affordable Care Act.[65] The policy requires that manufacturers of drugs, medical devices, and other medical supplies covered by federal health care programs collect and track financial relationships with physicians and report these data to the Centers for Medicare and Medicaid Services (CMS). As part of the Sunshine Act, CMS created the Open Payments data platform,[66] which

60. Robert J. Blendon et al., *Public Trust in Physicians – U.S. Medicine in International Perspective*, THE NEW ENG. J. OF MED. (Oct. 23, 2014), pnhp.org/news/improving-trust-in-the-profession.

61. *See* Mechanic, *supra* note 23, at 666.

62. 42 U.S.C. § 1320a-7b (2018).

63. *A Roadmap for Physicians: Fraud & Abuse Laws*, OFF. OF INSPECTOR GEN. U.S. DEP'T. OF HEALTH & HUM. SERV., https://oig.hhs.gov/compliance/physician-education/01laws.asp (last visited Apr. 14, 2019).

64. 42 C.F.R. §§ 403.900-403.914 (2019).

65. The Patient Protection and Affordable Care Act, Pub. L. No. 111-148, 124 Stat.119 (2010) (enacted).

66. *See* CTRS. FOR MEDICARE AND MEDICAID SYS. OPEN PAYMENTS DATA, https://openpaymentsdata.cms.gov (last visited Apr. 14, 2019).

makes details of physician-manufacturer financial relationships publicly available. Although effects of this policy on trust have yet to be empirically studied, patients may interpret such disclosure as a signal of honesty and thus increase trust in medical providers.[67] Moreover, it has been shown that patients believe that physicians who receive such payments are less honest and trustworthy.[68] By requiring disclosure of these types of payments, the Sunshine Act could limit outside financial incentives and influence, better aligning provider incentives with the best interest of their patients.

As a mechanism to help instill trust in the health care AI context, policymakers and regulators should work to ensure that safeguards like the AKS and Sunshine Act apply to health care AI systems. Similar to drug companies, powerful technology firms have their own strong incentives which could benefit from clearer rules and greater transparency. Furthermore, trust can be at risk if a patient perceives an AI system to be used primarily for economic efficiency at the expense of the patient's own interest or treatment effectiveness. Medical diagnostic tools using AI have the potential to increase efficiency in the health care space by reducing some of the resources and time needed by traditional diagnostic methods. But there are no guarantees that these efficiencies will actually lead to better health outcomes or be in a patient's best interests. In fact, health care systems could conceivably adopt a more efficient AI system because of cost savings while ignoring health outcomes and interests. As a response, a government body could monitor and investigate hospitals to study the relationship between cost-savings from the deployment of AI systems and any resulting changes in quality of care. When doctors engage in certain types of conflicts of interest—intentionally sacrificing patient safety for increased profit—they can face criminal actions by state licensing bodies, civil suits from patients, and the simple power of censure by their professional colleagues. Similar systems would need to be put in place for AI systems.

Information integrity. In health care, the quality of information and data being used to inform treatment is paramount to establishing patients' trust in their providers. To begin with, patients must be able to expect that their providers will ensure that the information used to guide the decision-making process is accurate. Furthermore, patients should be able to ensure that any information about their own health will be used only in ways that are appropriate. Finally, patients must have access to both sets of information in order to correct mistakes and to fully benefit from knowledge about themselves. These are tall orders that other information fields have yet to conquer. Despite the challenges, there are some models in health care that provide an initial starting point. We must emphasize, however, that the two

67. Alison Hwong & Lisa Lehmann, *Putting the Patient at the Center of the Physician Payment Sunshine Act*, HEALTH AFF. BLOG (June 13, 2012), https://www.healthaffairs.org/do/10.1377/hblog20120613.020227/full.

68. Alison R. Hwong et al., *The Effects of Public Disclosure of Industry Payments to Physicians on Patient Trust: A Randomized Experiment*, 32 J. GEN. INTERNAL MED. 1186, 1188 (2017).

examples discussed here are the barest of starting points. The integrity of data, what we are calling information integrity, is a challenge that all fields—public and private, AI-related and not—will have to master in the Digital Age.

Nevertheless, the two examples worth contemplating in the current health care field are: (1) electronic source data in clinical investigations and (2) data integrity in current good manufacturing practice (CGMP) for pharmaceuticals. Both of these involve the FDA in some role.[69]

As computerization allows for more and more clinical data (such as electronic lab reports, digital media from devices, and electronic diaries completed by study subjects) to be captured electronically, the FDA has taken a role in ensuring the integrity of clinical investigation data by publishing guidance for the industry on electronic source data in clinical investigations.[70] This guidance included concrete expectations for clinicians handling electronic data, including the creation of data element identifiers to facilitate examination of the audit trail of data sources and changes, as well as outlining the responsibilities of clinical investigator(s) to review and retain electronic data throughout a clinical trial. Although the guidance is not binding, this document is useful because it sets forward good industry standards and practices. In the pharmaceutical context, regulation is more concrete as a result of the Food, Drug, and Cosmetic Act (FD&C Act)[71] which requires that drugs meet baseline standards of safety and quality. Examples of these more concrete requirements in the FD&C Act include:

- requiring that "backup data are exact and complete" and "secure from alteration, inadvertent erasures, or loss" and that "output from the computer . . . be checked for accuracy".[72]
- requiring that data be "stored to prevent deterioration or loss."[73]
- requiring that production and control records be "reviewed"[74] and that laboratory records be "reviewed for accuracy, completeness, and compliance with established standards."[75]

The FDA also has provided industry guidance in this area to ensure that drug companies take concrete steps to protect the integrity of data.[76] The FDA's

69. U.S. FOOD AND DRUG ADMIN., GUIDANCE FOR INDUSTRY: ELECTRONIC SOURCE DATA IN CLINICAL INVESTIGATIONS (2013), https://www.fda.gov/downloads/drugs/guidances/ucm328691.pdf; U.S. FOOD AND DRUG ADMIN., FACTS ABOUT CURRENT GOOD MANUFACTURING PRACTICES https://www.fda.gov/drugs/developmentapprovalprocess/manufacturing/ ucm169105.htm (last updated June 25, 2018).

70. Data Integrity and Compliance with Drug CGMP: Questions and Answers, 83 Fed. Reg. 64, 132 (Dec. 13, 2018).

71. Federal Food, Drug, and Cosmetic Act, 21 U.S.C. §§ 301-399 (2018).

72. 21 CFR § 211.68

73. 21 CFR § 212.110(b)

74. 21 CFR §§ 211.22, 211.192

75. 21 CFR § 211.194(a)

76. Data Integrity and Compliance with Drug CGMP: Questions and Answers, 83 Fed. Reg. 64,132.

recommendations cover a variety of topics including data workflows within a CGMP computer system, how audit trails for data should be reviewed, and the restriction of access to CGMP computer systems within a company to only necessary operators.

In the AI context, researchers have proposed standards to document a dataset's purpose, its intended use, potential misuse, and areas of ethical or legal concern.[77] In the same way it is standard in the electronics industry to accompany components with a datasheet containing important information, a AI dataset could be accompanied with similar types of standards such as those listed above. Documentation in this vein could be either codified in regulation or created as an industry standard.

Ensuring the integrity of information is especially important in the context of health care AI systems. If AI can be seen as a furnace of innovation, data is its primary fuel. Machine learning systems thrive on data, and their effectiveness is determined both by the sheer amount of data received, as well as the data's quality. In 2017, the U.S. Department of Health and Human Services (HHS), with support from the Robert Wood Johnson Foundation, asked JASON—an independent group of elite scientists who advise the United States government on matters of science and technology, mostly of a sensitive nature—to consider how AI will shape the future of public health, community health, and health care delivery. JASON's report emphasized that in addition to issues around the completeness and interoperability of the data in health care systems, the inherent quality of the data is also important. For example, health data collected for the purpose of research studies would presumably be of higher quality than data collected as from a fitness bracelet or cell phone. As to this point, the JASON report suggests, "if EHR data are to be used to support AI applications, understanding this quality, and how AI algorithms react given the quality issues will be important. To date, very little research has looked at this issue."[78]

We advocate generally for more open and frank discussions in this area between computer science, health care, and policy experts on the industry and regulatory steps which can and should be taken to ensure the integrity of the data used by health care AI systems. These discussions should include whether common data formats and a single open data standard need to be created. AI systems don't necessarily need to be interpretable and transparent, but if their data is, society may see resulting increases in trust. Furthermore, consultation from members of the computer science and health care communities, can help

77. Timnit Gebru et al., *Datasheets for Datasets*, PROCS. OF THE 5TH WORKSHOP ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACHINE LEARNING, PROCS. OF MACHINE LEARNING 80 (2018), https://arxiv.org/pdf/1803.09010.pdf.

78. JASON, *Artificial Intelligence for Health and Health Care*, THE MITRE CORPORATION 43 (2017), https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf.

establish what constitutes information with high or low integrity, an essential starting point for any approach to the thorny issue of information integrity.

CONCLUSION

Significant challenges exist in establishing trust in AI systems in the health care context—a primary one being the black box nature of AI. Moving toward more open AI systems that can more easily be interpreted and understood would be a valuable endeavor that can ultimately help establish such trust. However, it is unclear whether or when we might reach the degree of clarity that could help create such trust. In light of this reality, we need to focus on alternatives to clarity for enhancing safety and trust in AI systems. This is not to say that our approach would ignore the technical and policy challenges in this sphere, but merely that we don't have reason to halt innovation because it is possible to develop trust through other pathways.

Although counter-intuitive, health care may be the ideal place to establish trust in AI systems. Patients already take a leap of faith in trusting medicines and procedures that they do not fully understand, nor do some of their health care practitioners. As noted, creating private or public regulatory bodies that can ensure the competence of both technical systems and their users, requiring transparency for the motives behind a systems' usage, and creating standards that holds data up to appropriate quality standards have instilled trust in health care broadly. These same principles can be used to instill trust in health care AI. If we can harness those pathways of trust, such as allowing existing expert mediators to serve as conveyors of trust and providing public or private mechanisms for ensuring that systems are reliable and deserving of society's trust, we could take the first steps towards establishing trust in AI and fully integrating AI systems into society.

As with any medicine or medical technology, there is a tension between safety and innovation. Health care AI brings with it enormous benefits, but also risks. One cannot expect to arrive immediately at a system that perfectly mitigates this tension. Our comments have looked at how policy and regulatory bodies might be used to help build trust through established pathways already in use in the health care context, but we also recognize that further action will be required by public, private, and peer bodies. AI in Health Care, however, is not the scary monster some may think. The pathway is there for providing reliability and establishing trust. Now, we have to do the work of guiding society down that road.