CURRENT ISSUES IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR HEALTHCARE

Clint Akarmann, Zach Harned, Francis E. Lewis,
Alissa Orlando & James Rathmell

Stanford Law School | Law and Policy Lab

Current Issues in Artificial Intelligence & Machine Learning for Healthcare

The Future of Algorithms: Navigating Legal, Social, and Policy Challenges (Winter 2018-2019)

Project Leads

Zach Harned
JD/MS Candidate

James Rathmell
JD/MBA Candidate

Report Authors

Clint Akarmann
JD/MBA Candidate

Francis E. Lewis
BS/MS Candidate

Alissa Orlando
MBA Candidate

Client

Qiu Liu
Clinical Pharmacology Team Leader
Food and Drug Administration, Center for Drug Evaluation and Research

Research Team

Clint Akarmann
JD/MBA Candidate

Matt Agnew
JD/MBA Candidate

David Hoyt
JD/MBA Candidate

Ruthanne Soh
MA Candidate

Michal Totchani
JSM Candidate

Policy Practicum Instructors

Phil Malone
Professor of Law and Director, Juelsgaard Intellectual Property and Innovation Clinic

Bryan Casey
Lecturer

Acknowledgements

In January 2019, Qi Liu of the Food & Drug Administration's Center for Drug Evaluation and Research (CDER) contacted Professor Russ Altman to learn more about the cutting-edge issues that researchers in healthcare AI/ML were facing. Professor Altman put Qi in touch with Bryan Casey and Phil Malone, who agreed to have the students in their upcoming Policy Lab pull together materials for Qi's interest group within CDER. That work eventually led to a presentation to Qi's group on May 15, 2019, and this paper.

We want to first and foremost thank Qi and her team for their interest in the healthcare AI/ML. This field is rapidly and dynamically changing; researchers and innovators seek frameworks to help them address common challenges. Qi and her team's openness to learning more and working on these issues is a hopeful sign that collaboration is possible.

Bryan and Phil, along with Luci Herman, the Program Director of the Law and Policy Lab, provided valuable insights and feedback along the way. The authors also benefited from the incredible and extensive background research conducted from January to March 2019 by Clint Akarmann, Matt Agnew, David Hoyt, Ruthanne Soh, and Michal Totchani.

As will quickly become apparent to readers of this paper, many of the most important findings come from our interviews with Stanford researchers. All of our interviewees were incredibly generous with their time—not only talking to us, but answering follow-up questions and reading our case studies to check them for accuracy. We want to thank Albert Haque, Dr. Lance Downing, Sharon Zhou, Edward Chou, Janelle Tiulentino, Jia Li, Bingbin Liu, Eric Loreaux, Pranav Rajpurkar, and Dr. Matt Lungren for their valuable insights.

We are frankly inspired by the way interest that researchers, innovators, and policymakers have all demonstrated in ensuring that healthcare AI/ML succeeds in its promise to deliver better and more equitable outcomes. We hope that this paper contributes to further collaboration and discussion among all stakeholders.

About the Stanford Law School Policy Lab

Engagement in public policy is a core mission of teaching and research at Stanford Law School. The Law and Policy Lab (The Policy Lab) offers students an immersive experience in finding solutions to some of the world's most pressing issues. Under the guidance of seasoned faculty advisers, Law and Policy Lab students counsel real-world clients in an array of areas, including education, intellectual property, public enterprises in developing countries, policing and technology, and energy policy.

Policy labs address policy problems for real clients, using analytic approaches that supplement traditional legal analysis. The clients may be local, state or federal public agencies or officials, or private non-profit entities such as NGOs and foundations.

Typically, policy labs assist clients in deciding whether and how qualitative or quantitative empirical evidence can be brought to bear to better understand the nature or magnitude of their particular policy problem, and identify and assess policy options. The methods may include comparative case studies, population surveys, stakeholder interviews, experimental methods, program evaluation or big data science, and a mix of qualitative and quantitative analysis. Faculty and students may apply theoretical perspectives from cognitive and social psychology, decision theory, economics, organizational behavior, political science or other behavioral science disciplines. The resulting deliverables reflect the needs of the client with most resulting in an oral or written policy briefing for key decision-makers.

Directed by former SLS Dean Paul Brest, the Law and Policy Lab reflects the school's belief that systematic examination of societal problems, informed by rigorous data analysis, can generate solutions to society's most challenging public problems. In addition to policy analysis, students hone the communications skills needed to translate their findings into actionable measures for policy leaders and the communities they serve. The projects emphasize teamwork and collaboration, and many are interdisciplinary, giving law students the opportunity to work with faculty and colleagues from across the university with expertise in such fields as technology, environmental engineering, medicine, and international diplomacy, among others.

Executive Summary

Recent advances in artificial intelligence and machine learning (AI/ML) represent a substantial opportunity for healthcare delivery. At the same time, researchers and innovators are grappling with challenging issues in three key areas: cybersecurity and privacy, algorithmic transparency and explainability, and fairness.

It is crucial for industry, policymakers, regulators, researchers, and innovators to collaborate on approaches to these challenges. The goal of this report is to improve stakeholders' understanding of the current state-of-play across these three areas.

This report was informed by research in the Stanford Law & Policy Lab, as well as interviews with over a dozen leading healthcare AI/ML researchers. Common themes emerging from our research include:

- AI/ML researchers and innovators are aware of the ethical, social, and legal concerns with developing AI/ML systems for healthcare. They endeavor to make systems that are safe and protect user data, which integrate with practitioner workflows rather than supplanting practitioners, and which are fair.

- However, there are no one-size-fits-all solutions to the challenges we identified, and researchers face practical constraints along with uncertainty about how best to proceed with their research to protect the interests of all stakeholders.

- Novel AI/ML techniques demonstrate the vulnerabilities of systems (e.g. adversarial attacks, re-identification of data) as well as promise solutions to common challenges (e.g. privacy-preserving techniques, tools for post-hoc explanation).

- Generally, healthcare AI/ML tools have the potential to improve on the status quo, by making healthcare safer and more equitable.

In this report, we provide background on challenges and opportunities faced by cutting-edge healthcare AI/ML research, and present case studies of three healthcare AI/ML applications where these themes arise.

Table of Contents

Introduction

Recent advances in artificial intelligence and machine learning (AI/ML) are poised to radically change the healthcare delivery landscape. Accenture estimates that AI/ML could create $160 billion in annual healthcare savings by 2026.[1] Use cases range from disease prevention and patient adherence to population health management and improving clinical workflows.[2]

At the same time, researchers and innovators are grappling with a number of pressing challenges that could impede the success of healthcare AI/ML applications. Important questions have yet to be resolved, and this necessitates a collaborative approach from policymakers, regulators, researchers, innovators, and industry.

Working in the Stanford Law and Policy Lab, we researched these challenges and conducted interviews with over a dozen leading experts in the field. We found that three major areas present concerns: cybersecurity and privacy, algorithmic transparency and explainability, and fairness. Activity across these three areas is incredibly dynamic—every day, researchers are identifying new problems, while generating new solutions to old problems.

This report provides an overview of the cutting-edge research and theory in each area. We supplement this theoretical background with three case studies for the reader to understand how researchers are grappling with these issues.

Our hope is that stakeholders will use this report to get up-to-speed on cutting edge issues in this rapidly evolving and incredibly exciting field of research. From our conversations, we discovered that there is a real need for frameworks and guidance to ensure that healthcare AI/ML applications fulfill their promise to deliver better and more equitable healthcare globally. By becoming knowledgeable about the challenges researchers are facing, stakeholders can develop robust strategies and collaborative solutions to lower the barriers to innovation.

[1] Accenture, *Artificial Intelligence: Healthcare's New Nervous System* (2017), https://www.accenture.com/us-en/insight-artificial-intelligence-healthcare.
[2] Megan Zweig and Denise Tran, *The AI/ML use cases investors are betting on in healthcare*, ROCK HEALTH, https://rockhealth.com/reports/the-ai-ml-use-cases-investors-are-betting-on-in-healthcare/.

Current Issues in Cybersecurity & Privacy

Overview

State-of-the-art AI/ML techniques rely on massive quantities of data for training. In the healthcare context, this poses a clear dilemma: how do we best preserve privacy of patient data while enabling access to data for training purposes? We can think of privacy challenges as the set of issues that arise with how institutions with *authorized access* to data (e.g., a hospital that stores patient records) use and share that data, and for what purposes.

At the same time, healthcare AI/ML tools face a novel set of cybersecurity challenges. These are the issues that arise when malicious actors, who gain *unauthorized access* to patient data or AI/ML tools, attempt to use that data for their own ends, or exploit or undermine the proper functioning of the tool.

Cybersecurity

Healthcare AI/ML tools present a novel set of cybersecurity threats above-and-beyond those created by connected medical devices. First, the trove of data that these tools use to train on represents a honeypot for malicious actors. Second, researchers have developed techniques that undermine the performance of tools and can lead to adverse patient results. We primarily focus on the second set of threats below.

The tablet below presents a threat taxonomy for cybersecurity in healthcare AI/ML.[3] Healthcare AI/ML tools are uniquely susceptible to Class 2 and Class 4 threats. Two cutting-edge techniques, adversarial learning and synthetic data generation, present Class 4 threats to the integrity of AI systems, particularly in the healthcare context. Meanwhile, the reliance of AI/ML tools and the siloing of data to train them has created an attractive target for those seeking to steal data, creating the right environment for innovations in Class 2 threats.

[3] This framework was developed by David Hoyt for his final paper for this Policy Lab.

Threat Taxonomy for Healthcare AI/ML	
Class 1: Intentional Direct Physical Harm	Malicious actor takes advantage of a medical device security vulnerability and uses the device to directly cause physical harm to a patient, either by affirmatively delivering harm or by overriding the device such that care is withheld and the patient is physically harmed
Class 2: Intentional Indirect Physical Harm	Malicious actor manipulates a medical device such that the healthcare provider causes physical harm to the patient
Class 3: Threat of Force	Malicious actor holds a patient and/or healthcare provider hostage by disabling or threatening to disable a piece of critical infrastructure or medical device that is required for providing care unless the actor's demands are met
Class 4: Data Theft	Malicious actor steals patient data, provider data, and/or proprietary data from tool manufacturer
Class 5: Financial Manipulation/Crime	Malicious actor steals and sells medical records for financial gain

Adversarial Attacks

One of the hallmark recent developments in the AI field is the Generative Adversarial Network (GAN), popularized by Ian Goodfellow.[4] GANs involve two neural networks competing with each other: one generating some output (e.g. an image), while the other attempts to discern whether the output was generated or is instead authentic. Through training, each becomes more adroit at deceiving the other network, until the final product is a generated output that can fool even humans. The GAN gave rise to a larger field of adversarial learning, thereby creating a whole host of new cybersecurity threats.

Adversarial learning involves the deployment of GAN-like methods to fool or spoof certain AI systems. Instances of this abound across a number of domains, including the creation of visual images that fool

[4] Ian J. Goodfellow et al., *Generative Adversarial Nets*, ARXIV (Jun. 10, 2014), https://arxiv.org/pdf/1406.2661.pdf.

a machine learning classifier.[5] These attacks can be multimodal, including adversarial sounds that are inaudible to the human ear but perceived by a machine learning model (e.g., the hidden command to make Google browse evil dot com).[6] Adversarial techniques may involve the imperceptible alteration of an image, or merely the placing of an innocuous object (such as a sticker[7]) in the view of the machine learning classifier. These attacks have proven successful not only under white box conditions–where the adversary has access to the AI/ML model being attacked–but also under the more realistic black box conditions, where the adversary is not privy to the inner workings of the AI/ML model being attacked, but fools it nonetheless.[8]

> "Nefarious actors could seek to attack artificial intelligence systems by deliberately introducing bias into them, smuggled inside the data that helps those systems learn."
>
> DOUGLAS YEUNG, SCIENTIFIC AMERICAN

Such instances of spoofing can easily arise in the healthcare space. Recent work has detailed the ability to imperceptibly (to humans) change an image of a mole in order to fool a machine learning model into classifying the mole as cancerous rather than benign.[9] Similar techniques can also be used to manipulate text-based ML healthcare classifiers. They illustrate the need for AI machine vision systems to be not merely accurate, but also robust to perturbations and adversarial attacks. Misuse of these adversarial techniques can result in patient harm, as well as insurance fraud.

This threat is particularly challenging because it involves the use of AI to fool AI.[10] This creates an ever-evolving arms race, with each new iteration briefly outcompeting the other until adaptations are made. An illustrative instance of this involved a machine learning conference where 11 papers, detailing methods to defend against adversarial attacks, were presented. Shortly thereafter, an MIT

[5] Anh Nguyen, Jason Yosinki, and Jeff Clune, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, ARXIV (Dec. 5, 2014), https://arxiv.org/pdf/1412.1897v1.pdf.
[6] Audio Adversarial Examples, https://nicholas.carlini.com/code/audio_adversarial_examples/.
[7] Will Knight, *How to hide from the AI surveillance state with a color printout*, MIT TECH. REV. (Apr. 24, 2019), https://www.technologyreview.com/f/613409/how-to-hide-from-the-ai-surveillance-state-with-a-color-printout.
[8] Adam Conner-Simons, *Fooling Google's image-recognition AI 1000x faster* MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LAB (Dec. 20, 2017), https://www.csail.mit.edu/news/fooling-googles-image-recognition-ai-1000x-faster.
[9] Samuel G. Finlayson et al., *Adversarial attacks on medical machine learning*, 363 SCIENCE 1287, http://science.sciencemag.org/content/363/6433/1287.
[10] However it should be noted that some adversarial methods fool not only machines, but humans if their attention is time-limited. Gamaleldin F. Elsayed et al., *Adversarial Examples that Fool both Computer Vision and Time-Limited Humans*, ARXIV (May 22, 2018), https://arxiv.org/abs/1802.08195.

graduate student posted a website demonstrating means of defeating seven of the newly proposed methods.[11] However it should be noted that such methods are not all bad news. In fact, adversarial learning can be used for myriad good ends, including the promotion of fairness,[12] as well as the helpful creation of synthetic data, to which we now turn our attention.

Synthetic Data

The creation of synthetic data involves using GAN-like methods to create realistic data in a variety of modalities (e.g., images, sound, text). In order to maximize performance, AI/ML methods grow increasingly greedy when it comes to training data, and there is often a dearth of relevant healthcare training examples. Synthetic data provides a means to supplement the AI/ML model with additional training data to improve accuracy.[13]

Synthetic data can also be used to create a more diverse training set when natural limitations of such diversity are imposed (e.g., challenges in representative sampling at a hospital located in a high-socioeconomic-status region). Synthetic data can be helpful when use of authentic data creates privacy concerns. For example, making electronic health record (EHR) data open and available to machine learning researchers can be quite challenging given the privacy implications, but synthetic EHRs could be generated and shared, since there are no persons affiliated with those fictitious but accurate records whose privacy could be violated.[14]

However, the ability to create synthetic data can also be used for nefarious purposes. For example, researchers have been able to generate synthetic fingerprints that unlock a biometrically-secure device.[15] This new generation of AI-powered spoofing introduces novel problems—both definitional and practical—regarding hacking.

---

[11] Tom Simonite, *AI Has a Hallucination Problem That's Proving Tough to Fix*, WIRED (Mar. 9, 2018), https://www.wired.com/story/ai-has-a-hallucination-problem-thats-proving-tough-to-fix/.
[12] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, *Mitigating Unwanted Biases with Adversarial Learning*, AAAI (2018), http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_162.pdf.
[13] Hoo-Chang Shin et al., *Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks*, ARXIV (Sep. 13, 2018), https://arxiv.org/pdf/1807.10225.pdf.
[14] Scott H. Lee, *Natural language generation for electronic health records*, 1 NPJ DIGITAL MEDICINE 63 (2018), https://www.nature.com/articles/s41746-018-0070-0.
[15] Philip Bontrager, Aditi Roy, and Julian Togelius, *DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Latent Variable Evolution*, ARXIV (Oct. 18, 2018), https://arxiv.org/pdf/1705.07386.pdf.

The New "Hacking"

The Computer Fraud and Abuse Act (CFAA) and its state-level equivalents are the primary anti-hacking statutes. AI/ML, adversarial learning, and synthetic data introduce a challenging definitional issue regarding what exactly constitutes hacking. These methods blur the lines between the more traditional forms of cyberintrusion and modern adversarial attacks. This has prompted legal scholars to wonder whether tricking a robot or machine learning model would be considered hacking under CFAA, and noting that there is no clear answer on this front.[16]

Healthcare AI/ML systems require valuable health data, which is especially sensitive and hard-to-obtain relative to other kinds of data. The agglomeration of such data creates a veritable honeypot for malicious actors. This threat, and the means for protecting against it, will be discussed in further detail in the subsequent Privacy subsection. But this emphasis on data introduces yet another novel cybersecurity threat, that of data poisoning, which involves the subtle manipulation or poisoning of a dataset used for training AI/ML models, such that training a model on such data results in biased performance.[17]

Privacy

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) regulates the use and disclosure of healthcare information by covered entities. However, researchers have recently demonstrated that the anonymization techniques enshrined in HIPAA and similar regulations are not always sufficient to protect patients from being individually identified. Likewise, computer vision techniques—where a AI/ML tool is trained on images—are problematic for privacy in a way that techniques trained on other structured data are not. In light of this, new privacy-preserving techniques are being advanced.

AI and "Anonymization"

A common practice to combat privacy-related concerns is to anonymize healthcare data by removing personally-identifiable information. However, this method provides false comfort: researchers at the University of California Berkeley were recently able to identify individuals by correlating data from

[16] Ryan Calo et al., *Is Tricking a Robot Hacking?*, U. WASH. SCHOOL OF L. (Mar. 27, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3150530.
[17] Douglas Yeung, *When AI Misjudgment Is Not an Accident*, SCI. AM. (Oct. 19, 2018), https://blogs.scientificamerican.com/observations/when-ai-misjudgment-is-not-an-accident/.

activity trackers against demographic information.[18] The researchers concluded that the HIPAA privacy standards may be insufficient to properly protect patients from being individually identified.[19, 20]

For a researcher who takes precautions to scrub data of identifying information, the ability to easily re-identify data using powerful AI/ML tools is inherently problematic. On the one hand, this threat could make institutions less willing to share the data necessary to train a system, if the covered entity worries that patient data could easily be re-identified with limited demographic information. On the other hand, research may be stymied if researchers are unable to publish the dataset on which they trained a given system, since it will prevent future researchers from exploring and validating results, or improving upon the original system.

Privacy in Machine Vision

Many of the most promising applications of healthcare AI/ML are those that train on unstructured image data. In fact, all three of the case studies presented in this report fall into this category. Yet images are uniquely challenging from a privacy perspective: unlike structured data, which can theoretically be scrubbed of identifying information (even if that information can be later re-identified), images of patients and other individuals in the healthcare system are often identifiable. In the Stanford Smart Hospital case study, we discuss how researchers are coping with this tension.

Privacy-Preserving Techniques

Given the sensitivity of healthcare data and the recent rise to prominence of privacy concerns among the general public, researchers are advancing several techniques to enable AI/ML systems to train on patient data without infringing on patient privacy. Some of the most promising techniques are federated learning, split learning, homomorphic encryption, differential privacy, and visual privacy.

---

[18] Liangyuan Na et al., *Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning*, 1 JAMA NETWORK OPEN e186040 (Dec. 21, 2018), https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2719130.

[19] John Hickey, *Advancement of artificial intelligence opens health data privacy to attack*, BERKELEY NEWS (Dec. 21, 2018), https://news.berkeley.edu/2018/12/21/advancement-of-artificial-intelligence-opens-health-data-privacy-to-attack/.

[20] I. Glenn Cohen and Michelle M. Mello, *HIPAA and Protecting Health Information in the 21st Century*, 320 J. AM. MED. ASS'N. 231, https://jamanetwork.com/journals/jama/fullarticle/2682916.

- **Federated learning** was developed by Google researchers in 2017.[21] This technique entails a device or system downloading an AI/ML model that has already been trained on other data, then refining the model on locally-stored data before uploading the results in summary form to improve the original model. While promising, federated learning is challenging in healthcare applications, because health records are not standardized.

- **Split learning** offers an alternative to federated learning models, in which several or many models are partially trained on local datasets, and are then uploaded, where a master model finishes the training.[22] Unlike in federated learning, split learning does not require the system to share information about the master model with local devices

- **Homomorphic encryption** leverages cryptographic techniques to allow a system to train on encrypted data without first decrypting it.[23] Unfortunately, it is computationally burdensome and not commonly used in practical development.

- **Differential privacy** involves adding noise to data sets in order to ensure statistical privacy.[24] This active degradation of data preserves the *overall* characteristics of a population while making it very difficult to identify a single entry within the broader dataset, ensuring individual privacy. However there is a privacy budget, restricting the amount of queries that can be made of the dataset.

- **Visual privacy** involves the transformation of images, such as the blurring of faces or other major degradations, to remove sensitive information.[25] Though visual privacy is not specific to AI/ML, researchers are currently exploring ways to use AI/ML to more intelligently degrade input images. One advantage of these techniques over other methods is that they can degrade images without the risk of reversal. If camera hardware is designed to only take images at a low-resolution, then it is impossible for a malicious actor to retrieve a higher-resolution

---

[21] Brendan McMahan and Daniel Ramage, *Federated Learning: Collaborative Machine Learning without Centralized Training Data*, GOOGLE AI BLOG (Apr. 6, 2017), https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.
[22] Praneeth Vepakomma, et al., *No Peek: A Survey of private distributed deep learning*, ARXIV (Dec. 8, 2018), https://arxiv.org/pdf/1812.03288.pdf.
[23] Andy Greenberg, *Hacker Lexicon: What is Homomorphic Encryption?*, WIRED (Nov. 3, 2014), https://www.wired.com/2014/11/hacker-lexicon-homomorphic-encryption/.
[24] Matthew Green, *What is Differential Privacy*, CRYPTOGRAPHIC ENG. (Jun. 15, 2016), https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/.
[25] Interview with Edward Chou, Masters Student in Computer Science, & Yanan Sui, Postdoc, Partnership in AI-Assisted Care, in Stanford, Cal. (Feb. 21, 2019).

version of that image. By contrast, encryption models are both computationally and time expensive, and they may provide a level of security encryption in excess of what is required. As a result, visual privacy systems are likely to continue gaining traction in the healthcare setting due to the unstructured image data being captured.

Key Takeaways: Cybersecurity & Privacy

- Healthcare AI/ML tools are susceptible to novel attacks, including adversarial learning and generation of synthetic data, that can be deployed by malicious actors.
- These adversarial attacks are problematic, and there is currently no formal approach that guarantees protection, at least for the long-term.
- Data is a precious resource, and hence likely to be subject to attack. Maintaining good cybersecurity hygiene, proper database maintenance, and imposition of access controls will be crucial for defense against data poisoning attacks.
- AI/ML techniques are making it increasingly hard to have truly anonymized data, as re-identification is becoming increasingly easy to do.
- Computer vision, on which many of the most promising diagnostic and other healthcare AI/ML tools rely, poses unique and hard-to-solve-for privacy concerns, since visual data is necessarily intrusive to some degree.
- Researchers have developed several novel and high-potential privacy-preserving AI/ML techniques, such as split neural networks.
- Guidance or frameworks for approaching these new challenges could be essential for promoting innovation and research.

Case Study in Cybersecurity and Privacy: Stanford Smart Hospital Project

The Stanford School of Medicine and Computer Science department have collaborated to form the Partnership in AI-Assisted Care (PAC), which leverages computer vision and ML technologies to solve pressing problems in healthcare.[26] The partnership is focused on four core projects: (1) Hand Hygiene, (2) Senior Well-Being, (3) ICU Clinical Pathway Support, and (4) Surgical Support.[27]

We interviewed several of the researchers involved with the project, including Albert Haque, Janelle Tiulentino, Edward Chou, Yanan Sui, and Bingbin Liu. They gave us valuable insights into how cybersecurity and privacy concerns can impact the development and regulation of AI/ML healthcare tools.

The researchers we interviewed were largely focused on the Hand Hygiene project, which is PAC's first foray in the hospital setting and is intended to serve as a "foundation for future research initiatives."[28] The project uses a computer vision system to track missed hand hygiene events by physicians, nurses, and other hospital staff and intervene in real-time, with the goal of reducing the rate of hospital-acquired infections.[29] The system is currently being tested in two hospitals: Intermountain Healthcare in Utah and Lucile Packard Children's Hospital (LPCH) at Stanford.[30]

Privacy Challenges in the Development Phase

The Hand Hygiene system is not intended to supplant tasks performed by physicians or nurses. Rather, it functions as a support tool, alerting hospital staff when they have forgotten to wash their hands. As a result, the system does not face the same level of regulatory scrutiny that is typically applied to AI/ML diagnostic applications, whose outputs have a direct impact on the care that a patient receives. For one, the stakes are considerably different—failing to spot melanoma or misdiagnosing a chest X-ray is different from mistakenly reminding a physician to wash their hands. Furthermore, a world in which the hand hygiene system only spots a fraction of missed hand hygiene events is an improvement from the base rate where *all* such events go unnoticed.

---

[26] Stanford Partnership in AI-Assisted Care, https://aicare.stanford.edu/index.php.
[27] Projects, Stanford Partnership in AI-Assisted Care, https://aicare.stanford.edu/projects/index.php.
[28] Interview with Janelle Tiulentino, Software Engineer, Partnership in AI-Assisted Care, in Stanford, Cal. (Feb. 20, 2019).
[29] Intelligent Hand Hygiene Support, Stanford Partnership in AI-Assisted Care, https://aicare.stanford.edu/projects/hand_hygiene/.
[30] Intelligent Hand Hygiene Support, Stanford Partnership in AI-Assisted Care, https://aicare.stanford.edu/projects/hand_hygiene/.

Though the hand hygiene system's goals may have been uncontroversial, the system's underlying machine vision technology meant that privacy considerations would play a key role in the system's development and implementation. The system leverages machine vision technology for the "characterization of fine-grained motions such as handwashing or surgical procedures" through the use of high resolution videos."[31] Since these videos capture sensitive patient/provider interactions in the hospital setting, PAC researchers had to be careful to remain compliant with privacy regulations, such as HIPAA, in addition to addressing the privacy concerns of stakeholders like physicians, nurses, and patients.

The quality of the visual input data plays a crucial role in the training of a machine vision system. The better the quality, including resolution, of such data, the better a system is at performing a certain task, such as identifying missed hand hygiene events. As many researchers noted, absent privacy concerns, it would be ideal to train the system on full-color, high-resolution images.[32] Instead, "privacy regulations such as HIPAA and GDPR" limited the visual data that researchers could leverage when developing the system.[33] HIPAA restricts the use of cameras in healthcare settings, especially when patients can be identified, and hospitals have been fined for recording patients without their consent.[34] Furthermore, access to recordings is strictly controlled, a significant hurdle given that these recordings are needed to train the system.

The development of the Hand Hygiene system thus hinged on a delicate tradeoff: researchers sought to ensure the system was "non-intrusive and privacy-safe" without sacrificing too much of its accuracy.[35] The solution came in the form of deidentified depth images, which allow humans to "understand the semantics of [a] scene" even though the images are lacking in color.[36] By using depth

[31] Albert Haque et al., *Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance*, ARXIV (Apr. 24, 2018), https://arxiv.org/pdf/1708.00163.pdf.
[32] Interview with Edward Chou & Yanan Sui.
[33] Albert Haque et al., *Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance*, ARXIV (Apr. 24, 2018), https://arxiv.org/pdf/1708.00163.pdf.
[34] The Fairview Southdale Hospital in Minnesota was fined by the U.S. Centers for Medicare and Medicaid Services (CMS) for taping patients without their consent during psychiatric evaluations in the emergency department. Jeremy Olson, *Fairview Southdale cited for secret videotaping*, THE STAR TRIBUNE (Sept. 13, 2018), http://www.startribune.com/fairview-southdale-cited-for-secret-videotaping-during-psychiatric-evaluations/493187271/.
[35] Albert Haque et al., *Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance*, ARXIV (Apr. 24, 2018), https://arxiv.org/pdf/1708.00163.pdf.
[36] Albert Haque et al., *Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance*, ARXIV (Apr. 24, 2018), https://arxiv.org/pdf/1708.00163.pdf.

images, the researchers were able to develop a functioning privacy-preserving system that shielded the identities of patients and providers."[37]

Depth images, however, are not a panacea. While they certainly improve patient/provider privacy, they can also be circumvented. For example, if there is a fixed number of individuals being tracked by the sensor, individuals may be identified through other characteristics, such as their gait.[38]

The use of depth images also risks hindering the performance of the hand hygiene system.[39] Researchers acknowledged that high-resolution RGB images were "much better" for training the system as compared to depth images.[40] This mirrors the tradeoff that exists between performance and explainability—adopting more explainable systems may result in performance losses that far outweigh the social utility of individual explanations.[41] Rather than capturing high-resolution video and then later masking the footage on the back-end to ensure privacy, the system is instead designed to only capture low-resolution video.[42] As a result, the system is only able to track more pronounced movements, an example of the inherent tradeoff between privacy and performance.[43]

Depth images may also limit potential follow-on applications to existing technologies. One researcher described how it would be impossible, for example, to track which medicines a patient was taking without using high-resolution color images as inputs.[44] Given the undesirability of this current paradigm, researchers are actively exploring various privacy-preserving techniques (described in the following section) that ideally will permit the system to capture and be trained on higher-resolution images while also protecting the privacy of patients and providers.[45]

---

[37] Interview with Janelle Tiulentino.
[38] Interview with Albert Haque, Doctoral Student in Computer Science, Partnership in AI-Assisted Care, in Stanford, Cal. (Jan. 28, 2019).
[39] Interview with Albert Haque.
[40] Interview with Edward Chou & Yanan Sui.
[41] Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, BERKELEY TECH. L. J. (forthcoming 2019).
[42] Interview with Bingbin Liu, Masters Student in Computer Science, Partnership in AI-Assisted Care, in Stanford, Cal. (Apr. 4, 2019).
[43] Interview with Bingbin Liu.
[44] Interview with Albert Haque.
[45] Interview with Edward Chou & Yanan Sui.

Privacy-Preserving Techniques Used by the Researchers

The Privacy Group within PAC is exploring various privacy-preserving techniques to enable healthcare AI/ML tools to be deployed with privacy safeguards.[46] Researchers Edward Chou and Yanan Sui described how these techniques can be grouped into three categories: (1) homomorphic encryption, (2) differential privacy, and (3) visual privacy.

Chou and Sui discovered that homomorphic encryption was too computationally burdensome to be of practical use to them. They also noted that, even though differential privacy provides theoretical guarantees, it often "removes too much information from the data" in order to be useful for researchers. Differential privacy is characterized by a tradeoff between accuracy and privacy - the more that is "asked" of a database, the more noise needs to be included in order to prevent privacy leakage.[47] Furthermore, researchers determine how much privacy can be leaked by setting a "privacy budget." When this budget is set too low in an attempt to ensure privacy, the system may be filled with so much noise that it is no longer practical, or even safe, to use. This makes visual privacy one of the most promising avenue for researchers, which explains it use in computer vision applications such as the Hand Hygiene system.

One broader concern that researchers surfaced was that of regulatory uncertainty. Researchers described a general lack of guidance as to the reach of privacy regulations such as HIPAA. Take the case of an encryption-based system which reduces a high-resolution RGB video into a string of decimals: it is not clear whether researchers can subsequently share that decimal string without running afoul of existing privacy regulations. Without further guidance, researchers are left in the dark as how best to proceed.

Pushback During Implementation Phase

PAC researchers not only had to navigate privacy considerations in the development phase, but they also faced considerable privacy-related pushback from stakeholders during the implementation phase. One researcher described how "clinicians really did not like having the sensor aimed at them" and how senior citizens also disliked having sensors in their rooms at senior care facilities.[48] Some nurses at Stanford hospital refused to enter rooms that were equipped with a sensor, and the nurses'

[46] Except where otherwise noted, all information comes from Interviews with Albert Haque (Jan. 28, 2019) and Edward Chou & Yanan Sui (Feb. 21, 2019).
[47] Matthew Green, *What is Differential Privacy*, CRYPTOGRAPHIC ENG. (June 15, 2016), https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/.
[48] All information comes from Interviews with Albert Haque (Jan. 28, 2019) and Janelle Tiulentino (Feb. 20, 2019).

union protested the rollout of the system, worried that it would primarily serve as a tool for hospital management to monitor staff.

One researcher expressed uncertainty as to whether enough thought had been given to the design of the sensor: "In my opinion, I don't know if we've spent enough time on the design of [the sensor] because . . . maybe 85% of the time people are not comfortable with this idea of being surveyed." They mentioned that a less obtrusive design would help adoption. At least some of the pushback, they believed, stemmed from the fact that the system did not clearly indicate *what* data it was capturing and *how* it was using this data. Patients may have been more receptive to the system had they known that it was only capturing low-resolution depth images that could not later be used to identify the patients.

Collaboration Agreements

PAC's collaborations with outside hospitals have raised questions and concerns related to data sharing and privacy. Crafting and signing a collaboration agreement that satisfies both PAC and the hospital can be challenging, as exemplified by the question of which party retains ownership and use rights of data obtained through the partnership. Both parties have a vested interest in how the data is used and stored. Researchers want to leverage data from multiple institutions in order to build more robust systems. Hospitals, on the other hand, remain wary of allowing sensitive patient data to leave their control.

Data silos are a particular concern for PAC researchers. Collaboration agreements commonly specify that hospitals retain complete control over data captured by sensors installed at that hospital.[49] Researchers cannot download the data onto their own servers and must instead process the data on the hospital's servers. According to one researcher, this can make it "very difficult for [researchers] to label the data which is necessary for supervised training." While the researchers did not discuss federated learning or split learning techniques, this could be a promising avenue to deal with data sharing and access.

Cybersecurity

The sensor hardware of the hand hygiene system poses several challenges on the cybersecurity front, raising concerns that malicious actors could gain access to the sensitive video data captured by the system. A researcher described how the system's sensor can be tweaked remotely in order to lower the granularity of the depth and thermal modalities.[50] According to her, there is a credible risk that

[49] All information comes from Interview with Bingbin Liu (Apr. 4, 2019)
[50] All information comes from Interviews with Albert Haque (Jan. 28, 2019) and Janelle Tiulentino (Feb. 20, 2019).

hackers could gain control of the software that is used to adjust these sensor settings and, thus, have access to the entire system.

The procurement of sensor hardware has also emerged as a critical point of concern for PAC researchers. One researcher described how "there are no sensors on the mass market that fully meet [PAC's] desired specifications." In the past, researchers have resorted to creating their own sensors using off-the-shelf components. While potentially adequate for small-scale, proof-of-concept use cases, having researchers develop their own hardware is not a viable workaround in the case of the Hand Hygiene system - the video data captured by the system is highly sensitive, and the researchers' lack of cybersecurity expertise means that there can be no assurances that proper safeguards are in place to prevent a malicious attack.

As a result, PAC, in conjunction with Stanford Health Care, has solicited proposals from external vendors to develop a sensor that meets its specifications for the hand hygiene project. This process has further elevated cybersecurity considerations associated with the system hardware. Sensors supplied by U.S. vendors were "significantly" more expensive than those that could be obtained overseas, especially from China. A Chinese vendor, for example, can custom-build a sensor that fully meets PAC's desired specifications at "a fraction of the price" of what it would cost in the U.S. Though cost is an important factor when choosing a vendor, researchers are mindful of the many legal and political roadblocks that may thwart a potential deal with a Chinese vendor.

When soliciting proposals, PAC was also looking for a vendor that could help it implement the sensor's data pipeline, which is meant to securely funnel sensor data over to cloud storage. "It's challenging to fully vet any vendor on their ability to accomplish this securely, especially when spec'ing a system that hasn't been widely implemented before," one researcher described. "There's only so much a vendor can show us at the RFP [request for proposal] stage." Though it remains to be seen which vendor PAC ultimately chooses to supply its sensors, cybersecurity considerations will have played a major role in the PAC's decision.

Privacy-protection techniques are also susceptible to malicious actors. Though deidentified depth images help shield the identities of patients and providers, adversaries may still find a way to uncover sensitive information. "Even with these depth sensors," one researcher explained, "there's so many things unrelated to [image] color [such as gait pattern, walking speed, hip sway, shoulders, and hair style] that an adversary, if they actually wanted to identify these people, they could." Differential privacy has been touted as a potential solution, though questions remain as to whether researchers can set the right privacy budget and minimize the amount of noise that is introduced to the system while also ensuring privacy.

Finally, one researcher described how the balkanization of data is a “very scary prospect” and a potential security risk. They said that “it’s impossible to know what and who is working on” other projects ongoing at Stanford. While Stanford may have the resources to ensure that its IT systems are secure, such balkanization may lead to vulnerabilities at institutions which do not have adequate resources dedicated to cybersecurity hygiene and countermeasures.

Key Takeaways: Stanford Smart Hospital Project

- Stanford researchers have developed a machine vision system to spot missed hand hygiene events in hospitals and intervene in real-time, helping curb the spread of hospital-acquired infections.
- Due to privacy regulations such as HIPAA, the system was trained on deidentified depth images, which lack color and are lower-resolution.
- A tradeoff exists between the performance of the system and protecting patient and provider privacy, with researchers constrained as to the images they can use to train the system.
- Researchers are exploring a variety of privacy-preserving techniques, including homomorphic encryption, differential privacy, and visual privacy, to allow the system to be trained on higher-resolution images.
- Researchers also faced pushback from patients and providers during the implementation stage.

Current Issues in Algorithmic Transparency & Explainability

Overview

Decisions issued by traditional medical algorithms, such as clinical decision support systems, are typically self-evidently interpretable. Given an input data point, it is usually possible for an auditor to trace the program execution and understand the branching logic that contributes to the final outcome. In contrast, decisions from state-of-the-art AI/ML techniques can involve millions of parameters tuned for prediction. This complexity makes precise understanding of an output in response to an input difficult, leading the the characterization of such techniques as "black box".[51]

In the healthcare context, interpretability is of particular importance, given that human lives literally lie in the balance. Black-box methods can be more accurate than traditional clinical support tools for some tasks, even surpassing human physician performance. How then can the benefits of these black box-methods be retained while improving understanding about how they make decisions? Fortunately there are a number of methods that have been developed that provide interpretability for these otherwise opaque models.

There are four common definitions of interpretability: *incomplete objectives*, *model transparency*, *interpretable evaluation*, and *post-hoc explanation*.[52] Post-hoc explanation, i.e. explanation for why a model arrived at a result after the fact, is most similar to how clinicians are able to explain their own decisions. It is most widely applicable across black-box models, and so will be our focus for the section. We then discuss four classes of methods for post-hoc explainability in wide use by healthcare AI/ML researchers today, namely *input space analysis*, *feature activation analysis*, *ablation studies*, and *additional model output*. We then close with a brief discussion of where interpretability fits within a larger cost-benefit framework, noting that the importance of interpretability is proportional to costs of mistakes, but could be balanced out by high expected benefits of healthcare AI/ML systems.

Interpretability as Post-Hoc Explanation

If we can't know the precise process by which a prediction was generated, we can instead gain confidence that a model understands the meaning of a prediction, by analyzing it after training. This is what we refer to as *post-hoc explanation* or *post-hoc interpretability*. There are a variety of methods for

[51] W. Nicholson Price, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421 (2017).
[52] Zachary C. Lipton, *The Mythos of Model Interpretability*, arXiv.org (2017), https://arxiv.org/abs/1606.03490.

doing this, several of which we examine below. Note that, as far as human decisions can be said to be interpretable, they are of this form of post-hoc explanations. We cannot peer inside the "black box" of a human brain to see how a decision is made. However, we can make inquiries after the fact, probing how well the human understood the relevant factors in making her decision. This form of explanation is most common in clinical practice. When a doctor is asked to justify a decision, she will point to the relevant features (e.g. point to an area on a scan, cite figures from the patient chart, etc.) in order to explain her decision making process. Therefore, we should arguably demand a similar standard of post-hoc explainability from healthcare AI/ML systems.

Measuring Post-Hoc Interpretability

Many methods exist for increasing the interpretability of black-box models. In this section, we group these methods by four types: Input Space Analysis, Feature Activation Patterns, Ablation Studies, and Other Outputs. We provide a non-exhaustive list of interpretability algorithms that fit within each method, and supply a brief explanation of the common features of each group.

Input Space Analysis

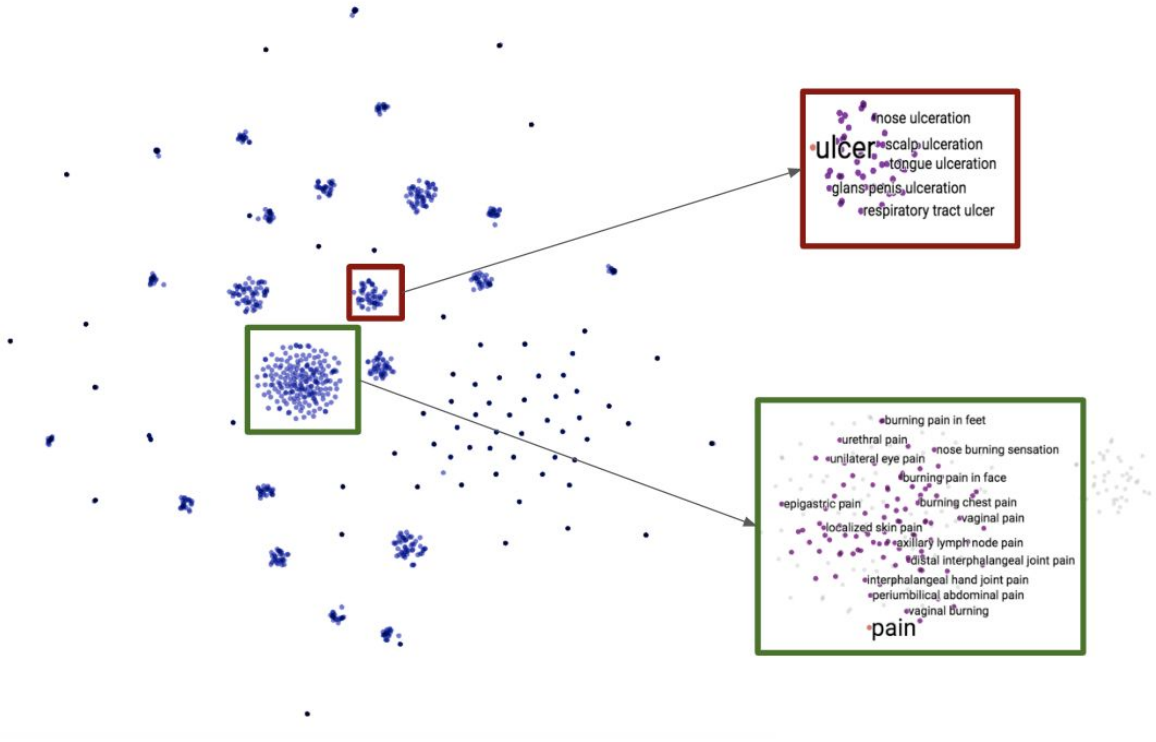
In Input Space Analysis, the aim is to examine which data points or inputs are similar to each other (viz. which data points are clustered together). If the data is clustered in a way that comports with a human perspective, that is, the data is organized in a way that makes sense on an ex post basis, then we may trust the model. For instance, if a classifier is designed to distinguish cars from trucks, when we conduct an input space analysis, it would be nice to see that the classifier has clustered certain car types together (e.g. sports cars, sedans, minivans, etc.). This would bespeak the model having a strong understanding of the task. If on the other hand it were clustering the data points into bizarre categories (e.g. vehicles with three holes in the hubcaps), we would be less confident that the model is correctly performing the task, even if the accuracy is high (since it may be overfit to the test set and thus not generalize well).

Examples of Input Space Analysis techniques include:

- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**: Developed by Laurens van der Maaten and Geoffrey Hinton.[53] Typically the data used to train machine learning has thousands of features, sometimes more. Data with two features can be directly plotted in two dimensions, but data with a thousand features can only be plotted in one thousand dimensional space, which of course we as humans cannot visualize. The process of t-SNE allows for the projection of this high-dimensional data points into a low-dimensional space,

[53] Laurens van der Maaten and Geoffrey Hinton, *Visualizing Data using t-SNE*, J. MACHINE LEARNING RESEARCH 9 (2008), http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

such as 2D or 3D so that we can plot and visualize it. The crucial part is that t-SNE performs this dimensionality reduction while still preserving the clustering properties of the high-dimensional data. This allows us to visualize whether and how the data is clustering, and analyze if we think this is intuitive, and hence whether the system is making meaningful classifications. This t-SNE method is commonly applied to neural network hidden layer representations of data to see what input data leads to similar neural representations.



2D t-SNE graph derived from Natural Language Processing of medical texts[54]

- **Explanation by Examples:** This was introduced in the medical context by Caruana et al.[55] This method is just what is sounds like, using the comparison to prior examples as a means of justifying a decision. For instance, a doctor might justify a diagnosis by noting that the presentation, scan, and EHR information are just like that of a prior patient that had the same condition. AI/ML algorithms achieve this via examining how closely the data representing certain cases is clustered together, with the idea that similar cases are clustered closely

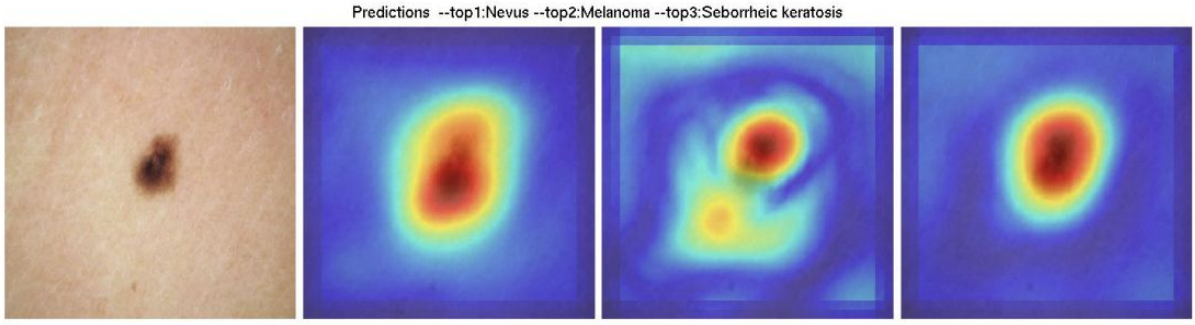
[54] Xavier Amatriain, NLP & Healthcare: Understanding the Language of Medicine, MEDIUM (Nov. 5, 2018), https://medium.com/curai-tech/nlp-healthcare-understanding-the-language-of-medicine-e9917bbf49e7.
[55] Rich Caruana et al., *Case-based explanation of non-case-based learning methods*, PROC. AMIA SYMPOSIUM (1999), https://www.ncbi.nlm.nih.gov/pubmed/10566351.

together, and hence are likely to have similar classifications. The more features that are used in the model, the less likely it is that such clustering is mere coincidence, and the more likely the cases have genuine connections between each other. Caruana et al. use this technique to explain decisions by neural networks trained to predict pneumonia and decision trees trained to predict Cesarean sections.

Feature Activation Patterns

Analysis of Feature Activation Patterns seeks to make machine learning models describable by illuminating how portions of the model respond to various inputs, in much the same way that neuroscientists look to fMRI scans of neural responses to stimuli. When it is observed that clusters of neurons respond uniformly to specific stimuli, this indicates some relation between them, and likely a specific function achieved by those neurons. Examples of Feature Activation Pattern techniques include:

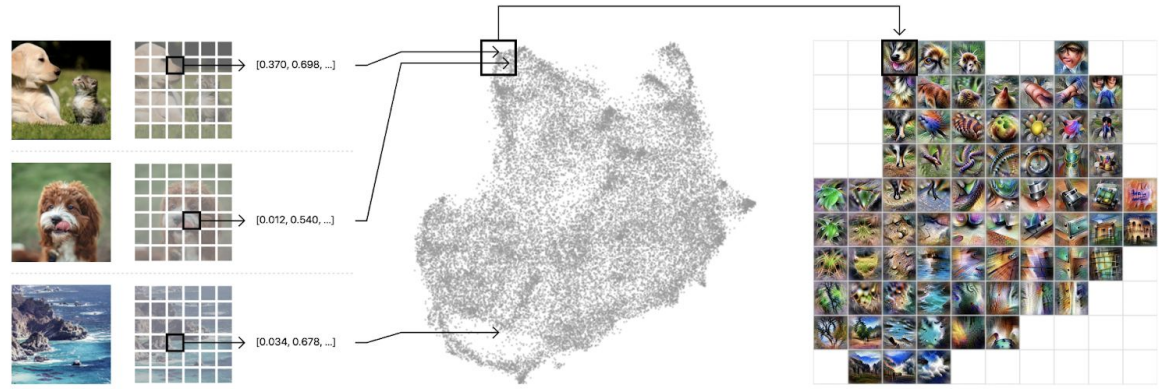
- **Class Activation Maps (CAMs):** This technique was developed by researchers at MIT.[56] Class Activation Maps take the final layers of a Convolutional Neural Network,[57] and determine which pixels of the image are most relevant for the classification decision. This produces a heat map, which gives some notion of what input image portions excite the artificial neurons the most. Use of CAMs in the medical context can be seen in the CheXNet project, which is the subject of our case study below.[58]



Three CAMs on a skin lesion image[59]

[56] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, *Learning Deep Features for Discriminative Localization*, CVPR (2016), https://arxiv.org/pdf/1512.04150.pdf
[57] Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, NIPS (2012), https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
[58] Pranav Rajpurkar, Matthew Lungren, Andrew Ng et al., *ChexNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*, ARXIV (2017), https://arxiv.org/abs/1711.05225
[59] Xi Jia and Linlin Shen, Skin Lesion Classification using Class Activation Map, ARXIV (Mar. 3, 2017), https://arxiv.org/abs/1703.01053.

- **Activation Atlases:** This technique was developed by a collaboration between OpenAI and Google Brain, [60] and it involves a combination of tools. Activation Atlases give ways to peer into the inner workings of a complex neural network trained to classify images. This allows users to see what images the neural network things are closely related to each other, hence noting that water pictures are clustered with pictures of land, as are different animal pictures, often flowing into each other in meaningful ways. This is done using the t-SNE method mentioned above, as well as using feature visualization processes.[61] These visualizations of the neural network as a whole are complex, and hence are often made to be interactive in order to facilitate exploration of the model.



Process for creating an Activation Atlas[62]

Ablation Studies

Ablation Studies analyze the importance of various factors contributing to a model decision by removing or ablating those factors and then noting any drop in model performance or confidence. If removing a portion of an input (e.g. blocking out certain portions of an image) causes model performance to degrade severely, then we've gained some understanding that the removed piece plays an important role in producing the final decision. Conversely, if removing a portion causes little to no change in output confidence, the removed piece is likely inconsequential to the model. Ablation

[60] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, Chris Olah, *Exploring Neural Networks with Activation Atlases*, DISTILL (2019), https://distill.pub/2019/activation-atlas/
[61] Matthew Zeiler and Rob Fergus, *Visualizing and Understanding Convolutional Networks*, ECCV (2014), https://arxiv.org/pdf/1311.2901.pdf
[62] Introducing Activation Atlases, OPENAI (Mar. 6, 2019), https://openai.com/blog/introducing-activation-atlases/.

can be applied to both inputs and model weights, making it a versatile interpretability technique. Examples of Ablation Study techniques include:.

- **Saliency Maps:** This technique was developed by researchers at University of Oxford.[63] Similar to CAMs, it creates a heatmap, but rather than looking to neuronal excitation, it occludes certain parts of the image to determine which pixels if changed, would have the most dramatic effect on model performance.[64] If occluding or covering up certain pixels results in significant degradation in model performance, than these pixels are important to the model's classification decision. Conducting such a procedure across the entire image allows for the creation of a heatmap.

- **Textual Input Ablation:** Not only pixel can be ablated, but any form of data. As an example from the medical context, researchers at Stanford[65] trained a neural network to predict 3-12 month mortality for a patient from Electronic Health Records. To provide interpretability of the model, they ablated or temporarily removed certain codes (e.g. diagnoses, conditions, procedures) from the EHR, and looked for the greatest degradation in performance. This allowed for the creation of a top ten list of the codes most relevant to the model's prediction.

Additional Outputs

Additional Output involves a model producing some form of information adjunct to its classification that is meant to explain that particular decision. These additional outputs can take many forms, with varying degrees of interpretability. Examples include producing visual heat maps on an input image or producing a report in natural language. If the additional output is plausibly related to the prediction, we gain confidence that the model is operating effectively. Examples of Additional Output techniques include:

- **Localization:** The localization technique takes in limited information and then attempts to make predictions to fill in the gaps of that limited information. This has been used by researchers from Google Brain[66] in the medical context to predict whether they think the existence of a cancerous tumor is likely, and if so, where.

- **Generated Text:** Another possible output a model can produce is natural language text. This

[63] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, *Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps*, ARXIV (2014), https://arxiv.org/pdf/1312.6034.pdf
[64] Matthew Zeiler and Rob Fergus, *Visualizing and Understanding Convolutional Networks*, ECCV (2014), https://arxiv.org/pdf/1311.2901.pdf
[65] Anand Avati et al., *Improving Palliative Care with Deep Learning*, BIBM (2017), https://arxiv.org/pdf/1711.06402.pdf
[66] Yun Liu et al., *Detecting Cancer Metastases on Gigapixel Pathology Images*, GOOGLE AI (2017), https://arxiv.org/pdf/1703.02442.pdf

technique was used in the deep neural network context by Andrej Karpathy at Stanford.[67] In such cases, a complex neural network for images can make a classification decision, but also produce a textual explanation of why it made that particular decision. In the medical context, researchers are exploring how to create physician-level natural language reports of scan pathologies.

When is Interpretability Less Important?

Arguments against the need for interpretability of black box algorithms typically focus on special situations of cost-benefit analysis. Elizabeth Holm argues that there are two circumstances in which the need for interpretability is less: situations with either low costs or high benefits.[68]

Low Costs

The risk of poorly understood models stems from the unknown chance of the model making a mistake when deployed. Thus, when the cost of a mistake is low, there is little risk when deploying a model, sometimes resulting in a low need to interpret the model.

High Benefits

This is certainly the more controversial of the two cases. However an argument can be made that, as possible benefits increase (or consequences of incorrect decisions become dire), accuracy rises to supremacy above interpretability. Holm notes that certain checks must still be in place (e.g. physician oversight) in order for this to be acceptable.

Interpretability is always desirable, and often it can be had without any detriment to performance. However there are circumstance in which a trade off is required to be made, and if certain criteria are met (e.g. very low costs, especially high benefits concatenated with appropriate physician oversight), these may be instances in which trading accuracy for performance is acceptable.

[67] Andrej Karpathy and Fei-Fei Li, *Deep Visual-Semantic Alignments for Generating Image Descriptions*, CVPR (2015), https://arxiv.org/pdf/1412.2306.pdf
[68] Elizabeth Holm, *In Defense of the Black Box,* Science (2019), https://science.sciencemag.org/content/364/6435/26

Key Takeaways: Algorithmic Explainability & Interpretability

- Post-hoc interpretability, explanation of why a given result was achieved, is likely the optimal and sufficient for healthcare AI/ML systems, since this is the same kind of interpretability we get from clinical decision makers today.
- Many methods exist for improving post-hoc black-box interpretability, and are in frequent use by practitioners and researchers for algorithm improvement and to facilitate user/physician involvement.
- Interpretability is not always needed, especially in scenarios where the costs of mistakes are low or the degree of success is high.

Case Study in Interpretability: Pathology Project

The Pathology Team in Stanford Professor Andrew Ng's Lab is currently developing a detection system with the potential to revamp the workflow of pathologists. The system's goals are twofold: (1) decrease physicians' task loads by helping them pinpoint the most likely areas to find a given target, such as a bacterium, on a stain; and (2) fully replace the task of finding such target, an application that is particularly relevant for populations that lack access to medical support. The second goal of fully replacing the task has not yet been implemented, though the team is currently "training an additional deep learning model" in hopes of eventually launching such a system.

> "Broadly, we try to involve the physicians as much as possible in looking at [the system's] results and seeing whether it is helpful. Ultimately, the tool is for them. We want to keep them in-the-loop of evaluating it."
>
> SHARON ZHOU

We spoke to Sharon Zhou, a PhD student in Andrew Ng's lab, to get a better sense of the researchers' thought process around interpretability when developing and implementing the system. Zhou described how the researchers strive to ensure that the end human stakeholders—the pathologists, in this case—are satisfied by the model's predictions and related metrics: "Broadly, we try to involve the physicians as much as possible in looking at [the system's] results and seeing whether it is helpful.[69] Ultimately, the tool is for them. We want to keep them in-the-loop of evaluating it." The researchers solicited physician feedback both on the model's development and its resulting outputs but also on the physicians' satisfaction with how those outputs were ultimately delivered to them along with their thoughts on the overall workflow.

Researchers relied on two key interpretability techniques in the feature activation pattern category —saliency maps and Class Activation Maps (CAMs)—to understand, broadly speaking, where the model is "looking" when it is examining cell stains. Saliency maps and CAMs are used to examine model weights, including patterns or locations that stimulate a given neuron. In this case, such techniques could shed light on how the model is placing weights on pixels within a stain.

Zhou went on to describe how her team paid special attention to "huge performance boosts but also dips" in the system's performance. For example, if researchers notice that the system's performance is cratering, they then examine whether the system is looking at an unexpected area of the stain. Paying close attention to the system's output, Sharon described, "is extremely important" because even if

[69] Interview with Sharon Zhou, Ph.D. student, Stanford Machine Learning Group, in Stanford, Cal. (Apr. 22, 2019).

the raw score between two different methods may be the same, the system "might be looking at different areas." She gave the hypothetical of one system that may be very good at detecting a given target on the edge, or surface, of cells while another may be more proficient at spotting the target at the center, or lumen, of the tissue. Though both methods may result in the same raw score, it is crucial for researchers to be able to understand *how* a score is derived.

Using interpretability techniques such as saliency maps and CAMs, researchers probe how helpful the system's outputs will be for pathologists, the end-users of the system. "We evaluate exactly how good it is . . . we get a sense that this is going to be very helpful for the pathologist if we give her the top ten rankings but nothing more than that." Thus, the researchers assess what role the system's outputs will play in the pathologist's workflow.

Integrating a model into an actual human workflow is "extremely important" and "something that people often overlook." Zhou believes that "the competitive advantage of a machine-learning startup is not going to be the model, it's not going to be the data, it's going to be the *design* of the interface . . . that is where your competitive advantage is." She mentioned that researchers often spend too much time seeking small boosts in their model's performance rather than thinking about how the model will fit in the end user's workflow.

Zhou finds that while some interpretability techniques such as confidence scores are widely used by technologists, they are often "extremely confusing" for end-users. As a result, Zhou's team is thinking of incorporating "more interpretable thresholding" into its system's design. One possibility is to group confidence scores into buckets (e.g., "Highly Likely," "Likely," "Unlikely," or "Highly Unlikely") to guide end-users as to whether or not a stain patch contains the target. Another option is to highlight and threshold CAMs in different colors to show how the model is placing weights on different features. One potential holdup, however, is that such thresholding is "highly variable for the type of application" that is being built - there is no one-size-fits-all approach to interpretability.

Key Takeaways: Pathology Project

- Stanford Professor Andrew Ng's lab is developing a pathology detection system with the potential to revamp the workflow of pathologists
- Researchers solicited physician feedback when developing the pathology system and sought to determine how physicians would use the system's outputs as part of their workflow
- Researchers relied on two key interpretability techniques —saliency maps and Class Activation Maps (CAMs)—to understand how the pathology model weighed different portions of the cell stains in reaching its outputs
- These techniques were also used to create more interpretable thresholding for pathologists, the end-users of the system. Such techniques included highlighting CAMs to show what features the system was placing more weight on

Current Issues in Algorithmic Fairness

Overview

AI/ML researchers are beginning to focus more on the task of creating algorithms that produce equitable outcomes, an area often referred to as algorithmic fairness.[70] According to IBM, even though 82% of enterprises are considering AI deployments, 60% fear bias-related liability issues and 63% lack the in-house talent to confidently manage the technology.[71] Within the medical context, 83% of medical students believe that AI will improve medicine,[72] but one common concern involves the representativeness of AI/ML researchers (*viz.* 15% female, 3% black) and how that might introduce bias into the models they create.[73] In this section, we will explore four primary issues: (1) what causes fairness issues, (2) different ways to measure fairness, (3) opportunities created by equitable AI, and (4) strategies to improve algorithmic fairness.

What Causes Fairness Issues

There are several features of machine learning that may cause bias in the healthcare setting. In this section, we will explore three common causes: sampling bias, training group size disparity, and biased or incorrect labelling. AI/ML is different from randomized control trials in that results improve over time, as the system learns. Therefore, algorithmic biases may become exacerbated over time. Bias can be introduced in the following ways:[74]

- **Sampling bias**. Training data may be selected from a non-representative population. Initial bias may compound over time: future observations confirm predictions and fewer opportunities are given to make observations that contradict predictions. For instance, when

[70] Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need?*, CHI (May 4, 2019), https://arxiv.org/pdf/1812.05239.pdf.
[71] Natasha Lomas, *IBM Launches Cloud Tool to Detect AI Bias and Explain Automated Decisions,* TECHCRUNCH (Sep. 19, 2018), https://techcrunch.com/2018/09/19/ibm-launches-cloud-tool-to-detect-ai-bias-and-explain-automated-decisions/.
[72] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen, *Machine Learning in Medicine: Addressing Ethical Challenges,* 15 PLOS MEDICINE 11 (Nov. 6, 2018), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6219763/.
[73] Sara Myers West, Meredith Whittaker, and Kate Crawford, *Discriminating Systems: Gender, Race, and Power in AI*, AI NOW INSTITUTE (Apr. 2019), https://ainowinstitute.org/discriminatingsystems.pdf.
[74] Ziyuan Zhong, *A Tutorial on Fairness in Machine Learning*, MEDIUM (Oct. 21, 2018), https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

an algorithm is trained on Apple Watch health data, it is disproportionately collecting data from upper-class white male subjects of average weight.

- **Training group size disparity**. If the training data coming from the minority group is much less than that coming from the majority group, it will perform with lower accuracy on the minority group. Unlike in randomized control trials, where the training set representativeness is ideally proportional to the representativeness we see in the total population, the algorithmic performance is dependent on the number of entries in the training data set. This means that for an algorithm to perform as accurately on different races or genders, the number of data points collected must be equal, not just proportionate, to the number of entries for the dominant group. This is especially concerning given the lower cancer screening rates for people of color. For instance, African American women are less likely than white women to have follow-up pap smears, resulting in a smaller data set for detecting cervical cancer in black women.[75]

- **Biased or incorrect labelling.** In supervised learning, data may be mislabelled either because of bias on the part of the human labelers or because of deficient label categories. An example of biased labelling could be looking at an image of a stern woman and labelling her as angry, when she may just be thinking. Another example would be labeling a photo of a non-gender-binary person and only being able to select whether that person is a man or a woman.

AI/ML practitioners are aware that improved data sets can make outcomes fairer. In one survey of ML researchers, out of the 21% of respondents whose teams had previously tried to address fairness issues found in their products, the most commonly attempted strategy (73%) was "collecting more training data."[76]

Fairness Metrics

In this section, we will explore three common fairness metrics and the challenges of using each. There are three predominant classifications for algorithmic fairness metrics/definitions:[77]

---

[75] Cynthia Arvizo and Haider Mahdi, *Disparities in Cervical Cancer in African American Women: What Primary Care Physicians Can Do*, 84 CLEVELAND J. MEDICINE 10, (Oct. 2017), https://www.mdedge.com/ccjm/article/147719/oncology/disparities-cervical-cancer-african-american-women-what-primary-care.

[76] Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need?*, CHI (May 4, 2019), https://arxiv.org/pdf/1812.05239.pdf.

[77] Sam Corbett-Davies and Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* (Sep. 11, 2018), https://5harad.com/papers/fair-ml.pdf

- **Anti-classification**, meaning that protected attributes—like race and gender—are not explicitly used by the model.

  Example: The algorithm reviewing a chest X-ray is unable to tell the race or the gender of the patient because it has been explicitly removed from the dataset.

- **Classification parity**, meaning that false positive and false negative rates are equal across groups defined by their protected attributes.

  Example: False negative rate and false positive rate of pneumonia detection in chest X-rays is the same across men and women.

- **Calibration**, meaning that output scores correspond with the ground truth at the same rate across groups, thus ensuring that similarly situated individuals are treated similarly.

  Example: An AI model's score indicating the immediacy of a surgery recommendation reflects the clinical urgency of that patient, regardless of the patient's sensitive attributes, so as to promote fairness for populations that physicians tend to deprioritize.

Originally, researchers and practitioners thought that anti-classification would result in fair results. However, machines quickly picked up on proxies for demographic variables, resulting in equally biased results. Human too are susceptible to proxy variables, such as during blinded orchestra auditions where judges could still tell if an instrumentalist was a woman based on the click of her heels.[78]

Therefore, many in the AI community switched to focusing on both classification parity and calibration. Two researchers, Jon Kleinberg and Alexandra Chouldechova, showed that except in rare cases that are essentially never instantiated in the real world, calibration and classification parity for false positive rate and false negative rate are mathematically incompatible.[79,80] One can only have two of these three fairness properties (e.g. calibration and balance in the false positive rate but not the

[78] Curt Rice, *How Blind Auditions Help Orchestras to Eliminate Gender Bias,* THE GUARDIAN (Oct. 14, 2013), https://www.theguardian.com/women-in-leadership/2013/oct/14/blind-auditions-orchestras-gender-bias.
[79] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, *Inherent trade-offs in the fair determination of risk scores*, ARXIV (Nov. 17, 2016), https://arxiv.org/abs/1609.05807.
[80] Alexandra Chouldechova, F*air Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction instruments*, 5 BIG DATA 2 (2016), https://arxiv.org/abs/1610.07524.

false negative rate, or balance in the false positive rate and false negative rate but not calibration, etc.).

Therefore, selecting the correct fairness metric is context-specific. If a false positive is less harmful than a false negative, one may choose to promote fairness across groups with calibration and balance of the false negatives. This means the false positives are not balanced well across the groups, but because a false positive is not as harmful, this unavoidable unfairness is the best it can be. Failure to balance for false negative rates may be more appropriate for conditions like Celiac's disease. In this case, going gluten-free would be a serious change in lifestyle for the patient and incur greater expenses buying gluten-free products, but a false negative would only ead to continued gastrointestinal discomfort and nothing life threatening. Failure to balance for false positive rates may be more important for diseases like malaria, which have near certain mortality rates if the disease goes untreated, but the consequences of a false positive are minor since treatment only costs $0.10 per course and does not have strong side effects.

There is also a tradeoff between optimizing accuracy for the majority and adding additional constraints to the model that will ensure sufficient performance for minority groups. While this may lower the overall accuracy of the model with regard to the test data, it may lead to improved accuracy for the population the technology is deployed on in the real world, especially if the training set is not representative of the population who use the product. Researchers are building tools that optimize algorithms for target fairness metrics without compromising on accuracy.[81]

Opportunities for New Technologies to Improve Fairness in Healthcare

In the previous sections we focused on the risks posed by the rise of AI in healthcare. But this technology also presents several opportunities to improve equitable healthcare by improving rural access, patient convenience, contextualized decision-making, and awareness of cutting-edge knowledge.

Improve Rural Access

The scarcity of primary care physicians and specialists in rural regions means that several conditions, such as diabetic retinopathy, go undiagnosed for a long period of time. More than 50% of Americans

[81] Alekh Agarwal et al., *A Reductions Approach to Fair Classification,* ICML (2018), https://arxiv.org/pdf/1803.02453.pdf.

with diabetes don't have annual eye exams, which are undertaken by ophthalmologists.[82] A 2015 study found that 24% of US counties had no ophthalmologists or optometrists.[83] Therefore, tools like FDA-approved IDx-Dr, which uses AI to screen eye images for diabetic retinopathy, are improving detection rates in under resourced areas. IDx-Dr has an 87.2 percent chance of accurately protecting a more-than-mild case of diabetic retinopathy.

Pneumonia, the leading cause of hospitalization and death worldwide, is another case where AI/ML tools can improve access to underserved populations. Chest X-rays are currently the best available method for detecting if someone is infected with pneumonia,[84] but, according to some estimates, two thirds of the global population has "inadequate access to diagnostic imaging specialists." Even if populations have access to radiology diagnostics, there is frequently a shortage of experts who can interpret the X-rays once they are taken. AI/ML tools such as CheXNet (discussed later as a case study) can play a key role in delivering medical imaging expertise to these underserved, often rural, communities.

Improved Convenience

Especially with the rise of mobile diagnostic tools and chatbots, patients can remotely access medical information at any time without a doctor's visit, and 80% of patients agree that remote consultations are more convenient and equally as reliable as in-person visits.[85]

Improved Contextualized Decision-Making

As mentioned, AI/ML diagnostic tools pull from large datasets, ideally with diverse training examples. Physicians may not always have the same luxury to learn from diverse populations. For instance, dermatologists in majority-white communities may have trouble diagnosing skin cancer for patients of color. Whereas an AI diagnostic device could pool training data from other locations where the data set is more representative.

---

[82] Emily Mullin, *The First AI Approved to Diagnose Disease is Tackling Blindness in Rural Areas*, QUARTZ (Sep. 6, 2018), https://qz.com/1371580/can-ai-deliver-on-its-promise-to-close-the-gap-between-rural-and-urban-health-care/.
[83] Diane Gibson, *The Geographic Distribution of Eye Care Providers in the United States: Implications for a National Strategy to Improve Vision Health*, 73 PREVENTATIVE MEDICINE 30 (Apr. 2015), https://www.sciencedirect.com/science/article/pii/S0091743515000109?via%3Dihub.
[84] Pranav Rajpurkar et al., *CheXNet: Radiologist- Level Pneumonia Detection of Chest X-Rays with Deep Learning*, ARXIV (Nov. 14, 2017), https://arxiv.org/abs/1711.05225.
[85] Melissa Rohman, *Virtual Visits May Improve Patient Convenience Without Compromising Quality of Care, Communication*, HEALTH IMAGING (2019), https://www.healthimaging.com/topics/practice-management/virtual-video-visits-convenience-quality-patient-care.

Improved Knowledge of Cutting-Edge Best Practices

Even where doctors are present, they can be wrong, especially if they have had less advanced training. In the US, about 20 percent of mammograms result in a false-negative, and more than 50 percent of those women who get an annual mammogram for 10 years in a row will have a false-positive result during those 10 years.[86] Algorithms can quickly absorb cutting edge research and new information, whereas human radiologists may be reliant on training or tools that are out of date, occasionally leading to inaccurate results or sub-par care plans.

Strategies to Improve Fairness

This section will explore common tools used to improve fairness, including datasheets and third-party audit tools. Datasheets or model cards,[87] are routinely used in other engineering professions, and now are being applied to data sets. Datasheets show information including the objective function set for the algorithm (e.g., maximize accurate detections vs. minimize lab costs), model type, learning method (e.g., supervised vs. unsupervised), source of data set (e.g., specific hospital network, national insurance provider), number of data entries, number of data entries by demographics, where applicable, time period data set covers, variables considered, and accuracy metrics.[88] Datasheets are common place in nearly every scientific discipline, such as mechanical engineering, and are now becoming increasingly common in computer science.

The purpose of these datasheets would be to better inform consumers of AI/ML tools, of the differences between various data sets and AI/ML models. The Food and Drug Administration (FDA) could play a role in standardizing the disclosures required by AI/ML software providers and database curators, just as FDA has standardized disclosures for pharmaceutical and medical device companies.

There is also a proliferation of third-party tools, such as independent audit firms or automated screening tools developed by players including IBM and Accenture, to screen for bias.[89] Researchers are working actively in this space as well to develop new debiasing tools and pioneering novel AI

---

[86] Hope Reese, *The Way We Use Mammograms is Seriously Flawed but AI Could Change That*, QUARTZ (Sep. 6, 2018), https://qz.com/1367216/mammograms-are-seriously-flawed-the-way-we-use-them-now-ai-could-change-that/.
[87] Karen Hao, *This is How AI Bias Really Happens - And Why It's So Hard to Fix*, MIT TECH. REV. (Feb. 4, 2019), https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.
[88] Timnit Gebru et al., *Datasheets for Datasets*, AI NOW INSTITUTE (Apr. 16, 2019), https://arxiv.org/pdf/1803.09010.pdf.
[89] Natasha Lomas, *IBM Launches Cloud Tool to Detect AI Bias and Explain Automated Decisions*, TECHCRUNCH (Sep. 19, 2019), https://techcrunch.com/2018/09/19/ibm-launches-cloud-tool-to-detect-ai-bias-and-explain-automated-decisions/.

model development.[90] For example, Jieyu Zhou implemented a constrained interface network, known as Reducing Bias Amplification (RBA), which attempts to debias word embeddings.[91] However, other leading researchers argue that these debiasing techniques merely hide rather than fully remove the bias,[92] so further research in this space is required.

[90] Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems*, CHI (2019), https://arxiv.org/pdf/1812.05239.pdf.
[91] Jieyu Zhao et al., *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-Life Constraints,* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (2017), https://arxiv.org/pdf/1707.09457.pdf.
[92] Hila Gonen and Yoav Goldberg, *Lipstick on a Pig: Debiasing MEthods Cover Up Systemic Gender Biases in Word Embeddings but do not Remove Them*, ARXIV (Mar. 9, 2019), https://arxiv.org/pdf/1903.03862.pdf.

Key Takeaways: Fairness

- Datasets can be biased by (1) sampling bias, (2) training group size disparity, and/or (3) biased or incorrect labelling.
- The three most common fairness metrics are anti-classification, classification parity, and calibration.
- AI promotes fairness by improving access to care in rural regions, increasing patient convenience, contextualizing decision making, and promoting knowledge of cutting-edge best practices.
- Standardization of AI/ML datasets with datasheets can help improve transparency and fairness.

Case Study in Fairness: CheXNet

From skin cancer classification to arrhythmia detection, healthcare AI/ML tools are increasingly outperforming medical professionals at specific tasks. One such example is the diagnostic imaging tool CheXNet.[93]

Developed by Stanford researchers from the Departments of Computer Science, Medicine, and Radiology, CheXNet is a 121-layer convolutional neural network that uses deep learning to detect pneumonia from chest X-rays, consistently exceeding the performance of practicing radiologists. The system uses chest X-ray images as inputs and then "outputs the probability of pneumonia along with a heatmap localizing the areas of the image most indicative of" infection." CheXNet was trained on the robust Chest X-ray14 dataset, a public dataset from the NIH which contains over 100,000 individually-labeled chest X-ray images. In addition to pneumonia, the system has been shown to outperform published results for thirteen other thoracic diseases, including emphysema and pneumothorax, that were also found in the dataset.[94]

The development and implementation of CheXNet raises a host of issues related to algorithmic fairness including the following: (1) Practical limitations of gathering data and training a representative model; (2) practical limitations of testing that model in a representative way; (3) deployment bias and developing an AI/ML medical tool for use in underserved populations; and (4) concerns specific to developing tools for pediatric patients.

We interviewed Pranav Rajpurkar, one of the researchers who developed CheXNet. He identified two main motivations for building an AI/ML diagnostic medical imaging tool: (1) improving healthcare delivery by addressing diagnostic error, including intervariability among radiologists and (2) increasing access to medical imaging expertise and radiology services globally, especially in underserved populations.[95]

Practical Limitations of Building and Training a Representative Model

One weekend a few months back, Rajpurkar sent a text to one of his colleagues in the lab. The NIH had announced that it was publicly releasing a large database of chest X-rays. The timing could not have

---

[93] Pranav Rajpurkar et al., *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*, (Dec. 25, 2017), https://arxiv.org/pdf/1711.05225.pdf.
[94] Emily Matchar, *Can an Algorithm Diagnose Pneumonia*, SMITHSONIAN (Nov. 28, 2017), https://www.smithsonianmag.com/innovation/algorithm-diagnose-pneumonia-180967327/.
[95] Interview with Pranav Rajpurkar, Ph.D. student in Computer Science, in Stanford, Cal. (Apr. 26, 2019).

been better, as CheXNet's development hinged on finding a large enough data set. Historically, there had been a "lack of datasets with strong radiologist-annotated ground truth and expert scores against which researchers can compare their models."[96] Come Monday, Rajpurkar and his colleagues went straight to work and began training the CheXNet on this new dataset. By Tuesday or Wednesday, the model was already "doing better than the best published results." CheXNet's potential was now apparent to the researchers.

> "This tool should not be used on children until we are validating that this works on children and it's able to pick up pneumonia."
>
> PRANAV RAJPURKAR

Though crucial in training CheXNet, the NIH dataset does not ensure that the model is free from bias. Rather, it is through such input, or training, data that bias often creeps into an AI/ML model. As described earlier, bias comes from a variety of sources, including sample size bias and incorrect labeling of data.

When looking at the NIH dataset that was used to train CheXNet, Rajpurkar and his team had much to be pleased about. The data set was robust, containing over 100,000 X-ray images, and labeled with fourteen pathologies.[97] There were certain features, however, that the researchers felt they could improve upon in order to make the training data more representative and, thus, less likely to introduce bias into CheXNet.[98]

One issue had to do with the labeling of the data set. The NIH data set of chest X-rays contained labels for fourteen pathologies, including "infiltrate." Infiltrate, however, is a term that has fallen out of use in the radiology community. Infiltrate was seen in some of the radiology reports despite the fact that "consolidation," which refers to the exact same condition, is also a label. Rajpurkar mentioned several other example of confusing labels in reports which may skew the training and accuracy of CheXNet.

Rajpurkar and his team began by looking at the labels attached to the input data, examining how useful each label was and the prevalence of the various pathologies. In order to address the inaccurate labeling found in the NIH dataset, the researchers introduced the option to label the presence of a pathology as "uncertain" rather than a binary "yes" or "no." This gave researchers the flexibility to use uncertainty labels and then impute the actual values for each of the pathologies. One

[96] CheXpert, Stanford ML Group, https://stanfordmlgroup.github.io/competitions/chexpert/.
[97] Pranav Rajpurkar et al., *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*, ARXIV (Dec. 25, 2017), https://arxiv.org/pdf/1711.05225.pdf.
[98] Interview with Pranav Rajpurkar.

such method involves training the model without any labels and then using the model to re-label the data set.

This effort to build a data set with accurate labels culminated with the release of CheXpert, a data set consisting of 224,316 chest radiographs of 65,240 patients at Stanford Hospital. It includes an automated rule-based labeler that extracts "observations from the free text radiology reports to be used as structured labels for the images."[99] There are three key advantages to CheXpert's labeler over the method employed by the NIH: (1) CheXpert's labeler does not use automatic mention extractors which were found to produce weak extractions; (2) Several rules were added to better capture negation and uncertainty; and (3) CheXpert's labeler resolves cases of uncertainty rules double matching with negation rules in reports (e.g., mistakenly labeling the phrase "cannot exclude pneumothorax" as negative when it should be instead classified as uncertain).

One particularly vexing challenge that the researchers encountered was that, at some point, the CheXNet model becomes more accurate than the "ground truth" of the training data. Unlike the testing data described in the subsequent section, the report labels in the training data come from a number of radiologists and have not been scrubbed for consistency or accuracy.

Practical Limitations of Testing the Model

CheXNet's development was not complete. Though the system had been trained on over a hundred thousand chest X-rays, it had not yet been evaluated on a separate testing set. Given that this step is crucial to determining the performance of CheXNet, researchers had to be careful that such testing was conducted in a representative manner.

However, representative testing, as Rajpurkar described, does not mean that "CheXNet should apply to any chest X-ray in the world." This is because certain pathologies, such as tuberculosis, appear in X-rays from underserved populations but were not part of the training dataset, which was composed of X-rays from U.S. patients. Rather, representative testing refers to how well CheXNet "generalizes to the Stanford population," specifically a "randomly mixed set of such patients who were not part of the training."

The CheXNet testing set was "very expensive" to develop, as the researchers cannot rely on the existing X-ray report labels which are "very noisy" and are not tagged with an uncertainty label. Instead, Rajpurkar described how the researchers had subspecialty cardiothoracic radiologists

[99] Jeremy Irvin et al., *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison* ARXIV (Jan. 21, 2019), https://arxiv.org/pdf/1901.07031.pdf.

examine every image in the test set and give their interpretation—any disagreements among the radiologists were resolved by majority vote.

Improving Access to Underserved Populations

As described earlier, CheXNet has the potential to play a key role in democratizing access to medical imaging expertise in underserved, often rural, communities. The goal, as Rajpurkar sees it, is for CheXNet to be able to identify pathologies such as tuberculosis that are common in underserved communities but that currently cannot be identified by CheXNet since such pathologies are absent in the U.S-based training set.

It may be possible to identify pathologies such as tuberculosis using some of the labels present in the existing training set. But, it is unclear whether the way the model identifies such features in the Stanford population will be the same as to the way it does so in underserved populations where tuberculosis is actually present. This sampling bias is best solved by collecting and training the system on a data set that is representative of underserved populations. This is particularly challenging, however, since the labels on X-rays from these populations may not be trustworthy given the lack of medical imaging expertise. Mistakes in diagnosis, which may be more frequent in underserved populations, can make it difficult for researchers to obtain a reliable ground truth. Having experienced radiologists in the U.S. determine this ground truth may also prove difficult since such radiologists are likely not as adept in identifying pathologies more commonly seen in underserved areas.

Rajpurkar believes that one useful intermediate step would be for CheXNet to identify populations that it has not yet seen before. The communication from CheXNet to the end-user would go something like this: “Hey this doesn’t look like a chest X-ray I’ve seen before. Don’t trust me on it.” Though this does not allow CheXNet to be rolled out in underserved regions, it does allow developers of CheXNet to better understand gaps in the system’s reach across populations.

Developing Tools for Pediatric Patients

Millions of children worldwide, many without access to medical imaging expertise, are impacted by pneumonia. Despite this pressing need, researchers did not include pediatric X-rays in the CheXNet dataset. This is because “there are a few pathologies that show up in [children] but not adults.” This raises the issue of sample size disparity—unless CheXNet is trained on a large enough number of pediatric chest X-rays, its performance will lag relative to adult chest X-rays. “This tool,” as Rajpurkar described, “should not be used on children until we are validating that this works on children and it’s able to pick up pneumonia.”

Key Takeaways: CheXNet

- CheXNet is a convolutional neural network that uses deep learning to detect thoracic diseases, such as pneumonia, from chest X-rays, consistently exceeding the performance of practicing radiologists.
- In order to make CheXNet's training data more representative, researchers scrubbed a publicly-available chest X-ray dataset from the NIH for incorrect labels and introduced the option to label the presence of a pathology as "uncertain" rather than a binary "yes" or "no."
- Researchers tested ChexNet by examining how well the system generalized to a randomly mixed set of patients who were not part of the training set.
- CheXNet has the potential to play a key role in democratizing access to medical imaging expertise in underserved, often rural, communities, though the system must first be trained on data sets that are representative of pathologies seen in such populations.
- Since certain pathologies show up in children but not adults, CheXNet will not be used in the pediatric setting until the system has been trained on a sufficiently large corpus of pediatric chest X-rays.

Conclusion

There is no doubt that advances in AI/ML create unparalleled promise for medicine and healthcare. However, and as we have discussed throughout this whitepaper, there are a number of challenges that researchers and innovators face in this field. The good news is that researchers are attuned to such concerns. Even more than so, they seek to develop frameworks for best practices.

In this paper, we have explored three of the most pressing areas: cybersecurity and privacy; interpretability and explainability; and, finally, the myriad opportunities and challenges around fairness. What we have seen throughout our research and our case studies is that the field is rapidly evolving: new threats and problems are emerging every day, as are promising solutions to old challenges.

This paper is an attempt to level-set: to get all stakeholders on the same page. It is meant to be informative, if not comprehensive. Where we had strong recommendations, we have been sure to note them. But, more often, we have noted that one-size-fits-all solutions are the exception, not the rule. Because of practical constraints and uncertainty, stakeholders must develop flexible and dynamic frameworks that can keep up with the rapidity of innovation. This is not to say that frameworks are futile. Rather, researchers have unequivocally conveyed the need for collaborative guidance for the work they are doing, to reduce uncertainty and to ensure that healthcare AI/ML systems are both safe and equitable.