AMA Journal of Ethics®

February 2019

Volume 21, Number 2: E119-197

Artificial Intelligence in Health Care

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E121-124

FROM THE EDITOR

Ethical Dimensions of Using Artificial Intelligence in Health Care

Michael J. Rigby

An artificially intelligent computer program can now diagnose skin cancer more accurately than a board-certified dermatologist.[1] Better yet, the program can do it faster and more efficiently, requiring a training data set rather than a decade of expensive and labor-intensive medical education. While it might appear that it is only a matter of time before physicians are rendered obsolete by this type of technology, a closer look at the role this technology can play in the delivery of health care is warranted to appreciate its current strengths, limitations, and ethical complexities.

Artificial intelligence (AI), which includes the fields of machine learning, natural language processing, and robotics, can be applied to almost any field in medicine,[2] and its potential contributions to biomedical research, medical education, and delivery of health care seem limitless. With its robust ability to integrate and learn from large sets of clinical data, AI can serve roles in diagnosis,[3] clinical decision making,[4] and personalized medicine.[5] For example, AI-based diagnostic algorithms applied to mammograms are assisting in the detection of breast cancer, serving as a "second opinion" for radiologists.[6] In addition, advanced virtual human avatars are capable of engaging in meaningful conversations, which has implications for the diagnosis and treatment of psychiatric disease.[7] AI applications also extend into the physical realm with robotic prostheses, physical task support systems, and mobile manipulators assisting in the delivery of telemedicine.[8]

Nonetheless, this powerful technology creates a novel set of ethical challenges that must be identified and mitigated since AI technology has tremendous capability to threaten patient preference, safety, and privacy. However, current policy and ethical guidelines for AI technology are lagging behind the progress AI has made in the health care field. While some efforts to engage in these ethical conversations have emerged,[9-11] the medical community remains ill informed of the ethical complexities that budding AI technology can introduce. Accordingly, a rich discussion awaits that would greatly benefit from physician input, as physicians will likely be interfacing with AI in their daily practice in the near future.

This theme issue of the *AMA Journal of Ethics* aims to tackle some of the ethical dilemmas that arise when AI technology is used in health care and medical education. Some of the most exigent concerns raised in this issue include addressing the added risk to patient

privacy and confidentiality, parsing out the boundaries between the physician's and machine's role in patient care, and adjusting the education of future physicians to proactively confront the imminent changes in the practice of medicine. Additionally, dialogue on these concerns will improve physician and patient understanding of the role AI can play in health care, helping stakeholders to develop a realistic sense of what AI can and cannot do. Finally, anticipating potential ethical pitfalls, identifying possible solutions, and offering policy recommendations will be of benefit to physicians adopting AI technology in their practice as well as the patients who receive their care.

One major theme to be addressed in this issue is how to balance the benefits and risks of AI technology. There is benefit to swiftly integrating AI technology into the health care system, as AI poses the opportunity to improve the efficiency of health care delivery and quality of patient care. However, there is a need to minimize ethical risks of AI implementation—which can include threats to privacy and confidentiality, informed consent, and patient autonomy—and to consider *how* AI is to be integrated in clinical practice. Stakeholders should be encouraged to be flexible in incorporating AI technology, most likely as a complementary tool and not a replacement for a physician. In their commentary on a case of implementing an artificially intelligent computer algorithm into a physician's workflow, Michael Anderson and Susan Leigh Anderson emphasize the importance of user technical expertise in interpreting AI-guided test results and identify potential ethical dilemmas. In a similar case regarding the use of IBM Watson™ as a clinical decision support tool, David D. Luxton outlines benefits, limitations, and precautions in using such a tool. Furthermore, in an empirical study, Irene Y. Chen, Peter Szolovits, and Marzyeh Ghassemi demonstrate that machine learning algorithms might not provide equally accurate predictions of outcomes across race, gender, or socioeconomic status. Finally, in responding to a case that considers the use of an artificially intelligent robot during surgery, Daniel Schiff and Jason Borenstein affirm the importance of proper informed consent and responsible use of AI technology, stressing that the potential harms related to the use of AI technology must be transparent to all involved.

A second major theme in this issue revolves around the role AI can play in medical education, both in preparing future physicians for a career integrating AI and in directly using AI technology in the education of medical students. Steven A. Wartman and C. Donald Combs contend that, given the rise of AI, medical education should be reframed from a focus on knowledge recall to a focus on training students to interact with and manage artificially intelligent machines; this reframing would also require diligent attention to the ethical and clinical complexities that arise among patients, caregivers, and machines. In a related article, C. Donald Combs and P. Ford Combs explore the use of artificially intelligent, virtual patients (VPs) in medical education. With their exciting applications in teaching medical history taking, such as in psychiatric intake evaluation, VPs offer a readily accessible platform with several benefits over traditional

standardized patients; however, the disadvantages and shortcomings are equally important, emphasizing the need for clarity about the role of VPs in medical education.

A final theme addressed in this issue elucidates the legal and health policy conflicts that arise with the use of AI in health care. Hannah R. Sullivan and Scott J. Schweikart unveil legal issues such as medical malpractice and product liability that arise with the use of "black-box" algorithms because users cannot provide a logical explanation of how the algorithm arrived at its given output. Additionally, Nicole Martinez-Martin uncovers a policy gap governing the protection of patient photographic images as they apply to facial recognition technology, which could threaten proper informed consent, reporting of incidental findings, and data security. Finally, Elliott Crigger and Christopher Khoury report on the American Medical Association's recent adoption of policy on AI in health care, which calls for the development of thoughtfully designed, high-quality, and clinically validated AI technology, which can serve as a prototypical policy for the medical system.

There is no doubt that AI will have widespread ramifications that revolutionize the practice of medicine, transforming the patient experience and physicians' daily routines. The use of AI in health care can even extend into unexpected areas such as artistic practice, as investigated by Sam Anderson-Ramos, with new dilemmas emerging from the rise of thinking machines in previously human pursuits. Additionally, Elisabeth Miller visually depicts the potential impact of AI on mechanized human bodies. Nonetheless, there is much work to do in order to lay down the proper ethical foundation for using AI technology safely and effectively in health care. This theme issue of the *AMA Journal of Ethics* intends to provide such a foundation with an in-depth view of the AI-induced complexities of black-box medicine, exploring patient privacy and autonomy, medical education, and more. Ultimately, patients will still be treated by physicians no matter how much AI changes the delivery of care, and there will always be a human element in the practice of medicine.

References

1. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.
2. Ramesh AN, Kambhampati C, Monson JRT, Drew PJ. Artificial intelligence in medicine. *Ann R Coll Surg Engl*. 2004;86(5):334-338.
3. Amato F, López A, Peña-Méndez EM, Vaňhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. *J Appl Biomed*. 2013;11(2):47-58.
4. Bennett CC, Hauser K. Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach. *Artif Intell Med*. 2013;57(1):9-19.
5. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep*. 2014;16(1):441.

6. Shiraishi J, Li Q, Appelbaum D, Doi K. Computer-aided diagnosis and artificial intelligence in clinical imaging. *Semin Nucl Med*. 2011;41(6):449-462.
7. Luxton DD. Artificial intelligence in psychological practice: current and future applications and implications. *Prof Psychol Res Pr*. 2014;45(5):332-339.
8. Riek LD. Healthcare robotics. *Commun ACM*. 2017;60(11):68-78.
9. Luxton DD. Recommendations for the ethical use and design of artificial intelligent care providers. *Artif Intell Med*. 2014;62(1):1-10.
10. Luxton DD. *Artificial Intelligence in Behavioral and Mental Health Care*. San Diego, CA: Elsevier Academic Press; 2016.
11. Peek N, Combi C, Marin R, Bellazzi R. Thirty years of artificial intelligence in medicine (AIME) conferences: a review of research themes. *Artif Intell Med*. 2015;65(1):61-73.

Michael J. Rigby is a fifth-year student in the Medical Scientist Training Program (MSTP) at the University of Wisconsin School of Medicine and Public Health in Madison. He is currently a PhD candidate in molecular neuroscience and is studying the mechanisms that underlie neurodegenerative diseases. He earned a BS in molecular and cellular biology at the University of Illinois at Urbana-Champaign and is interested in pursuing a career as a physician-scientist in neurology.

Citation
AMA J Ethics. 2019;21(2):E121-124.

DOI
10.1001/amajethics.2019.121.

Acknowledgements
I would like to thank everyone involved that turned a passing idea into this theme issue. Most importantly, I thank the authors for their time and dedication to make stimulating contributions. I also want to thank my mentor, Dr. David D. Luxton, for his guidance and support as well as the editorial staff at the *AMA Journal of Ethics*. Finally, I thank my sister and brother-in-law, Teresa and Ryan Westfall, for their constant encouragement to learn more about mathematics, computer science, and, most importantly, artificial intelligence.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

ISSN 2376-6980

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E125-130

CASE AND COMMENTARY

How Should AI Be Developed, Validated, and Implemented in Patient Care?

Michael Anderson, PhD and Susan Leigh Anderson, PhD

Abstract

Should an artificial intelligence (AI) program that appears to have a better success rate than human pathologists be used to replace or augment humans in detecting cancer cells? We argue that some concerns—the "black-box" problem (ie, the unknowability of how output is derived from input) and automation bias (overreliance on clinical decision support systems)—are not significant from a patient's perspective but that expertise in AI is required to properly evaluate test results.

Case

Dr A is a pathologist who has been working at the Community Hospital for several years. She begins her day by evaluating biopsy specimens from Ms J, a 53-year-old woman who underwent a lumpectomy with sentinel lymph node biopsy for breast cancer (a procedure to determine whether a primary malignancy has spread). The surgeon removed 4 lymph nodes that were submitted for biopsy. The lymph node samples were processed and several large (multiple gigabytes), high-resolution images were uploaded to the patient's health record for Dr A's evaluation and review.

While searching an image for abnormal-appearing cells, Dr A recalled reading about a new artificial intelligence (AI) computer program published by the Google Brain project. This computer program is based on an AI algorithm that is able to scan images at a fraction of a second and find possible cancerous cells. The authors reported that this program's sensitivity for detecting breast cancer was 92.4%, better than the human eye's 73.2% sensitivity.[1] As the program is based on a neural network algorithm that is able to learn from its mistakes, it is designed to have the ability to improve its scan sensitivity over time with continued use.

Dr A hopes her team can use this algorithm as a screening tool to help patients like Ms J. This algorithm could help by facilitating the pathology team's capacity to (1) identify potentially cancerous cells—or positive "hits"—in biopsied tissue images and (2) further evaluate those hits to see whether they truly are metastatic cancer cells. Dr A's colleagues agree that this technology could increase their power to identify abnormalities in patients' samples, but some hesitate to advise Community Hospital labs to invest in this technology due to the algorithm's yet unknown risk for generating false-

positive or false-negative hits. Some are also concerned about *automation bias*, or overreliance on output from a clinical decision support system after users become accustomed to easier workflow facilitated by the program.[2,3] Automation bias could result in a human user overlooking potential cancerous cells in an image. The team plans to consider these items at their next pathology department meeting.

Commentary

An AI program's higher reported success rate for spotting cancer cells compared with a human success rate would support using it to at least augment human pathologists' identification of cancer cells. It could even be argued that using the AI program could help train pathologists, since it has a higher success rate than the human eye alone at this stage. Why would one not see if the program identifies some cancer cells that might have been missed otherwise? This is a reason for introducing the program. We could imagine a hospital being sued if it were known that such a program existed and would have spotted cancer cells missed by a pathologist and a patient died as a result. However, 2 concerns about predictive algorithms deserve further ethical consideration.

The black-box issue. Since the AI program is based on a neural network algorithm, we don't know how it spots cancer cells (the "black- box" issue). That is, the AI program cannot identify the actual features it noticed to make a determination that certain cells are cancerous. Nevertheless, if the program is used to augment rather than replace the work of a pathologist, couldn't it help a pathologist become more knowledgeable about which cells are cancerous? A pathologist's job could include trying to identify which visual features cancerous (or normal) cells have in common and verbalizing what it is about cancerous cells that prompts the AI program to identify them as cancerous. Doing this job would illuminate the black box, and the transparency achieved could enable pathologists and patients to feel more comfortable relying on the AI program.

But now let us imagine that the AI program is widely used and improves its success rate from 92% to virtually 100%. Would we care so much about how it does it? There certainly would be curiosity among medical personnel, but would it matter to patients—those whose lives are affected by the presence or absence of cancer cells? We think not. This is because the job this program does, in contrast to other AI programs, is factually black or white. It finds cancer, if it is there. All we care about is how successful it is in spotting cancer cells. If it achieved a 100% success rate, with humans still lagging far behind, the program could replace pathologists doing this job, freeing them to do other jobs in which their expertise would be critical. After all, the AI program works much more quickly and isn't prone to making an incorrect diagnosis because of human weaknesses, like being tired.

We have maintained that the black-box problem isn't ethically or clinically critical here, especially if the program has 100% accuracy for cancer detection. But when there is a

possibility of an AI program leading to actions that affect human lives in ways that could be thought to be controversial, the black-box problem could still be important. Consider self-driving cars, for example: lives are at stake when a car "decides" whether to swerve left or right when a pedestrian suddenly appears on either side of a narrow road and there's no time to stop, or when a car must "choose" between braking abruptly to avoid hitting a pedestrian but could injure a passenger. A black box here has to do with understanding why a self-driving car does what it does, and not knowing why would not be ethically acceptable. That is, many of us want the basis for "decisions" made by artificially intelligent agents or devices to be well understood before we tend to feel comfortable using them.

Automation bias. Automation bias refers generally to a kind of complacency that sets in when a job once done by a health care professional is transferred to an AI program. We see nothing ethically or clinically wrong with automation, if the program achieves a virtually 100% success rate. If, however, the success rate is lower than that—92%, as in the case presented—it's important that we have assurances that the program has quality input; in this case, that probably means that the AI program "learned" from a cross section of female patients of diverse ages and races. With diversity of input secured, what matters most, ethically and clinically, is that that the AI program has a higher cancer cell-detection success rate than human pathologists.

If the goal is to ensure the most accurate diagnosis, as long as neither the AI program nor human diagnosticians have 100% accuracy, it is possible that the highest accuracy could be achieved when human pathologists' knowledge and skill is augmented by AI. Perhaps when mistakes are made, the program makes different mistakes than humans do; thus, using both methods would seem to yield the most accurate diagnosis.

Hypothetical or Actual?

If the case is to be read as a hypothetical scenario, the facts presented can simply be taken at face value; we've taken this approach in the preceding commentary. If the case is to be read as an actual scenario, however, greater care should be taken to verify that the facts presented are indeed correct before any analysis is done. As might be expected, the devil is in the details, and technical details in particular can be difficult for those not well versed in AI. Before considering the inner workings of and evaluation of an AI algorithm, for example, Dr A should consider some facts related to the Google Brain project.

First, it is not the case that Google Brain has published or intends for use at this time the algorithm described for detecting cancer. As stated in the *Google AI Blog*,

> Training models is just the first of many steps in translating interesting research [in]to a real product. From clinical validation to regulatory approval, much of the journey from "bench to bedside" still lies ahead—but

we are off to a very promising start, and we hope by sharing our work, we will be able to accelerate progress in this space.[4]

Given the natural inclination of developers to showcase their work in the best possible light, it seems prudent to regard their evaluation as an upper bound of our expectations.

Second, it should be made clear that the 73.2% tumor-level sensitivity reported for the human eye is, in fact, a result achieved by a single pathologist on one particular set of slides and, therefore, might not in fact be as indicative of the limits of human performance as the figure suggests.[1] Determining such a limit is likely to be an important agenda item for this technology's developers, as it will certainly influence how we consider the clinical and ethical appropriateness of its uses. As tumor-level sensitivity for the human eye increases, the persuasiveness of the case for replacing its use with an algorithm seems to diminish.

Third, although it would be possible for the AI program to improve through use over time, it does not currently do so as claimed in the case (Y. Liu, written communication, May 2018). Liu states that "having specific 'frozen' trained models are better in health care where stability and reproducibility of results are critical." That is, the reliability of the status quo (ie, the "frozen" AI program) has clinical and ethical value that should be considered upon reviewing the facts of the case presented. However, given that the program does not in fact learn over time, no ethical value should be attached to the alleged training process.

Lastly, information regarding false positives (in which cancer is falsely detected) and false negatives (in which cancer is present but not detected) was in fact reported,[1] although the case represents this information as unknown. The program's false positives are negligible: less than 0.0001% per slide, reported as 8 false positives in each of the 100 000 image patches into which each slide is divided.[1] Information about false negatives in the case can be gleaned from the reported sensitivity of 92.4%,[1] which implies that the AI system misses 7.6% of the tumors. Given that this data exists, Dr A can unhesitatingly use it to determine if these percentages are sufficient to recommend the program's use.

From an ethical perspective, there is some concern that the AI program in the case has not gone through a complete validation process. For example, the AI program's success rate has been compared to that of just one pathologist, with just one set of slides; this comparison doesn't offer convincing proof that the success rate of the AI program would be better than that of a particular hospital's diagnosticians. On the other hand, the 92.4% success rate, using a well-established data set, is a high bar to surpass, particularly given that human pathologists are bound to have days when they have difficulty concentrating. Thus, using the AI program would seem to be advisable.

In light of possible misunderstandings about facts of the case or AI program research, we suggest that, just as medical and ethical expertise should be represented in review board decisions regarding medical practice, so AI expertise should be represented in review board decisions regarding use of AI programs in health care. To expect those versed only in medicine to have the understanding required to determine whether such technology should be used seems overly optimistic given its complexities. It is a brave new world unlikely to be successfully negotiated without brave new approaches.

References

1. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. Arxiv. https://arxiv.org/abs/1703.02442. Updated March 8, 2017. Accessed August 3, 2018.
2. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc*. 2012;19(1):121-127.
3. Lyell D, Magrabi F, Raban MZ, et al. Automation bias in electronic prescribing. *BMC Med Inform Decis Mak*. 2017;17(1):28.
4. Stumpe M, Peng L. Assisting pathologists in detecting cancer with deep learning. *Google AI Blog*. March 3, 2017. http://ai.googleblog.com/2017/03/assisting-pathologists-in-detecting.html. Accessed August 3, 2018.

Michael Anderson, PhD is a professor emeritus of computer science at the University of Hartford in West Hartford, Connecticut. With Susan Leigh Anderson, he has been instrumental in establishing machine ethics as a field of study by co-chairing the Association for the Advancement of Artificial Intelligence fall 2005 symposium on machine ethics, co-authoring an *IEEE Intelligent Systems* special issue on machine ethics, and co-authoring an invited article for *AI Magazine* on the topic. He is a co-editor, with Susan Leigh Anderson, of *Machine Ethics* (Cambridge University Press, 2011). He earned a PhD in computer science and engineering at the University of Connecticut.

Susan Leigh Anderson, PhD is a professor emerita of philosophy at the University of Connecticut in Storrs, Connecticut. With Michael Anderson, she has been instrumental in establishing machine ethics as a field of study by co-chairing the Association for the Advancement of Artificial Intelligence fall 2005 symposium on machine ethics, co-authoring an *IEEE Intelligent Systems* special issue on machine ethics, and co-authoring an invited article for *AI Magazine* on the topic. She is a co-editor, with Michael Anderson, of *Machine Ethics* (Cambridge University Press, 2011). She earned a PhD in philosophy at the University of California, Los Angeles.

Editor's Note
The case to which this commentary is a response was developed by the editorial staff.

Citation
AMA J Ethics. 2019;21(2):E125-130.

DOI
10.1001/amajethics.2019.125.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The people and events in this case are fictional. Resemblance to real events or to names of people, living or dead, is entirely coincidental. The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E131-137

CASE AND COMMENTARY

Should Watson Be Consulted for a Second Opinion?

David D. Luxton, PhD, MS

Abstract

This article discusses ethical responsibility and legal liability issues regarding use of IBM Watson™ for clinical decision making. In a case, a patient presents with symptoms of leukemia. Benefits and limitations of using Watson or other intelligent clinical decision-making tools are considered, along with precautions that should be taken before consulting artificially intelligent systems. Guidance for health care professionals and organizations using artificially intelligent tools to diagnose and to develop treatment recommendations are also offered.

Case

Ms L is a 63-year-old woman who visits her primary care physician, Dr R, with new-onset fatigue and gum bleeding. After a thorough history and physical examination, Dr R orders a complete blood count, the results of which show anemia, thrombocytopenia, and leukocytosis. Dr R urgently refers Ms L to Dr O in hematology-oncology. Following more testing, Dr O concludes that Ms L has acute myeloid leukemia, admits her to the hospital, and schedules her for induction chemotherapy.

After several weeks of therapy, Ms L seems to be benefitting little from treatment and her condition has worsened since admission. Stumped by what once seemed like a routine case, Dr O recommends to Ms L genetic testing of her cancerous cells, which could offer additional information that could potentially lead to a different diagnosis or treatment plan. Dr O also considers consulting colleagues at the University of Tokyo, who recently used the IBM Watson™ artificial intelligence (AI) system to correctly diagnose a rare form of leukemia in a very similar case.[1] Dr O recalls Watson being able to sift through millions of pages of clinical literature as well as being able to incorporate a patient's genetic background and clinical history to come up with a diagnosis and treatment plan, all in a fraction of the time that it would take a physician to do.

Dr O wonders: *Perhaps Watson would come to the same conclusion I did. Or maybe Watson would find something I missed and help save Ms L's life. Or maybe Watson would be totally unhelpful and waste my and Ms L's time*. Dr O also wonders whether to explain Watson to Ms L and invite her to consider these questions.

Commentary

Watson is an advanced question-answering computer system developed by IBM that can be used as a clinical decision support system (CDSS) to assist health care professionals in making decisions about diagnoses and treatment options.[2] The system uses a variety of artificial intelligence (AI) approaches including natural language processing, information retrieval, semantic analysis, automated reasoning, and machine learning.[2] IBM calls its software architecture DeepQA, where QA stands for "question and answering."[2] The DeepQA system famously beat 2 Jeopardy game show contestants in a televised exhibition match in 2011.[2]

Watson is an example of augmented intelligence, whereby normal human intelligence is supplemented through use of technology in order to help people become faster and more accurate at the tasks that they're performing.[2] The system works essentially like this: massive amounts of unstructured and semistructured data such as that from the clinical literature, health records, and test results (eg, pathology reports) are fed into the Watson system database. A physician poses a query to the system describing symptoms of a specific patient and other related factors. Watson first parses the input to identify the most important pieces of information and then mines a patient's data to find relevant facts about the patient's clinical and hereditary history. The system then examines available data (that were previously inputted) to form and test hypotheses and finally provides a list of individualized, confidence-scored recommendations, such as a patient's eligibility for specific treatments. The system uses numerous scoring methods and sophisticated algorithms to determine the degree of certainty that retrieved evidence supports the candidate answers. The system can then describe the supporting evidence in text form for its ranked responses.[3] Because information is constantly being fed to Watson, the system can learn over time to optimize its recommendations.

IBM Watson Health™ presently offers commercialized applications of the Watson system for genomics, drug discovery, health care management, and oncology.[4] IBM has partnered with several academic and private institutions to apply Watson to patient care research and treatment. For example, in 2013, IBM partnered with a company called WellPoint to train Watson in utilization management and partnered with Memorial Sloan Kettering Cancer Center to train Watson in extracting and interpreting data related to lung cancer.[5] And, in 2015, IBM and Manipal Hospitals (a large hospital system in India) announced the launch of IBM Watson for Oncology, which sorts through information and provides insights to physicians and cancer patients to help them identify personalized, evidence-based cancer care options. The service is also made available directly to patients through Manipal Hospitals' website as a physician-mediated expert second opinion.[6]

Given the potential that Watson and any other intelligent CDSS has for clinical care and research, it's essential that physicians such as Dr L consider the ethical (and legal)

ramifications regarding their use. Some essential questions are these: (1) Should Watson ever replace the clinical judgment of a physician? (2) What are the liability concerns of professionals who use Watson? (3) What are the limitations of Watson and their ethical implications?

Watson's Role

According to IBM, Watson is intended to assist and enhance the decision making of health care professionals by giving them greater confidence in their diagnostic and treatment decisions for their patients.[1] Thus, the system is not intended to replace the judgment of health care professionals, nor should it be viewed as any kind of authoritative decision-making tool. In the United States, the Food and Drug Administration (FDA) regulates the safety and effectiveness of devices and drugs. Because Watson is considered a management tool under the control of physicians (like a peer-reviewed publication) and not a device, the system does not presently require regulatory oversight.[7] However, regulatory requirements could change as Watson and other emerging AI systems are used to make diagnoses or treatment decisions with little or no supervision from physicians.[8]

Liability

While Watson aims to assist the accuracy of clinical judgment and improve health outcomes, use of the system as an assistant also has potential to increase liability for health care professionals and organizations. As noted by Jacobson,[9] technological innovations create opportunities for error in diagnosis and treatment, and those errors could result in more visible and potentially detrimental outcomes than what might have happened without the new technology. As a hypothetical example, Watson could recommend a particular medication regimen that a physician decides to pursue while ignoring other contraindicating patient data because of the physician's assumption that Watson (or any other CDSS like it) had evaluated that information. Such a scenario could result in a malpractice claim against the physician.

While we can hope that advances in technologies such as Watson can improve outcomes for patients, they also have the potential to prematurely contribute to a higher legal standard of care that could put health care professionals at greater risk for negligence.[5] This is because expectations of the standard of care can shift while the impact of the technology on health outcomes is not yet fully known. For example, if Watson is shown to improve diagnostic accuracy and treatment recommendations for leukemia, then expectations that clinicians who consult Watson will get the diagnosis and treatment recommendation "right" could be raised to a higher level.

Unfortunately, a paucity of clinical trials evaluating every possible diagnosis and treatment approach can limit the reliability and usefulness of Watson. That is, recommendations provided by the system might not be supported by sufficient research

to instill confidence in health care professionals, who could be found liable if their diagnoses or treatment recommendations are shown to be incorrect or possibly prove harmful. Thus, health care professionals who use Watson, such as Dr O in the case example, should do so with an awareness of potential harm that overreliance on the system could cause in the individual case, but also with appreciation for how the system can also improve their decision making.

Understanding Watson's Limitations

There are precautions that should be taken into consideration before consulting Watson. First, it's important for physicians such as Dr O to understand the technical challenges of accessing quality data that the system needs to analyze in order to derive recommendations. Idiosyncrasies in patient health care record systems is one culprit, causing missing or incomplete data. If some of the data that is available to Watson is inaccurate, then it could result in diagnosis and treatment recommendations that are flawed or at least inconsistent. An advantage of using a system such as Watson, however, is that it might be able to identify inconsistencies (such as those caused by human input error) that a human might otherwise overlook. Indeed, a primary benefit of systems such as Watson is that they can discover patterns that not even human experts might be aware of, and they can do so in an automated way. This automation has the potential to reduce uncertainty and improve patient outcomes.

It is possible, however, that Watson might make a recommendation that is inconsistent with current clinical standards or that contradicts what a physician considers to be the appropriate decision. For example, a clinical standard might be always to prescribe a particular medication with a particular diagnosis, but an intelligent system such as Watson could recommend an alternative (eg, a nonstandard medication or no medication at all). In such a scenario, physicians must be able to support their decision to follow or not to follow the alternative and to understand the potential clinical and legal consequences. However, systems such as Watson can and should be designed to use a rule base that can limit recommendations to current clinical standards, thus ensuring that recommendations are consistent with treatment guidelines and currently accepted practices.

Inconsistency is associated with another consideration sometimes referred to as the "black-box" problem, whereby developers and users are unable to demonstrate how the system operates or derived its decisions for a particular course of action.[8] For example, Watson's machine learning algorithms can derive conclusions that are not consistent with a physician's judgment regarding the diagnosis or prognosis, yet why it derived particular solutions might not be obvious. From an ethical point of view, it is therefore essential that both the developers and the users of AI systems understand (or at least be able to explain) the basis of how the algorithms work to reduce risk of harm to patients.

Requirements for an audit trail with a minimum level of detail to describe the decision process might be one way to address the black-box issue and help ensure public trust.[4]

Technologies such as IBM Watson also have potential to create unrealistic patient expectations regarding outcomes. For example, a patient such as Ms L might be overly optimistic about her new treatment because her physician consulted Watson, a presumably superintelligent machine that can do things better than humans. Because the foundation of the patient-physician relationship is trust, patients should be informed—by the health care professionals who treat them—about the tools and tests used to make decisions about their health. It is therefore of great importance that clinical computing tools be presented as decision assistants, rather than as decision makers, and that their limitations be communicated effectively.

Meeting Challenges of Clinical Decision Support Systems

Historically, the deployment and adoption of technological clinical decision-making tools has been met with some challenges.[10,11] Some of the challenges are due to technological limitations (such as those associated with the data problems mentioned in this article), technology adoption issues (eg, usability and workflow integration), and physicians' perception of the technology when assured capabilities and timelines have not been achieved.[10] For example, IBM Watson Health has been criticized for not living up to promises of the system's ability to transform cancer treatment and outcomes.[12,13] Regardless, the need for intelligent automated systems such as Watson is evident given the exponential expansion and complexity of clinical data. For example, IBM has suggested that a person will generate 1 million gigabytes of health-related data in a lifetime—which is equivalent to more than 300 million books.[14] Given the amount and complexity of patient data, physicians would be remiss not to consult intelligent systems such as Watson. In the future, it may very well be considered unethical (and create liability) not to consult Watson or intelligent systems like it for a second opinion, assuming that such systems prove effective in what they purport to do.

In conclusion, the emergence of innovative technologies raises familiar and sometimes new legal and ethical ramifications for the health care profession. Health care organizations must educate and train their staff on the capabilities and limitations of technological tools while ensuring that patients are adequately informed of how these tools are used to make decisions about their care. Many of the challenges regarding the adoption and deployment of systems such as Watson have solutions that can be addressed in time. When new technologies become available, it inevitably requires time for the study of their safety and clinical effectiveness. It's unlikely that intelligent systems such as Watson will one day displace health professionals, but instead they will advance patient care and clinical research beyond its present limits.

References

1. Monegain B. IBM Watson pinpoints rare form of leukemia after doctors misdiagnosed patient. *Healthcare IT News*. August 8, 2016. https://www.healthcareitnews.com/news/ibm-watson-pinpoints-rare-form-leukemia-after-doctors-misdiagnosed-patient. Accessed November 16, 2018.
2. Luxton DD. An introduction to artificial intelligence in behavioral and mental health care. In: Luxton DD, ed. *Artificial Intelligence in Behavioral and Mental Health Care*. San Diego, CA: Elsevier Science; 2015:1-26.
3. Ferrucci D, Brown E, Chu-Carroll J, et al. Building Watson: an overview of the DeepQA Project. *AI Magazine*. 2010;31(3):59-79.
4. IBM. IBM Watson health: empowering heroes, transforming health. https://www.ibm.com/watson/health/. Accessed November 20, 2018.
5. IBM Watson hard at work: new breakthroughs transform quality care for patients [press release]. New York, NY: Memorial Sloan Kettering Cancer Center; February 8, 2013.
6. Manipal Hospitals. Watson for Oncology. https://watsononcology.manipalhospitals.com/. Accessed November 2, 2018.
7. US Food and Drug Administration; Federal Communications Commission; Office of the National Coordinator for Health Information Technology. FDASIA health IT report: proposed strategy and recommendations for a risk-based framework. https://www.fda.gov/downloads/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDRH/CDRHReports/UCM391521.pdf. Published April 2014. Accessed December 4, 2018.
8. Luxton DD, Anderson SL, Anderson M. Ethical issues and artificial intelligence technologies in behavioral and mental health care. In: Luxton DD, ed. *Artificial Intelligence in Behavioral and Mental Health Care*. San Diego: Elsevier Science; 2015:255-276.
9. Jacobson PD. *Medical Liability and the Culture of Technology*. Pew Project on Medical Liability. http://www.pewtrusts.org/-/media/legacy/uploadedfiles/wwwpewtrustsorg/reports/medical_liability/medmal092204pdf.pdf. Published 2004. Accessed August 7, 2018.
10. McCullagh LJ, Sofianou A, Kannry J, Mann DM, McGinn TG. User centered clinical decision support tools: adoption across clinician training level. *Appl Clin Inform*. 2014;5(4):1015-1025.
11. Garg AX, Adhikari NK, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*. 2005;293(10):1223-1238.
12. Ross C, Swetlitz I. IBM pitched Watson as a revolution in cancer care. It's nowhere close. *STAT*. September 5, 2017. https://www.statnews.com/2017/09/05/watson-ibm-cancer/. Accessed November 205, 2018.

13. Freedman DH. A reality check for IBM's AI ambitions. *MIT Technology Review*. June 27, 2017. https://www.technologyreview.com/s/607965/a-reality-check-for-ibms-ai-ambitions/. Accessed August 8, 2018
14. IBM and Partners to Transform Personal Health with Watson and Open Cloud [press release]. Armonk, NY: IBM; April 13, 2015. https://www-03.ibm.com/press/uk/en/pressrelease/46609.wss. Accessed on August 15, 2018.

David D. Luxton, PhD, MS is an affiliate associate professor in the Department of Psychiatry and Behavioral Sciences at the University of Washington School of Medicine in Seattle. His research focuses on the development and study of health care technology, artificial intelligence, and ethics.

Editor's Note
The case to which this commentary is a response was developed by the editorial staff.

Citation
AMA J Ethics. 2019;21(2):E131-137.

DOI
10.1001/amajethics.2019.131.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The people and events in this case are fictional. Resemblance to real events or to names of people, living or dead, is entirely coincidental. The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E138-145

CASE AND COMMENTARY

How Should Clinicians Communicate With Patients About the Roles of Artificially Intelligent Team Members?

Daniel Schiff, MS and Jason Borenstein, PhD

Abstract

This commentary responds to a hypothetical case involving an assistive artificial intelligence (AI) surgical device and focuses on potential harms emerging from interactions between humans and AI systems. Informed consent and responsibility—specifically, how responsibility should be distributed among professionals, technology companies, and other stakeholders—for uses of AI in health care are discussed.

Case

Mr K is a 54-year-old man referred to Dr L's outpatient spine neurosurgery clinic because he has a 6-week history of left-sided lower back pain, left leg weakness, and shooting pain. Prior to Mr K's appointment, Dr L reviewed the MRI of Mr K's lumbar spine, noting herniation of disc between the fifth lumbar vertebra and the first sacral vertebra (L5-S1), which is compressing Mr K's sacral (S1) nerve root.

"What a classic case," Dr L murmurs to herself. She grabs her reflex hammer and walks down the hall to exam room 3. After performing a brief evaluation and reviewing Mr K's MRI with him, Dr L recommends surgery to relieve compression of the S1 nerve.

"Isn't that a dangerous procedure? Could I end up paralyzed?" Mr K asks.

"There are certain risks, but with the help of the *Mazor Robotics* Renaissance® Guidance System technology, the procedure is relatively safe." Dr L explains the surgical planning using the Mazor system: "It employs artificially intelligent software to analyze your images and plan placement of my surgical tools. I've been using this technology for about a year now, and I've done over 30 surgeries—just like the one I'm recommending for you—with this technology."

Mr K looks uncomfortable. "I don't want a robot doing my surgery. I want you to do it all."

Dr L wonders how to respond.

Commentary

In this commentary, we examine a hypothetical case involving an assistive surgical device that is in use today, the Mazor Robotics Renaissance Guidance System.[1] It can assist surgeons like Dr L in performing procedures such as spinal fixation.[2,3] With a complex technology like the Renaissance System, a series of policies and procedures are important for ensuring its ethical use. These measures include well-designed clinical trials; creation and implementation of procedures before, during, and after surgery, especially concerning complications, errors, and robustness measures; training on the technology's characteristics, uses, and limitations; and how to inform patients about such information. Depending on the type of technology, approval by the US Food and Drug Administration or other regulatory entities might be required.

While these considerations might be relevant to any complex device, several more specific challenges emerge with respect to artificial intelligence (AI) technologies. AI can refer to a range of techniques including expert systems, neural networks, machine learning, and deep learning.[4] Medical ethics has begun to highlight concerns about uses of AI and robotics in health care, including algorithmic bias, the opacity and lack of intelligibility of AI systems, patient-clinician relationships, potential dehumanization of health care, and erosion of physician skill.[5,6] In response, members of the medical community and others have called for changes to ethical guidelines and policy and for additional training requirements for AI devices.[6]

Given the potential of AI to augment human medical care, the proper role of health care professionals vis-à-vis their digital counterparts is particularly relevant. First, the "black-box" problem—the mystery of how the system derives its outputs—is an issue for any complex and opaque medical technology. It raises questions about how to communicate possible biases, risks, and error rates during the informed consent process.[6,7] Second, as Mr K's concerns demonstrate, informed consent can be complex given uncertainties, fears, or even overconfidence about uses of AI. Finally, assigning responsibility and liability when errors occur is also complicated by the technical complexity and opacity of AI and the challenge of distributing responsibility across many parties. We address each of these ethical concerns below.

Informed Consent and the Black-Box Problem

One ethical challenge emerging from interactions between Mr K and Dr L in the case pertains to the difficulty of obtaining consent to use a novel AI device. As Appelbaum notes, "Valid informed consent is premised on the disclosure of appropriate information to a competent patient who is permitted to make a voluntary choice."[8] As is commonly known, relevant information includes the purpose of the treatment, its potential benefits and risks, and possible alternative treatment options. Yet the novelty and technical sophistication of an AI device places additional demands on the informed consent process. When an AI device is used, the presentation of information can be complicated

by possible patient and physician fears, overconfidence, or confusion. Moreover, for an informed consent process to proceed appropriately, it requires physicians to be sufficiently knowledgeable to explain to patients how an AI device works, which is rendered difficult by the black-box problem.

The black-box problem emerges for at least a subset of AI systems, including neural networks, which are trained on massive data sets to produce multiple layers of input-output connections.[9] The result can be a system largely unintelligible to humans beyond its most basic inputs and outputs.[10] In other words, those interacting with an AI system might not understand to any appreciable degree how it works (ie, its functioning seems like a black box). This challenge pertains not only to neural networks but also to any informationally or technically complex system that may be opaque to those who interact with it, such as Mazor's advanced and proprietary image recognition algorithms.[3]

The opacity of an AI system can make it difficult for health care professionals to ascertain how the system arrived at a decision and how an error might occur. For instance, can physicians or others understand why the AI system made the prediction or decision that led to an error, or is the answer buried under unintelligible layers of complexity? Will physicians be able to assess whether the AI system was trained on a data set that is representative of a particular patient population? And will physicians have information about comparative predictive accuracy and error rates of the AI system across patient subgroups? In short, if physicians do not fully understand (yet) how to explain an AI system's predictions or errors, how could this knowledge deficit impact the quality of an informed consent process and medical care more generally?

Ongoing conversations within many professional communities will be needed to grapple with these issues, but recommendations are already emerging. For example, Char et al. state,

> Physicians who use machine-learning systems can become more educated about their construction, the data sets they are built on, and their limitations. Remaining ignorant about the construction of machine-learning systems or allowing them to be constructed as black boxes could lead to ethically problematic outcomes.[6]

Moreover, professional societies are recommending that AI systems be "transparent."[11]

Assuming Dr L is well informed about the Renaissance Guidance System, she should seek to explain to Mr K the core technologies used, such as the basic nature of the image recognition algorithm. She should clearly distinguish between the roles human caregivers will play during each part of the procedure and the roles the AI/robotic system or device will play. For example, she should explain that she is responsible for the preoperative plan, whereas the Renaissance Guidance System will manually guide

placement of tools or implants.[3] Also, Dr L should clearly state the potential harms that might result from either human or robotic missteps.

Patient Perceptions of AI

Interconnected with lack of knowledge about AI systems—including how errors could occur—are varied perceptions patients and health care professionals have about AI technology. Computing experts offer wide-ranging visions of where AI is going, from utopian views in which humanity's problems are largely solved to dystopian scenarios of human extinction.[12] These visions can influence whether patients, such as Mr K in the case, and physicians embrace AI (perhaps too quickly) or fear it (even though it might improve health outcomes). For example, a 2016 survey of 12 000 people across 12 European, Middle-Eastern, and African countries found that only 47% of respondents would be willing to have a "robot perform a minor, non-invasive surgery instead of a doctor," with that number dropping to 37% for major, invasive surgeries.[12,13] These findings indicate that a sizeable proportion of the public has uneasiness about medical AI.

How should a physician respond to patients like Mr K who express concerns about the use of AI? In addition to delineating the role of the AI system, the physician can address the patient's fears or overconfidence by describing the risks and potential novel benefits attributable to the AI system. For example, beyond merely sharing that she has used this procedure in the past, Dr L should describe studies comparing the Renaissance Guidance System to human surgeons.[2] In this way, the patient's inaccurate perceptions of AI can be countered with a professional assessment of the benefits and risks involved in a specific procedure. While these 2 recommendations are important for proper informed consent, understanding and responding to patients' fears is also essential to good patient engagement and medical care. These 2 recommendations are not intended to be an exhaustive list; rather, they are a starting point for addressing sources of serious clinical and ethical concern about AI.

Medical Errors, AI, and the Problem of Many Hands

Suppose that Dr L uses the AI device to treat Mr K and a medical error occurs. How might one begin to assign responsibility for the error? Determining who is morally responsible and perhaps legally liable for a medical error involving use of a sophisticated technology is often complicated by the "problem of many hands."[14] This problem refers to the challenge of attributing moral responsibility when the cause of a harm is distributed among multiple persons—and perhaps organizations—in a way that obfuscates blame attribution. As Harris et al. state, individuals might use a many hands argument in an attempt "to evade personal responsibility for wrongdoing."[15] Given that many parties are involved in the design, sale, procurement, and use of AI systems in health care, identifying the primary locus of responsibility for a medical error can be difficult.[16] Moreover, the opacity of some AI systems compounds this challenge in new ways. Yet

transparency and clarity about roles and responsibilities can help ensure that the responsibility net is cast neither too narrowly nor too broadly.

A first step towards assigning responsibility for medical errors (thus hopefully minimizing them in the future) is to disentangle which people and professional responsibilities might have been involved in committing or preventing the errors. In the context of health care and AI, we suggest the following as a subset of the actors who could in principle be held ethically responsible for a medical error.

- *Coders and designers*. Coders and designers should be responsible for documenting what they created and, insofar as possible, implementing strategies for making explainable the technology and its underlying processes, such as how the AI is learning from training data.

- *Medical device companies*. Companies should clearly articulate prerequisites for successful application of an AI technology, such as the quality of diagnostics, imaging, and preparation for surgical procedures. Moreover, given black box concerns with AI systems, physicians might require additional information and training. Companies should therefore detail types of errors and side effects, their likelihood and severity, and differences in predictive accuracy and error rates across demographic subgroups, health conditions, and patient histories. Given uncertainties and risks surrounding complex, novel AI technologies in health care, companies should be responsible for providing meaningful information to hospitals and physicians, even if doing so surpasses what the law strictly requires.

- *Physicians and other health care professionals*. Physicians should be responsible for acquiring basic understanding of the AI devices they use and the types and likelihood of errors across subgroups, insofar as this information is available. Physicians should also be responsible for communicating relevant information to patients and health care teams and for adhering to use standards provided by device companies. Thus, if a medical error occurs because instructions for using an AI device were not followed, the primary responsibility could lie with the physician (or team); however, if a medical error occurs because adequate instructions or training were not provided by the company, the primary responsibility could lie elsewhere.

- *Hospitals and health care systems*. Hospitals are key to ensuring proper development, implementation, and monitoring of protocols and best practices for use of AI systems in health care. This organizational responsibility includes providing training, protocols, and best practices related to AI use and properly informing patients about the technology. Hospitals should also be involved in

developing robustness measures (including simultaneous diagnosis and crosschecking by physicians and AI). Best practice standards are also needed for error assessment and mitigation in cases of complications and for quality improvement.

Other actors, including regulators, insurance companies, pharmaceutical companies, and medical schools, also have important responsibilities. Each actor can take steps to ensure safe, ethical use of AI systems and encourage others to do so, too. These actions can help promote coordination among the various stakeholders about the use of AI in health care and contribute to a clearer sense of how to assign responsibility for successes as well as errors.

Challenges of AI in Health Care

While the challenges of integrating AI into the health care arena involve variations of familiar ethical issues, AI nevertheless presents new possibilities and concerns that deserve renewed attention. We suggest that companies provide detailed information about AI systems, which can help ensure that physicians—and subsequently their patients—are well informed. By explaining to patients the specific roles of health care professionals and of AI and robotic systems as well as the potential risks and benefits of these new systems, physicians can help improve the informed consent process and begin to address major sources of uncertainty about AI. Hopefully, the health care community will collectively meet these goals by encouraging open and robust dialogue about evaluating new AI technologies and integrating them into training and patient care.

References

1. Garrity M. Who is Mazor Robotics' biggest competitor in the spine market? *Becker's Spine Review*. November 27, 2017. https://www.beckersspine.com/orthopedic-a-spine-device-a-implant-news/item/39026-who-is-mazor-robotics-biggest-competitor-in-the-spine-market.html. Accessed June 6, 2018.
2. Joseph JR, Smith BW, Liu X, Park P. Current applications of robotics in spine surgery: a systematic review of the literature. *Neurosurg Focus*. 2017;42(5):e2.
3. Mazor Robotics. Renaissance: how it works. https://mazorrobotics.com/en-us/product-portfolio/mazor-x/mazorx-how-it-works. Published May 29, 2018. Accessed May 31, 2018.
4. Brookfield Institute for Innovation + Entrepreneurship; Policy Innovation Hub; Ontario. Intro to AI for policymakers: understanding the shift. https://brookfieldinstitute.ca/wp-content/uploads/AI_Intro-Policymakers_ONLINE.pdf. Published March 2018. Accessed July 13, 2018.
5. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219.

6. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983.
7. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318(6):517-518.
8. Appelbaum PS. Clinical practice. Assessment of patients' competence to consent to treatment. *N Engl J Med*. 2007;357(18):1834-1840.
9. Castelvecchi D. Can we open the black box of AI? *Nature*. 2016;538(7623):20-23.
10. Knight W. The dark secret at the heart of AI. *MIT Technology Review*. April 11, 2017. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/. Accessed June 20, 2018.
11. American Medical Association. Augmented intelligence in health care H-480.940. https://policysearch.ama-assn.org/policyfinder/detail/augmented%20intelligence?uri=%2FAMADoc%2FHOD.xml-H-480.940.xml. Modified 2018. Accessed December 7, 2018.
12. Müller VC, Bostrom N. Future progress in artificial intelligence: a survey of expert opinion. In: Müller VC, ed. *Fundamental Issues of Artificial Intelligence*. Cham, Switzerland: Springer; 2016:555-572.
13. PricewaterhouseCoopers. What doctor? Why AI and robotics will define new health. https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/ai-robotics-new-health.pdf. Published April 2017. Updated June 2017. Accessed October 15, 2018.
14. PricewaterhouseCoopers. What doctor? Why AI and robotics will define new health: data explorer. https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/data-explorer.html#!/D/18/stackedbars?cut=Territory&Tecf=0. Accessed December 6, 2018.
15. Thompson DF. Moral responsibility of public officials: the problem of many hands. *Am Polit Sci Rev*. 1980;74(4):905-916.
16. Harris CE, Pritchard MS, Rabins MJ. *Engineering Ethics: Concepts and Cases*. 4th ed. Belmont, CA: Wadsworth Cengage Learning; 2009.
17. Dixon-Woods M, Pronovost PJ. Patient safety and the problem of many hands. *BMJ Qual Saf*. 2016;25(7):485-488.

Daniel Schiff, MS is a PhD student at the Georgia Institute of Technology in Atlanta, where he studies artificial intelligence and its intersection with social policy. He earned an AB in philosophy from Princeton University and an MS in social policy from the University of Pennsylvania.

Jason Borenstein, PhD is the director of Graduate Research Ethics Programs and the associate director of the Center for Ethics and Technology at the Georgia Institute of Technology in Atlanta. His appointment is divided between the School of Public Policy and the Office of Graduate Studies.

Editor's Note
The case to which this commentary is a response was developed by the editorial staff.

Citation
AMA J Ethics. 2019;21(2):E138-145.

DOI
10.1001/amajethics.2019.138.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The people and events in this case are fictional. Resemblance to real events or to names of people, living or dead, is entirely coincidental. The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E146-152

MEDICAL EDUCATION

Reimagining Medical Education in the Age of AI

Steven A. Wartman, MD, PhD and C. Donald Combs, PhD

Abstract

Available medical knowledge exceeds the organizing capacity of the human mind, yet medical education remains based on information acquisition and application. Complicating this information overload crisis among learners is the fact that physicians' skill sets now must include collaborating with and managing artificial intelligence (AI) applications that aggregate big data, generate diagnostic and treatment recommendations, and assign confidence ratings to those recommendations. Thus, an overhaul of medical school curricula is due and should focus on knowledge management (rather than information acquisition), effective use of AI, improved communication, and empathy cultivation.

Natural illnesses are cured,
but never those which medicine creates,
for it knows not the secret of their cure.
Marcel Proust[1]

Information Overload

The system for educating medical students is approaching a crisis driven by 2 compelling forces: growing externalization of available medical knowledge outside the minds of physicians and stress-induced mental illness among learners.[2-5]

Classically, a physician is defined as a professional who possesses special knowledge and skills derived from rigorous education, training, and experience,[6] but the amount of available medical knowledge now exceeds the organizing capacity of the human mind.[7] What's known colloquially as "information overload" is caused not only by the volume of biomedical and clinical knowledge, but also by the rapidity of its increase and pressures on learners to achieve board scores high enough on the 3 United States Medical Licensing Examinations® to be chosen for competitive residency positions.[4,8] Medical practice today requires both high productivity and delivering on expectations for health outcomes—demands that can negatively impact learners' mental health.

Demands on students within the existing information-based curricula are taking a profound toll on their well-being. Students are confronted with information overload and concerns about never knowing enough. Recently, considerable attention has been paid to the deteriorating mental health of students.[9,10] While wellness and resilience programs have arisen in many medical schools[11] as a result, these programs implicitly focus on shortcomings of students and do not adequately address the root of the problem: the demanding learning environment.

Despite broad awareness of these trends, medical education continues to be largely information based, as if physicians are still the only source of medical knowledge. The reality of this web-enabled era is different. Patients readily garner more information, both correct and incorrect, to bring to clinical encounters and expect meaningful discussions with their physicians. These expectations challenge physicians not only to keep current but also to be able to communicate options to patients in a language that speaks meaningfully to their individual concerns and preferences. To do so requires specific training in effective communication as well as gaining deep understanding of the basis of patient decision making, including how patients' understanding of medical information is influenced by their inherent values and biases.

In addition, the skills required of practicing physicians will increasingly involve facility in collaborating with and managing artificial intelligence (AI) applications that aggregate vast amounts of data, generate diagnostic and treatment recommendations, and assign confidence ratings to those recommendations.[12,13] The ability to correctly interpret probabilities requires mathematical sophistication in stochastic processes, something current medical curricula address inadequately. In part, the need for more sophisticated mathematical understanding is driven by the analytics of precision and personalized medicine, which rely on AI to predict which treatment will work for a particular disease in a particular subgroup of patients. The long-standing approach of basing diagnostic or treatment choices on the "average patient" in a large population is no longer precise enough to meet the standards of personalized medicine. As a result, treatments for patients with different physical, cultural, and genetic attributes will vary in personalized medicine. As more practicing physicians use AI to support clinical decisions, they will need to be highly skilled in explaining treatment options to their patients. Merely expanding the current curricula to address this shortcoming will not be sufficient.

Medicine and AI

As we pointed out earlier, the increasing incongruence between the organizing and retention capacities of the human mind and medicine's growing complexity should compel significant re-engineering of medical school curricula. Curricula should shift from a focus on information acquisition to an emphasis on knowledge management and communication.[14] Nothing manifests this need for change better than the observation that every patient is becoming a big data challenge.[12] For clinicians, the need to

understand probabilities—such as confidence ratings for diagnostic or therapeutic recommendations generated by an AI clinical decision support system—will likely increase as personalized medicine continues to enlarge its role in practice. The ability to interpret these probabilities clearly and sensitively to patients and their families represents an additional—and essential—educational demand that speaks to a vital human, clinical, and ethical need that no amount of computing power can meet.

Good communication requires in-depth understanding of the psychology of choice, as the pioneering work of Tversky and Kahneman makes clear.[15] These authors explored how different phrasing affected participants' responses to a hypothetical life-and-death decision. The importance of the so-called "framing effect" has been demonstrated in a wide variety of settings, including health care.[16] Importantly, when potential patients were asked about the role of evidence in medical decision making, personal choice could eclipse medical evidence, and evidence of harm could be perceived as more compelling than evidence of effectiveness.[17] As patients become increasingly knowledgeable about medical information, physicians must be able to assess and respond to the heuristics of decision making. A key point is that the heuristics and biases of both physicians and patients need to be regarded as important parts of clinical encounters that must be skillfully managed to achieve optimal diagnosis and treatment, which is not taught in medical education today.

In 1991, Charles Van Doren wrote: "We have become a nation of passive recipients of services, most of them provided by complex machines whose operations we do not understand."[18] We agree and further argue that the psychology of choice should be front and center in the reimagined medical school curriculum. This is not to say that basic medical information should be eliminated from the curriculum. Rather, it should be integrated with teaching probability, communication, and empathetic skills.

Stewardship and Ethics

A fresh approach to teaching ethics is also called for in this new era, one that focuses on helping students respond to complexities that arise among patients, caregivers, and AI applications. Ethical challenges posed by this phenomenon are not new, but the emphasis on knowledge acquisition and technical competence seems to have diminished the prominence of empathy in curricula. No matter how high the confidence rating for the diagnosis or therapy recommended by an AI program, humans and their reactions to therapy are infinitely variable at the individual level. Physicians must therefore strengthen their capacity to respond to patients' suffering and express compassion.[19] The late Paul Kalanithi wrote in his book, *When Breath Becomes Air*, "the physician's duty is not to stave off death or return patients to their old lives, but to take into our arms a patient and family whose lives have disintegrated and work until they can stand back up and face, and make sense of, their own existence."[20] Anatole Broyard, reflecting on the clinical encounter, stated: "Not every patient can be saved, but illness may be eased by

the way caregivers respond."[21] Broyard and other writers recognize the importance of expressing abiding concern for others by respecting patients' rights to make choices according to their values and understanding how those values influence decisions. Doing so means having real, tested abilities to provide the uniquely human services patients need—to go beyond probabilities by addressing the complexities of caring for other humans. Perhaps offloading some biomedical and clinical knowledge onto AI applications will provide curricular space for restoring an emphasis on empathy.

Challenges to Curricular Reform

The history of medical education reform amply demonstrates that curricular change has been incremental, reactive, and mostly around the margins.[22] Changes that have occurred, such as earlier clinical experiences, more problem-based learning, and clinical skills testing, have not fundamentally altered learning environments and information-retention expectations imposed by medical school curriculum committees, the Liaison Committee on Medical Education, the Accreditation Council for Graduate Medical Education, and the National Board of Medical Examiners testing program.[23-26] Given the curricular needs addressed above, changes in 21st-century medical education must be radical, not incremental. The current learning environment, with its excessive information-retention demands, has proven to be toxic and in need of complete overhaul. The speed of technological innovation means that the skills of some faculty members are outdated compared to those of their students. In a recent visit to a medical school by one of the authors (SAW), when students were asked if they were "being taught in the manner in which they prefer to learn," no student said that this was the case.

Accordingly, we advocate new curricula that respond to the challenges of AI while being less detrimental to learners' mental health. These curricula should emphasize 4 major features:

1. Knowledge capture, not knowledge retention;
2. Collaboration with and management of AI applications;
3. A better understanding of probabilities and how to apply them meaningfully in clinical decision making with patients and families; and
4. The cultivation of empathy and compassion.

Barriers to such curricular changes are substantial and include long-standing faculty practices and funding streams, university policies and procedures, and a history of incremental reform by regulatory and accreditation bodies. It is our opinion that significant reform cannot take place within the existing regulatory structure. Perhaps changing the accreditation and licensing framework should be foremost among our considerations in reimagining medical education for the 21st century.

References

1. Proust M. *Remembrances of Things Past*. New York, NY: Vintage Books; 1982. *The Captive, the Fugitive, and Time Regained*; vol 3. Quoted by: Callahan D. Proust on treating chronic illness. *Over 65 Blog*. December 2, 2013. http://www.over65.thehastingscenter.org/proust-on-treating-chronic-illness. Accessed April 6, 2018.
2. Obermeyer Z, Lee TH. Lost in thought—the limits of the human mind and the future of medicine. *N Engl J Med*. 2017;377(13):1209-1211.
3. Slavin SJ. Medical student mental health. *JAMA*. 2016;316 (21):2195-2196.
4. Quintero GA. Medical education and the healthcare system—why does the curriculum need to be reformed? *BMC Med*. 2014;12:213.
5. Ishak W, Nikravesh R, Lederer S, Perry R, Ogunyemi D, Bernstein C. Burnout in medical students: a systematic review. *Clin Teach*. 2013;10(4):242-245.
6. Funder JW. Medicine as a profession. *Clin Med (Lond)*. 2010;10(3):246-247.
7. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. *Acad Med*. 2018;93(8):1107-1109.
8. Gauer JL, Jackson JB. The association of USMLE Step 1 and Step 2 CK scores with residency match and location. *Med Educ Online*. 2017:22(1):1358579.
9. Dyrbye L, Shanafelt T. A narrative review on burnout experienced by medical students and residents. *Med Educ*. 2016:50(1):132-149.
10. Dyrbye LN, West CP, Satele D, et al. Burnout among US medical students, residents, and early career physicians relative to the general US population. *Acad Med*. 2014:89(3):443-451.
11. Slavin SJ, Schindler DL, Chibnall JT. Medical student mental health 3.0: improving student wellness through curricular changes. *Acad Med*. 2014:89(4):573-577.
12. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med*. 2017;376(26):2507-2509.
13. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983.
14. Cabitza F, Rasoni R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318(6):517-518.
15. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science*. 1981;211(4481):453-458.
16. Mezzio DJ, Nguyen VB, Kiselica A, O'Day K. Evaluating the presence of cognitive bias in health care decision making: a survey of US formulary decision makers. *J Manag Care Spec Pharm*. 2018;24(11):1173-1183.
17. Carman KL, Maurer M, Mangrum R, et al. Understanding an informed public's views on the role of evidence in making health care decisions. *Health Aff (Millwood)*. 2016;35(4):566-574.
18. Van Doren C. *A History of Knowledge*. New York, NY: Ballantine Books; 1991.
19. Campbell J, Moyers BD, Flowers BS. *The Power of Myth*. New York, NY: Anchor Books; 1991.

20. Kalanithi P. *When Breath Becomes Air*. New York, NY: Random House; 2016.
21. Broyard A. *Intoxicated by My Illness and Other Writings on Life and Death*. New York, NY: Ballentine Books; 1992.
22. Skochelak SE. A decade of reports calling for change in medical education: what do they say? *Acad Med*. 2010;85(9)(suppl):S26-S33.
23. Liaison Committee on Medical Education. Functions and structure of a medical school: standards for accreditation of medical education programs leading to the MD degree. http://lcme.org/wp-content/uploads/filebase/standards/2019-20_Functions-and-Structure_2018-09-26.docx. Published March 2018. Accessed October 29, 2018.
24. Accreditation Council for Graduate Medical Education. ACGME Common program requirements. https://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf. Revised February 2017. Accessed October 29, 2018.
25. United States Medical Licensing Examination® website. https://www.usmle.org/. Accessed December 7, 2018.
26. AMA passes first policy recommendations on augmented intelligence [press release]. Chicago, Illinois: American Medical Association; June 14, 2018. https://www.ama-assn.org/ama-passes-first-policy-recommendations-augmented-intelligence. Published June 14, 2018. Accessed October 25, 2018.

Steven A. Wartman, MD, PhD is president emeritus of the Association of Academic Health Centers (AAHC) and a sociologist, board-certified internist, and master of the American College of Physicians. He served as president of the AAHC from 2005 to 2018, and, before joining the association, he was executive vice president for academic and health affairs and dean of the school of medicine at the University of Texas Health Science Center at San Antonio. He was also a Robert Wood Johnson Clinical Scholar at Johns Hopkins University and a Henry Luce Scholar in Indonesia. He received his AB degree from Cornell University and his MD and PhD degrees from Johns Hopkins University.

C. Donald Combs, PhD is the vice president and founding dean of the School of Health Professions at Eastern Virginia Medical School (EVMS) in Norfolk, Virginia. He also is a fellow of the Society for Simulation in Healthcare and holds senior faculty appointments in the EVMS School of Health Professions; the Department of Modeling, Simulation and Visualization Engineering at Old Dominion University; Paris Descartes University; and Taipei Medical University. He holds degrees from South Plains College, Texas Tech University, and the University of North Carolina at Chapel Hill.

Citation
AMA J Ethics. 2019;21(2):E146-152.

DOI
10.1001/amajethics.2019.146.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E153-159

MEDICAL EDUCATION

Emerging Roles of Virtual Patients in the Age of AI

C. Donald Combs, PhD and P. Ford Combs, MS

Abstract

Today's web-enabled and virtual approach to medical education is different from the 20th century's Flexner-dominated approach. Now, lectures get less emphasis and more emphasis is placed on learning via early clinical exposure, standardized patients, and other simulations. This article reviews literature on virtual patients (VPs) and their underlying virtual reality technology, examines VPs' potential through the example of psychiatric intake teaching, and identifies promises and perils posed by VP use in medical education.

Virtual Patients in Medical Education

Over the past 20 years, a revolution has taken place in the use of health care simulation. Technological advances in computational power, graphics, display systems, tracking, interface technology, haptic devices, authoring software, and artificial intelligence (AI) have supported creation of low-cost, user-friendly virtual reality (VR) technology and virtual patients (VPs). VPs are defined by the Association of American Medical Colleges as "a specific type of computer-based program that simulates real-life clinical scenarios; learners emulate the roles of health care providers to obtain a history, conduct a physical exam, and make diagnostic and therapeutic decisions."[1-3] VPs represent a fusion of simulation technologies and VR, which is generally defined as "a three-dimensional, computer-generated environment which can be explored and interacted with by a person."[4]

Research has been conducted documenting many settings where VR and VPs add value in education and in clinical practice. No longer merely a prop in a virtual world, VPs are designed to interact in 2D and 3D virtual worlds and to engage in face-to-face dialogues with users.[5,6] Artificially intelligent VPs interact verbally and nonverbally, and the most sophisticated VPs approach verisimilitude by engaging in rich conversations, recognizing nonverbal cues, and reasoning about social and emotional factors.[7] Learners interact with avatars (computer representations of patients that can speak and answer learner questions) in ways that mimic real and standardized patients. VPs provide a safe, effective means by which learners practice clinical skills before interacting with patients.

This article reviews literature on VPs and their underlying VR technology, examines the application of VPs in teaching psychiatric intake, and identifies promises and perils posed by VPs in medical education. Research suggests that VPs can successfully facilitate learners' acquisition of core knowledge in psychiatry and help develop their skills in history taking, interviewing, clinical reasoning, decision making, and assessing suicide risk.[8,9] We use psychiatric intake as an example because psychiatric issues such as opioid overuse, posttraumatic stress syndrome, and suicidality are pervasive. Shortages of mental health professionals and services limit learners' exposure to these clinical problems,[10] so VPs could play an especially useful role.[11]

The Promise of VPs

Technologically savvy learners have expectations about learning methods that differ from those of previous generations, and some faculty been slow to respond to this change. Devices such as laptops, tablets, and smartphones linked with sensors and applications through ubiquitous Wi-Fi networks are no longer merely peripheral to learners' experiences: they have become indispensable elements of education and practice. This evolution coincides with significant changes in the health care sector. An aging, often chronically ill population demands increasing attention from health care practitioners, and stringent clinical productivity expectations can reduce the time available for clinician-educators to participate in traditional teaching models. These trends are exacerbated by patients' decreasing lengths of stay in hospitals that further limit opportunities for students to participate in longitudinal treatments.[12] One response to these constraints is broader use of VPs.

Use of VPs offers several advantages when compared to traditional methods of teaching clinical skills. Online learning materials, such as VPs, are accessible any time and almost anywhere there is a computer with an internet connection. Once the VP software is developed, it can be reused without additional cost and VP "knowledge" can be updated quickly.[13] VPs also have advantages over standardized patients. VPs are more uniform than standardized patients because there is no variation in VP behavior once the software is completed and, unlike standardized patients, VPs do not need to be physically present with a learner. VPs might also convey more didactic information than standardized patients, who rely on recall. Additionally, VPs combine images, animations, video, and audio clips, which digital natives find more stimulating than textbooks.[14] VPs can help students learn clinical and ethical decision making, basic practitioner-patient communication, and history-taking skills.[15,16]

Nevertheless, it is important to acknowledge that VPs are not equivalent to real patients and cannot replace traditional clinic-based teaching. Modern state-of-the-art simulations are still limited compared to the reality of symptoms exhibited by patients. Additionally, learning through VPs outside a classroom requires substantial self-discipline, and enthusiasm for learning could deteriorate due to a lack of face-to-face

feedback from teachers and fellow students. Those observations aside, VR and VPs have uses in patient care (including exposure therapy, autism treatment, and responding to phantom limb pain) and uses in learning anatomical analysis, team training, surgical management, expressing empathy, and facilitating patient wellness.[17,18]

Using VPs to replicate clinical conditions and settings can provide a useful context for learning. Psychiatric intake, for example, is important because information elicited during the intake process can either be a prelude to accurate diagnosis and appropriate treatment or, to varying degrees, lead to mistakes, misunderstandings, and inappropriate care. A typical intake process includes gathering information on patient characteristics: address, sex, family, income, education, primary care practitioner, other clinicians, past and current health problems, relationship information, current functioning, and mental and physical symptoms.[19] This information can influence the quality of clinician-patient interactions, the accuracy of a diagnosis, and the effectiveness of a treatment plan. Research suggests that VPs can successfully facilitate learners' acquisition of core knowledge in psychiatry and help develop their skills in history taking, interviewing, clinical reasoning, decision making, and assessing suicide risk.[8,9]

The Peril of VPs

VPs have proven useful, but they have shortcomings. Current VPs might not represent the diversity of a population and, when racial or ethnic diversity is represented, VPs with darker skin tone could trigger learners' unconscious bias.[20,21] Stigmatizing language used in health records also influences learners' attitudes towards patients and their medication prescribing behavior.[22] People who are sincere in renouncing prejudice can remain vulnerable to biased habits of mind. Intentions are not good enough. Studies demonstrate bias affecting nearly every group of people.

> If you are Latino, you will get less pain medication than a white patient. If you're an elderly woman, you will receive fewer life-saving interventions than an elderly man.... If you are an obese child, your teacher is more likely to assume you're less intelligent than if you were slim.[23]

Bias is, no doubt, reflected in VP construction as well. Joanna Bryson, an expert in AI, notes that sexist AI could be a consequence of AI programming being done predominantly by "white, single guys from California"[24] and that it might be addressed, at least partially, by diversifying the software development workforce. According to Bryson, it should come as no surprise that machines express opinions of the people who program them: "When we train machines by choosing our culture, we necessarily transfer our own biases. There is no mathematical way to create fairness. Bias is not a bad word in machine learning. It just means that the machine is picking up regularities."[24] This concern about bias applies not only to AI but also to VPs.

In addition to concerns about bias in VP creation and use, there is significant potential for malicious intent in their programming. One example is a virtual human, named Norman, created at the Massachusetts Institute of Technology.[24] Norman illustrates that the data used to teach a machine learning algorithm can significantly influence what is learned and how a VP "behaves." When the output of an AI algorithm is biased and unfair, the culprit is usually not the algorithm but biased data used in training. Norman was subjected to extended exposure to the darkest corners of Reddit and is called the "world's first psychopath AI."[25] He represents a case study of AI gone wrong when biased data is used in machine learning algorithms.

For most users of AI, VR, and VPs, what goes on in the "black box" of programming is unknown and assumed to be trustworthy. As noted by Marc Goodman, a law enforcement agency adviser, "The thing people don't get is that cybercrime is becoming automated and it is scaling exponentially."[26] The most dangerous type of AI system and the one most difficult to defend against is an AI system made malevolent on purpose.[27] The easiest method to compromise a user immersed in VR and VPs is for the software programmer to subject the user, unknowingly, to content designed to change, persuade, or influence a user's decisions in harmful ways.[28] Additionally, any software—including VPs—can be hacked. In 2016, the number of reported data breaches increased by 40% over the previous year.[29,30]

Malicious intent, embedded bias, and mistaken connections among patient characteristics are all perils of VPs, and they raise interesting legal questions. The implications of using VPs in teaching psychiatric intake, for example, are frighteningly broad. Opportunities to amplify or distort a broad range of variables represented in intake records are numerous and could negatively influence students' learning.

Evaluating Promises and Perils of VPs

The Nuffield Council on Bioethics released a briefing note in 2018 on what it sees as big ethical questions about uses of AI in health care.[31] Modified to apply to VPs, we ask the following questions:

- What is the danger of VPs providing incorrect feedback?
- Who is responsible when the feedback is flawed?
- What is the potential for the malicious use of VPs?
- Will VPs diminish in-person interactions among teachers and learners?
- What impact does the growing use of VPs have on teaching and learning?

These questions can help clarify our thinking about the appropriate roles of VPs in education, how they should be constructed and used, and how we might increase the promise and decrease the peril of their use.

References

1. Association of American Medical Colleges Institute for Improving Medical Education. Effective use of educational technology in medical education. Colloquium on educational technology: recommendations and guidelines for medical educators. https://members.aamc.org/eweb/upload/Effective%20Use%20of%20Educational.pdf. Published March 2007. Accessed August 8, 2018.
2. Kononowicz AA, Zary N, Edelbring S, Corral J, Hege I. Virtual patients—what are we talking about? A framework to classify the meanings of the term in healthcare education. *BMC Med Educ*. 2015;15(11):1-7.
3. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. *Med Educ*. 2009;43(4):303-311.
4. Virtual Reality Society. What is virtual reality? https://www.vrs.org.uk/virtual-reality/what-is-virtual-reality.html. Accessed July 17, 2018.
5. Parsons TD. Virtual standardized patients for assessing the competencies of psychologists. In: Khosrowpour M, ed. *Encyclopedia of Information Science and Technology*. Vol 9. 3rd ed. Hershey, PA: IGI Global; 2015:6484-6493.
6. Maichen K, Danforth D, Price A, et.al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. *Simul Healthc*. 2017;12(2):124-131.
7. Rizzo A, Parsons T, Buckwalter JG, Lange B, Kenny P. A new generation of intelligent virtual clinical patients for clinical training. http://ict.usc.edu/pubs/A%20New%20Generation%20of%20Intelligent%20Virtual%20Patients%20for%20Clinical%20Training-ABS.pdf. Accessed December 6, 2018.
8. Pantziaras I, Fors U, Ekblad S. Training with virtual patients in transcultural psychiatry: do the learners actually learn? *J Med Internet Res*. 2015;17(2):e46.
9. Foster A, Chaudhary N, Murphy J, Lok B, Waller J, Buckley PF. The use of simulation to teach suicide risk assessment to health profession trainees—rationale, methodology, and a proof of concept demonstration with a virtual patient. *Acad Psychiatry*. 2015;39(6):620-629.
10. Weiner S. Addressing the escalating psychiatrist shortage. *AAMC News*. February 13, 2018. https://news,aamc.org/patient-care/article/addressing-escalating-psychiatrist-shortage/. Accessed October 17, 2018.
11. Doolen J, Giddings M, Johnson M, Guizado de Nathan G, O Badia L. An evaluation of mental health simulation with standardized patients. *Int J Nurs Educ Scholarsh*. 2014;11(1):55-62.
12. *American Hospital Association. Trends affecting hospitals and health systems.* https://www.aha.org/guidesreports/2018-05-22-trendwatch-chartbook-2018. Published 2018. Accessed October 17,2018.
13. Triola M, Feldman H, Kalet AL, et al. A randomized trial of teaching clinical skills using virtual and live standardized patients. *J Gen Intern Med*. 2006;21(5):424-429.
14. Prensky M. Digital natives, digital immigrants. *On the Horizon*. 2001;9(5):1-6.

15. Kenny P, Parsons T, Gratch J, Leuski A, Rizzo A. Virtual patients for clinical therapist skills training. *Lect Notes Comput Sci*. 2007;4722:197-210.
16. Stevens A, Hernandez J, Johnsen K, et al. The use of virtual patients to teach medical students history taking and communication skills. *Am J Surg*. 2006;191(6):806-811.
17. Craig E, Georgieva M. VR and AR: driving a revolution in medical education and patient care. *Educause Review*. August 30, 2017. https://er.educause.edu/blogs/2017/8/vr-and-ar-driving-a-revolution-in-medical-education-and-patient-care. Accessed July 8, 2018.
18. Hsieh MC, Lee JJ. Preliminary study of VR and AR applications in medical and healthcare education. *J Nurs Health Stud*. 2018;3(1):1-5.
19. Virginia mental health intake and evaluation. https://www.apadivisions.org/division-31/publications/records/virginia-intake-form.docx. Accessed July 17, 2018.
20. Urresti-Gundlach M, Tolks D, Kiessling C, Wagner-Menghin M, Härtl A, Hege I. Do virtual patients prepare medical students for the real world? Development and application of a framework to compare a virtual patient collection with population data. *BMC Med Educ*. 2017;17(1):174.
21. Zipp SA, Krause T, Craig SD. The impact of user biases toward a virtual human's skin tone on triage errors within a virtual world for emergency management training. *Proc Hum Factors Ergon Soc Annu Meet*. 2017;61(1):2057-2061.
22. P Goddu A, O'Conor KJ, Lanzkron S, et al. Do words matter? Stigmatizing language and the transmission of bias in the medical record. *J Gen Intern Med*. 2018;33(5):685-691.
23. Nordell J. Is this how discrimination ends? *Atlantic*. May 7, 2017. https://www.theatlantic.com/science/archive/2017/05/unconscious-bias-training/525405/. Accessed June 4, 2018.
24. Wakefield J. Are you scared yet? Meet Norman, the psychopathic AI. *BBC*. June 2, 2018. https://www.bbc.com/news/technology-44040008. Accessed July 17, 2018.
25. Massachusetts Institute of Technology. Norman—world's first psychopath AI. http://norman-ai.mit.edu. Accessed July 17, 2018.
26. Markoff J. As artificial intelligence evolves, so does its criminal potential. *New York Times*. October 23, 2016. https://www.nytimes.com/2016/10/24/technology/artificial-intelligence-evolves-with-its-criminal-potential.html. Accessed June 4, 2018.
27. Yampolskiy RV. Taxonomy of pathways to dangerous AI. Paper presented at: 30th AAAI Conference on Artificial Intelligence; February12-13, 2016; Phoenix, AZ. https://arxiv.org/ftp/arxiv/papers/1511/1511.03246.pdf. Accessed June 4, 2018.
28. Andrasik AJ. Hacking humans: the evolving paradigm with virtual reality. SANS Institute. https://www.sans.org/reading-room/whitepapers/testing/hacking-

humans-evolving-paradigm-virtual-reality-38180. Published November 2017. Accessed July 17, 2018.
29. Donaldson S. Virtual and augmented reality: transforming the way we look at the internet and data security. *Fossbytes*. March 19, 2017. https://fossbytes.com/virtual-and-augmented-reality-security/. Accessed July 8, 2018.
30. CyberScout. *Identity Theft Resource Center Data Breach Reports: 2016 End of Year Report*. https://www.idtheftcenter.org/images/breach/2016/DataBreachReport_2016.pdf. Accessed November 19, 2018.
31. Nuffield Council on Bioethics. Artificial intelligence (AI) in healthcare and research. http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf. Published May 2018. Accessed June 4, 2018.

C. Donald Combs, PhD is the vice president and founding dean of the School of Health Professions at Eastern Virginia Medical School (EVMS) in Norfolk, Virginia. He also is a fellow of the Society for Simulation in Healthcare and holds senior faculty appointments in the EVMS School of Health Professions; the Department of Modeling, Simulation and Visualization Engineering at Old Dominion University; Paris Descartes University; and Taipei Medical University. He holds degrees from South Plains College, Texas Tech University, and the University of North Carolina at Chapel Hill.

P. Ford Combs, MS completed a master's degree in bioinformatics and computational biology at George Mason University in Fairfax, Virginia. He holds an undergraduate degree from the University of North Carolina at Chapel Hill and has taken courses at the Conservatori Superior de Musica del Liceu in Barcelona, Spain.

Citation
AMA J Ethics. 2019;21(2):E153-159.

DOI
10.1001/amajethics.2019.153.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®
February 2019, Volume 21, Number 2: E160-166

HEALTH LAW

Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?

Hannah R. Sullivan and Scott J. Schweikart, JD, MBE

Abstract

As capabilities of predictive algorithms improve, machine learning will become an important element of physician practice and patient care. Implementation of artificial intelligence (AI) raises complex legal questions regarding health care professionals' and technology manufacturers' liability, particularly if they cannot explain recommendations generated by AI technology. The limited literature on liability for innovation provides opportunities to consider possible implications of AI for medical malpractice and products liability and new legal solutions for addressing liability issues surrounding "black-box" medicine.

Liability When Patients Are Injured Through New Technologies

Artificial intelligence (AI) is widely employed in health care, with a recent report showing that 86% of provider organizations, technology vendors, and life science companies use some form of AI.[1] AI can be broadly defined as machine intelligence that "performs tasks that normally require human intelligence"[2] or "that work[s] to achieve goals."[3] Among the most compelling applications of AI is the use of predictive algorithms in precision medicine. Algorithms in precision medicine guide care by predicting patient risks, making accurate diagnoses, selecting drugs, and even prioritizing patients to preserve or assign limited health resources.[4] Significantly, the mechanisms behind such recommendations are unknown and currently undiscoverable; an algorithm that cannot demonstrate the path to its conclusion is ultimately a black box.[5,6] The unknowable reasoning of "black-box" AI, often referred to as its opacity, stems from "deep neural networks," with their "reasoning … embedded in the behavior of thousands of simulated neurons, arranged into dozens or even hundreds of intricately interconnected layers."[5] When provided with input data, for example, such as an MRI brain scan, a neural network trained on a large data set can find a "complex underlying pattern in the data"[7] and produce an output, such as a tumor classification, but is incapable of explaining the reasoning that led to its conclusion.[7-9] Modeled after the human brain, the neural network also learns in similar ways, including through self-teaching. When given additional data, the neural network can modify its decision-making process for a more accurate response, without any explanation of how it has done so. Becoming more autonomous with each improvement,

the algorithms by which the technology operates become less intelligible to users and even the developers who originally programmed the technology.[10]

Given the opaque nature of black-box AI, key legal questions emerge when confronted with possible medical malpractice caused by such technology. For example, consider a situation in which a black-box AI system assists in detection of breast cancer using mammography data and suggests an erroneous diagnosis, resulting in injury to a patient. Are our legal doctrines of tort liability sufficient to handle medical malpractice resulting from the use of black-box AI? If not, what modifications to traditional tort law might be required to address AI systems involved in medical malpractice?

Traditional Tort Liability

Liability for medical errors falls under tort law. A tort is a civil claim in which a party requests damages for injuries caused by a harmful, wrongful act of another. Patients may recover compensatory and punitive damages from physicians, health care organizations, pharmaceutical companies, and medical device manufacturers if they are injured as a result of the party's failure to meet judicially accepted standards. Typical tort claims in the realm of medicine and health include medical malpractice (negligence), *respondeat superior* (vicarious liability), and products liability.

Physician liability: malpractice (negligence). Liability for medical errors falls under a negligence framework, the "most publicly visible legal mechanism" for protecting quality of care, which requires physicians to compensate patients for injuries for which the physician is responsible.[11] The legal definition of negligence is "conduct which falls below the standard established by law for the protection of others against unreasonable risk of harm."[12] In judicial determinations, a physician's actions are judged not against those of a reasonable *man*, but rather against those of a reasonable *physician*—with the same knowledge, skills, and expertise—under like circumstances.[13] However, courts do not purport to possess the knowledge necessary to determine sound medical judgment. Thus, expert testimony of qualified physicians is required to establish the standard of care or what is "reasonable to expect of a professional given the state of medical knowledge at the time of the treatment in issue."[14] Given the nature of medical practice, custom is largely dispositive. Expert testimony may be based upon available clinical literature, statements by the Food and Drug Administration (FDA), practice guidelines issued by medical societies (providing a ready-made standard), the *Physicians' Desk Reference*, and expert reliance on research findings.[11] Standards of care evolve over time with advances in medical knowledge and technology, and hence new developments in technology might create uncertainty for physicians about what is the current standard of care.

Health care organizations: respondeat superior (vicarious liability). In addition to physician liability, the doctrine of *respondeat superior* places vicarious liability on employers for the

negligent acts of employees acting within the scope of their employment.[15] Under this doctrine, "hospitals can be held vicariously liable for the acts of their employees, including physicians, who commit malpractice."[16] Alternatively, hospitals and other health care providers may be held separately negligent for failing to exercise due care in hiring, training, or supervising employees, or for failing to maintain adequate facilities and equipment.[17]

Manufacturers and pharmaceutical companies: products liability. Under products liability theory, patients are entitled to recovery when they are injured by products that are "not reasonably safe" due to defective design, manufacture, or warning. The relevant law states that manufacturers of prescription drugs and medical devices, those "that may be legally sold or otherwise distributed pursuant only to a health-care provider's prescription," are liable for harm to persons caused by defects.[15] A product is defectively designed "if the foreseeable risks of harm posed by the drug or medical device are sufficiently great in relation to its foreseeable therapeutic benefits" such that reasonable providers would not prescribe it to "any class of patients."[18] Warnings or instructions are inadequate if they fail to reasonably disclose risks "to prescribing and other health-care providers who are in a position to reduce the risks of harm."[18] The law reflects the FDA's determination that prescription medical products have inherent and unavoidable risks and thus require physician approval before use. It also emphasizes that the physician plays an important role in patients' choices.

Thus, a key difference arises when the products liability doctrine is applied to cases involving medicine and health care, in that such cases are typically subject to the *learned intermediary doctrine*. The learned intermediary doctrine addresses how patient-focused liability doctrines apply to the use of pharmaceuticals and medical devices, wherein physicians intervene between the manufacturer and the ultimate consumer.[19] Essentially, the learned intermediary doctrine "prevents plaintiffs from suing medical device manufacturers directly," as the manufacturer has no duty to the patient directly.[16] Under this doctrine, the "physician, rather than the patient, is considered the end consumer of medical devices because the health care provider is in the best position to weigh the risks against the possible benefits of using the device."[16] The physician as end consumer means that manufacturers may fulfill their duty to warn about the potential dangers of their products by providing warnings to the physicians who will be using them. If a physician subsequently fails to properly warn a patient and adequately disclose the risks and benefits associated with the product, it is the physician who will face liability.

Applying Current Liability Doctrines to AI

Applying the aforementioned tort liability schemes to AI technologies is difficult because, as Yavar Bathaee notes, the law "is built on legal doctrines that are focused on human conduct, which when applied to AI, may not function."[20] Matthew Scherer explains that a

large source of this difficulty stems from the opaque nature and unforeseeable results of black-box AI. For example, if the designers of AI cannot foresee how it will act after it is released in the world, how can they be held tortiously liable? And if the legal system absolves designers from liability because AI actions are unforeseeable, then injured patients may be left with fewer opportunities for redress.[3]

One problem with black-box AI's fitting into current liability schemes is its increased autonomy. According to Mark Chinen, "The more autonomy machines achieve, the more tenuous becomes the strategy of attributing and distributing legal responsibility for their behavior to human beings."[21] As the AI system becomes more autonomous, fewer parties (ie, clinicians, health care organizations, and AI designers) actually have control over it, and legal standards founded on agency, control, and foreseeability collapse—directly impacting opportunities for recovery of damages based on legal theories of negligence and vicarious liability. Additionally, it is challenging to find a responsible party, as so many different entities—software developers, hardware engineers, designers, and corporations—go into the creation of AI systems. As Scherer notes, it may be unfair to "assign blame to the designer of a component whose work was far-removed in both time and geographic location from the completion and operation of the AI system."[3]

Also, there are problems in applying the standard products liability model to AI. One is that, as discussed earlier, an injured patient cannot sue a manufacturer directly because of the learned intermediary doctrine. Additionally, products liability claims in the health care context require that the injuring product be deemed a "medical device."[2,4] The "hardware components" of the AI system would be deemed the "device" for products liability purposes, not the software.[16] The legal reasoning of not allowing products liability to extend to software is that software, as opposed to hardware, is "technology that helps healthcare providers make decisions by providing them with information or analysis" and that the final decision of care rests with the health care professional,[4] while "blatant hardware defects" would instead be subject to products liability suit against the manufacturer.[16] As AI becomes further integrated into medicine and health care, it becomes clear that current legal standards and doctrines regarding medical malpractice are insufficient. The innovations are unprecedented and solutions to the problems they present are necessary.

Possible Legal Solutions to Address AI Liability

In light of significant challenges in applying the current tort framework to AI, legal and computer science experts have offered possible solutions that involve modifications to the current law or the creation of new legal doctrines.

AI personhood. One possible solution is to confer "personhood" on the artificially intelligent machine itself, viewing the machine as an independent "person" under the law. Viewing the machine itself as a person resolves agency questions, which are

important for analysis of vicarious liability claims (ie, *respondeat superior*), as the machine will be viewed as the "principal" and no longer as an agent.[22] The machine, deemed a principal under this model of personhood, will have burdens and duties of its own and will then be sued directly for any negligence claims. In such instances, the AI system will be required to be insured (similar to how physicians possess medical malpractice insurance themselves) and such claims will be paid out from the insurance; the AI system will be deemed a quasi-juridical person and treated the "same as any other physician."[16] Funding for such insurance may come from users of the AI technology, allowing for a "different form of cost-spreading" that promotes fairness, as its focus extends beyond the technology's creators and encourages users of such technology to also bear some cost.[22]

Common enterprise liability. A common enterprise theory of liability is another possible solution to harm caused by AI. David Vladeck notes that, instead of assigning fault to a specific person or entity (or trying to determine if there was a fault at all), if some injury is caused by an AI system, then all groups involved in the use and implementation of the AI system should jointly bear some responsibility.[22] The benefit of this solution is that all parties involved share the burden and that no finding of fault (which may be impossible because of the black-box nature of AI) is required. Instead, an inference of liability is shared among all relevant parties, thus allowing injured parties to be made whole.

Modify the standard of care. Another possible solution is to simply modify the duties and standard of care of health care professionals using black-box AI. Nicholas Price suggests a standard that would require facilities and health care professionals to exercise "due care in procedurally evaluating and implementing black-box algorithms."[7] Under this standard of care, facilities and clinicians would have a duty to evaluate black-box algorithms and to validate the algorithmic results.[7] Under this model, health care professionals are responsible for harm if they did not take adequate measures in properly evaluating the black-box AI technologies used in caring for the patient.

Conclusion

The rise of black-box AI and its use in medicine complicates application of existing tort law when trying to resolve claims of malpractice. If a patient becomes injured by use of an AI technology (black-box AI in particular), current legal models are insufficient to address the realities of these innovations. New legal solutions that craft novel legal standards and models that address the nature of AI, such as AI personhood or common enterprise liability, are necessary to have a fair and predictable legal doctrine for AI-related medical malpractice.

References

1. Tata Consultancy Services. Getting smarter by the sector: how 13 global industries use artificial intelligence. https://sites.tcs.com/artificial-intelligence/. Accessed November 27, 2018.
2. Chung J, Zink A. Hey Watson—can I sue you for malpractice? Examining the liability of artificial intelligence in medicine. *Asia Pac J Health Law Ethics*. 2018;11(2):51-80.
3. Scherer MU. Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harv J Law Technol*. 2016;29(2):353-400.
4. Price WN. Artificial intelligence in health care: applications and legal implications. *The SciTech Lawyer*. 2017;14(1):10-13.
5. Knight W. The dark secret at the heart of AI. *MIT Technology Review*. April 11, 2017. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/. Accessed August 15, 2018.
6. Pande V. Artificial intelligence's "black box" is nothing to fear. *New York Times*. January 25, 2018. https://www.nytimes.com/2018/01/25/opinion/artificial-intelligence-black-box.html. Accessed August 15, 2018.
7. Price WN. *Medical Malpractice and Black-Box Medicine. Big Data, Health Law, and Bioethics*. Cambridge, UK: Cambridge University Press; 2018.
8. Moshen H, El-Dahshan EA, El-Horbaty EM, Salem AM. Classification using deep learning neural networks for brain tumors. *Future Comput Inform J*. 2018;3(1):68-71.
9. Paul JS, Plassard AJ, Landman BA, Fabbri D. Deep learning for brain tumor classification. *Proc SPIE Int Soc Opt Eng*. 2017;10137(1013710).
10. Bleicher A. Demystifying the black box that is AI. *Scientific American*. August 9, 2017. https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/. Accessed November 15, 2018.
11. Furrow BR, Greaney TL, Johnson SH, Jost T, Schwartz R. *Health Law: Cases, Materials, and Problems*. 8th ed. St. Paul, MN: West Academic Publishing; 2018.
12. Restatement (Second) of Torts §282 (Am Law Inst 1965).
13. Restatement (Third) of Torts §12 (Am Law Inst 2010).
14. *Nowatske v Osterloh*, 543 NW2d 265, 272 (Wis 1996).
15. 27 Am Jur 2d Employment Relationship §356 (Thomson Reuters 2002).
16. Allain JS. From jeopardy to jaundice: the medical liability implications of Dr. Watson and other artificial intelligence systems. *LA Law Rev*. 2013;73(4):1049-1079.
17. Meera T, Phanjoubam M, Nabachandra H. Hospital's liability in malpractice suits. *J Med Soc*. 2016;30(1):1-2.
18. Restatement (Third) of Torts §6 (Am Law Inst 1998).
19. *Marcus v Specific Pharmaceuticals, Inc*, 77 NYS2d 508 (1948).
20. Bathae Y. The artificial intelligence black box and the failure of intent and causation. *Harv J Law Technol*. 2018;31(2):889-938.

21. Chinen MA. The co-evolution of autonomous machines and legal responsibility. *Va J Law Technol*. 2016;20(2):338-393.
22. Vladeck DC. Machines without principles: liability rules and artificial intelligence. *Wash Law Rev*. 2014;89(1):117-150.

Hannah R. Sullivan is a legal scholar for the American Medical Association Council on Ethical and Judicial Affairs in Chicago, Illinois. She received her bachelor's degree from Indiana University Bloomington. Currently, she is a second-year law student at DePaul University College of Law, where she is a Jaharis Health Law Institute Fellow and serves as a teaching assistant and research assistant. Her legal interests include health law and policy.

Scott J. Schweikart, JD, MBE is a senior research associate for the American Medical Association Council on Ethical and Judicial Affairs in Chicago, Illinois, where he is also the legal editor for the *AMA Journal of Ethics*. Previously, he worked as an attorney editor and reference attorney at Thomson Reuters and practiced law in Chicago. Mr Schweikart earned his MBE from the University of Pennsylvania, his JD from Case Western Reserve University, and his BA from Washington University in St. Louis. He has research interests in health law, health policy, and bioethics.

Citation
AMA J Ethics. 2019;21(2):E160-166.

DOI
10.1001/amajethics.2019.160.

Acknowledgements
Hannah R. Sullivan and Scott J. Schweikart contributed equally to this work.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

ISSN 2376-6980

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E167-179

ORIGINAL RESEARCH

Can AI Help Reduce Disparities in General Medical and Mental Health Care?

Irene Y. Chen, Peter Szolovits, PhD, and Marzyeh Ghassemi, PhD

Abstract

Background: As machine learning becomes increasingly common in health care applications, concerns have been raised about bias in these systems' data, algorithms, and recommendations. Simply put, as health care improves for some, it might not improve for all.

Methods: Two case studies are examined using a machine learning algorithm on unstructured clinical and psychiatric notes to predict intensive care unit (ICU) mortality and 30-day psychiatric readmission with respect to race, gender, and insurance payer type as a proxy for socioeconomic status.

Results: Clinical note topics and psychiatric note topics were heterogenous with respect to race, gender, and insurance payer type, which reflects known clinical findings. Differences in prediction accuracy and therefore machine bias are shown with respect to gender and insurance type for ICU mortality and with respect to insurance policy for psychiatric 30-day readmission.

Conclusions: This analysis can provide a framework for assessing and identifying disparate impacts of artificial intelligence in health care.

Bias in Machine Learning Models

While health care is an inherently data-driven field, most clinicians operate with limited evidence guiding their decisions. Randomized trials estimate average treatment effects for a trial population, but participants in clinical trials often aren't representative of the patient population that ultimately receives the treatment with respect to race and gender.[1,2] As a result, drugs and interventions are not tailored to historically mistreated groups; for example, women, minority groups, and obese patients tend to have generally poorer treatment options and longitudinal health outcomes.[3-9]

Advances in artificial intelligence (AI) and machine learning offer the potential to provide personalized care by taking into account granular patient differences. Machine learning using images, clinical notes, and other electronic health record (EHR) data has been

successful in several clinical tasks such as detection of diabetic retinopathy[10] and distinguishing between malignant and nonmalignant skin lesions in dermatoscopic images.[11] Prior research has established that machine learning using clinical notes to supplement lab tests and other structured data is more accurate than an algorithm using structured data alone in classifying patients with rheumatoid arthritis[12] and in predicting mortality[13] and the onset of critical care interventions[14] in intensive care settings.

This same ability to discern among patients brings with it the risk of amplifying existing biases, which can be especially concerning in sensitive areas like health care.[15,16] Because machine learning models are powered by data, bias can be encoded by modeling choices or even within the data itself.[17] Ideally, algorithms would have access to exhaustive sources of population EHR data to create representative models for diagnosing diseases, predicting adverse effects, and recommending ongoing treatments.[18] However, such comprehensive data sources are not often available, and recent work has demonstrated bias in critical care interventions. For example, recent Canadian immigrants are more likely to receive aggressive care in the ICU than other Canadian residents.[19]

In contrast to critical care, psychiatry relies more heavily on analysis of clinical notes for patient assessment and treatment. Text is a rich source of unstructured information for machine learning models, but the subjective and expressive nature of the data also makes text a strong potential source of bias.[20,21] Racism has established impacts on chronic and acute health,[22] which would affect EHR data. In addition, mental health problems of racial groups often depend heavily on the larger social context in which the group is embedded,[22] which would also influence clinical prediction based on EHR data.

In prior work, the first author and colleagues formalized a framework for decomposing sources of unfairness in prediction tasks, including an analysis of racial bias for prediction of hospital mortality from clinical notes.[23] In contrast to human bias, algorithmic bias occurs when an AI model, trained on a given data set, produces results that may be completely unintended by the model creators. The authors used the publicly available Medical Information Mart for Intensive Care (MIMIC-III) v1.4,[24] which contains de-identified electronic health record data from 53 423 intensive care unit (ICU) admissions for 38 597 adult patients from Beth Israel Deaconess Medical Center from 2001 to 2012. After restricting the data set to ICU admissions lasting over 48 hours and excluding discharge summaries, the researchers created a final data set of 25 879 patient stay notes and demonstrated that prediction errors for patient mortality differ between races.[23]

In this paper, we explore the potential impacts of bias in 2 algorithms, one for predicting patient mortality in an ICU and the other for predicting 30-day psychiatric readmission in an inpatient psychiatric unit. We expand on the first author's previous research, discussed above, on bias in ICU patient mortality prediction using the same MIMIC-III

data set cohort with gender and insurance type in addition to race as demographic groups. We also analyzed potential bias in 30-day psychiatric readmission prediction for the same demographic groups.

Because unstructured clinical notes from the EHR contain valuable information for prediction tasks—including information about the patient's race, gender, and insurance type—we focus on clinical narrative notes in EHR data available for each stay. We examine bias, as measured by differences in model error rates in patient outcomes between groups, and show that in the ICU data set, differences in error rates in mortality for gender and insurance type are statistically significant and that in the psychiatric data set, only the difference in error rates in 30-day readmission for insurance type is statistically significant.

Data and Methods

Data. We analyze prediction error in psychiatric readmissions at a New England hospital in a data set containing 4214 deidentified notes from 3202 patients, collected from stays between 2011 and 2015. We extracted notes, patient race, gender, insurance payer type, and 30-day psychiatric readmission from every patient stay. The data set is racially imbalanced but has relative gender parity. We use the insurance payer type—public, private, and other insurance—as a proxy for socioeconomic status. (See Supplementary Appendix Table S1 for demographic information.)

We also examine prediction error in ICU mortality using the MIMIC-III v1.4 data set with the cohort selection explained earlier. (See Supplementary Appendix Table S2 for demographic information.) For race, gender, and insurance payer type, we compare error rates for psychiatric readmission with error rates for ICU mortality in order to examine unfairness across different data sets and the clinical generalizability of our methods.

Methods. We use topic modeling with latent Dirichlet allocation[25] (LDA) to uncover 50 topics (eg, depression, pulmonary disease; see Supplementary Appendix Tables S3 and S4 for example topics) and corresponding enrichment values for race, gender,[17,26] and insurance type. We used 1500 iterations of Gibbs sampling to learn the 50 topics of the LDA for each data set. For the psychiatric data set, topics were learned using the LDA Python package[27] whereas for the ICU clinical notes, topics were learned using Mallet.[28] (This difference in software arose from restrictions on the servers hosting the respective data sources.) Following prior work on enrichment of topics in clinical notes,[13,26] we computed enrichment values for topics for race, gender, and insurance type.

We predict hospital mortality with ICU notes and 30-day psychiatric readmission with psychiatric notes using logistic regression with L1 regularization (implemented by Python package scikit-learn[29] with a hyperparameter of C = 1) using an 80/20 split for training and testing data over 50 trials. For both hospital mortality and psychiatric

readmission, we report the error rate (zero-one loss) of the learned model for each demographic group and the 95% confidence interval. Text was vectorized using TF-IDF[30] on the most frequent 5000 words for each data set. We report the area under the receiver operator curve (AUC)[31] for overall model performance as well as the generalized zero-one loss as a performance metric.[32] Following prior work,[23] we use the Tukey range test,[33] which allows for pairwise comparisons among more than two groups, to test whether differences in error rates between groups are statistically significant. All Tukey range test error rate comparisons were performed using the Python package statsmodels.[34]

Our cohort selection code for MIMIC-III v1.4 and our analysis code are made publicly available to enable reproducibility and further study.[35]

Results: Enrichment of Topic Modeled Notes

Psychiatric note topics. White patients had higher topic enrichment values for the anxiety[36] and chronic pain topics, while black, Hispanic, and Asian patients had higher topic enrichment values for the psychosis topic.[37] Male patients had higher topic enrichment values than female patients for substance abuse (0.024 v 0.015), whereas female patients had higher topic enrichment values than male patients for general depression (0.021 v 0.019) and treatment resistant depression (0.025 v 0.015), reflecting known clinical findings.[38,39] Previous work has shown that those with serious mental illness are more likely to have public insurance than private[39]; we similarly find that private insurance patients have higher topic enrichment values than public insurance patients for anxiety (0.029 v 0.0156) and general depression (0.026 v 0.017). However, public insurance patients have higher topic enrichment values than private insurance patients for substance abuse (0.022 v 0.016).

ICU note topics. Intensive care unit clinical notes have a different range of topics (see Supplementary Appendix Table S3) and more refined topics than psychiatric notes due to the larger data source (25 879 v 4214 patients). As in the psychiatric data set, male patients have higher topic enrichment values for substance use than female patients (0.027 v 0.011), whereas female patients have higher topic enrichment values for pulmonary disease than male patients (0.026 v 0.016), potentially reflecting known underdiagnosis of chronic obstructive pulmonary disease in women.[40,41] Verifying known clinical trends, Asian patients have the highest topic enrichment values for cancer (0.036), followed by white patients (0.021), other patients (0.016), and black and Hispanic patients (0.015).[42] Black patients have the highest topic enrichment values for kidney problems (0.061), followed by Hispanic patients (0.027), Asian patients (0.022), white patients (0.015), and other patients (0.014).[42] Hispanic patients have the highest topic enrichment values for liver concerns (0.034), followed by other patients (0.024), Asian patients (0.023), white patients (0.019), and black patients (0.014).[43] Finally, white patients have the highest topic enrichment values for atrial fibrillation (0.022), followed

by other patients (0.017), Asian patients (0.015), black patients (0.013), and Hispanic patients (0.011).[44]

Public and private insurance patients vary mainly in the severity of conditions they are being treated for. Those with public insurance often have multiple chronic conditions that require regular care.[45] In particular, compared with private insurance patients, public insurance patients have higher topic enrichment values for atrial fibrillation (0.24 v 0.013), pacemakers (0.023 v 0.014), and dialysis (0.023 v 0.013). However, compared with public insurance patients, private insurance patients have higher topic enrichment values for fractures (0.035 v 0.012), lymphoma (0.030 v 0.015), and aneurysms (0.028 v 0.016).

In sum, our results for gender and race reflect known specific clinical findings, whereas our results for insurance type reflect known differences in patterns of ICU usage between public insurance patients and private insurance patients.

Results: Quantifying Disparities in Care With AI

After establishing that findings from the clinical notes reflect known disparities in patient population and experience, we evaluated whether predictions made from such notes are fair. There are multiple definitions of algorithmic fairness[46-49]; here we compare differences in error rates in ICU mortality and 30-day psychiatric readmission for race, gender, and insurance type.

Prediction error in the ICU model. Unstructured clinical notes are a powerful source of information in predicting patient mortality—our models achieve an AUC[31] of 0.84 using only the ICU notes. Adding demographic information (age, race, gender, insurance type), improves AUC slightly, to 0.85. As shown in Figures 1 and 2, error rates for gender and insurance type all have nonoverlapping confidence intervals. For gender, female patients have a higher model error rate than male patients; for insurance type, public insurance patients have a much higher model error rate than private insurance patients. All results are statistically significant at the 95% confidence level.

Figure 1. 95% Confidence Intervals for Error Rate (Zero-One Loss) in ICU Mortality for Gender

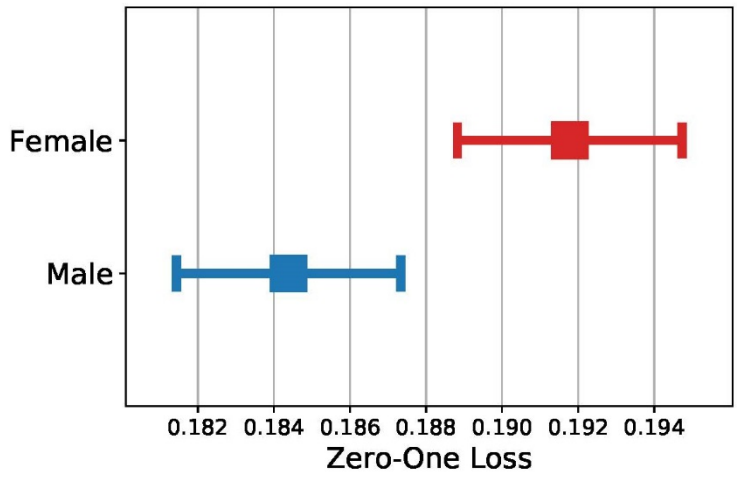
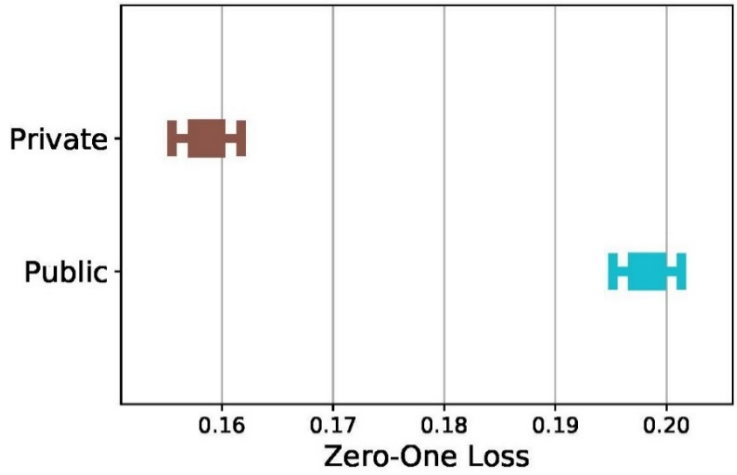


Figure 2. 95% Confidence Intervals for Error Rate (Zero-One Loss) in ICU Mortality for Insurance Type

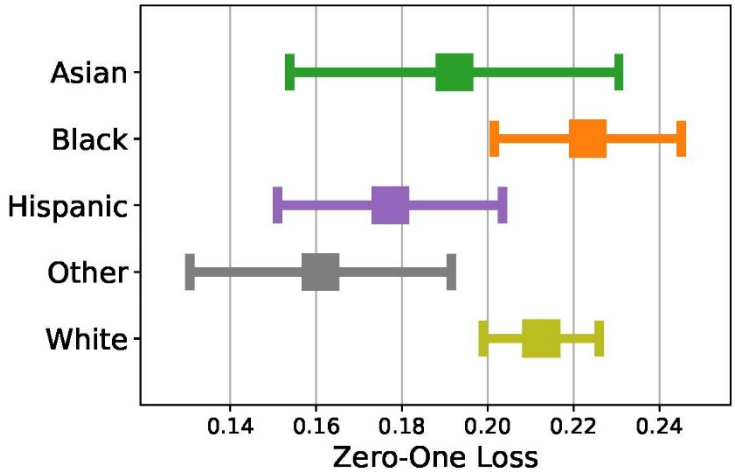


Prediction in the psychiatric setting. In contrast to ICU mortality, predicting 30-day psychiatric readmission is significantly more challenging, leading to lower model accuracy.[50] One potential cause could be the importance of unmeasured residential, employment, and environmental factors in predicting short-term psychiatric readmission.[51] Another factor could be the level of hospital intervention, such as outpatient appointments.[52]

Comparison of prediction errors in ICU and psychiatric models. We compare differences in error rates in 30-day psychiatric readmission and ICU mortality for race, gender, and insurance type. Figure 3 shows differences in error rates in psychiatric readmission

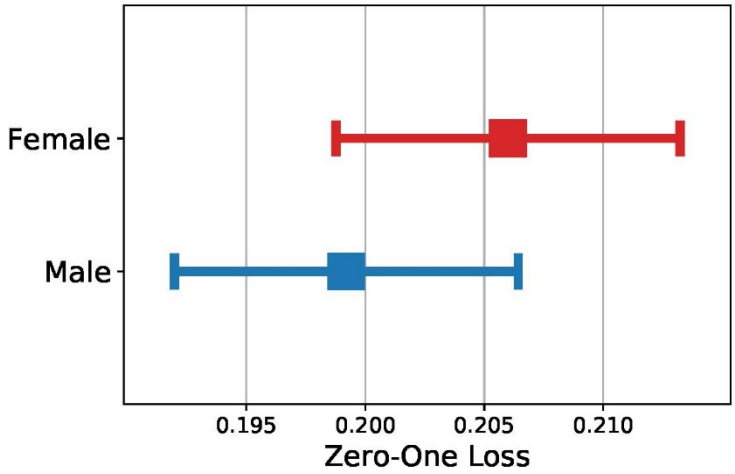
between racial groups, which were not statistically significant, with black patients having the highest error rate for psychiatric readmission. Differences in error rates in ICU mortality were also observed between racial groups.[23]

Figure 3. 95% Confidence Intervals for Error Rate (Zero-One Loss) in Psychiatric Readmission for Racial Groups



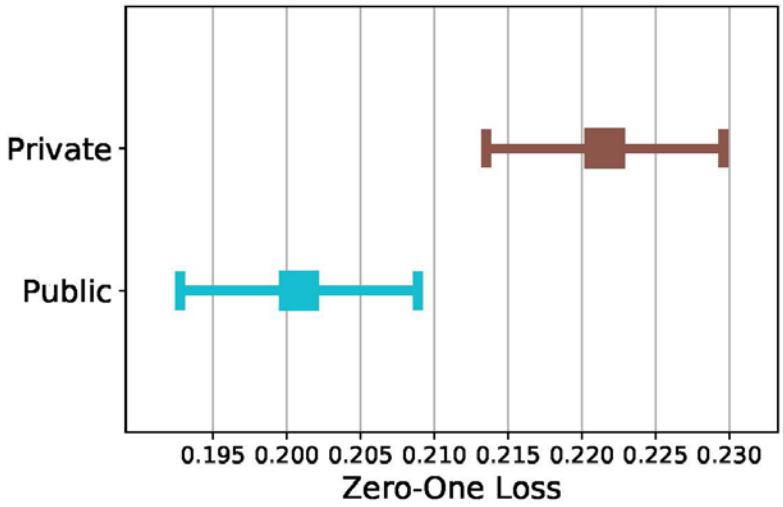
We show consistent gender differences across data sets in Figures 1 and 4, with the highest error rates for female patients, although the difference in error rates between genders was only statistically significant for ICU mortality. Note that because of the smaller size of the psychiatric notes data set, the confidence intervals overlap; however, the heterogeneity in topic enrichment values aligns with the higher error rates for female patients.

Figure 4. 95% Confidence Intervals for Error Rate (Zero-One Loss) in Psychiatric Readmission for Gender



Interestingly, model prediction errors for insurance type were statistically significant for both data sets (Figures 2 and 5), but the group with highest error rate changes. While public insurance patients have the highest error rate for ICU mortality, private insurance patients have the highest error rate for psychiatric readmission.

Figure 5. 95% Confidence Intervals for Error Rate (Zero-One Loss) in Psychiatric Readmission for Insurance Type



These differences in error rates for insurance type may indicate that insurance type affects patient care in ICU and psychiatric settings differently. We note that public insurance patients have higher baseline hospital mortality rates, whereas private insurance patients have higher baseline 30-day psychiatric readmission (see Supplementary Appendix Table S1). Such variation in baseline rates could be due to the previously noted prevalence of chronic conditions in public insurance patients,[45] making these patients more likely to need the ICU for regular care of multiple chronic conditions. Public insurance patients are also more likely to have serious mental illness than private insurance patients,[39] indicating that they may not come into a psychiatric hospital unless the situation is dire. In both data sets, predictions are better captured by notes for patients in the group that uses the care setting more regularly (ie, public insurance patients in the ICU and private insurance patients in the psychiatric hospital).

Responding to Algorithmic Biases in Machine Learning

AI and machine learning may enable faster, more accurate, and more comprehensive health care. We believe a closely cooperative relationship between clinicians and AI—rather than a competitive one[53]—is necessary for illuminating areas of disparate health care impact.[51] For example, a clinician should be able to provide feedback for the algorithm to implement, and the algorithm could actively query the clinician about

uncertain cases. Indeed, algorithmic scrutiny is vital to both the short-term and long-term robustness of the health care system.

In this paper, we have considered questions related to the disparate impact that AI may have in health care—in particular, on ICU mortality and 30-day psychiatric readmissions. Based on clinical notes, we demonstrated heterogeneity in the topics emphasized across race, gender, and insurance type, which tracks with known health disparities. We also showed statistically significant differences in error rates in ICU mortality for race, gender, and insurance type and in 30-day psychiatric readmission for insurance type.

In light of known clinical biases, how can AI assist in improving patient care? With increasing involvement of machine learning in health care decisions, it is crucial to assess any algorithmic biases introduced[54] by comparing prediction accuracy between demographic groups. Once algorithmic bias is uncovered, clinicians and AI must work together to identify the sources of algorithmic bias and improve models through better data collection and model improvements.

References

1. Oh SS, Galanter J, Thakur N, et al. Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS Med*. 2015;12(12):e1001918.
2. Mary Horrigan Connors Center, Brigham and Women's Hospital. Ten years of health advancements for women of all ages, ethnicities and nations. https://www.brighamandwomens.org/assets/BWH/womens-health/pdfs/connors-center-ten-year-report.pdf. Accessed August 9, 2018.
3. Johnson KS. Racial and ethnic disparities in palliative care. *J Palliat Med*. 2013;16(11):1329-1334.
4. Phelan SM, Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. *Obes Rev*. 2015;16(4):319-326.
5. Calderone KL. The influence of gender on the frequency of pain and sedative medication administered to postoperative patients. *Sex Roles*. 1990;23(11-12):713-725.
6. Bartley EJ, Fillingim RB. Sex differences in pain: a brief review of clinical and experimental findings. *Br J Anaesth*. 2013;111(1):52-58.
7. Hoffmann DE, Tarzian AJ. The girl who cried pain: a bias against women in the treatment of pain. *J Law Med Ethics*. 2001;29(1):13-27.
8. Tucker MJ, Berg CJ, Callaghan WM, Hsia J. The black-white disparity in pregnancy-related mortality from 5 conditions: differences in prevalence and case-fatality rates. *Am J Public Health*. 2007;97(2):247-251.
9. Howell EA. Reducing disparities in severe maternal morbidity and mortality. *Clin Obstet Gynecol*. 2018;61(2):387-399.

10. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.
12. Liao KP, Cai T, Gainer V, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)*. 2010;62(8):1120-1127.
13. Ghassemi M, Naumann T, Doshi-Velez F, et al. Unfolding physiological state: mortality modelling in intensive care units. *KDD*. 2014;2014:75-84.
14. Ghassemi M, Wu M, Hughes MC, Szolovits P, Doshi-Velez F. Predicting intervention onset in the ICU with switching state space models. *AMIA Jt Summits Transl Sci Proc*. 2017;2017:82-91.
15. Ferryman K, Pitcan M. Fairness in precision medicine. Data & Society. https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf. Published February 26, 2018. Accessed August 9, 2018.
16. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. *ProPublica*. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed August 9, 2018.
17. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. *Nature*. 2018;559(7714):324-326.
18. Ghassemi M, Naumann T, Schulam P, Beam AL, Ranganath R. Opportunities in machine learning for healthcare. arXiv. https://arxiv.org/abs/1806.00388. Published June 1, 2018. Updated June 5, 2018. Accessed August 9, 2018.
19. Yarnell CJ, Fu L, Manuel D, et al. Association between immigrant status and end-of-life care in Ontario, Canada. *JAMA*. 2017;318(15):1479-1488.
20. Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci USA*. 2018;115(16):e3635-e3644.
21. Williams DR, Costa M, Leavell JP. Race and mental health: patterns and challenges. In: Scheid TL, Brown, TN, eds. *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems*. 3rd ed. New York, NY: Cambridge University Press; 2017:281-304.
22. Priest N, Williams DR. Racial discrimination and racial disparities in health. In: Major B, Dovidio JF, Link BG, eds. *The Oxford Handbook of Stigma, Discrimination, and Health*. New York, NY: Oxford University Press; 2017:163-182.
23. Chen I, Johansson FD, Sontag D. Why is my classifier discriminatory? arXiv. https://arxiv.org/abs/1805.12002. Published May 30, 2018. Accessed August 9, 2018.
24. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.

25. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
26. Marlin BM, Kale DC, Khemani RG, Wetzel RC. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; January 28-30, 2012; Miami, FL:389-398.
27. Riddell A. lda: topic modeling with latent Dirichlet allocation. GitHub. https://github.com/lda-project/lda. Accessed August 6, 2018.
28. McCallum AK. Mallet: a machine learning for language toolkit. University of Massachusetts Amherst. http://mallet.cs.umass.edu/. Published 2002. Accessed August 6, 2018.
29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
30. Jones KS. A statistical interpretation of term specificity and its application in retrieval. *J Doc*. 1972;28(1):11-21.
31. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145-1159.
32. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On fairness and calibration. In: Proceedings of the 31st International Conferences on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA:5684-5693.
33. Tukey JW. Comparing individual means in the analysis of variance. *Biometrics*. 1949;5(2):99-114.
34. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference; June 28-July 3, 2010; Austin, TX:57-61.
35. Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. In: Proceedings of the 2nd Machine Learning for Healthcare Conference; 2017; Boston, MA:361-376.
36. Smith K. Gender differences in primary substance of abuse across age groups. In: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, US Department of Health and Human Services. *The CBHSQ Report*. Rockville, MD: Substance Abuse and Mental Health Services Administration; 2014:1-18. https://www.ncbi.nlm.nih.gov/books/NBK384845/. Accessed August 9, 2018.
37. Kuehner C. Why is depression more common among women than among men? *Lancet Psychiatry*. 2017;4(2):146-158.
38. Leiknes KA, Jarosh-von Schweder L, Høie B. Contemporary use and practice of electroconvulsive therapy worldwide. *Brain Behav*. 2012;2(3):283-344.
39. Rowan K, McAlpine DD, Blewett LA. Access and cost barriers to mental health care, by insurance status, 1999-2010. *Health Aff (Millwood)*. 2013;32(10):1723-1730.

40. Chapman KR, Tashkin DP, Pye DJ. Gender bias in the diagnosis of COPD. *Chest*. 2001;119(6):1691-1695.
41. Han MK, Postma D, Mannino DM, et al. Gender and chronic obstructive pulmonary disease: why it matters. *Am J Respir Crit Care Med*. 2007;176(12):1179-1184.
42. Thompson CA, Gomez SL, Hastings KG, et al. The burden of cancer in Asian Americans: a report of national mortality trends by Asian ethnicity. *Cancer Epidemiol Biomarkers Prev*. 2016;25(10):1371-1382.
43. Carrion AF, Ghanta R, Carrasquillo O, Martin P. Chronic liver disease in the Hispanic population of the United States. *Clin Gastroenterol Hepatol*. 2011;9(10):834-841.
44. Shen AY, Contreras R, Sobnosky S, et al. Racial/ethnic differences in the prevalence of atrial fibrillation among older adults—a cross-sectional study. *J Natl Med Assoc*. 2010;102(10):906-913.
45. Fox MH, Reichard A. Disability, health, and multiple chronic conditions among people eligible for both Medicare and Medicaid, 2005-2010. *Prev Chronic Dis*. 2013;10:e157.
46. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference; January 8-10, 2012; Cambridge, MA:214-226.
47. Hardt M, Price E, Srebro N, et al. Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems; December 5-10, 2016; Barcelona, Spain:3323-3331.
48. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: Proceedings of the 30th International Conference on Machine Learning; June 16-21, 2013; Atlanta, GA:325-333.
49. Kearns MJ, Neel S, Roth A, Wu ZS. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: Proceedings of the 35th International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden:2564-2572.
50. Rumshisky A, Ghassemi M, Naumann T, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry*. 2016;6(10):e921.
51. Schmutte T, Dunn CL, Sledge WH. Predicting time to readmission in patients with recent histories of recurrent psychiatric hospitalization: a matched-control survival analysis. *J Nerv Ment Dis*. 2010;198(12):860-863.
52. Nelson EA, Maruish ME, Axler JL. Effects of discharge planning and compliance with outpatient appointments on readmission rates. *Psychiatr Serv*. 2000;51(7):885-889.
53. AI versus doctors [news]. *IEEE Spectr*. 2017;54(10):13.
54. Miller AP. Want less-biased decisions? Use algorithms. *Harvard Business Review*. July 26, 2018. https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms. Accessed August 9, 2018.

Irene Y. Chen is a doctoral student in electrical engineering and computer science at the Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts. She is also pursuing a graduate education in medical sciences certificate in the Harvard-MIT Program in Health Sciences and Technology. She received a bachelor of arts degree in applied math-economics and computer science and a master of science degree in computational science and engineering from Harvard University.

Peter Szolovits, PhD is a professor of computer science and engineering and the head of the Clinical Decision-Making Group within the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts. He is also an associate member of the MIT Institute for Medical Engineering and Science and on the faculty of the Harvard-MIT Program in Health Sciences and Technology.

Marzyeh Ghassemi, PhD is an assistant professor of computer science and medicine at the University of Toronto and a faculty member at the Vector Institute, both in in Ontario, Canada. She previously served as a visiting researcher at Alphabet Inc. within its life sciences research organization, Verily, and as a postdoctoral fellow at the Massachusetts Institute of Technology, where she earned a PhD in electrical engineering and computer science.

Citation
AMA J Ethics. 2019;21(2):E167-179.

DOI
10.1001/amajethics.2019.167.

Acknowledgements
The authors thank Willie Boag and Tristan Naumann at MIT for help wrangling the data. This work was supported in part by a grant from the National Institute of Mental Health (1R01MH106577).

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®
February 2019, Volume 21, Number 2: E180-187

POLICY FORUM

What Are Important Ethical Implications of Using Facial Recognition Technology in Health Care?

Nicole Martinez-Martin, JD, PhD

Abstract

Applications of facial recognition technology (FRT) in health care settings have been developed to identify and monitor patients as well as to diagnose genetic, medical, and behavioral conditions. The use of FRT in health care suggests the importance of informed consent, data input and analysis quality, effective communication about incidental findings, and potential influence on patient-clinician relationships. Privacy and data protection are thought to present challenges for the use of FRT for health applications.

Promises and Challenges of Facial Recognition Technology

Facial recognition technology (FRT) utilizes software to map a person's facial characteristics and then store the data as a face template.[1] Algorithms or machine learning techniques are applied to a database to compare facial images or to find patterns in facial features for verification or authentication purposes.[2] FRT is attractive for a variety of health care applications, such as diagnosing genetic disorders, monitoring patients, and providing health indicator information (related to behavior, aging, longevity, or pain experience, for example).[3-5]

FRT is likely to become a useful tool for diagnosing many medical and genetic conditions.[6,7] Machine learning techniques, in which a computer program is trained on a large data set to recognize patterns and generates its own algorithms on the basis of learning,[8] have already been used to assist in diagnosing a patient with a rare genetic disorder that had not been identified after years of clinical effort.[9] Machine learning can also detect more subtle correlations between facial morphology and genetic disorders than clinicians.[4] It is thought that FRT can therefore eventually be used to assist in earlier detection and treatment of genetic disorders,[10,11] and computer applications (commonly known as apps) such as Face2Gene have been developed to assist clinicians in diagnosing genetic disorders.[12]

FRT has other potential health care applications. FRT is being developed to predict health characteristics, such as longevity and aging.[13] FRT is also being applied to predict behavior, pain, and emotions by identifying facial expressions associated with depression

or pain, for example.[14,15] Another major area for FRT applications in health care is patient identification and monitoring, such as monitoring elderly patients for safety or attempts to leave a health care facility[16] or monitoring medication adherence through the use of sensors and facial recognition to confirm when patients take their medications.[17]

As with any new health technology, careful attention should be paid to the accuracy and validity of FRT used in health care applications as well as to informed consent and reporting incidental findings to patients. FRT in health care also raises ethical questions about privacy and data protection, potential bias in the data or analysis, and potential negative implications for the therapeutic alliance in patient-clinician relationships.

Ethical Dimensions of FRT in Health Care

Informed consent. FRT tools that assist with identification, monitoring, and diagnosis are expected to play a prominent role in the future of health care.[6,18] Some applications have already been implemented.[13,19] As FRT is increasingly utilized in health care settings, informed consent will need to be obtained not only for collecting and storing patients' images but also for the specific purposes for which those images might be analyzed by FRT systems.[20] In particular, patients might not be aware that their images could be used to generate additionally clinically relevant information. While FRT systems in health care can de-identify data, some experts are skeptical that such data can be truly anonymized[21]; from clinical and ethical perspectives, informing patients about this kind of risk is critical.

Some machine learning systems need continuous data input to train and improve the algorithms[22] in a process that could be analogized to quality improvement research, for which informed consent is not regarded as necessary.[23] For example, to improve its algorithms, FRT for genetic diagnosis would need to receive new data sets of images of patients already known to have specific genetic disorders.[2] To maintain trust and transparency with patients, organizations should consider involving relevant community stakeholders in implementing FRT and in decisions about establishing and improving practices of informing patients about the organization's use of FRT. As FRT becomes capable of detecting a wider range of health conditions, such as behavioral[24] or developmental disorders,[25] health care organizations and software developers will need to decide which types of analyses should be included in a FRT system and the conditions under which patients might need to be informed of incidental findings.

Bias. As with any clinical innovation, FRT tools should be expected to demonstrate accuracy for specific uses and to demonstrate that overall benefits outweigh risks.[26] Detecting and evaluating bias in data and results should also receive close ethical scrutiny.[27] In machine learning, the quality of the results reflects the quality of data input to the system[28]—an issue sometimes referred to as "garbage in, garbage out." For example, when images used to train software are not drawn from a pool that is

sufficiently racially diverse, the system may produce racially biased results.[29] If this happens, FRT diagnostics might not work as well for some racial or ethnic groups as others. One recent example that gained notoriety was an FRT system used to identify gay men from a set of photos that may have simply identified the kind of grooming and dress habits stereotypically associated with gay men.[30] The developers of this FRT system did not intend it to be used for a clinical purpose but rather to illustrate how bias can influence FRT findings.[30]

Thankfully, potential solutions for addressing bias in FRT systems exist. These include efforts to create AI systems that explain the rationale behind the results generated.[31] Clinicians can also be trained to consider and respond to limitations and biases of FRT systems.[32] In addition, organizations such as the National Human Genome Research Institute have sought to diversify the range of people whose images are included in their image databases.[33]

Patient privacy. FRT raises novel challenges regarding privacy. FRT systems can store data as a complete facial image or as a facial template.[34] Facial templates are considered biometric data and thus personally identifiable information.[35] The idea that a photo can reveal private health information is relatively new, and privacy regulations and practices are still catching up. A few states, such as Illinois, have regulations that limit uses for which consumer biometric data can be collected.[36] The Health Insurance Portability and Accountability Act (HIPAA) governs handling of patients' health records and personal health information and includes privacy protections for personally identifiable information. More specifically, it protects the privacy of biometric data, including "full-face photographs and any comparable images," which are "directly related to an individual."[37] Thus, facial images used for FRT health applications would be protected by HIPAA.[38] Entities covered by HIPAA, including health care organizations, clinicians, and third-party business associates, would need to comply with HIPAA regulations regarding the use and disclosure of protected health information.[38] However, clinicians should advise patients that there may be limited protections for storing and sharing data when using a consumer FRT tool.

Some statutes that protect health information might not apply to FRT. The Genetic Information Nondiscrimination Act (GINA) of 2008, for example, does not apply to FRT for genetic diagnosis, as FRT does not fit GINA's definition of genetic testing or genetic information.[39] The Americans with Disabilities Act of 1990, which protects people with disabilities from discrimination in public life (eg, schools or employment),[40] would also likely not apply to FRT used for diagnostic purposes if the conditions diagnosed are currently unexpressed. Employers might also be interested in using FRT tools to predict mood or behavior as well as to predict longevity, particularly for use in wellness programs to lower employers' health care costs.

Broader influence of FRT. There will need to be careful thought and study of the broader impact of FRT in health care settings. One potential issue is that of liability. For example, if FRT diagnostic software develops to the point that it is used not just to augment but to replace a physician's judgment, ethical and legal questions may arise regarding which entity appropriately has liability.[41] Or if FRT is used to monitor compliance, track patients' whereabouts, or assist in other kinds of surveillance, patients' trust in physicians could be eroded, undermining the therapeutic alliance. It is therefore important to weigh the relative benefits and burdens of specific FRT uses in health care and to conduct research into how patients perceive its use. On the one hand, the use of FRT to monitor the safety of dementia patients could be perceived as having benefits that outweigh the burdens of surveillance. On the other, FRT medication adherence monitoring might not be sufficiently effective in improving adherence to outweigh the risk of undermining trust in the patient-physician relationship.[42]

As considered here, numerous applications of FRT in health care settings suggest the ethical, clinical, and legal importance of informed consent, data input and analysis quality, effective communication about incidental findings, and potential influence on patient-clinician relationships. Privacy and data protections are key to advancing FRT and making it helpful.

References

1. Gates KA. *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. New York, NY: New York University Press; 2011.
2. Parmar DN, Mehta BB. Face recognition methods and applications. arXiv. http://arxiv.org/abs/1403.0485. Published March 3, 2014. Accessed July 24, 2018.
3. Stephen ID, Hiew V, Coetzee V, Tiddeman BP, Perrett DI. Facial shape analysis identifies valid cues to aspects of physiological health in Caucasian, Asian, and African populations. *Front Psychol*. 2017;8:1883.
4. Chen S, Pan ZX, Zhu HJ, et al. Development of a computer-aided tool for the pattern recognition of facial features in diagnosing Turner syndrome: comparison of diagnostic accuracy with clinical workers. *Sci Rep*. 2018;8(1):9317.
5. Hossain MS, Muhammad G. Cloud-assisted speech and face recognition framework for health monitoring. *Mob Netw Appl*. 2015;20(3):391-399.
6. Sandolu A. Why facial recognition is the future of diagnostics. *Medical News Today*. December 9, 2017. https://www.medicalnewstoday.com/articles/320316.php. Accessed May 24, 2018.
7. Loos HS, Wieczorek D, Würtz, RP, von der Malsburg C, Horsthemke B. Computer-based recognition of dysmorphic faces. *Eur J Hum Genet*. 2003;11(8):555-560.
8. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *JAMA*. 2016;315(6):551-552.

9. Molteni M. Thanks to AI, computers can now see your health problems. *Wired*. January 9, 2017. https://www.wired.com/2017/01/computers-can-tell-glance-youve-got-genetic-disorders/. Accessed April 8, 2017.
10. Kosilek RP, Schopohl J, Grunke M, et al. Automatic face classification of Cushing's syndrome in women—a novel screening approach. *Exp Clin Endocrinol Diabetes*. 2013;121(9):561-564.
11. Schneider HJ, Kosilek RP, Günther M, et al. A novel approach to the detection of acromegaly: accuracy of diagnosis by automatic face classification. *J Clin Endocrinol Metab*. 2011;96(7):2074-2080.
12. Basel-Vanagaite L, Wolf L, Orin M, et al. Recognition of the Cornelia de Lange syndrome phenotype with facial dysmorphology novel analysis. *Clin Genet*. 2016;89(5):557-563.
13. Mack H. FDNA launches app-based tool for clinicians using facial recognition, AI and genetic big data to improve rare disease diagnosis and treatment. *MobiHealthNews*. March 21, 2017. http://www.mobihealthnews.com/content/fdna-launches-app-based-tool-clinicians-using-facial-recognition-ai-and-genetic-big-data. Accessed March 28, 2017.
14. Bahrampour T. Can your face reveal how long you'll live? New technology may provide the answer. *Washington Post*. July 2, 2014. https://www.washingtonpost.com/national/health-science/can-your-face-reveal-how-long-youll-live-new-technology-may-provide-the-answer/2014/07/02/640bacb4-f748-11e3-a606-946fd632f9f1_story.html. Accessed July 25, 2018.
15. Shakya S, Sharma S, Basnet A. Human behavior prediction using facial expression analysis. In: Proceedings of the 2016 International Conference on Computing, Communication and Automation (ICCCA); April 29-30, 2016; Noida, India:399-404.
16. Bina RW, Langevin JP. Closed loop deep brain stimulation for PTSD, addiction, and disorders of affective facial interpretation: review and discussion of potential biomarkers and stimulation paradigms. *Front Neurosci*. 2018;12:300.
17. Hossain MS, Muhammad G. Cloud-assisted framework for health monitoring. In: Proceedings of the 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE); May 3-6, 2015; Halifax, Canada:1199-1202.
18. Baum S. Using facial recognition and AI to confirm medication adherence, AiCure raises $12.25M. *MedCity News*. January 12, 2016. https://medcitynews.com/2016/01/aicure-fundraise/. Accessed July 25, 2018.
19. Wicklund E. EHR provider touts mHealth access by Apple's facial recognition app. mHealthIntelligence. https://mhealthintelligence.com/news/ehr-provider-touts-mhealth-access-by-apples-facial-recognition-app. Published November 9, 2017. Accessed October 26, 2018.

20. Balthazar P, Harri P, Prater A, Safdar NM. Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. *J Am Coll Radiol*. 2018;15(3, pt B):580-586.
21. Mohapatra S. Use of facial recognition technology for medical purposes: balancing privacy with innovation. *Pepperdine Law Rev*. 2016;43(4):1017-1064.
22. Watson M. Keeping your machine learning models up-to-date: continuous learning with IBM Watson machine learning (part 1). *Data Lab*. March 2018. https://medium.com/ibm-watson-data-lab/keeping-your-machine-learning-models-up-to-date-f1ead546591b. Accessed October 26, 2018.
23. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Aff (Millwood)*. 2014;33(7):1139-1147.
24. Wen L, Li X, Guo G, Zhu Y. Automated depression diagnosis based on facial dynamic analysis and sparse coding. *IEEE Trans Inf Forensics Secur*. 2015;10(7):1432-1441.
25. Borsos Z, Gyori M. Can automated facial expression analysis show differences between autism and typical functioning? *Stud Health Technol Inform*. 2017;242:797-804.
26. Ghaemi SN, Goodwin FK. The ethics of clinical innovation in psychopharmacology: challenging traditional bioethics. *Philos Ethics Humanit Med*. 2007;2(1):26.
27. Tunkelang D. Ten things everyone should know about machine learning. *Forbes*. September 6, 2017. https://www.forbes.com/sites/quora/2017/09/06/ten-things-everyone-should-know-about-machine-learning. Accessed January 13, 2018.
28. McCullom R. Facial recognition technology is both biased and understudied. *Undark*. May 17, 2017. https://undark.org/article/facial-recognition-technology-biased-understudied/. Accessed July 31, 2018.
29. Researchers flag up facial recognition racial bias [news]. *Biom Technol Today*. 2016;2016(5):2-3.
30. Agüera y Arcas B, Todorov A, Mitchell M. Do algorithms reveal sexual orientation or just expose our stereotypes? *Medium*. January 11, 2018. https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477. Accessed July 17, 2018.
31. Knight W. There's a big problem with AI: even its creators can't explain how it works. *MIT Technology Review*. April 11, 2017. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/. Accessed March 12, 2018.
32. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *New Engl J Med*. 2018;378(11):981-983.
33. National Human Genome Research Institute, National Institutes of Health. Atlas of human malformation syndromes in diverse populations.

https://research.nhgri.nih.gov/atlas/. Updated October 19, 2016. Accessed March 28, 2017.
34. Mazura JC, Juluru K, Chen JJ, Morgan TA, John M, Siegel EL. Facial recognition software success rates for the identification of 3D surface reconstructed facial images: implications for patient privacy and security. *J Digit Imaging*. 2012;25(3):347-351.
35. Brostoff G. 3D facial recognition gives healthcare data a new look. *Clinical Informatics News*. October 6, 2017. http://www.clinicalinformaticsnews.com/2017/10/06/3d-facial-recognition-gives-healthcare-data-a-new-look.aspx. Accessed July 17, 2018.
36. Hughes N. Google takes aim at controversial, stringent Illinois biometric privacy law. One World Identity. https://oneworldidentity.com/google-takes-aim-controversial-stringent-illinois-biometric-privacy-law/. Published April 25, 2018. Accessed August 1, 2018.
37. US Department of Health and Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed December 10, 2018.
38. Health Insurance Portability and Accountability Act of 1996, Pub L No. 104-191, 110 Stat 1936.
39. Genetic Information Nondiscrimination Act of 2008, Pub L No. 110-233, 122 Stat 881.
40. Americans with Disabilities Act of 1990, Pub L No. 101-336, 104 Stat 327.
41. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983.
42. Martinez-Martin N, Char D. Surveillance and digital health. *Am J Bioeth*. 2018;18(9):67-68.

Nicole Martinez-Martin, JD, PhD is a postdoctoral fellow at the Stanford Center for Biomedical Ethics in Stanford, California. She earned a JD from Harvard Law School and a PhD from the University of Chicago in comparative human development. Her research focuses on neuroethics as well as the ethics of digital health technology and machine learning with a focus on mental health issues and special populations.

Citation
AMA J Ethics. 2019;21(2):E180-187.

DOI
10.1001/amajethics.2019.180.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E188-191

MEDICINE AND SOCIETY

Making Policy on Augmented Intelligence in Health Care

Elliott Crigger, PhD and Christopher Khoury, MSc, MBA

Abstract

In June 2018, the American Medical Association adopted new policy to provide a broad framework for the evolution of artificial intelligence (AI) in health care that is designed to help ensure that AI realizes the benefits it promises for patients, physicians, and the health care community.

Responding to Industry Activity on Artificial Intelligence

In June 2018, the American Medical Association (AMA) adopted a new policy, H-480.940, "Augmented Intelligence in Health Care,"[1] to provide a broad framework for the evolution of artificial intelligence (AI) in health care that is designed to help ensure that AI realizes the benefits it promises for patients, physicians, and the health care community. In parallel, a wave of scientific and investment activity is cresting, focused on AI and its applications in health care and medicine. Advances in computing power, storage, sensors, and multidisciplinary research have laid the groundwork for the increased development and use of AI techniques in health care.

AI is the ability of a computer to function appropriately and with foresight in its environment, that is, to complete tasks in a manner typically associated with a rational human being. *Augmented intelligence* is an alternative conceptualization that focuses on AI's assistive role, emphasizing a design approach and implementation that enhances human intelligence rather than replaces it. Collectively, these areas of scientific research and health care industry activity represent thousands of peer-reviewed studies and, by one estimate, over $2.7 billion in investment across 121 digital health companies and 206 funding deals between 2011 and 2017, just within the United States.[2]

The background report that informs this new policy focused on 2 fundamental normative conditions for appropriate integration of AI into health care.[3] First, health care AI should be understood as a tool to augment professional clinical judgment, not a technology to replace or override it. Second, the development of health care AI tools should attend carefully to the design and evaluation of individual applications, to issues of patient privacy, and to thoughtful clinical implementation.

Ethical Dimensions of AI in Health Care Practice

Design challenges loom large. An AI-derived algorithm is only as good as the data with which it works.

> The research, patient care, and insurance records available as training data sets for health care AI can be highly variable.... Clinical trials systematically include or exclude participants with certain characteristics; patient charts and insurance records capture information only from those individuals who have access to the health care system.[3]

Rarely do such records contain information about social determinants of health, for example. AI systems can, invisibly and unintentionally, reproduce or magnify the biases of their human designers or training data sets in ways that risk exacerbating existing health disparities—such as when data reflect only the conditions of individuals who have access to health care to begin with.[4]

Conversely, algorithms properly designed and deployed can help compensate for or minimize human bias. AI algorithms must also be evaluated using criteria that are "clinically relevant and evaluation should be representative of how the algorithm will be applied in practice."[3] Predictive algorithms, for example, should be able to predict events sufficiently in advance to meaningfully influence care decisions and patient outcomes.

Addressing concerns about privacy and security are 2 further challenges for the evolution of health care AI. Existing practices of notifying patients and obtaining consent for data use are not adequate, nor are strategies for de-identifying data effective in the context of large, complex data sets when machine learning algorithms can re-identify a record from as few as 3 data points.[5] Algorithms can also be vulnerable to cyberattack.[6] Evolving technical responses to the challenge of ensuring data security and integrity, such as blockchain-style technologies,[7] are promising, but traditional expectations for health care privacy might no longer be attainable. Rigorous oversight of data use and transfer will be critical to protecting patients' interests.

To realize its promise, health care AI must be deployed in ways that promote quality of care and minimize potentially disruptive effects. Physicians will need to learn to work effectively with AI systems,[3] just as medical students and residents are now trained to work effectively with electronic health records.[8] If physicians are to base clinical recommendations on AI, "they will need to understand AI methods and systems sufficiently to be able to trust an algorithm's predictions."[3] Even as technical solutions to the problem of trust evolve, such as algorithms that "explain" to users why a particular prediction has been made, the health care organizations that implement AI systems should vigilantly monitor the operation of those systems to identify and address adverse consequences.

Legal experts and commercial developers of AI tools that aid in diagnosis must also begin to address questions of liability when incorrect diagnoses are made either by humans using augmented intelligence tools or by AI tools directly. Questions also remain about the evolving role of the patient-physician relationship and fiduciary compact in an algorithm-enabled health care environment.[9]

The AMA's adoption of H-480.940 suggests the ethical importance of these questions in calling for development of thoughtfully designed, high-quality, clinically validated health care AI that does the following[1]:

a) is designed and evaluated in keeping with best practices in user-centered design, particularly for physicians and other members of the health care team;
b) is transparent;
c) conforms to leading standards for reproducibility;
d) identifies and takes steps to address bias and avoids introducing or exacerbating health care disparities including when testing or deploying new AI tools on vulnerable populations; and
e) safeguards patients' and other individuals' privacy interests and preserves the security and integrity of personal information.

Values of ethical relevance considered in this policy include professionalism, transparency, justice, safety, and privacy.

References

1. American Medical Association. Augmented intelligence in health care H-480.940. https://policysearch.ama-assn.org/policyfinder/detail/augmented%20intelligence?uri=%2FAMADoc%2FHOD.xml-H-480.940.xml. Modified 2018. Accessed November 14, 2018.
2. Zweig M, Tran D. 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/. Accessed November 30, 2018.
3. American Medical Association. Augmented intelligence in health care: report 41 of the AMA Board of Trustees. https://static1.squarespace.com/static/58d0113a3e00bef537b02b70/t/5b6aed0a758d4610026a719c/1533734156501/AI_2018_Report_AMA.pdf. Accessed November 14, 2018.
4. AI Now Institute. The AI Now report: the social and economic implications of artificial intelligence technologies in the near-term. New York, NY: AI Now Institute; 2016. https://ainowinstitute.org/AI_Now_2016_Report.pdf. Updated September 22, 2016. Accessed January 26, 2018.
5. Osoba O, Welser W IV. An intelligence in our image: the risks of bias and errors in artificial intelligence. Santa Monica, CA: Rand Corporation; 2017. https://www.rand.org/pubs/research_reports/RR1744.html. Accessed February 19, 2018.

6. Finlayson SG, Won Chung H, Kohan IS, Beam AL. Adversarial attacks against medical deep learning systems. arXiv. https://arxiv.org/abs/1804.05296. Published April 15, 2018. Updated May 21, 2018. Accessed August 9, 2018.
7. JASON. Artificial intelligence for health and health care. McClean, VA: MITRE Corporation; 2017. https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf. Accessed February 19, 2018.
8. Miliard M. AMA, Regenstrief launch EHR training platform for medical students. *Healthcare IT News*. April 19, 2017. https://www.healthcareitnews.com/news/ama-regenstrief-launch-ehr-training-platform-medical-students. Accessed December 5, 2018.
9. Char Ds, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983.

Elliott Crigger, PhD is director of ethics policy and secretary to the Council on Ethical and Judicial Affairs at the American Medical Association in Chicago, Illinois.

Christopher Khoury, MSc, MBA is vice president of the Environmental Intelligence and Strategic Analytics unit at the American Medical Association in Washington, DC. The unit focuses on assessing and interacting with emerging elements across the health care, business, and policy sectors. He holds degrees in electrical engineering, biomedical engineering, and business.

Citation
AMA J Ethics. 2019;21(2):E188-191.

DOI
10.1001/amajethics.2019.188.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E192-195

ART OF MEDICINE

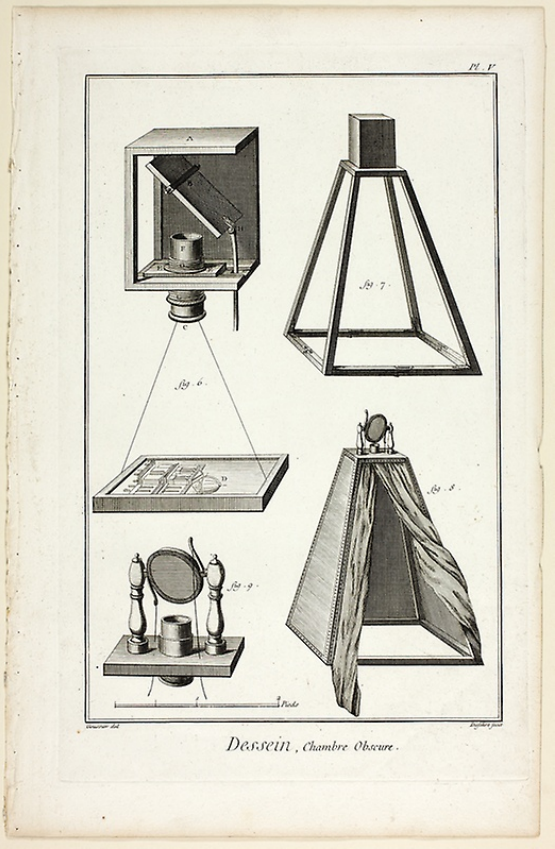
What Do Warhol, Pollock, and Murakami Teach Us About AI in Health Care?

Sam Anderson-Ramos, MFA

Abstract

As with medicine, artistic practice has a historical relationship with technologies. As technology advances, artists and medical practitioners will struggle with the complexities of introducing artificial intelligence into pursuits that have long been defined as fundamentally human. How will intelligent mechanization continue to aid efforts in art and medicine, even as it complicates them? Which new dilemmas will arise as essentially human pursuits are ever more deeply aligned with the rise of thinking machines?

Figure. *Design: Camera Obscura, from Encylopédie* (1762-1777), by A. J. Defehrt, after Louis-Jaques Goussier



Media
Engraving on cream laid paper.

The romantic myth of genius artists toiling away in a garret, painting or sculpting purely from their luminous imagination, has been embraced in the Western tradition at least since the 15th and 16th centuries. However, it was a misleading narrative even then. Not only did artists often not work alone, but they didn't always use their own hands, much less their own imaginations. The camera obscura harnessed light to project traceable images of nearby objects[1] and was used to great effect by 17th-century Dutch masters. In the 19th century, the photographic camera revolutionized the way artists understood and created images; we are only beginning to see the artistic potential of computers, 3D printers, the internet, and artificial intelligence (AI).

Take, for example, Andy Warhol's *Big Electric Chair*, a screen print made from a found news photograph of the execution chamber at Sing Sing Correctional Facility in New York State. If you look closely, you notice that the green image doesn't line up with its linen surface. Instead, it is skewed down and to the right, making for conspicuous green absences on all 4 sides and, in particular, at the bottom left, where the corner is sliced away. What looks like an accident speaks to the mechanized process by which the image was made. What better way to emphasize the artificiality of the screen printing process—anathema to painterly "geniuses" like Jackson Pollock in the 1950s—than to flaunt a byproduct of an art practice more akin to mass production than spontaneous creation? It does not take much to shift from considering the repetitive, assembly line printing process to the violent delivery of death implied by the electric chair itself. What we end up with is a layering of tools or technologies, a continuum that includes the law itself, the prison-industrial complex, the chair that destroys the condemned flesh it touches, the camera that captures the image, and the (commodified) art object that viewers encounter in the gallery. As each layer in this continuum is concerned with justice, a virtue presumably informed by not only judgment but also compassion, Warhol ultimately confronts the reality of a modern industrial process operating as mechanism for suffering and death. *Big Electric Chair* asks: Is American justice ethical?

Warhol did not shy away from embracing mechanization and its attendant technologies—he named his Manhattan studio the Factory, no less—and neither have artists in the 21st century. The question, "What is art?" has been prodded to the point of meaninglessness since Warhol's time, and the importance of technology in the creative process is largely taken for granted. However, as it does for the medical community, AI poses fresh challenges for the arts. An art object is typically understood as the product of a series of creative choices planned and executed by an individual or group. But what if the entity making the creative choices is an artificial one, such as a highly developed algorithm designed to make judgments based on what it has learned? In this case we are not only asking "What is art?" but also "What is an artist?" and "What is creativity?" The

ensuing cascade of doubts and conundrums is as daunting as any of our most lingering metaphysical dilemmas.

Artistic practice, as with medicine, is a human endeavor, based ultimately on person-to-person communication. AI will permanently complicate that dynamic. Can a computer make artwork that expresses and teaches the human experience? Thankfully, the art world has grown steadily more primed for just such existential debates. In fact, some great art has resulted from them. The Japanese artist Takashi Murakami is notable for combining factory production, commercial distribution, and computer design tools to create work that synthesizes 21st-century pop culture and the arts of classical Japan. Artists like Murakami have been increasingly willing to remove their hands from the creative process, implicitly or explicitly critiquing the role of an artist's control in their own creative output.

However, the health sector does not deal in hypotheticals. Should a surgeon rely on AI to determine where to make her first incision? If a life is lost as a consequence of utilizing AI, who (or what) should be held accountable? As technologies evolve and become more capable of making their own choices, these issues will only grow more complex, more urgent, and more consequential. We might be on a path toward a future dangerously dependent on intelligent software, a scenario that suggests cause for skepticism, if not resistance. On the other hand, we might be destined for something brighter: a courageous future teeming with brilliant, as yet unimagined, innovations in art and medicine driven by compassion and aided by machines that think.

References

1. Photography History Facts. History of camera obscura—who invented camera obscura? http://www.photographyhistoryfacts.com/photography-development-history/camera-obscura-history/. Accessed July 23, 2018.

Sam Anderson-Ramos, MFA is the assistant director for college and professional learning in the Department of Learning and Public Engagement at the Art Institute of Chicago. His fiction, essays, and art criticism have appeared in numerous online and print publications. His gallery and classroom teaching emphasize social justice issues and the politics, history, making, and interpretation of art.

Editor's Note

Visit the Art Institute of Chicago website or contact Sam Anderson-Ramos at sramos@artic.edu to learn more about the museum's medicine and art programming. Browse the *AMA Journal of Ethics* Art Gallery for more Art of Medicine content and for more about the journal's partnership with the Art Institute of Chicago.

Citation
AMA J Ethics. 2019;21(2):E192-195.

DOI
10.1001/amajethics.2019.192.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

AMA Journal of Ethics®
February 2019, Volume 21, Number 2: E196-197

ART OF MEDICINE

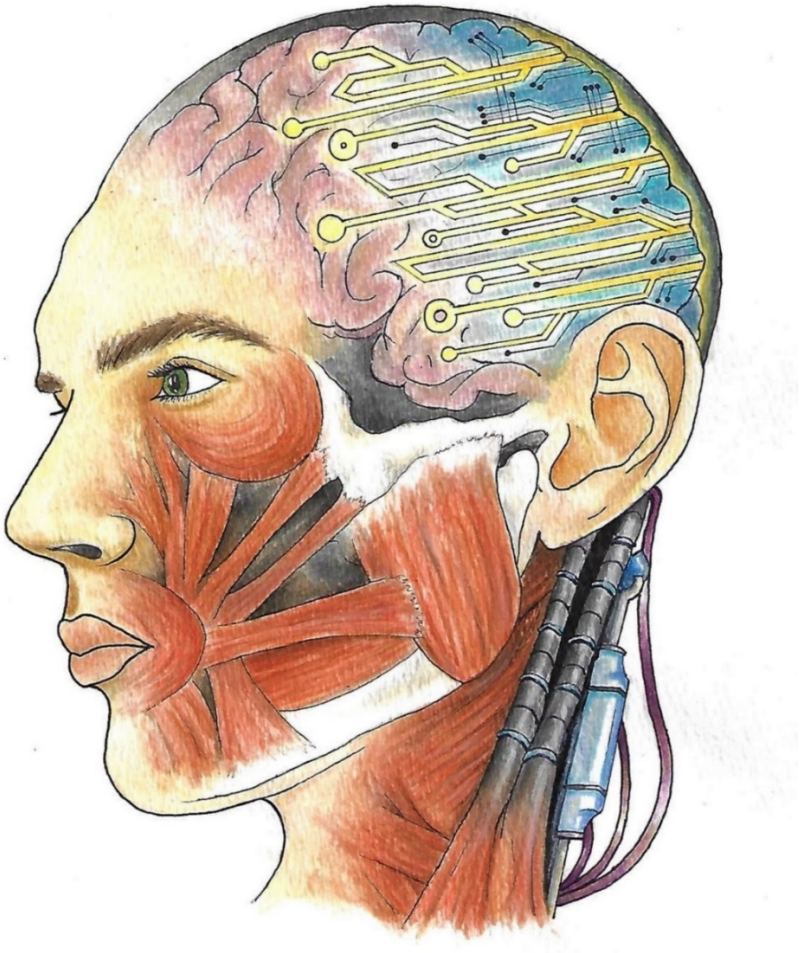
Technological Transformation

Elisabeth Miller

Abstract

Technology has enabled bionics and artificial intelligence, each of which can have important applications in health care. As we continue to substitute body parts with machinery, however, we might wonder, "What makes us human?" This drawing interrogates the relationship between humanity and embodiment, specifically in neck and facial musculature and brain structures.

Figure. *Technological Transformation*



Media
Water color pencils and black pen on paper.

This image represents humankind's union with technology. It shows the brain turning into a collection of integrated computer circuits and the neck muscles evolving into mechanization-ready cables, pumps, and wires. In artificial intelligence (AI), boundaries distinguishing life and technology are challenged. We wonder, "Is it possible for machines to think? Are our own brains just complex organizations of biological microchips?" Medical students are well positioned to appreciate how intimately technology is becoming part of human life. From wheelchairs and artificial limbs to new antibiotics and imaging, innovations are constantly growing in number and playing larger roles in our existence. If science unlocks the origins of thought, therapies for patients with neurocognitive or psychiatric problems could be enabled. Progress in AI will generate the need in medicine to explore ontological and ethical relationships among brains, minds, selves, and healing.

Elisabeth Miller is a third-year medical student at the University of Washington in Seattle. She earned an undergraduate degree in biology from Carroll College in Helena, Montana.

Citation
AMA J Ethics. 2019;21(2):E196-197.

DOI
10.1001/amajethics.2019.196.

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980