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