Explainable AI: How Do We Know What AI Is 'Thinking'?

ALAN M. REZNIK, MD, MBA, FAAOS

EDITOR'S NOTE: This is the sixth article in a series on artificial intelligence (AI) applications in orthopaedics and the future of medicine. This article discusses the ability to understand the "thought" process of AI and strategies required to make AI decision-making more transparent. The prior articles have reviewed the basics of AI, how AI can read an X-ray, natural language processing, medical ethics, and the differences between shallow and deep AI as it applies to the practice of medicine. Visit www.aaosnow.org to read more.

AI 'thinking'

Scholars have pointed to Stonehenge and 900 similar, smaller, manmade rock formations as the earliest forms of artificial intelligence (AI). The stone formations use rock positions and photons from the sun to accurately time winter and summer solstices. The photon-rock "machines" helped early societies to "predict" the best timing for planting. Other "thought" machines followed. From clocks to DaVinci's autotoms to Charles Babbage's mechanical calculating machine (circa 1830), mechanical devices continued to advance. When Ada Lovelace created the first program "algorithm" for Babbage's machine, the possibilities truly expanded. Still, mechanical devices were limited by speed and a very small number of memory states. The limitations changed when electronic circuits appeared.

Using the first electronic computing machines in 1950, Alan Turing wrote "Computing Machinery and Intelligence." He noted that Mr. Babbage's mechanical calculator was limited to solving equations with up to only seven terms or a power of seven (10^7). In Turing's day, electronic machines had 10^{50} possible states, far more than Babbage's, yet far less than 100 pieces of paper with 50 lines containing 30 base 10 digits per line $10^{(100x50x30)}$ (or $10^{150,000}$) possible states. In 1969, NASA engineers for the Apollo Moon Mission used "larger" computers costing millions to make critical computations for the spacecraft. Today's smartphones fit in your pocket, cost merely hundreds of dollars, and can process instructions at a rate of more than three billion per second. We know that this is more than 120,000 times faster than the Apollo machines, and today's cell phones have gigantic memories with $2^{1,000,000,000}$ (or $10^{301,029,996}$) possible states.

Understanding that massive memory was possible in the future, Turing believed that someday machines would "imitate" human thinking. He invented the "Turing test," during which a human blinded to the source of the interaction could not tell the difference between a human response and a machine response. To this end, in the late 1950s, the perceptron, an electronic representation of a human neuron, was built. The electronic model of a human nerve led to the idea of creating a digital brain. In 1956, at Dartmouth University, this included artificial general intelligence (AGI) and artificial super intelligence (ASI). AGI would be a machine that could solve any problem without programming, and ASI would surpass human thinking. AGI and ASI, in fact, overpromised and could not deliver. Like the dotcom bubble, many investors lost their money as companies went bankrupt.

As noted in a prior article in this series, the failure of early AI could be blamed on the cost of a megabyte. In the 1960s, one megabyte cost $2 million. Now, one megabyte costs less than one hundredth of a penny. Coincidently, advanced Nvidia video chips for gaming are optimized for the matrix algebra needed for today's AI brains to work. With the combination of massive memory, the cloud, complex calculations, and programing languages like Python and TensorFlow, we are moving cars, predicting items on Amazon, directing search preferences, discovering new treatments, reading medical papers, checking retinas for diabetic changes, and predicting outcomes.

Man versus machine

In general, humans are very good at extrapolating small amounts of data to come up with the correct answer, and AI is far better at extrapolating from very large amounts of data. For example, part of Memorial Sloan Kettering's 100,000 prostate slide library is being used to teach AI to find prostate cancer. It finds consistencies across more than 40,000 slides using unsupervised learning with high specificity. In contrast, it has been shown that after 39 hours of reading slides, human pathologist error rates increase. The liability for humans as pathologists is fatigue (Table 1). In contrast, AI algorithms working 24/7 just don't get tired. The AI logic inferences, once trained, remain constant. At the same time, these networks based on thousands of perceptrons using supervised or unsupervised learning do not explain their decisions. Moreover, we don't know whether the process for the decisions being made are based on tried-and-true medical reasoning or convoluted logic using factors we cannot fully appreciate or even understand.

Errors per week	39 hours	45 hours	50 hours	60 hours
None	78%	43%	32%	22%
At least one	22%	57%	68%	78%
At least two	18%	45%	50%	60%
At least three	12%	30%	35%	43%
At least four	5%	12%	16%	18%

Table 1 The percentage of pathologists experiencing adverse events increases significantly with a workload greater than 39 hours per week

ADAPTED FROM MAUNG R: *ALL IN A DAY'S WORK*. AVAILABLE AT: HTTPS://THEPATHOLOGIST.COM/OUTSIDE-THE-LAB/ALL-IN-A-DAYS-WORK. ACCESSED MAY 28, 2020.

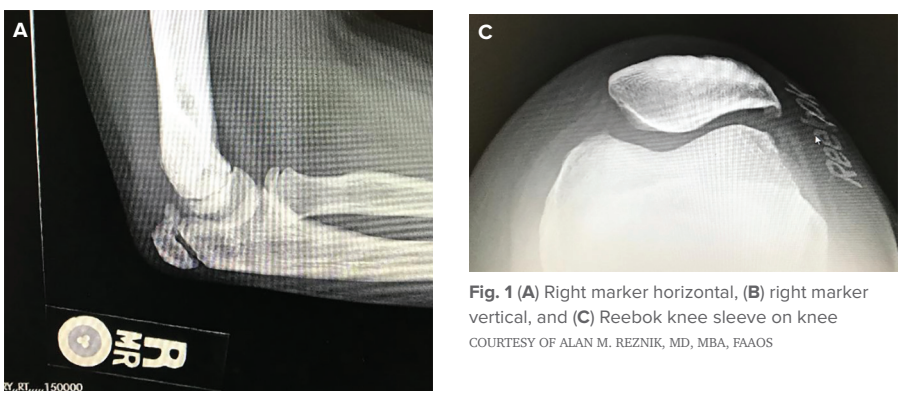

Fig. 1 (**A**) Right marker horizontal, (**B**) right marker vertical, and (**C**) Reebok knee sleeve on knee

COURTESY OF ALAN M. REZNIK, MD, MBA, FAAOS

Logic gone awry

As infallible as AI may seem, an earlier article in this series covering ethics and AI discusses how data can contain bias. This may be hidden by proxies for demographic, racial, or socioeconomic factors. Zip code, age, sex, and even surname can induce biased analyses. Very large datasets are full of unknown biases. Therefore, using them safely for medical decisions creates concerns about ethics and fairness. In 2019, a major insurance carrier applied an AI algorithm to direct medical care to those in whom its own outcome predictions revealed greatest benefit. Although that was seemingly a logical and cost-saving move, later review found the new allocations to be racially discriminatory. In retrospect, the plan was deemed unacceptable and stopped. One could argue that the harm was already done.

The need for explainable AI

The problem for perceptron-derived AI is in the beauty of how it works. Unlike Stonehenge, DaVinci, and Mr. Turing, modern neural networks are trained to solve a problem by looking at massive amounts of data. The network of perceptrons "learn" relationships between the training dataset and the desired outputs. The AI "brain" predicts the best action based on what it has learned, yet there is no explanation given and the learning algorithms generally remain hidden. Here, we need to look to examples to best understand the implications of this point.

Once trained, a neural network algorithm looking for lung nodules can separate X-rays into normal and not normal. At first, it did not show an abnormality's location. A clever programmer decided to place a blank square on the film and move it around in increments. When the AI reading converted back to normal, the logic dictated that the blank must be covering the abnormality (in the lung, a potential tumor, calcification, or pneumonia). The abnormalities found also included rotation of the right and left X-ray marker (Fig. 1a and 1b), a change in marker type, initials of the person taking the films, or an item of clothing overlapping on the film (Fig. 1c)—all distractions a

human reader would ignore. In that the AI did not explain its own "learned" definition of abnormal findings, the sorting process could lead radiologists into looking for tumors that were not there.

Similarly, an AI algorithm looking for X-rays containing a fracture was successful and accurate. It sorted films into those with a fracture present and those without a fracture. Looking deeper, machine type and location of the X-ray unit were found to be factors in diagnosing a fracture. The algorithm "understood" that emergency department X-ray units are more likely to be positive for a fracture. When location data were stripped from the decision process, the algorithm was barely better than random selection.

In contrast to these mistakes in logic, there are times AI can get it completely right. In the 100,000-slide training set used by AI to detect prostate cancer, when peeling back the layers of the neural network for an explanation of the diagnostic method, researchers found many of the AI pathology findings used by the algorithm to be the very same features that human pathologists use to make the same diagnosis.

These examples point out unexpected pitfalls in AI's ability to read X-rays and the importance of having AI explain how it got a result. Even when AI is on the mark, the result should be validated. In Europe, AI must explain itself to be used in medical treatment, and in the United States, AI must be approved by the Food and Drug Administration (FDA) and has to stop all additional learning after approval in order to be used. The term for this type of AI and the logical validation requirement is explainable AI, or XAI.

How do we get AI to explain itself?

AI uses natural language processing (NLP) to read articles, uses image processing to look at pathology slides, and aggregates population data to sort medical treatments to maximize benefits. Each requires very different strategies to explain an algorithmic outcome. Moving a blank square on an X-ray is only one example. AI was "looking" at films and, because of artifacts, it knew where the X-ray was taken better than the humans viewing the same films and used that information (i.e., X-ray machine type and location). Discovering the information being used and correcting for it is a much harder problem to solve. Certainly, the algorithm did not tell us what it was doing; the extra data it gleamed from the films just "worked" well, even if not medically justified.

Other strategies for XAI include decision trees and the so-called "forest of trees." Decision trees are used frequently in well-accepted treatment algorithms. For XAI, we start with a treatment option tree, and AI can use it as the backbone for all the data it processes. The algorithm can adjust the tree or add branches, change the questions at the nodes, and make decisions at each node. Then the program can report back the path of nodes and the decision at each one. Seeing the actions at the nodes of the tree, we have a window on the AI "thought" process.

In a forest of trees, we let the AI create a large random set of decision trees from the data. It uses no preconceived notions about the data. It creates questions at each node and then uses a training set to find the best single "tree" for the desired decision-making. Going forward, we can see the "best" tree and the values at each node in order to understand how the AI reasoned out the diagnosis or treatment plan. Again, the tree and nodes become the explanation of the logic being used.

The problem can be more challenging for NLP. One current solution is used in a medical diagnosis application. It uses NLP to "read" hundreds of thousands of patient charts. After asking sequential questions about a patient's symptoms, the NLP engine finds charts of patients with similar symptoms. It selects more questions to narrow down the relevant charts to make a diagnosis. Eventually, it gives the patient the odds of a given treatment choice being correct for him or her, not a single diagnosis. For example, it will indicate that 62 percent of similar patients with "chest pain" had gastrointestinal reflux and 53 percent of those patients used omeprazole for treatment (Fig. 2). The program avoids explanation and hence liability concerns by using the patient's own answers to find a potential diagnosis. It goes further by not giving an exact treatment directive, giving only probabilities. It also adds how many patients with the same symptoms saw a doctor for treatment. It suggests doing so if the odds indicate you should see a doctor. It is not a single diagnosis-maker, it's an odds-maker, and the patients are given the odds, not an exact plan.

For images, explainable AI may take other forms. Diagnosis may be limited to a menu, and the logical method (like the location of the tumor and radiographic parameters) may be required as part of the report. In one case, a newly approved AI program gives relative Kellgren and Lawrence (KL) grading for osteoarthritis based on an AI evaluation of plain films of the knee (Fig. 3). The FDA approval requires showing factors that explain the AI function. In this case, the software uses joint space narrowing, sclerosis, and osteophytosis, as well as the actual AI measurements of the joint space in millimeters, to find the KL value. Although the software is capable of reporting the actual grade level, the FDA prohibits it from giving a specific KL grade of zero through four. Instead, it reports KL grade less than or equal to 2 or greater than 2, as well as the presence or absence of the three factors. The FDA's decision was based on the understanding that clinical cutoffs for KL value are more useful and more forgiving relative to the values themselves (Fig. 3b, available in the online version). In contrast, the European Union (EU) permits the full KL grade and a numerical grade for each factor (Fig. 3c, available in the online version). The EU presentation is a more granular representation of the data. Of note, in my own beta testing of the program, having seen both presentations, I preferred to have the more granular, real KL values with the factors displayed combined with my own judgments. The FDA does not agree.

Fig. 2 K Health diagnosis process after a series of symptom-related questions
WITH PERMISSION FROM K HEALTH INC.

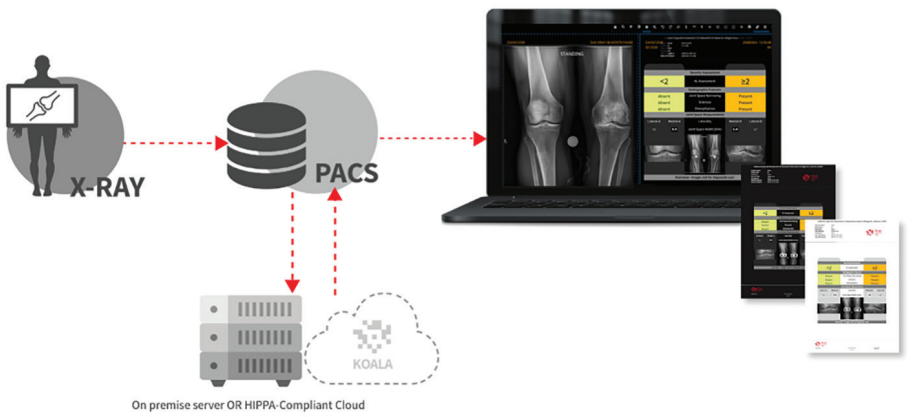

Fig. 3 Workflow for ImageBiopsy KOALA™ artificial intelligence analysis of plain films
COURTESY OF IMAGEBIOPSY INC.

The future

From the examples given, we can see how AI must be able to explain itself; in medical applications, that varies greatly pending the data sources and the AI algorithms used. Governing bodies such as the FDA may have the final word on what explanations are needed and limit outcome data type and form based on the approval process. Many times, the explainable AI solution may not be obvious or easy to implement. Each AI algorithm will present its own unique challenges when it comes to explaining its results. Lastly, AI may use factors and proxies for those factors as part of making its predictions. We must be aware that the outcome, when AI does explain itself using XAI standards, may deliver a novel approach to a problem, yield a biased outcome, or just plain surprise us.

References for the studies cited can be found in the online version of this article, available at www.aaosnow.org.

Alan M. Reznik, MD, MBA, FAAOS, specializes in sports medicine and arthroscopic surgery and serves on the *AAOS Now* Editorial Board. He was a prior member of the AAOS Communications Cabinet and Committee on Research and Quality. Dr. Reznik is chief medical officer of Connecticut Orthopaedics, assistant professor of orthopaedics at Yale University School of Medicine, and a consultant.