"But Why?" Understanding Explainable Artificial Intelligence

Opaque algorithms get to score and choose in many areas using their own inscrutable logic. To whom are said algorithms held accountable? And what is being done to ensure explainability of these algorithms?

By Tim Miller

DOI: 10.1145/3313107

Imagine the following. You finish as one of the top students in your class and are looking for that all-important first step in your career. You see an exciting graduate role at a great organization, which matches your own skills, experience, and characteristics. Excited, you tailor your curriculum vitae to the role, and a couple of days before the deadline, you submit it. Just 30 seconds later, you see an email from the company. You open it and it says: "Thank you for your application to our graduate program. Unfortunately, you have not been successful in this round. We hope you will consider us for future roles."

Puzzled, you wonder how they could have made a decision so quickly. Some digging on the Internet shows that the company uses an automated algorithm to filter out most applicants before ever being seen by a human. "No problem", you think, "I'll just find out why I could not pass the filter, update my application, and re-apply in the next round." However, your request to the company is met with the response: "We use advanced machine learning algorithms to make these decisions. These algorithms do not offer any reasons for their output. It could be your work experience, but it could be simpler the terminology used in your application. We really cannot tell." You have missed out on your dream job and have no idea why nor what you could have done differently.

MODERN DECISION MAKING

This scenario may sound somewhat dystopian, but it is precisely what many recruitment organizations do right now. Automated algorithms are being used to assess and classify applicants' potential, while only the highest-ranked applications make it into human hands. Typically, these filters

Image by Vdant85 / Shutterstock.com

are machine-learning algorithms. This means their decision processes are written not by a person, but are trained using data. The training process is that a number of features of importance are selected, such as degree name, institute name, grade-point average, work experience, etc., but also the terminology and phrasing used in the application. Then, taking a huge stack of prior applications and the final decisions of these applications (hire or not hire), a mathematical model is automatically derived, using a machine learning algorithm that predicts the likelihood of a person being hired. It does this by discovering patterns in the underlying data. For example, certain institutes produce more suitable graduates than others, but also certain styles of writing do too. This model is then used to predict future applicants' chances of being hired, filtering out those with a low rating.

These are what are sometimes called "black-box" algorithms. This means for recruiters and applicants, some input (the features for the particular individuals) are fed into the algorithm, and then some output appears. There is no indication what is happening inside.

These algorithms are not limited to recruitment. They are used to make sensitive and important decisions, such as estimating the probability of a prisoner re-offending if released, estimating the risks of children being neglected by their parents, and estimating the likelihood of someone defaulting on a bank loan; as well as in much more mundane tasks, such as in voice-based interaction with smartphones and recommending movies on streaming services based on what you have watched previously.

Worryingly, many of these models have shown bias against certain groups of people, precisely because the data used to train them were also biased. The most salient example is of Amazon's recruiting tool, which was trained on past applications and learnt to prefer male applicants over female applicants. However, gender was not one of the features used. Instead, the algorithm discovered a pattern in which applications that used more "masculine language" were more likely to be hired. Ultimately, Amazon downgraded the use of the tool, using it only as a recommendation to a human recruiter. This does nothing to solve the problem however, because it is the recruiters who fed the machine-learning algorithm the biased data in the first place.

In particular, explanations are social processes; when explainers give explanations, explainees argue or ask follow-up questions.

ANSWERING "WHY?" FOR TRUSTED AND ETHICAL ARTIFICIAL INTELLIGENCE

Studies such as the Amazon case are far from isolated. Recent books such as Cathy O'Neil's *Weapons of Math Destruction* and Virginia Eubanks *Automating Inequality* show the impact that poor automated decision-making algorithms have on real people.

One step to improving people's trust in these algorithms, and ultimately, to produce ethical artificial intelligence, is to produce artificial intelligence that can explain why it made a decision.

The field of explainable artificial intelligence (XAI) aims to address this problem. Given some output from an algorithm, an explainable algorithm can provide justifications or reasons why this output was reached. This can be as simple as noting which inputs were the most important, to providing some details of the inner workings of the model, or to informing people what they would need to do differently to get a certain other output, such as to get an interview.

DO WE NEED EXPLAINABLE ARTIFICIAL INTELLIGENCE?

There are some people in artificial intelligence who reject the need for XAI. For example, Geoff Hinton, creator of the backpropagation algorithm widely used in deep learning and highly respected AI researcher, was recently quoted [1] as saying that asking algorithms to explain their decisions or beliefs would be "a complete disaster." He continued, "People can't explain how they work, for most of the things they do. When you hire somebody, the decision is based on all sorts of things you can quantify, and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decision, you are forcing them to make up a story."

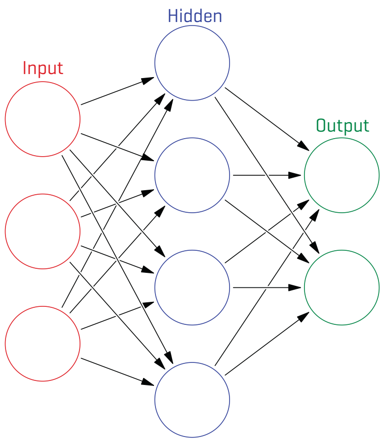
Hinton is referring to the concept of post-hoc explanations; human explanations are constructed after our reasoning and these explanations can be inaccurate. His claim appears to be that because neural networks are a metaphor for the human brain and are hard to understand, to explain their decisions would require neural network architectures to "make up stories" too. These would be incorrect or incomplete and therefore "a complete disaster." Hinton and others argue in order to trust these systems; we should instead regulate them based on their outputs. That is, based on their performance over many tasks.

I think this reasoning is entirely incorrect.

First, it is contradictory. Hinton's own words are an explanation of how he reached his opinion on XAI. Is he making up a story about this? I imagine he would claim it is based on careful reasoning. But in reality, his words are abstract summaries about the neurons in his brain firing in a particular way that nobody understands. The ability to produce and communicate such summaries to others is a strength of the human brain. Philosopher Daniel Dennett claims consciousness itself is simply our brain creating an "edited digest" of our brains inner workings for precisely the purpose of communicating our thoughts and intentions (including explanations) to others. There is no theoretical barrier I know that would prevent similar digests for machine learning models.

Second, these arguments against

Figure 1. An artificial neural network



explainability are on typically on the grounds of regulation. But what about ethics? Is it ethical to make important decisions about individuals without being able to explain these decisions? For example, in some U.S. states, parole judges use algorithms for predicting the likelihood of a prisoner re-offending if released from jail, influencing parole decisions. Is it ethical to keep someone in jail on the basis of a decision from a black box without them knowing why, or knowing what they need to do before their next hearing? Similarly, is it ethical to reject an individual's application because a black box with 95 percent accuracy says so, with no way for the applicant to find out how to improve? Many people will claim it is not. Hinton and many AI experts argue we should merely run experiments and see if algorithms are biased or safe, effectively ignoring the small percentage of wrong decisions. These types of arguments seem to be born from the privilege of white middle-class males such as myself, who are unlikely to be adversely affected by such decisions.

Third, at the individual level, there is also an issue of trust. Would you trust an algorithm that advises you to have invasive surgery if it could not give any reasons why? Would a doctor trust the advice of this algorithm? With the Watson Health project, IBM is finding out that earning trust of medical professionals is a difficult barrier to break. People do not develop trust based only on statistics; they develop trust based on their own interactions and experiences, and of those around them. This is a mistake that the artificial intelligence community continues to make, to our detriment.

XAI: THE CHALLENGES

Explainability is not a new challenge in artificial intelligence. The first research on XAI dates back to the mid-1980s, with approaches to explain rule-based expert systems. However, the modern techniques employed in artificial intelligence, in particular, machine-learning techniques that use deep neural networks, have recently lead to an explosion of interest on the topic. Further, these techniques present new challenges to explainability that were not present 1980s expert systems.

Challenge 1: Opaqueness. Modern AI models, in particular deep neural networks, are not particularly easy to understand. If we consider an algorithm written by a software engineer, we could give that software engineer a sample input and they would typically be able to tell you what the output would be. With most techniques in artificial intelligence, this is not the case. For example, given a logistic regression model or a heuristic search algorithm, the output of the algorithm is harder to predict. In the logistic regression case, this is because the regression equation was not derived manually, but from the underlying data, and is not in a simple if/then rule format that facilitates easy understanding. For heuristic search, this is because it derives a small 'program' on the fly from the search algorithm's inputs and goal, and it will find one of possible many solutions to the task. Models such as these are still somewhat understandable, to the point where given the actual output, the engineer could probably trace this back quite easily and see why it was made.

Deep neural networks, on the other hand, are highly opaque. Not only will the engineer have a hard time predicting the output, but they will have a hard time tracing the reasoning back and debugging. This is because the power of deep neural networks comes from the so-called hidden layers, where the learning algorithms find important correlations between variables and "store" them in nodes that have no meaningful labels attached to them, as shown in Figure 1. I believe, rather than being driven by

Figure 2. Saliency map highlighting important pixels of an image of a moka pot (original image on left, saliency map on the right).

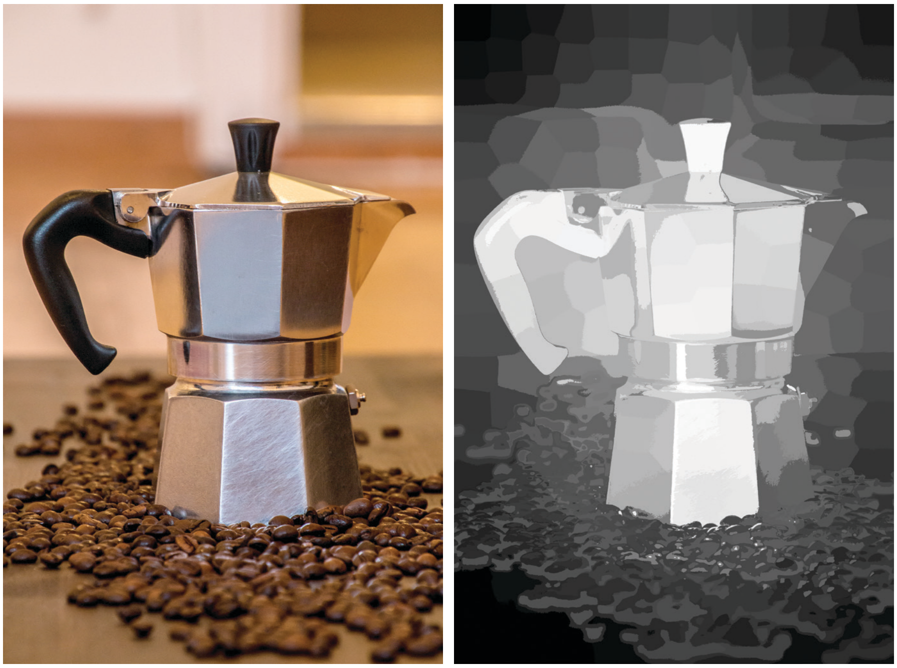
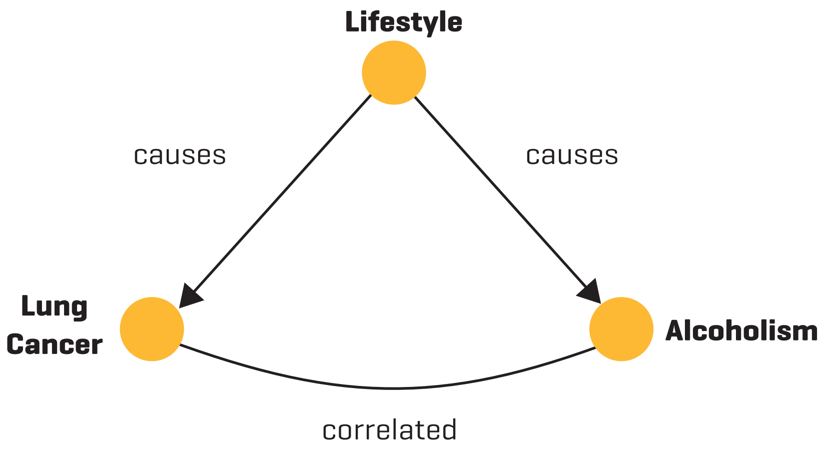


Figure 1 by Glosser.ca – Own work, Derivative of File: Artificial neural network.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24913461

Figure 3. Causal model.



ethics and trust, most XAI is driven by AI researchers simply wanting to understand their own models better to improve them.

Researchers are attempting to overcome opaqueness in mainly three different ways:

1. **Importance weighting.** Many techniques reverse engineer which parts of an input were "important" for a decision; for example, by highlighting which pixels in an image were important in recognizing a particular object. This could be illustrated, for example, with the saliency map from an image of a moka pot presented in Figure 2.

2. **Interpretable models.** Other techniques aim to extract or learn a less opaque (more interpretable) model; for example, given a particular decision from a deep neural network, produce a decision tree or linear equation that approximates that decision and is easier to understand.

3. **Discard deep neural network models.** More recently, some machine learning researchers advocate the idea of using deep neural networks to discover important features in the hidden layers, debugging the neural network to find what those hidden features mean, and then discarding the deep neural network model, just using the discovered features to learn a more interpretable model, such as random forests or linear regression.

In my view, none of these is the solution to XAI; they offer little insight to anyone other than experts in the field. However, they will be useful for those experts and importantly, could be used to form the basis of explainability to end users.

Challenge 2: Causality. Causality between two events indicates one event was the result of the other; for example, smoking causes lung cancer. Correlation is a statistical measure that indicates a relationship between two variables; for example, smoking is highly correlated with alcoholism. However, smoking does not cause alcoholism and alcoholism does not cause smoking. There is some confounding variable, such as lifestyle, that causes both, so they are positively correlated. Figure 3 shows a causal model of this example.

Machine-learning algorithms excel at finding correlations between things.

Would you trust an algorithm that advises you to have invasive surgery if it could not give any reasons why? Would a doctor trust the advice of this algorithm?

For example, correlations between age, gender, and previous purchases correlate with future purchases; or the types of volunteer work people do with their likelihood of being hired at a particular organization. However, using statistics to find such relationships does not uncover causes.

This leads to issues in explainability because explanations that refer to causes, known as causal explanations, are easier for people to understand. If I asked an algorithm why it predicted a particular person had a high chance of lung cancer, the answer "Because they are an alcoholic" would be an unsatisfying answer. It is the reason why the prediction was made, but it is not why they may have lung cancer. A better explanation would be "Because people who drink a lot tend to lead a lifestyle in which they smoke a lot, and smoking causes cancer".

There are two ways to get around this issue:

1. **Learn causal models.** Instead of learning about correlations, machine-learning models can learn about causes. This would be ideal but has major challenges. First, we often do not have access to these confounding variables such as "lifestyle," meaning that finding the causes can be challenging. Second, machine learning is built on the field of statistics, and for decades statisticians have ignored causality, meaning that the foundations of learning causes are only in their infancy.

2. **Exploiting human strengths.** Humans are hard wired to extract causes from series of events. For example, if we see a cartoon character push over another character, we identify the push as the cause. However, this is not true—the illustrator "caused" the fall. If people were not hard-wired to extract causes from events, we would not be able to watch cartoons, or even televisions or movies. XAI can exploit this by giving users enough information to see the correlations and let them determine the causes (perhaps incorrectly) themselves, the same way we do with cartoons.

Challenge 3: Human-centeredness. The final challenge for XAI is that it is ultimately humans with little knowledge of AI who will need to understand decisions.

Is it ethical to make important decisions about individuals without being able to explain these decisions?

As discussed earlier, I believe most XAI is driven by AI experts' desire to better understand, debug, and improve their own models. In my view, this has led to a situation in which "explainable" AI is giving explanations for other experts, rather than for non-experts. Such systems will not offer job seekers or prisoners up for parole sufficient insight into why they did not get the decision that they hoped for.

To achieve truly XAI, we will need to attack both the technical and the human challenges. The starting point for this is to determine what users (and others affected by decisions) would like to understand about AI-based decisions, and what society thinks are the ethically important questions that need to be answered. This is in contrast to the orthodoxy, in which AI experts think about their complex models and determine what they think are important parts of the model that need to be explained, and then hope that these satisfy users.

Recently, I published an article that surveyed more than 200 papers from philosophy, cognitive psychology/science and social psychology on what an explanation is and how people generate, selected, present, and evaluate explanations [2]. The key finding is human-to-human explanations are context sensitive, which means they do not just provide causes for decisions, but that explanations differ based on the particular question asked and the people to whom the explanation is presented. In particular, explanations are social processes; when explainers give explanations, explainees argue or ask follow-up questions.

While this finding may seem obvious, research in XAI is only now starting to frame the problem like this. Most prior research treats an explanation as a set of statements, not a process, simply highlighting important causes while providing no chance of follow-up if the explainee is not satisfied. While some of the prior research can form part of an ongoing interaction between AI and humans, much more research is needed to determine what questions people want to ask, how to elicit these questions, how to answer them, how to present them, and how to determine whether someone understands the explanation.

DISCUSSION

As AI becomes more prevalent in our world, it will continue to make important decisions that have real impact on people's lives. Ethical concerns and lack of trust in these technologies will continue to limit their adoption. In my view, XAI will be one piece of this solution. Ultimately, I believe this is a multi-disciplinary problem that will need to combine computer science, social science, and human-computer interaction.

XAI is not a panacea for all ethical concerns and problems of distrust of artificial intelligence. However, I believe explanatory systems that interact naturally with non-experts, using the non-expert's language and their concept of explanation, is a necessary but not sufficient requirement to address ethical and trust issues of artificial intelligence.

References

[1] Simonite, T. Google's AI guru wants computers to think more like brains. *Wired*, December 12, 2018; https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/

[2] Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1•–38.

Biography

Tim Miller is an associate professor of computer science at the University of Melbourne. His primary research interest lies in the area of artificial intelligence, in particular human-agent collaboration and explainable AI. His work lies at the intersection of artificial intelligence, human-computer interaction, and cognitive science/psychology.