Published in *Artificial Intelligence for Better or Worse* in 2019 by Future Leaders (www.futureleaders.com.au)

Explainable artificial intelligence: What were you thinking?

Tim Miller

Decisions that affect our lives in both trivial and important ways are increasingly being made by algorithms. These algorithms, especially those derived using artificial intelligence (AI), are often inscrutable — at least, they are for now. If we want to hold people and organisations to account for their decisions, and if they want people to have trust in their decisions, we will need to be able to explain why a decision was made. How will we achieve explainable decision making? What will explainability look like? This chapter looks to answer these questions in the context of our future lives.

Automated decision making

Imagine receiving a diagnosis from a medical specialist who advises you that the best treatment is to have invasive surgery that will keep you off work for several weeks. When you ask "Why?", they simply say, "I don't know, it just feels right". Would you take the surgery, or get a second opinion? My guess is that you would do as I do and get a second opinion, because it is strange to even consider a medical specialist who cannot explain their decisions. However, replace this medical specialist with an automated decision-making tool, and this is the situation that many people will face if or when automated recommendation tools are used in medicine.

This is not limited to medicine. Automated tools that make decisions affecting peoples' lives are increasingly being rolled out in many domains, such as finance, law, recruitment and security. Their ability to explain why decisions are made is limited.

The idea of making decisions automatically is nothing new. People have tried ways to understand decisions for centuries using probabilistic models. Contemporary job roles such as financial actuaries and weather forecasters use statistical models to make predictions about financial behaviour and weather.

However, the gap between how contemporary AI "works" and how people work is getting wider. Unlike handcrafted statistical models that use human-relatable concepts, *machine learning models* are derived automatically from data and are not designed for human consumption. In particular, techniques such as deep neural networks have what are known as 'hidden' variables, which are concepts that the neural network itself learns, such as the novel combination of several other variables. These hidden nodes have no human-interpretable label attached to them. As such, even experts cannot look at a deep neural network and understand why a decision was made. But it is not just neural networks that are difficult to understand. Most contemporary AI models are intertwined combinations of thousands or even millions of variables. Experts understand how the underlying mechanisms work and can implement them, but paradoxically, while they understand the algorithms, often they cannot understand why particular decisions are made by those algorithms. It is easy to write a

program that takes a million variables and combines them together to make a decision, but when looking for a reason for a particular decision, anything more than 5–7 variables is too cognitively demanding for our working memory (Cowan, 2010).

The complexity of contemporary AI models is entirely deliberate: adding more complexity to models seems to give them better accuracy for many tasks. Some future Turing award winner could find simpler types of models that are both accurate and easy to understand, but for now, many people talk of how a trade-off is often made between explainability and accuracy (Gunning, 2017).

The need for explainability

There is no agreed-upon definition of "explainability", either in philosophy, psychology or computer science. For this chapter, we take a pragmatic definition that a decision is explainable if someone can understand why the decision was made and how they could manipulate the inputs to get a different answer. In short, a person should be able to construct an explanation as to why the decision was made.

Much of the focus on current research is on explanation in AI — a concept that is quite well understood in the social sciences (Miller, 2019). The assumption here is that rather than asking people to construct their own explanations of why an automated decision was made, we construct computer programs that give explicit explanations, making it easier for the person to understand. However, other forms of explainability are also getting traction, such as making decisions that are simpler to understand in the first place. A system with high

explainability gives people actionable insight into what they need to do in with future interactions with the system.

The question is: do we even need explainability? Some people in the AI community argue not — if we insist on explainability we will need to make our models simpler and they will lose their "magic juice". Instead, they argue, we should just make more accurate models, and everything will be fine. However, I disagree with this stance. There are two reasons why explainability is useful: trust and ethics.

Trust

Experience shows that giving people an automated decision-making tool and hoping that they use it is bound to fail for anything other than the lowest of low-stakes decisions. In my experience, it is better to start out with the assumption that people do *not* trust a model/decision, and work from there. This is typically the default stance and so it should be. Despite its hype, AI models do not have common sense and do not understand what or why they are doing things. They do not question their decisions, they cannot understand that they are making decisions outside of what they were built to do. The onus is on the creators of the model to show that it is not brittle.

Simply giving some accuracy statistics from previous experience is not sufficient. Would you have invasive surgery just because an algorithm with a reported 99% accuracy told you to? What if you were in the 1%? Even accuracy is not enough. If one in a million people actually required invasive surgery, but we test one million people with a model that is 99% accurate, we would find that about 10,000 people would be incorrectly told they needed invasive surgery! This is known

as the base rate fallacy. Being able to check why would help a medical specialist to identify many of those incorrect cases.

I am certain more accurate models would lead to more trust in general; however, a person being told that they require surgery may have just that one interaction with the automated decision-making system. Is it reasonable to give them no insight into why they need surgery before they go? We can fall back on the medical specialist, but they may find the system just as opaque.

Even for people who will interact with systems many times, I am not convinced with the argument that accuracy is all we need for trust. If they do not trust it to begin with, they will never use it and will never build up trust. If they do start to use it and it sometimes makes mistakes, we know that trust is much easier to lose than to gain, especially for machines (Hoffman, 2013). So, just a few mistakes is all it takes for the tool to be shelved.

Ethics

Second, a reason to desire explainability is ethics. Is it ethical to make decisions that affect individuals without being able to give them understand why?

> *You have not been shortlisted for a job that you applied for due to an automated CV matching system, but trust us, the model is accurate.*
>
> *You will not be gaining parole today, but trust us, the model is accurate.*

I think any person would object to being told these statements in such a position.

In my experience, people who argue that accuracy alone is not enough such as this tend to have white male privilege (as do I), and would be higher unlikely to ever be in a position to be disadvantaged by automated decisions. This brings us to the issue of *fairness*.

A prevailing view was that algorithms cannot be biased because they are cold, hard calculating machines. However, recently there have been many examples of machine learning models that are biased because the data that is used to train them is biased itself. Amazon used over 10 years of previous hiring decisions to train a machine-learning model to predict the likelihood of an applicant being hired, based on their CV, and used this to cut people from the application process automatically. A useful tool if implemented correctly. However, quickly Amazon realised that women were almost never shortlisted, even though gender was not even used as part of the decision. They found that mentioning the words "women" or "female" seemed to be enough for someone to be cut. The problem: Amazon's hiring decisions in the past were biased, and its model had learnt these biases too.

Explainability would not eliminate or even mitigate bias; however, it would allow developers, recruiters, and applicants to show that a model is biased. In the Amazon case, an explanation to an applicant that they missed out because the phrase "President of the Women in Computer Science Club" appeared on their CV would arouse suspicion.

Humans in the loop

One answer is simply to always have people overseeing decisions. For high-stakes decisions that are not time pressured,

this is surely one aspect to help with trust and ethical issues: the humans are accountable so the AI model just needs to provide its answers. The human operator can overrule a decision.

Human judgment is better than AI in many situations, so human oversight is important, but overruling an automated decision requires the human to understand why it is wrong. Can this be done without explainability when there are thousands of variables to consider? In some cases perhaps, but not in many. Early expert systems from the 1990s that were great at medical diagnosis were rejected because medical specialists could not understand their decisions. The specialists made the decisions and were accountable for them, while the expert systems were intended to be an aid. Putting a human in the loop was a given, but if the human in the loop cannot understand, then the automated tools are ignored because nobody who is accountable for an important decision would accept an answer they could not reason about.

What could explainability look like in the future?

I have argued why I think explainability is important, but what will it look like in the future, and what are some of the advantages and pitfalls? In this section, I will analyse explainability through the lens of some existing systems and see how it could look, using both low-stakes and high-stakes decision making.

Low stakes: recommendations in online retail

As a customer, online retail may seem largely transactional: you browse products, select some, and then (hopefully for the retailer!) buy. However, these systems are some of the most

successful instances of AI in existence. They monitor and record every item we view, our searches and purchases over time, the items we have rated and reviewed, and even our locations. This data is used for several purposes, most notably for pricing, but even for anticipatory purchasing in which companies like Amazon ship items to local warehouses having predicted a surge in purchases of those in the near future. However, we will look at perhaps one of the lowest stakes decisions: recommending new products and services.

If you have ever been on the internet at all, you will have seen recommendations. Search engines recommend products on the sidebar that are related to the search terms, but also, in many search engines, these recommendations are customised for you based on the data they have about you. For online retailers who sell products and services via the web, you will likely have seen recommendations when you add items to your cart: "People who purchased this item also purchased ..." Or "Top picks for you!" These are personalised for you!

Personalised recommendations are quite harmless, and many people prefer them to blanket advertising, so it may seem like an area where trust and explainability are not so useful. However, let's look at a simple case in which trust in a retailer could be destroyed without explainability.

Imagine that you have just been to a local book store and purchased Daniel Dennett's book *From Bacteria to Bach and Back: The Evolution of Minds* using your credit card. If you have not read this already, I strongly recommend it (no, I am not an algorithm). When you get home, you go to an online retailer called Rainforest, one of the largest online retailers in

the world, log in, and you see "Recommended for you: *From Bacteria to Bach and Back: The Evolution of Minds* by Daniel C Dennett!". Wait, how did they know this? You just purchased this book at a bookstore and now a powerful retailer recommends it to you. This is too much of a coincidence! You conclude that your credit card company must be selling your purchase information to Rainforest!

This may seem like a strawman, but it is based on a situation that a colleague of mine encountered; although his taste in literature is not as good as mine so he did not buy a Daniel Dennett book. My colleague was and probably still is strongly suspicious that the large retailer he used purchased data from his credit card company, but he had no way to find out, short of phoning them, which gave him no answers either. My question is: if he had access to an explanation about why this book was recommended, he may have found something far less sinister. Then again, he may be right, but this chapter is not about selling data.

Returning to you, the protagonist in our story, imagine if the recommendation came with a link: "Find out why we recommended this". You click the link to find something reassuring:

> *We recommend "From Bacteria to Bach and Back: The Evolution of Minds" by Daniel C Dennett because on Thursday you viewed "The Book of Why: The New Science of Cause and Effect" by Judea Pearl and Dana MacKenzie, and customers who bought this item also bought "From Bacteria to Bach and Back: The Evolution of Minds".*

Such an explanation would likely allay your concerns about privacy — it would have for my colleague. It is reasonably simple — just showing the correlations between purchases of customers.

Of course, this does not explain the decision in its entirety, and also, it may not satisfy everyone. In particular, it does not mean this is the real reason. The real reason could be because the seller has an oversupply of Daniel Dennett's book and is trying to promote it. Nonetheless, it should cause many people to eliminate the possibility that their credit card purchases were shared. So, even for low-stakes decisions, explainability can be useful; and also, the quality of explanations need not be as high as for high-stakes decisions.

Higher stakes: automated CV matching

In recent years, we have seen the application of machine learning to assist with job recruitment. This technology promises some wonderful efficiencies: instead of manually assessing each applicant for a job, their CV is automatically ranked by an algorithm. The human recruiter needs only to look at the top fraction. These algorithms are trained by taking a long list of previous applications, with three basic types of information: the CVs of the applicants, the job description they applied for, and the outcome of that hiring decision. A machine-learning algorithm is then trained to look for patterns in how previous hiring managers have made decisions.

At its core, this is a useful idea. I have been on recruitment panels that have received 400–500 applicants. It would have been wonderful to only have to look at, for example, the top 20

of those. From that 20, we could manually inspect to see who was going to be shortlisted.

However, the algorithms are just trained to detect patterns in previous hiring decisions. The human analogy would be to just give the CV a cursory glance and decide based on key phrases. Surely, hiring panels of people do not just scan for patterns on and discard people who do not fit with their intuitive notion of a good applicant? Unfortunately, it seems that they do, and this is where the problem lies! We saw earlier the example of Amazon's automated CV scanner learning the hiring biases made by previous recruitment panels. In many cases, biases in hiring decisions against minorities are based solely on cognitive biases of which the panel would be largely unaware.

Can explainability help with this? It cannot remove bias — the bias is in the data. But it can help people determine whether their algorithm is making biased decisions.

Consider a situation of a company trying to decide whether it should use an automated CV scanner. A reliable CV scanner would have value for the human resources (HR) team, avoiding having to sift through hundreds of CVs each time a recruitment round closes. The HR team adopt a tool called Perfect Match,[1] which promises to do the job. How can the HR team use explainability to check that Perfect Match does not discriminate?

The HR team decide to use an explainability system that their data science team recommend. It includes two useful algorithms: (1) one for determining the importance of features in the input; and (2) another for *counterfactual analysis*.

Feature importance is a technique that determines which of the inputs provided to an algorithm most heavily influenced its decision. In the case of the CV scanning, this would be the information on the CV and job description. Feature importance techniques usually operate by randomly varying inputs, such as information on a CV, to see whether it changes the output. If we do this many times for many inputs, it might see that the original outcome (e.g. an application not shortlisted) is always the output of the algorithm when a particular input is randomly changed. For example, it would take a CV and generate 100 copies in which it varied just the name of the university from which the applicant graduated, and then run this through the CV scanning algorithm. If the outcome was "shortlisted" for all of these variations, it would conclude that, *for this applicant*, the university they attended was not important, meaning that some other factors, such as experience, were weighing higher.

The HR team go back through a series of recent hires and run their CVs and the job application through the matching algorithm. The explainability algorithm gives them say, the top 10 items that influenced a decision to shortlist or not shortlist. The algorithm also gives the top 10 negative influence features: those things on the CV that made the application weaker; for example, a weak grade-point average may lead to a lower ranking.

They run one CV through: a male software engineer, recently graduated from a good university. The explainability algorithm shows the top 10 features, which include: the title of the degree, the grade-pointed average, the name of the univer-

sity, that the applicant was treasurer for his university's computer science student society for a year, and so on. All things that one would expect are signs of a good application. One strange thing comes from the negative influence features: the applicant included "gymnastics" as part of his hobbies. The team brush this off as an anomaly: nothing is perfect.

The second applicant is a female computer science and statistics graduate, with three years' experience in a similar role. She does not make the shortlist. According to the explainability algorithm, her strong points are her degree title, and her three years' experience, especially the job title "data scientist". However, for the features leading to her not being hired, the HR team again sees "gymnastics" as a hobby, and more worryingly, weighing the applicant down is that she was the president of the Women in Tech Society at her university and also a member of IEEE Women in Engineering.

The HR team become suspicious, so they try out the other explainability feature: counterfactual analysis. This gives them the ability to take a CV and construct a new, fictitious CV that looks similar to the first, but for which the outcome is different; for example, shortlisted versus not shortlisted. That is, it allows the team to determine what are the things the applicant would have had to have done differently to be shortlisted. They run the second applicant's CV through the counterfactual analysis. Unsurprisingly, her degree, experience, and job role do not change much, but "gymnastics" becomes "chess" and her volunteering work is removed on the fictitious CV.

As a final confirmation, the company uses another part of the counterfactual analysis tool: an algorithm that classifies

gender based on CV — something specifically used to determine bias in algorithms. The fictitious CV is classified as Male, while the original is classified as Female. So, the way for the second applicant to be shortlisted is to present themselves as being "more male" — that is, remove volunteering from groups for females and have hobbies that are more stereotypically male.

The HR team agree that Perfect Match is not for them, despite its catchy name.

High stakes and real time: air traffic control

The amount of air traffic in the world is increasing every year. More people are flying, more flights are in air space, requiring more fine-grained control of air space. As a result, more air traffic controllers are needed. However, air traffic controllers are becoming increasingly harder to find. The result: a move towards assisting air traffic controllers so that each controller could control a larger airspace.

In this context, one important set of decisions is to detect conflicts (two or more flights passing less than the minimum separation distance), redirect one or more of the flights to another airspace, and assess the impact of this move to ensure it does not create another conflict. With safety being paramount in air travel, these are decisions that need to be correct; and more importantly, they are real-time decisions, meaning that there is a hard deadline of 3–5 minutes that decisions need to be made to prevent conflicts. The stakes are high, and time is constrained. It is no wonder that air traffic controllers are required to do extensive training and, at least in

many countries, are only able to control a small set of airspaces on which they are trained.

The conflict-detection problem is well known, and it may seem quite easy. However, it is not just the congested airspace that makes this a problem. First, controllers are responsible for only a small part of airspace, and they need to consider flights coming in from neighbouring airspaces that they will be responsible for in the near future. Second, visual displays of airspace are useful, but they are two-dimensional — the controllers have an overhead view of the airspace. This is not enough information to make the call on potential conflicts, because two flights heading towards the same longitude and latitude may be separated safely by height. As such, controllers need to do mental calculation of the height to avoid rerouting too many flights.

Such a decision is ripe for assistance from automated decision making. A system that alerts controllers reliably (always safe, not too many false positives) would be useful. In fact, such systems already exist, but the lack of explainability means they are either not deployed or are never used by controllers. The time taken for controllers to understand why recommendations are made is often more than making the decision themselves. Often, the controllers are distracted by false positive alerts, and this diverts their attention away from the remaining airspace. Further, controllers use heuristics to solve these issues, and optimal solutions proposed by algorithms are often counter-intuitive for controllers, even if they are correct.

This is a situation in which explainability could help. Any explanation would need to be brief, summarised, and meaningful to controllers. They also need to be interrogable — if an unsafe situation arises, aviation authorities will need to be able to trace what happened and why the combination of human and system made certain decisions, which means more elaborate explanations must be available that need not be consumed in real time.

Now, imagine a situation in which an air traffic controller is alerted to a potential conflict in a congested airspace. Two flights are heading towards the same airspace and will cross paths at an unsafe distance in four minutes. The controller askes for an explanation, and is given the following details:

1. Flight 1 from Sydney to Melbourne is travelling at 380 km/h straight through region 54 at altitude of 14,200 metres above sea level.
2. Flight 2 from Melbourne to Sydney is travelling at 340 km/h towards region 54 at an altitude of 14,550 metres above sea level.
3. Both flights will occupy region 54 at 16:29.08, with vertical separation of just 350 metres and horizontal separation of 3 km.

Combined with the visual overhead, the controller makes a decision that while 3 kilometres is too close horizontally, 350 metres is a safe enough distance vertically, as it is slightly above the threshold of minimum separation and the algorithm tends to be conservative in its estimates due to noisy data. However, the controller knows that in the areas the flights are

travelling, location estimates are highly accurate. The controller dismisses the alert, happy that they have avoided re-routing one of the flights unnecessarily. A success for explainability!

Or is it? Ninety seconds later, the controller receives another alert about the same two aircraft, informing that flights will cross paths at an unsafe distance in two and a half minutes. Puzzled, the controller revisits and is given the explanation:

1. Flight 1 from Sydney to Melbourne is travelling at 380 km/h straight through region 54 at altitude of 14,320 metres above sea level.
2. Flight 2 from Melbourne to Sydney is travelling at 340 km/h towards region 54 at an altitude of 14,550 metres above sea level
3. Both flights will occupy region 54 at 16:29.08, with vertical separation of just 230 metres and horizontal separation of 3 km.

This time, the vertical separation is just 230 metres, because Flight 1's height is now 14,320 metres above sea level. What has happened? Looking into the situation, the controller sees that Flight 1 was ascending at the time of the initial alert. The two-dimensional display does not capture this, and in order to be brief, the explanation system included only the most highly-relevant details, omitting many other factors. The controller asks Flight 1 to descend to 14,100 metres, and the conflict is resolved safely.

This scenario is intended to illustrate a potential flaw in explainability systems: in most cases, they will need to omit

information that is used as part of a decision. The strength of automated decision-making systems is the number of variables that they can consider as part of a decision — they do not have the same cognitive limitations as humans. However, explainable systems, in most cases, are limited by human cognitive abilities, and almost by definition, they will need to omit some information that is used as part of the decision so that the human can understand. This is particularly in time-constrained environments like air traffic control. In this scenario, had the explanation not been given, the controller may not have chosen to focus on the reported distance, which is what they could see, but may have also investigated other variables, such as whether the flights were ascending or descending.

Clearly, this particular scenario is a strawman — no explanatory system for air traffic control would ever ignore that a flight is ascending or descending. Such information is too important to be left out. However, it is illustrative that some factors that play a role in decision will be omitted, and it is human tendency to ignore factors that we cannot see. As a reader, did you question whether other information was relevant? Of course, air traffic controllers are highly trained to look for such factors, but one focus of current explainability research is how we can determine which factors are important for explainability and which can be omitted.

Discussion

In this chapter, I discussed what explainability is in AI and why it is important, and hypothesised what it may look like to have explainable algorithms in the near future — both how it can help and hinder.

However, it is important to note two things. First, the ideas of what explainability may look like in these three domains is based entirely on a guess of what I think would be useful, and are not based on any serious study of how people understand the decisions made in these domains. Research is required that not just implements the algorithms of explainability, but determines what different people in a particular domain want to know (the job applicant may want different explanations from the recruiter), how they interpret it, how they evaluate it, whether it helps them or just takes up cognitive load, and whether it gives them an actionable insight into future decisions. If we simply guess at these things, explainability algorithms will be ignored just as much as the algorithms they are expected to explain.

Second, some of the explainability techniques discussed in this chapter are ideas already being researched and deployed, such as feature important and counterfactual analysis. However, currently, these techniques are not at the stage that they can be deployed widely. The algorithms can be slow, they can give counter-intuitive and plainly wrong explanations, and there has been very little work to validate that they help people understand decisions.

These two points together imply that more research is required in explainability to bring about something like the future I paint. Importantly, this research must be multi-disciplinary, bringing together researchers from computer science, cognitive and social psychology, interaction design, and domain experts to explore these interesting questions. If we fail at this, I predict it is likely that the promise of using AI to

improve high-stakes decisions will not make it out of the laboratory.

Endnote

1 The name Perfect Match is made up for the purposes of the article. I could not find a CV scanning tool of this name, but if you do start a CV matching company and use this name, I want a cut!

References

Cowan N (2010). The Magical Mystery Four: How is working memory Capacity limited, and why? *Current Directions in Psychological Science*, *19*, 51–57.

Gunning D (2017). Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Hoffman et al. (2013). Trust in automation. *IEEE Intelligent Systems*, *28*, 84–88.

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