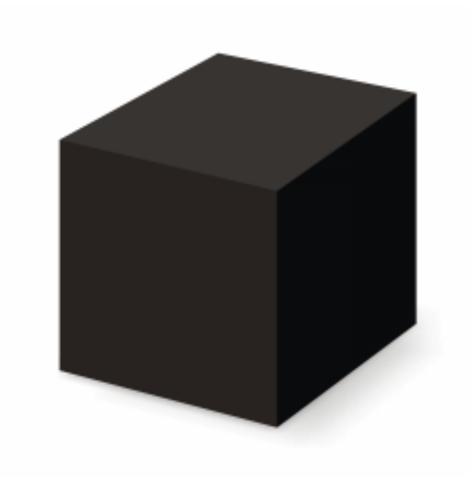
The “black box” metaphor in machine learning

Dallas Card Follow
Jul 5, 2017 · 14 min read



It has become quite common these days to hear people refer to modern machine learning systems as “black boxes”. As an example, consider a recent episode of the Sam Harris podcast in which he interviewed AI pioneer Stuart Russell. Harris asks:

> *“So, if I’m not mistaken, most, if not all of these deep learning approaches, or even more generally machine learning approaches are, essentially black boxes, in which you can’t really inspect how the algorithm is accomplishing what it is accomplishing.”*

Although this metaphor is appropriate for some particular situations, it is actually quite misleading in general, and may be causing a considerable amount of confusion. As we’ll see, a deep learning system is not a black box; even the development of such a system need not be a black box. The real challenge, however, is that both of these things are complex, and not necessarily well understood. Here, I want to try to clarify some of these ideas, and at the same time think through what we mean by *explanations*.

As I'll explain below, I believe the confusion arises at least in part from the misconceptions people have about how these systems work. When people reach for the black box metaphor, what they seem to be expressing is the fact that it is difficult to make sense of the *purpose* of the various components in a machine learning model. Although this is indeed difficult, I want to argue that it is also an unrealistic expectation. Along the way, I'll try to explain the difference between models and how they are trained, discuss scenarios in which the black box metaphor *is* appropriate, and suggest that in many ways, humans are the real black boxes, at least as far as machine learning is concerned.

1. Explanations

To begin, it is useful to reflect upon what people mean when they talk about explanations. This is by no means a trivial question, but there seem to be at least two particularly relevant ways of thinking about this.

When we ask someone for an explanation of why they did something (*"Why did you do X?"*), we're operating on a certain set of background assumptions. In the case of a decision that was carefully made, we are typically assuming that they had some *good reason* for acting as they did, and we are basically asking for the reasoning process they used to make the decision. For example, we might expect that they weighed the pros and cons and chose a course of action based on the expectation of it leading to some particular outcome.

When asking about why something went wrong, by contrast, we are instead asking for a kind of post-hoc explanation of failure. For example, after a car accident, we might want an explanation of what caused the accident. Was the driver distracted? Did another car cause them to swerve? Rather than a process of reasoning, we are asking, more or less, for the critical stimulus that caused a particular reaction outside of normal behaviour.

When people think about artificial intelligence, they typically seem to have in mind the first kind of explanation. The expectation is that the system made a deliberation and chose a course of action based on the expected outcome. Although there are cases where this is possible, increasingly we are seeing a move towards systems that are more similar to the second case; that is, they receive stimuli and then they just *react*.

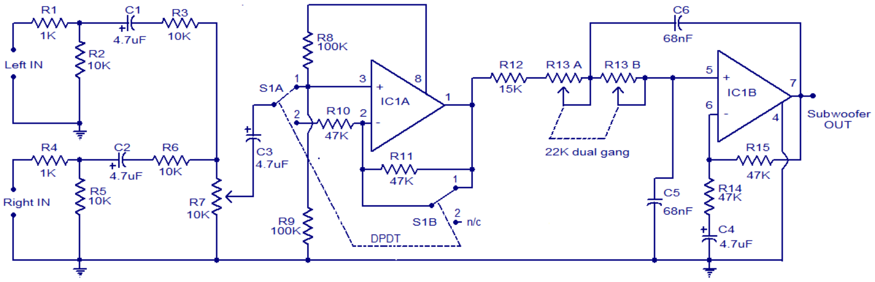
There are very good reasons for this (not least because the world is complicated), but it does mean that it's harder to understand the *reasons* for why a particular decision was made, or why we ended up with one model as opposed to another. With that in mind, lets dig into what we mean by a model, and the metaphor of the black box.

2. Boxes and Models

The black box metaphor dates back to the early days of cybernetics and behaviourism, and typically refers to a system for which we can only observe the inputs and outputs, but not the internal workings. Indeed, this was the way in which B. F. Skinner conceptualized minds in general. Although he successfully demonstrated how certain learned behaviours could be explained by a reinforcement signal which linked certain inputs to certain outputs, he then famously made the mistake of thinking that this theory could easily explain all of human behaviour, including language.

As a simpler example of a black box, consider a thought experiment from Skinner: you are given a box with a set of inputs (switches and buttons) and a set of outputs (lights which are either on or off). By manipulating the inputs, you are able to observe the corresponding outputs, but you cannot look inside to see how the box works. In the simplest case, such as a light switch in a room, it is easy to determine with great confidence that the switch controls the light level. For a sufficiently complex system, however, it may be effectively impossible to determine how the box works by just trying various combinations.

Now imagine that you are allowed to open up the box and look inside. You are even given a full wiring diagram, showing what all the components are, and how they are connected. Moreover, none of the components are complex in and of themselves; everything is built up from simple components such as resistors and capacitors, each of which has behaviour that is well understood in isolation. Now, not only do you have access to the full specification of all the components in the system, you can even run experiments to see how each of the various components responds to particular inputs.



You might think that with all this information in hand, you would now be in a position to give a good explanation of how the box works. After all, each individual component is understood, and there is no hidden information. Unfortunately, complexity arises from the interaction of many simple components. For a sufficiently complex system, it is unlikely you'd be able to predict what the output of the box will be for a given input, without running the experiment to find out. The only explanation for why the box is doing what it does is that all of the components are following the rules that govern their individual behaviour, and the overall behaviour emerges from their interactions.

Even more importantly, beyond the *how* of the system, you would likely be at a loss to explain *why* each component had been placed where it is, even if you knew the overall purpose of the system. Given that the box was *designed* for some purpose, we assume that each component was added for a reason. For a particularly clever system, however, each component might end up taking on multiple roles, as in the case of DNA. Although this can lead to a very efficient system, it also makes it very difficult to even think about summarizing the *purpose* of each component. In other words, the *how* of the system is completely transparent, but the *why* is potentially unfathomable.

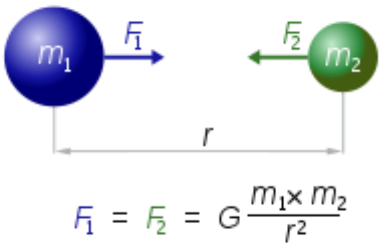
This, as it turns out, is a perfect metaphor for deep learning. In general, the entire system is open to inspection. Moreover, it is entirely made up of simple components that are easily understood in isolation. Even if we know the purpose of the overall system, however, there is not necessarily a simple explanation we can offer as to *how* the system works, other than the fact that each individual component operates according to its own rules, in response to the input. This, indeed, is the true explanation of how they system works, and it is entirely transparent. The tougher question of course is *why* each component has taken on the role that it has. To understand this further, it will be helpful to separate the idea of a *model* from the *algorithm* used to train it.

3. Models and Algorithms

To really get into the details, we need to be a bit more precise about what we're talking about. Harris refers to "how the algorithm is accomplishing what its accomplishing", but there are really two parts here: a model—such a deep learning system—and a learning algorithm—which we use to fit the model to data. When Harris refers to "the algorithm", he is presumably talking about the model, not necessarily how it was trained.

What exactly do we mean by a model? Although perhaps somewhat vague, a statistical *model* basically captures the assumptions we make about how things work in the world, with details to be learned from data. In particular, a model specifies what the inputs are, what the outputs are, and typically how we think the inputs might interact with each other in generating the output.

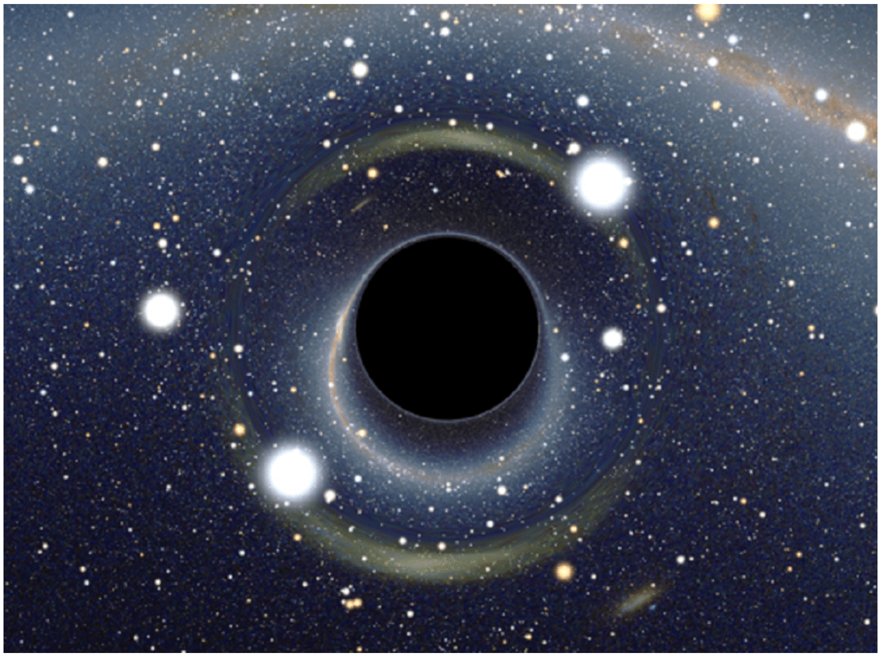
A classic example of a model is the equations which govern Newtonian gravity. The model states that the output (the force of gravity between two objects) is determined by three input values: the mass of the first object, the mass of the second object, and the distance between them. More precisely, it states that gravity will be proportional to the product of the two masses, divided by the distance squared. Critically, it doesn't explain *why* these factors should be the factors that influence gravity; it merely tries to provide a parsimonious explanation that allows us to predict gravity for any situation.



Of course, even if this were completely correct, in order to be able to make a prediction, we *also* need to know the corresponding scaling factor, *G*. In principle, however, it should be possible to learn this value through observation. If we have assumed the correct (or close to correct) model for how things operate in reality, we have a good chance of being able to learn the relevant details from data.

In the case of gravity of course, Einstein eventually showed that Newton's model was only approximately correct, and that it fails in extreme conditions. For most circumstances, however, the Newtonian model is good enough, which is why people were able to learn the constant *G*= 6.674×10^(−11) N · (m/kg)^2, and use it to make predictions.

Einstein's model is much more complex, with more details to be learned through observation. In most circumstances, it gives approximately the same prediction as the Newtonian model would, but it is more accurate in extreme circumstances, and of course has been essential in the development of technologies such as GPS. Even more impressively, the secondary predictions of relativity have been astounding, successfully predicting, for example, the existence of black holes before we could ever hope to test for their existence. And yet we know that Einstein's model too, is not completely correct, as it fails to agree with the models of quantum mechanics under even more extreme conditions.



Gravitation, of course, is deterministic (as far as we know). In machine learning and statistics, by contrast, we are typically dealing with models that involve uncertainty or randomness. For example, a simple model of how long you are going to live would be to just predict the average of the population for the country in which you live. A better model might take into account relevant factors, such as your current health status, your genes, how much you exercise, whether or not you smoke cigarettes, etc. In pretty much every case, however, there will be some uncertainty about the prediction, because we don't know all the relevant factors. (This is different of course from the apparent true randomness which occurs at the sub-atomic level, but we won't worry about that difference here).

In addition to being an incredibly successful rebranding of neural networks and machine learning (itself arguably a rather successful rebranding of statistics), the term *deep learning* refers to a particular type of model, one in which the outputs are the results of a series of many simple transformations applied to the inputs (much like our wiring diagram from above). Although deep learning models are certainly complex, they are not black boxes. In fact, it would be more accurate to refer to them as glass boxes, because we can literally look inside and see what each component is doing.

The problem, of course, is that these systems are also complicated. If I give you a simple set of rules to follow in order to make a prediction, as long as there aren't too many rules and the rules themselves are simple, you could pretty easily figure out the full set of input-to-output mappings in your mind. This is also true, though to a lesser extent, with a class of models known as *linear models*, where the effect of changing any one input can be interpreted without knowing about the value of other inputs.

Deep learning models, by contrast, typically involve non-linearities and interactions between inputs, which means that not only is there no simple mapping from input to outputs, the effect of changing one input may dependent critically on the values of other inputs. This makes it very hard to mentally figure out what's happening, but the details are nevertheless transparent and completely open to inspection.

The actual computation performed by these models in making a prediction is typically quite straightforward; where things get difficult is in the actual learning of the model parameters from data. As described above, once we assume a certain form for a model (in this case, a flexible neural network); we then need to try to figure out good values for the parameters from data.

In the example of gravity, once we have assumed a "good enough" model (proportional to mass and inversely proportional to distance squared), we just need to resolve the value of one parameter (*G*), by fitting the model to observations. With modern deep learning systems, by contrast, there can easily be *millions* of such parameters to be learned.

In practice, nearly all of these deep learning models are trained using some variant of an algorithm called stochastic gradient descent (SGD), which takes random samples from the training data, and gradually adjusts all parameters to make the predicted output more like what we want. Exactly why it works as well as it does is still not well understood, but the main thing to keep in mind is that it, too, is transparent.

Because it is usually initialized with random values for all parameters, SGD can lead to different parameters each time we run it. The algorithm itself, however, is deterministic, and if we used the same initialization and the same data, it would produce the same result. In other words, neither the model nor the algorithm is a black box.

Although it is somewhat unsatisfying, the complete answer to why a machine learning system did something ultimately lies in the combination of the assumptions we made in designing model, the data it was trained on, and various decisions made about how to learn the parameters, including the randomness in the initialization.

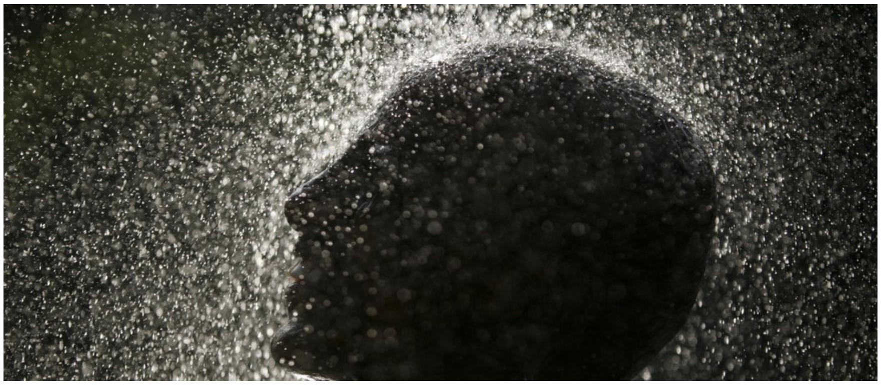
4. Back to black boxes

Why does all this matter? Well, there are at least two ways in which the concept of black boxes *are* highly relevant to machine learning.

First, there are plenty of algorithms and software systems (and not just those based on machine learning) that *are* black boxes as far as the user is concerned. This is perhaps most commonly the case in proprietary software, where the user doesn't have access to the inner workings, and all we get to see are the inputs and outputs. This is the sort of system that ProPublica reported on in it's coverage of judicial sentencing algorithms (specifically the COMPAS system from Northpointe). In that case, we know the inputs, and we can see the risk scores that have given to people as the output. We don't, however, have access to the algorithm used by the company, or the data it was trained on. Nevertheless, it is safe to say that *someone* has access to the details—presumably the employees of the company—and it is very likely completely transparent to them.

The second way in which the metaphor of black boxes is relevant is with respect to they systems we are tying to learn, such as human vision. In some ways, human behaviour is unusually transparent, in that we can actually ask people why they did something, and obtain explanations. However, there is good reason to believe that we don't always know the true reasons for the things we do. Far from being transparent to ourselves, we simply don't have conscious access to many of the internal processes that govern our behaviour. If asked to explain why we did something, we may be able to provide a narrative that at least conveys how the decision making process felt to us. If asked to explain how we are able to recognize objects, by contrast, we might think we can provide some sort of explanation (something involving edges and colours), but in reality, this process operates well below the level of consciousness.

Although there are special circumstances in which we can actual inspect the inner workings human or other mammalian systems, such as neuroscience experiments, in general, we are trying to use machine learning to mimc human behaviour using only the inputs and the outputs. In other words, from the perspective of a machine learning system, *the human is the black box*.



6. Conclusion

In conclusion, it’s useful to reflect on what people want when they think of systems that are *not* black boxes. People typically imagine something like the scenario in which a self-driving car has gone off the road, and we want to know why. In the popular imagination, the expectation seems to be that the car must have evaluated possible outcomes, assigned them probabilities, and chose the one with the best chance of maximizing some better outcome, where better is determined according to some sort of morality that has been programmed into it.

In reality, it is highly unlikely that this is how things will work. Rather, if we ask the car why it did what it did, the answer will be that it applied a transparent and deterministic computation using the values of its parameters, given its current input, and this determined its actions. If we ask why it had those particular parameters, the answer will be that they are the result of the model that was chosen, the data it was trained on, and the details of the learning algorithm that was used.

This does seem frustratingly unhelpful, and it is easy to see why people reach for the black box metaphor. However, consider, that we don’t actually have this kind of access for the systems we are trying to mimic. If we ask a human driver why they went off the road, they will likely be capable of responding in language, and giving some account of themselves—that they were drunk, or distracted, or had to swerve, or were blinded by the weather—and yet aside from providing some sort of narrative coherence, we don’t really know why they did it, and neither do they. At least with machine learning, we can recreate the same setting and probe the internal state. It might be complicated to understand, but it is not a black box.