Can Counterfactuals Provide AI Explainability?

Jonathan P. Chang

May 24, 2020

Abstract

The increasing use of autonomous AI systems to make decisions that directly affect people's livelihoods—e.g., loan approvals, job applications, and prison sentencing—has sparked interest in developing methods to *explain* the decisions made by such systems. One recent paper by Wachter et al. (2018) proposes the use of *counterfactuals* as explanations of AI decision making. In this paper, I critically examine this proposal by using two prominent interpretations of counterfactuals, possible worlds semantics and structural equation models, to analyze how well Wachter et al.'s counterfactual explanations meet the goals of explainability as defined in the literature.

1 Introduction

Recent years have seen a rapid increase in the public deployment of autonomous systems driven by artificial intelligence (AI)—that is, systems which can algorithmically make complex decisions with little to no human input. Although AI systems hold great promise for productivity and efficiency, incidents like a self-driving car failing to identify a pedestrian[1] and a facial recognition system displaying racial bias[2] have led to concerns about the fairness and correctness of such systems.

[1] https://www.theverge.com/2019/5/17/18629214/tesla-autopilot-crash-death-josh-brown-jeremy-banner

[2] https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender

In response, a growing body of literature on “explainable AI” now aims to develop formal methods for analyzing and understanding the decisions made by AI systems, enabling systematic evaluations of fairness and correctness. “Explainability” is hard to define precisely, since it can mean different things to different stakeholders, so the best way to understand it is in terms of its goals. One recent survey of the field (Adadi and Berrada, 2018) identifies four goals for explainability:

1. **Justification**: An explanation should reveal the reasons behind the system’s decision. This provides a level of transparency; for instance, such an explanation can provide assurance that the system is not secretly making use of a protected category like race or gender.

2. **Control**: An explanation should offer insight into how the system can be changed to yield a different outcome, allowing its designers to identify and correct errors.

3. **Improvement**: An explanation should highlight areas of weakness in the system that its designers could use as a starting point for making improvements.

4. **Discovery**: An explanation should reveal some facts about the world that might be helpful to people even in other contexts; for instance, an explainable chess AI could be used to discover new strategies.

Of course, this taxonomy only lays out what an explanation should *achieve*, not what it should concretely *look* like. Thus, the bulk of the explainable AI literature is dedicated to proposal and discussion of concrete methods for achieving these goals. Before I can explain these methods, however, I must first introduce some technical terminology that I will use throughout the paper to discuss AI systems at a more detailed level.

1.1 Technical Notation

As a precise technical explanation of how AI systems work is beyond the scope of this paper, throughout the discussion I will describe the functioning of an AI system merely at an abstract level using the following terminology. Let S denote an AI system. Then, we say that S makes a decision d when S is applied to some *subject* or *input* x; I denote this in shorthand as $S(x) = d$.

The exact nature of both d and x depends on what task S is designed to perform; for example, in a self-driving car x might be an image of the road, and d might correspond to a physical action like braking. However, one property that holds across nearly all tasks is that the input x is not a primitive, monolithic entity; rather, it is a complex object made up of simpler components. This can again be illustrated by the self-driving car example: the input image x is made up of pixels. In general, the individual components that comprise an input x are referred to as *features*. When needed, I will use the notation $x = \langle x_1, x_2, \ldots, x_n \rangle$ to explicitly "decompose" x into its component features; in this notation, each x_i (i denoting an integer from 1 to n) denotes an individual feature (e.g., one pixel).

1.2 Approaches to Explainability

While there have been a number of proposals for concrete approaches to explainability (surveyed in Adadi and Berrada (2018); Arrieta et al. (2019)), the leading approaches that have actually been put into real-world use can be roughly categorized into three types (Bhatt et al., 2019): *feature importance*, *adversarial perturbation*, and *counterfactual explanations*. This paper focuses on the latter approach, counterfactual explanations, and I define it in detail in Section 2. However, for the sake of comparison I will also briefly describe the other two approaches here at a high level, skipping the mathematical details of exactly how each approach is implemented.[3] After laying out my critique of counterfactual explanations, I will briefly return to these other approaches in Section 4 to compare the practical merits of counterfactual explanations compared to other techniques in light of my critique.

Feature importance is a well-established technique and is the most commonly used approach to AI explainability in practice. At the core of this technique are so-called "explanation functions": for a system S that makes decision d on input x, an explanation function produces an "importance score" ϕ_i for each feature x_i in x. As the name implies, each importance score is a measure of how important the corresponding feature was to S when making decision d; a higher score means higher importance. To illustrate, consider applying this technique to explain why a self-driving car stopped. If the self-driving car is making decisions similar to how a human driver would,

[3] My high-level descriptions are summaries of the technical definitions provided by Bhatt et al. (2019); readers interested in the full technical description of these methods may consult that paper.

then we would expect that red pixels get high importance scores (since such pixels might correspond to stop signs and red traffic lights), while blue pixels get low importance scores (since such pixels don't correspond to any relevant traffic signal).

In contrast to feature importance, which produces evaluations of individual inputs, the *adversarial perturbation* approach instead aims to evaluate the system S as a whole. Specifically, it characterizes the "robustness" of S, which can be loosely defined as a measure of how hard it is to trick S into making a wrong decision. For any input x for which S makes a correct decision d (where correctness is according to a human judge), adversarial perturbation searches for the smallest possible change to x that will cause S to make an incorrect decision. Intuitively, a small change should be one that would *not* fool a human. For example, some famous work in this direction showed that an AI system designed to identify different kinds of animals in an image can be tricked into misclassifying one animal as another (e.g., thinking a cat is a dog) by adding a miniscule amount of noise to the image, so small that a human would not notice the change (and would thus not be fooled). Robustness is then defined as the average amount of change needed to trick S in this way, capturing the intuition that a robust system should not be sensitive to small changes that a human would not notice.

Having introduced these other approaches to explainability, I can now turn to the focus of this paper: the third approach, *counterfactual explanations*, originally proposed by Wachter et al. (2018).

2 The Counterfactual Approach to Explainability

A *counterfactual explanation* of a decision $S(x) = d$ is a counterfactual sentence where the antecedent describes a change to the input x' and the consequent describes a change in the decision d'. Wachter et al. give the following example in the context of an AI system for processing loan applications:

> If your income had been $45,000, you would have been offered a loan [by the system].

According to Wachter et al., this kind of sentence serves as an explanation in the sense that it "describes a dependency on external facts that led to the decision." The key advantage of counterfactual explanations over other approaches, as touted by Wachter et al., is their intuitiveness to a general audience: whereas a layperson might not know what an importance score or robustness score means for them in practice, counterfactuals are a common feature of everyday language, and as such the layperson can much more naturally understand a counterfactual explanation. For instance, in a loan application setting, an applicant might want to know what actions they could take to improve their chances of getting a loan; importance scores and robustness scores do not immediately translate to any specific action, whereas the counterfactual explanation in Wachter et al.'s example points out a specific action in an easy to understand way.

Wachter et al. do not merely propose such counterfactual sentences as some abstract ideal of what an explanation should look like—they lay out a concrete technical methodology for automatically *generating* such explanations. As before, the mathematical details of how this method works are beyond the scope of this paper,[4] but at a high level, the method systematically simulates changes to the input x (e.g., changing the feature representing income from \$30,000 to \$45,000), then applies S to each changed input x' and checks whether S's decision changes. Every x' that leads to a changed decision d' can then be interpreted as a counterfactual explanation: if x' had been the case, S would have made decision d'. As a terminological note, henceforth I will adopt the terminology $A > B$ as a shorthand for the counterfactual "if A had been the case, B would have been the case", so the above counterfactual explanation can be succinctly expressed as $x' > d'$.

Bhatt et al. (2019) have observed that from a technical perspective, counterfactual explanations and adversarial perturbation are closely related, in that both involve systematically changing the input x to change the decision d. The core difference between the two lies in what *kind* of change they seek to make. Where adversarial perturbation looks for small changes that are unnoticeable to humans, counterfactual explanations seek the opposite: the change should be noticeable and meaningful to a human, as it is meant to capture the intuitive notion of "how things could have been." Furthermore, where adversarial perturbation focuses on tricking the system into making an incorrect decision, for counterfactual explanations the correctness of the changed decision is irrelevant: regardless of whether or not granting a loan to someone with an income of \$45,000 is the "right" decision, Wachter et al.'s

[4] Interested readers may consult Section III of Wachter et al. (2018)

example counterfactual still usefully reveals a property of the loan application system (namely its reliance on income) that both users and developers of the system might be interested in knowing.

3 Evaluating Counterfactual Explanations

Purported advantages notwithstanding, the use of counterfactuals to explain AI decisions runs into philosophical difficulties regarding what counterfactuals actually *mean*—that is, how to formally evaluate a counterfactual sentence. There are two leading approaches to interpreting counterfactuals: possible worlds semantics and structural equation models. I will apply both of these approaches to evaluate Wachter et al.-style counterfactual explanations, and show that while each method yields slightly different results, they both reveal potential problems for counterfactual explanations in terms of meeting the four goals of explainability.

3.1 Possible Worlds Semantics

Possible worlds semantics, exemplified by Lewis (1973b), holds that the meaning of the counterfactual $A > B$ is that given some set W of *possible worlds*, every world $w \in W$ where A holds is also a world where B holds. Specific theories differ in their definition of W; most notably, *strict* analyses treat W as the worlds that are "accessible" from the actual world, while *similarity* analyses treat W as the worlds that are "most similar" to the actual world (Starr, 2019). There has been much philosophical debate over how terms like "accessible" and "similar" should be precisely defined, but we need not delve into this. For our purposes, it suffices to use our intuitive, everyday understanding of "accessible" and "similar," and to treat accessibility and similarity as playing the same role of limiting the scope of possible worlds.

It is worth noting that Wachter et al.'s technical approach to generating counterfactual explanations is closely related to possible worlds. One way to understand a possible world is as a set of truth values Veltman (2005); Kratzer (2012). In other words, a possible world may be treated as a collection of statements like "John's income is \$45,000." In this sense, when Wachter et al. simulate changes to x, they are effectively enumerating possible worlds—each simulated change x' is an (incomplete) description of a possible world containing the fact that x' (instead of x) is the case. Likewise, running the system S on each input x' to obtain decision d' amounts to constructing the set W of accessible/most-similar worlds. Indeed, Wachter et al. themselves describe their method as discovering the most similar possible worlds.

3.2 Problems for Counterfactual Explanations Under Possible Worlds Semantics

Among the four goals of explainability, the one for which a problem most immediately appears under possible worlds semantics is **discovery**. If the counterfactual $x' > d'$ is merely making a statement about possible worlds where x' is the case, it is hard to see how that could translate to knowledge about the way our *actual* world works. To illustrate this concern, consider the discovery example used in the introduction: using a chess-playing AI to discover new chess strategies. Suppose that a chess AI looks at a particular board layout and decides to move its pawn, and Wachter et al.'s method generates the following counterfactual explanation of this decision:

> If the opponent's queen had been in play, the chess AI would have moved its knight (instead of moving its pawn).

According to possible worlds semantics, the formal meaning of the above sentence is:

> In the most similar world(s) where the opponent's queen was in play, the chess AI moved its knight.

Suppose now that a human player, in the same position as the AI system was (in the real world), is seeking guidance on what move to make. It is hard to see how the above statment, which makes some claim about a hypothetical alternate world where the opponent's queen is in play, could *on its own* be of any practical use to the human player, who is faced with the reality of the opponent's queen *not* being in play. To make such knowledge of the alternate world relevant to our decisions regarding the real world, we need an extra missing ingredient: namely, some kind of *relationship* between the alternate world and the real world. We know that the alternate world differs from the real world in (at least) two ways: the presence of the opponent's queen, and the chess AI's choice of which piece to move. If we had some additional knowledge of *how* those differences came about, then perhaps we could work backwards to construct a chain of reasoning corresponding to a concrete strategy. But possible worlds semantics, which operationalize the alternate world either as a primitive point in a space Lewis (1973b) or a collection of truth values Veltman (2005); Kratzer (2012) offers no such path to drawing connections between worlds.

A similar concern arises when considering **control** and **improvement**, which both involve finding ways to change the system's behavior. Let us return to Wachter et al.'s loan application example, repeated below:

> If your income had been $45,000, you would have been offered a loan [by the system].

According to possible worlds semantics, this translates to:

> In the most similar world(s) where your income is $45,000, the system offers you a loan.

Suppose that after a change in management, the bank becomes more risk-averse, and thus wants the system to use a higher income threshold for approving loans. Like in the chess case, the above sentence is arguably of little help to the system designer in determining what must be changed in the system to make it use a higher income threshold. This is because it makes a statement about the most similar possible world(s), whereas the designer needs a statement about the system itself. Specifically, while it may be true that the counterfactual explanation reveals *that* S depends on income, the designer needs to go one step further: they need to know *how* S depends on income; that is, what kinds of internal decision making processes the system is using when going from income to loan decision. But once again, possible worlds, being nothing more than primitive points or collections of truth values, do not obviously encode such information.

By contrast, **justification** may hold up better under possible worlds semantics. Adadi and Berrada explicitly define justification as *not* requiring any understanding of the inner workings of the system—it suffices to say that a feature x_i was a reason for d, not *why* this was the case. However, whether counterfactuals successfully do this depends on how one understands a "reason" for a decision. In particular, Wachter et al. consider "reason" to be synonymous with "cause." So for Wachter et al., to say:

> If your income had been $45,000, you would have been offered a loan by the system.

Is the same as to say:

> The system denied you a loan because your income was less than $45,000.

Thus, under this conceptualization of reason, counterfactual explanations reveal the reasons for a system's decision (and hence justify it) if and only if they reveal the causes of the system's decision. Certainly, some philosophers, notably Lewis (1973a), believe that counterfactuals do support causal reasoning: according to Lewis, causal dependence can be defined in terms of a counterfactual:

> e causally depends on c iff both e and c occurred, and if c hadn't occurred then e wouldn't have occurred.

Notably, the structure of this counterfactual directly mirrors those of Wachter et al.'s explanations, which take the form of describing a change to x (implying that x itself did not occur) and demonstrating a change in d (implying that d itself did not occur). Thus, if Lewis's account is correct, then Wachter et al.'s explanations constitute causal claims.

However, Lewis's account is subject to criticism, and one criticism that particularly affects Wachter et al. (2018)'s project is that of *preemption*. Preemption occurs when two events are individually sufficient to cause an outcome, but the fact that one event actually causes the outcome prevents the other cause from manifesting. This is illustrated in the following example from Hall (2004):

> Billy and Suzy throw rocks at a bottle. Suzy's rock hits first and shatters the bottle, and so Billy's ends up flying harmlessly past where the bottle once was.

In this example, our intuitive judgment is that Suzy throwing a rock caused the bottle to break. Lewis's account holds that the statement "Suzy's throw caused the bottle to break" is true if and only if had Suzy not thrown, the bottle would not have broken. The problem is that arguably, the most similar world in which Suzy did not throw is one where Billy still threw, and the known laws of physics still apply, so (because of Billy's rock) the bottle actually does end up breaking. Thus, Lewis's account seems to counterintuitively conclude that Suzy's throw was not a cause of the bottle breaking.

An analogous scenario can be described for an AI system, and causes a similar problem for Wachter et al. (2018):

> In AI system S, which is biased against older people, low income and old age are individually sufficient for S to deny a loan. Robert, who is both of old age and low income, applies for a loan, which S denies.

Consider an auditor investigating whether S has an age bias. They look for evidence that Robert's age was a reason for S's decision, and following Wachter et al. (2018) they ask "if Robert had been younger, would S have offered a loan?" The answer in this case is no, because Robert's income still triggers S's denial. Indeed, the problem is particularly vexing for Wachter et al. (2018) because it may affect their technical component as well: if S truly works as described, then presumably no amount of simulating changes to Robert's age alone would yield a different decision, so Wachter et al.'s would fail to reveal age as a reason for S's decision.

That being said, this criticism is not a complete nail in the coffin for possible worlds semantics. For one thing, there have been attempts to fix the Lewis (1973a) account to work for preemption cases. Furthermore, even if such fixes fail, all it would show is that causation is not fully reducible to counterfactuals. It may still be the case that counterfactuals work as a reasonable *heuristic* for causation, correctly describing causation in many common cases but failing on specific edge cases like those involving preemption. If so, then Wachter et al.'s method could still find practical use by likewise serving as a heuristic for justification, with the caveat that the justifications it uncovers may be incomplete and should thus be seen as a starting point, not a comprehensive picture.

In conclusion, we have found that under possible worlds semantics, counterfactual explanations may at least partly meet the goal of justification, but have difficulty meeting the remaining goals of discovery, control, and improvement. The problem of causality revealed itself to be relevant to whether counterfactual explanations provide justification.

In hindsight, causality could also be relevant to meeting the challenges posed by discovery, control, and improvement. When discussing control and improvement, I argued that possible worlds interpretations of counterfactual explanations are insufficient because they reveal factors that are relevant to a system's decision without describing *how* that factor is relevant; a causal statement of the form "x caused f_1, which caused f_2,. . .,which caused d" could provide this missing link. Similarly, in the case of discovery, I stated a need to understand how the differences between different worlds come about; a causal model that describes how changing one fact leads to changing another fact, and so on, seems like a prime candidate for providing this understanding. The apparent importance of causality thus motivates a look at the second approach to evaluating counterfactuals: structural equation models.

3.3 Structural Equation Models

Structural equation models (SEMs), introduced by Pearl (2000), are a way of formally representing causal relationships between facts in the world. At a high level, a SEM is a directed acyclic graph (DAG) where the nodes represent facts in the world, which can either be true or false, and an edge from $v_1 \rightarrow v_2$ says that v_2 depends on v_1. Dependency, in turn, is understood in terms of the titular structural equations: the value of a node v is set by a boolean equation (that is, an equation using logical connectives like AND ($\wedge$) and OR ($\vee$)) over all nodes with edges to v; formally:

$$v = f(v_1, v_2, \ldots, v_k), \text{where for each } v_i, \text{there exists an edge } v_i \rightarrow v$$

As a simple example, consider this basic loan approval model: the loan will be approved (denote the corresponding node as A) if the applicant is likely to repay it (P); in turn, the applicant is judged as likely to repay if their income is over \$45,000 (I) and their credit score is over 650 (C). This results in the following SEM (Figure 1):

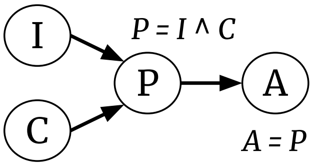

Figure 1: Visualization of the SEM for the basic loan approval example

To evaluate a counterfactual $A > B$ using an SEM, Pearl prescribes the following process known as *intervention*:

1. Remove all edges going into A (and their corresponding equations).

2. Set $A =$ True.

3. Fill in all other nodes according to remaining structural equations.

4. Check the resulting value of B.

For example, in the preceding simple loan model, to evaluate "if the model had judged you likely to repay the loan, the loan would have been granted" we remove the equation $P = I \wedge C$ and set $P =$ True, and according to the remaining equation $A = P$ we set A to True. Since A works out to be True, the counterfactual is true.

Compared to possible worlds semantics, SEMs look much more promising for **discovery**. Recall that in Section 3.2, I argued that the key missing ingredient for discovery was some notion of *relationships* between different possibilities; that is, systematic reasons that certain facts might be different in an alternate world. Such relationships between facts are exactly what SEMs are designed to encode! Consider an SEM interpretation of the counterfactual from the chess AI example:

> If the opponent's queen had been in play, the chess AI would have moved its knight (instead of moving its pawn).

Under the SEM approach, evaluating this counterfactual involves constructing an SEM that encodes relationships between board layouts and actions, and intervening on the node that represents the presence of the opponent's queen. This intervention, in turn, sets off a chain reaction in the rest of the DAG, in which subsequent nodes have their values updated to account for the presence of the opponent's queen. At a high level, tracing this chain reaction and the nodes involved could serve as a description of why the queen's presence led to the knight being moved. Such a description could be thought of as a strategy—or at least, it more closely resembles a strategy when compared to what possible worlds semantics gave us.

By similar reasoning, SEMs might also hold promise for **control** and **improvement**. For these goals, the key limitation of possible worlds semantics was that it merely states that S depends on some feature x_i, while control and improvement require the additional step of understanding *how* S depends on x_i. At the very least, SEMs take a step closer to this: like in the case of discovery, tracing the chain reaction that results from intervening on x_i might be thought of as tracing the intermediate processing steps that were involved in going from x_i to the decision d, thus giving some insight into the system's internal reasoning process.

The picture is slightly less straightforward for **justification**. As noted in Section 3.2, possible worlds semantics already partially achieves the goal of justification, in that when complications like preemption are not involved, the counterfactual $x' > d'$ accurately captures a causal connection between x' and d', meeting Wachter et al.'s criteria for justification. Indeed, the SEM approach can handle such simple cases just as well as possible worlds semantics did. The remaining question, then, is whether SEMs can do better on the hard cases, like those involving preemption.

The answer to this is somewhat complicated. Pearl (2000) and follow-up work do demonstrate that SEMs can correctly capture causation in the presence of preemption and other similar complications. However, they achieve this by slightly modifying the connection between causation and counterfactuals. Under Lewis's possible worlds account, "e causally depends on c" translates to the counterfactual "if c had not occurred, e would not have occurred." As noted in Section 3.2, this is important because it mirrors the structure of Wachter et al.'s counterfactual explanations.

By contrast, Pearl's theory stipulates that, to account for complications like preemption, the counterfactual must additionally hold fixed events that are not related to the intervention. For example, in Hall's bottle-breaking case, in reality Billy's rock did not hit the bottle (since Suzy's rock got there first). According to Pearl, this fact must be held fixed when evaluating whether Suzy's throw caused the bottle to break, and so the resulting counterfactual looks more like "If Suzy had not thrown, and Billy's rock (still) had not hit the bottle, the bottle would not have broken" (Menzies and Beebee, 2020). Thus the simpler counterfactual "If Suzy had not thrown, the bottle would not have broken" (which, again, is the form Wachter et al.'s counterfactual explanations take) is *not* equivalent to a causal claim when preemption is involved, and as such the SEM approach does not fill in the gaps left by possible worlds semantics when it comes to providing a full (causal) picture of justification.

3.4 Problems for Counterfactual Explanations Under Structural Equation Models

So far, we have seen evidence that compared to possible worlds semantics, the SEM approach is more promising for discovery, control, and improvement, and effectively equivalent for justification. As such, we might be tempted to claim that SEMs are a net improvement over possible worlds semantics. This apparent improvement, however, papers over a fundamental problem:

where does the SEM for a given counterfactual come from in the first place, and how do we know when an SEM is "complete"?

In other words, there is a problem of circularity, which is particularly salient in the setting of AI explainability: the SEM interpretation of counterfactuals presupposes that we know ahead of time what the relevant variables are and how they relate to each other, so that we can build the SEM to intervene on. But the very reason we need explanations for AI decisions is that we *don't* know what the relevant variables and relations are—if we had known, for instance, that system S uses income and credit score to determine likelihood to repay the loan, we wouldn't have needed Wachter et al.'s system to generate a counterfactual to tell us that! This circularity problem is not unique to the AI explainability setting; in general, SEMs have been criticized for offering a *nonreductive* account of causation (Schaffer, 2016).

That said, in the AI explainability setting there is a possible way out of the circularity, at least for a subset of AI systems based on a popular technique called *neural networks*. A neural network can be described in the following (grossly oversimplified, but sufficient for our purposes) way: it consists of interconnected *neurons* that can either be on (1) or off (0). A directed connection between two neurons, $n_1 \to n_2$, means that n_2 being on or off depends on whether n_1 is on or off; specifially, the value of a neuron n is a thresholded weighted sum of all neurons connected to n:

$$n = f(w_1 n_1 + w_2 n_2 + \ldots + w_k k_k)$$

Where the w's are numerical constants ("weights") and f is an "activation function" that determines when the sum is large enough to turn on neuron n. This can be visualized as follows (Figure 2):

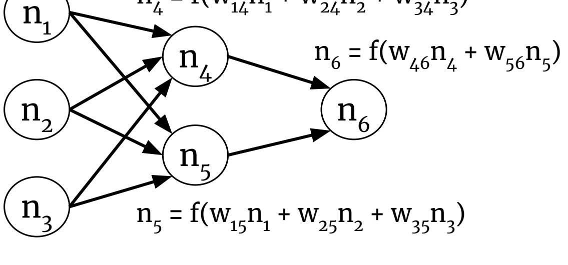
$n_4 = f(w_{14}n_1 + w_{24}n_2 + w_{34}n_3)$
$n_6 = f(w_{46}n_4 + w_{56}n_5)$
$n_5 = f(w_{15}n_1 + w_{25}n_2 + w_{35}n_3)$
n_1
n_2
n_3
n_4
n_5
n_6

Figure 2: Visualization of a neural network

There are clear parallels between Figure 1 (an SEM) and Figure 2 (a neural network): both are DAGs consisting of binary-valued nodes (on/off or true/false) whose values are determined as a function of their incoming connections. For AI systems based on neural networks, these parallels might offer a way around the circularity problem: instead of creating an SEM from scratch, we can directly convert the neural network into an SEM, with neurons becoming nodes and thresholded weighted sums becoming structural equations.

Of course, this proposal papers over technical details of how exactly this conversion would operate. Rather than delving into that though, I instead observe that even if we generously assume such a conversion is possible and works perfectly, it still does not completely circumvent the circularity problem. This is because oftentimes, the antecendents we are interested in are *not* directly stated in the input x, so if the system is to use such information it must first encode it as an intermediate step between x and d. For example, suppose we are testing a facial recognition system for gender bias. Nothing in an image directly states the person's gender, so to use gender as a factor the system would first need to use some intermediate neuron(s) to encode gender in terms of some combination of the pixels in the image. But the nature of neural networks is such that we *don't* know ahead of time what each intermediate neuron means; the point of explanations is to help discover this. Thus, we are right back at the circularity problem: if we knew ahead of time what each neuron means, we wouldn't need explanations to begin with.

In fact, the issue of interesting antecedents being higher-level concepts rather than direct features in the input raises additional concerns for the SEM approach beyond the circularity problem. Even if we set aside the circularity problem and assume that we know which combination of pixels to intervene on to evaluate the effect of gender, that still leaves the question of *how* we would go about intervening on multiple nodes, since Pearl (2000) defines intervention in terms of a *single* node. Indeed, Briggs (2012) shows that when a counterfactual involves logically complex antecedents (involving, as in our hypothetical gender scenario, some combination of nodes), intervention is ambiguous: there might be multiple valid interventions, which could lead to different conclusions. This is, of course, a problem for SEMs in general, and is not specific to the AI explainability setting. In the AI explainability setting, it suggests that an SEM interpretation would be limited only to the simplest explanations, greatly reducing its scope and practical utility.

4 Discussion and Conclusion

I have evaluated Wachter et al. (2018)'s counterfactual explanations of AI systems in the context of two interpretations of counterfactuals: possible worlds semantics and structural equation models. Structural equation models appear more promising in terms of meeting all four explainability goals identified by Adadi and Berrada (2018), but they are hampered by a problem of circularity which defeats the purpose of explanations in the first place. In comparison, possible worlds semantics are free from this circularity problem, but they arguably fail to meet the goals of **control**, **improvement**, and **discovery**, and even their ability to provide **justification** may be limited.

In fairness to Wachter et al., they are aware of and open about the limitations of their approach. In particular, they admit that counterfactual explanations are not sufficient for all purposes, and are not suited to applications that require insight into the internal workings of the system (like control and improvement). Instead, they focus exclusively on justification. However, while we have found that justification indeed fares best out of the four goals when it comes to counterfactual explanations, even it encounters some problems, as counterfactual explanations may not offer a sufficiently causal picture for justification.

Of course, this is not to say that the Wachter et al. (2018) approach is without merit. Indeed, we should now consider counterfactual explanations in the broader context of the explainability literature by briefly comparing them to the other existing approaches. As Wachter et al. have argued, the key advantage of counterfactual explanations is their accessibility to a general audience. This property makes them practically useful for justification, since the person seeking reasons for an AI system's decisions is very likely to be a layperson (e.g., a loan applicant) with little to no knowledge of the mathematics behind the system. By contrast, feature importance, which is more closely tied to the internal workings of the system, might be more useful for control and improvement, and likewise adversarial perturbation is almost purpose-built for the goal of improvement, as it centers around identifying weaknesses of the system.

As such, rather than thinking about the different approaches to explainability as being in competition, we should view them as complementing each other: each one is best suited to a different goal of explainability. Counterfactual explanations might fail to meet the goals of control and improvement, but people interested in those goals can still turn to feature importance or adversarial perturbation; conversely, feature importance and adversarial perturbation might be inaccessible to general audiences, but such audiences can (at least partially) find the justifications they seek in counterfactual explanations.

Of course, it remains the case that even in this broader-context picture, counterfactual explanations are somewhat held back by the problem of causation. That being said, it should be noted that feature importance and adversarial perturbation have their own weaknesses, some of which are closely related to the problems plaguing counterfactual explanations. For example, the problem of interesting antecedents being higher-level concepts rather than direct features in the input is also relevant when considering feature importance, as it implies that there is no single importance score that can reveal the importance of something like, for example, gender in an image. All this is not to say that today's approaches to explainability are a lost cause. Rather, the takeaway should be that users of any explainability technique—counterfactual or otherwise—should be aware of these limitations and account for them wherever possible, and also that the job of explainability researchers is far from done, with many questions—particularly surrounding causation—still to be answered.

References

Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 2018.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *arXiv:1910.10045 [cs]*, October 2019.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable Machine Learning in Deployment. *arXiv:1909.06342*, September 2019.

Rachael Briggs. Interventionist counterfactuals. *Philosophical Studies*, 160 (1), August 2012.

Nathan Hall. Two Concepts of Causation. In *Causation and Counterfactuals.* 2004.

Angelica Kratzer. *Modals and Conditionals: New and Revised Perspectives.* Oxford University Press, New York, 2012.

David Lewis. Causation. *Journal of Philosophy*, 70, 1973a.

David Lewis. *Counterfactuals.* Harvard University Press, Cambridge, MA, 1973b.

Peter Menzies and Helen Beebee. Counterfactual Theories of Causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, UK, 2000.

Jonathan Schaffer. Grounding in the Image of Causation. *Philosophical Studies*, 173, 2016.

William Starr. Counterfactuals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, 2019.

Frank Veltman. Making Counterfactual Assumptions. *Journal of Semantics*, 22(2), 2005.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 2018.