바둑학연구, 제17권 제2호, pp. 29-60
Journal of Go Studies
2023. Vol.17, No.2. pp. 29-60

# Explaining Go: Challenges in Achieving Explainability in AI Go Programs

Zack Garrett
Excelsior Classical Academy, USA

**Abstract:** There has been a push in recent years to provide better explanations for how AIs make their decisions. Most of this push has come from the ethical concerns that go hand in hand with AIs making decisions that affect humans. Outside of the strictly ethical concerns that have prompted the study of explainable AIs (XAIs), there has been research interest in the mere possibility of creating XAIs in various domains. In general, the more accurate we make our models the harder they are to explain. Go playing AIs like AlphaGo and KataGo provide fantastic examples of this phenomenon. In this paper, I discuss a non-exhaustive list of the leading theories of explanation and what each of these theories would say about the explainability of AI-played moves of Go. Finally, I consider the possibility of ever explaining AI-played Go moves in a way that meets the four principles of XAI. I conclude, somewhat pessimistically, that Go is not as imminently explainable as other domains. As such, the probability of having an XAI for Go that meets the four principles is low.

# I. Introduction

The field of AI research has seen remarkable developments in recent years. Current large language models (LLMs) are capable of writing high quality passages of natural language text. Visual art generators like Dall-E 2 and Midjourney can make impressive digital paintings. In less than twenty years, we have gone from Deep Blue defeating Garry Kasparov in Chess to AlphaGo defeating Lee Sedol and Ke Jie in Go. AIs are now used to make important decisions that affect humans. Insurance companies can use them to determine whom to insure (Nieva 2023). Banks can use them to determine to whom they should lend money (Arun, Ishan, and Sanmeet 2016). Finally, judges can use them to evaluate the risk of recidivism in parole decisions (Ghasemi et al 2021).

With the rapid development of AI technology, a number of ethical issues have emerged. Since AIs are beginning to play larger roles in human societies, it is important that we understand how they make their decisions. That is to say, we want our AIs to be explainable. If someone's loan application is rejected, it is important to know why the application was rejected. Did the AI reject the applicant because of their race or because of their history with creditors? The former reason would be unethical, but the latter reason would just be good business. Determining the answer to this question can often be incredibly difficult. This is because there is often a tradeoff between accuracy and explainability. The more accurate our AI models, the harder it is to explain how they make their decisions.

The focus of this paper is not the ethical issues involving the explainability of AI. Instead, I will only focus on the explainability of Go-playing AIs like AlphaGo (and its offspring), KataGo, and Leela. In this context, many of the ethical issues with explainability go away. That being said, exploring the

nature of explainability in the context of Go-playing AIs can shed light on the nature of explainable AIs (XAIs) in general.

In this paper, I explore how different theories of explanation apply to Go-playing AIs. In section 2, I briefly discuss the problem of XAI and its connections to Go. In section 3, I provide a brief introduction to some of the leading philosophical theories of explanation. In section 4, I consider the Deductive Nomological theory (DN) and the Inductive Statistical theory (IS), concluding that an explanation meeting the criteria of either view would not count as an XAI. In section 5, I consider how causal-counterfactual theories fair when dealing with Go-playing AIs. Causal-counterfactual theories offer better explanations of Go moves than DN explanations, but may be particularly difficult for non-experts to understand. In section 6, I take a brief detour to argue that the current explanatory capabilities of AI Go programs provide only inadequate explanations. Finally, I consider the pragmatic elements of explanation and how a Go-playing AI might differ from other AIs with regard to the possibility of giving generally accessible explanations.

## II. Explaining AI

Most of the concern about explaining AI decisions comes from ethical concerns about the use of AI technologies for making decisions that affect people. The European Union recently passed the General Data Protection Regulation (GDPR), which contains articles pertaining to the use of AI for automated decision making. Many believe that the way the GDPR is written guarantees Europeans a right to an explanation.[1] So, when an AI makes a decision about a loan or about a person's eligibility for insurance, the compa-

1) See, for example, Selbst and Powles 2018.

ny that uses the AI must be able to provide an explanation of this decision.

Much of the philosophical discussion on explanation, however, comes from the philosophy of science. Philosophers in this area are concerned with the metaphysical and epistemological properties of primarily scientific explanations. The problem of XAI comes at the crossroads between discussions in the philosophy of science and those in ethics. Since the goal of this paper is to discuss the explainability of Go-playing AIs, I will stick to the former rather than the latter. Of course, there are still some ethical issues at play in explaining the moves of Go-playing AIs. If we can get an AI that is capable of explaining its reasons, then human Go players will have an invaluable resource for furthering their personal study of the game. This provides a benefit to Go players but could severely hurt many professional Go players who depend on the income they get from teaching the game. Attila Egry-Nagi and Antti Törmänen put the possibility in the following way:

> Scholars of the game benefit more clearly from the existence of good AI engines, as the computer can just 'tell the truth' about a debated board position. Previously, players would have to pay for teaching to get the same effect, but now it is enough to simply have a strong computer—or, in fact, even just a modern smartphone. Consequently, many Go teachers are now facing the danger of losing their jobs, even though they can still provide a big value that AIs cannot: they can explain why particular moves are good or bad. (Egry-Nagi and Törmänen 2020, p. 7)

They claim that teachers still have the advantage of being able to explain why a move is good or bad. If an XAI for Go could explain its moves, then there would be little room left for Go teachers. In this sense, there is an in-

teresting ethical problem in the opposite direction. Normally we are driven to create explainable AIs because there is a moral obligation to know how the AI makes its decisions. In this case, there may be a moral reason to avoid knowing how Go-playing AIs make their decisions.

To make the discussion simpler, I will primarily focus on two particular examples of AI decisions throughout this paper: (i) a hypothetical scenario where an AI rejects a loan application and (ii) the 37th move of the 2nd game between AlphaGo and Lee Sedol.

Suppose that Sally has applied for a loan from a bank that uses an AI to evaluate loan applications. Sally's loan application is rejected, and she becomes curious why she was rejected. Was her credit score too low? Was her income too low? Or was it because of something out of her control like her race or her gender? The task of the bank is to provide Sally with an explanation that she can understand. Ideally, Sally should know what she could do to get a different result next time she applies for a loan.

Now, consider AlphaGo's move 37 depicted in Figure 1 below. The move stunned spectators when it was first played because it went against general wisdom about Go. Namely, playing a shoulder hit on the 5th line against a stone on the 4th line is inadvisable. Why did AlphaGo play its move at that point? Why, for example, did it play at P10 instead of D13—the move suggested by KataGo2)? In searching for an explanation of AlphaGo's move, we are looking for an explanation that can be understood by Go experts and the general Go playing community. Ideally, an explanation would allow us to understand the AI's moves well enough to potentially predict moves like it in the future and play them ourselves.

---

2) KataGo 1.12.4 using 40x256 network s11101.

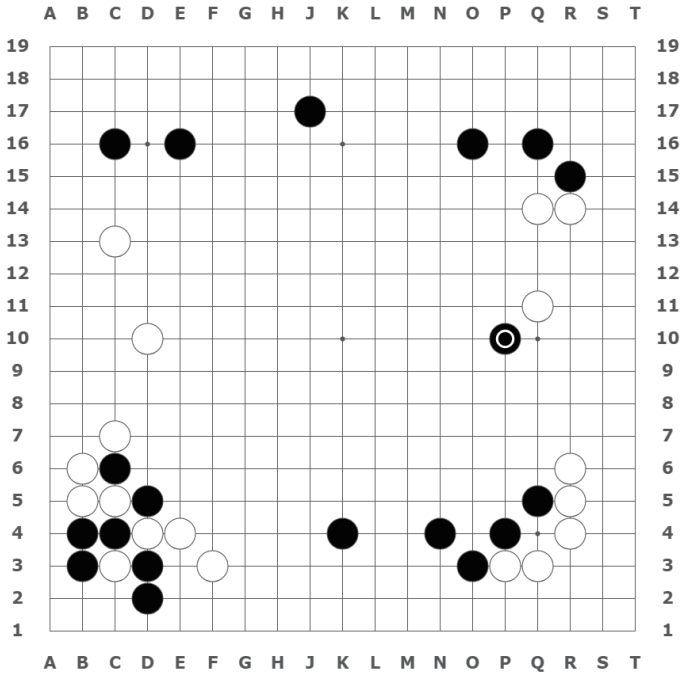


**Figure 1: AlphaGo vs. Lee Sedol 2nd Game 37th move**

Expert Go players have been using AI to train for a few years now. Doesn't this mean that there is no explainability problem with regard to Go? No. There are many things that humans can currently learn from AI. Many joseki have been discovered that, though they will make little impact in the games of beginner and intermediate players, can make a difference at the level of professional players. Even so, there are many situations where the AI picks a move that is minusculely better than alternative moves. The fact that the AI can do this repeatedly throughout a game leads to it playing at superhuman levels. Some moves made by the AI are apt for explanations. Maybe an AI plays at a vital point, or maybe it plays in an area that clearly negates the influence of the opponent's stones. These kinds of moves can

easily be explained. The problem of XAI in the context of Go isn't explaining any given move the AI makes. Instead, the problem is explaining the AI's justifications for playing one move over another seemingly equal value move, like the decision between P10, D13, and E12 shown in Figure 2 below. The difference between the values of these moves is small, but small differences can build up over the course of a game. The issue of the explainability of Go-playing AIs is clearest in this domain, in the gap between human play and perfect play.

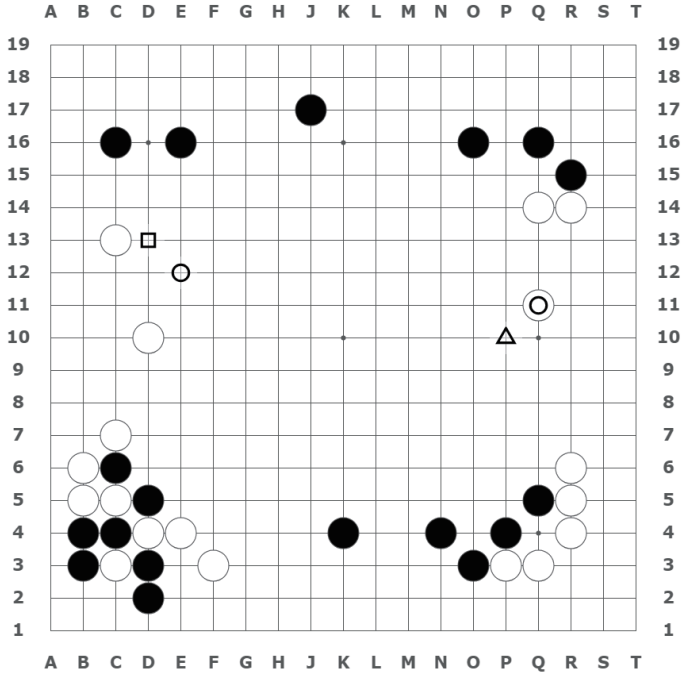


**Figure 2: AlphaGo vs. Lee Sedol 2nd Game 37th move with additional potential moves.**

That being said, the problem of XAI extends to beginners in Go as well. Supposing that we can make an explanation of AI-played Go moves that is understandable to experts, can we make one that is understandable to intermediate or beginner players? Consider Sally's loan application. If the bank gives Sally an explanation that contains tons of actuarial math in it, she won't be able to make heads or tails of it. But presumably the bank's duty is to explain their decision to Sally in terms she can understand. This paper will discuss explanations of AI-played moves given to experts, but it will also discuss the possibility of explaining AI-played moves to weaker players.

## III. Theories of Explanation

The Deductive-Nomological theory (DN) was one of the first well-defined theories of explanation. It was first presented by Carl Hempel and Paul Oppenheim who describe four adequacy conditions for explanations.

1. The explanandum[3] must be a logical consequence of the explanans[4].
2. The explanans must contain general laws, and these must actually be required for the derivation of the explanandum.
3. The explanans must have empirical content; i.e., it must be capable, at least in principle, of test by experiment or observation.
4. The sentences constituting the explanans must be true. (Hempel and Oppenheim 1948, p. 247-248)

3) The explanandum is the phenomenon or event to be explained.
4) The explanans is the facts that are meant to explain the explanandum.

According to the DN theory, then, an adequate explanation of a phenomenon or an event is a sound inference from at least two premises where the conclusion is the phenomenon to be explained. One premise is a general law and the other is a set of empirical claims. For example, consider the following explanation:

**Empirical Claims:** A thermometer made of glass and filled with mercury is submerged in heated water.

**General Law:** If a thermometer made of glass and filled with mercury is submerged in heated water both the glass and the mercury will expand, but the mercury will expand more, causing it to rise inside the thermometer.

**Conclusion:** The mercury in the thermometer rises.[5)]

The DN theory need not be limited to just deductive inferences. We can also consider explanations that are based on statistical inferences. In particular, explanations involving inductive statistical inferences can be adequate. Hempel calls these explanations "Inductive-Statistical Explanations" (Hempel 1965), and the resulting theory of explanation can be called the Inductive Statistical theory (IS). The important difference between a DN explanation and an IS explanation is that the IS explanation only has a statistical law—one that only states a statistical relationship rather than a necessary one. As such, the premises of the inference do not guarantee the conclusion. Here is an example of an IS explanation:

**Empirical Claims** Winston had a blood alcohol level of .2 and drove his car at high speeds.

---

5) This example comes from Hempel and Oppenheim 1948, p. 246.

**Statistical Law:** If someone has an elevated blood alcohol level and drives at high speeds, they will probably crash.

**Conclusion:** Winston crashed his car.

The motivation behind both the DN and IS versions of the theory is that we can explain phenomena and events by reference to the propositions that guarantee the phenomenon/event or, in the case of IS explanations, make the event more likely to happen. There is, however, a common complaint against the DN theory. Among the class of DN explanations are ones that include irrelevant details. For example, no man who takes birth control medicine will get pregnant, but this generalization along with the knowledge that John Jones is a man who takes birth control medicine does not explain why John Jones does not get pregnant. Of course, the relevant feature of John Jones is not the birth control but is instead his biological sex. However, a DN explanation involving Jones' use of birth control can be given for why he is not pregnant. The problem of irrelevancies prompted the creation of the statistical relevance theory (SR) of explanation.

The SR model of explanation attempts to capture the idea of a successful explanation by measuring the statistical relevance of various parameters.[6] Consider a window that was broken when an errant baseball hit it. Aunt Gertrude, finding the wreckage later that day, may ask why the window broke. Of course, the correct explanation and answer to her question is that the baseball hit by little Timmy hit the window. It would be wrong to tell her that the window broke because there were flowers on the kitchen table.

We can capture this case easily enough by measuring the statistical relevance of the baseball's hitting the window and the statistical relevance of the

6) For more on the SR theory, see Salmon 1971.

flowers' being on the table. We do this by calculating the conditional probability of the different parameters.

*P*(Window breaks | Baseball hits the window & Flowers on the table) =
*P*(Window breaks | Baseball hits the window)

Prior to the window's breaking, the probability of the window breaking given that it was hit by the baseball and the flowers were on the table is equal to the probability of the window breaking given that the baseball hit the window. This tells us that the flowers were statistically irrelevant, and hence their presence does not explain the breaking of the window.

Returning to John Jones, *P*(Jones doesn't get pregnant | Jones is a man) = *P*(Jones doesn't get pregnant | Jones is a man & he takes birth control). We can clearly see that Jones' taking birth control is statistically irrelevant to his not getting pregnant. So, the birth control is not part of an adequate explanation of why he doesn't get pregnant. His being a man, on the other hand, is statistically relevant, and does explain why he doesn't get pregnant. SR helps fill the hole left by DN by better connecting the explanans to the explanandum.

The SR theory is not the only way to capture the relevance between the explanans and the explanandum. As it turns out, many explanations do their explaining by identifying the causes of the explanandum. We can have, for example, a theory of explanation that states that ψ can be partially or wholly explained by φ if φ is a cause of ψ.[7] There are, of course, many different

7) In earlier years Nathan Salmon argued for the SR theory, but this changed in later years when he presented the causal mechanical theory, which says that explanation is not just a matter of statistical relevance, but it also requires causation. See Salmon 1984.

theories of causation. This paper cannot possibly cover every theory of causation and how those theories would interact with theories of explanation. Instead, I will focus here on just the counterfactual theory of causation. A counterfactual is a specific kind of conditional—one where we consider how the world would be different were some features changed. For example:

> If the allies had lost in WW2, then more people would speak German.

The sentence above is a counterfactual because it sets up a scenario that is counter to the facts (since the allies actually won), and then it draws some consequent about how the world would be. The counterfactual theory of causation identifies the causes of events with those things that cannot be changed without the event changing. Suppose that the cue ball in a game of billiards hits the 8 ball, causing it to drop into a pocket. We can generate a counterfactual like the following one:

> If the cue ball had missed the 8 ball, then the 8 ball would not have dropped into the pocket.

This counterfactual is true, and so we can determine that the cue ball was, at least, partially responsible for the 8 ball's dropping into the pocket—the cue ball was a partial cause of the event. David Lewis puts it succinctly as follows:

> We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without

> it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well. (Lewis 1973, p. 632)

Now, returning to a causal theory of explanation, how do we know that the baseball's hitting the window explains the window's breaking? Well, had the baseball not hit the window, then it wouldn't have broken. Had the baseball been lighter than it was, then the window would not have broken. Had the baseball been thrown with less force, then the window would not have broken.

The flowers, of course, are not part of the explanation of the window's breaking. This is because counterfactuals involving the flowers do not lead to the window's failing to break. Had the flowers been on the floor, had they been heavier, or had they been roses, the window still would have broken.

To get a complete explanation, we put together all or some important subset of the counterfactuals that result in a different event happening. So, to get a complete explanation of the breaking of the window, we need a list of all of the counterfactuals that result in the window remaining intact. Doing so will give us an understanding of what needs to change from the actual world to change the event with which we are concerned. In doing so, we understand why the event happened. The window broke because a ball carrying sufficient momentum hit the window.

Suppose that Sally's loan application has been rejected by the bank. The bank employs an AI to make decisions on loan applications. Sally reasonably expects an explanation from the bank. She worries that her application was denied because of her race or her gender. Sam Baron (2023) recently proposed one way that the bank can provide an explanation for its AI's decisions. The bank can make a list of counterfactuals of the form "If Sally's

income had been $50,000, then the AI would have accepted her application." and "If Sally's credit score had been 700, then the AI would have accepted her application." Once Sally has been given a full list of the changes in her application that would be sufficient for the AI to change its decision, then she can extrapolate which features of her application contributed to its rejection.

The counterfactual analysis of explanation gives a pretty reasonable explanation to Sally, but it doesn't work in all cases. Consider the case of Aunt Gertrude when she asks for an explanation for her broken window. Suppose that she is told a list of counterfactuals about the momentum of the ball or the tensile strength of the window. Will she be satisfied with this explanation? Of course not. She may be concerned with who broke the window, not with the physics of window breaking. She may even struggle to understand an explanation that involves a list of counterfactuals. Perhaps she knows very little about tensile strength and momentum. A much better explanation would simply state that Little Timmy broke the window while playing ball.

Explanation is a method of communication—a means by which we give others understanding. A chemist may explain their research very differently when talking to a colleague than when talking to their parents. In turn, the explanation they give to their parents will differ from the one given to a five-year-old. To each audience the chemist gives a different explanation, but each conversation they have includes an explanation.

For another situation, suppose that a student receives a C on an essay she wrote. She demands an explanation from her teacher. If the teacher tells the student that she made some interesting points throughout the paper, and so she didn't deserve a D, the student will not be happy with the explanation. The student does not want to know why she got a C rather than a D—she wants to know why she got a C rather than an A or a B. A less successful student may be pleasantly surprised by receiving a C and ask for an expla-

nation, hoping to find out what she did right. In the case of this second student, the teacher's explanation is accurate. She wants to know why she got a C rather than a D. The correct explanation is not merely a matter of giving understanding, it is a matter of giving the audience's desired understanding. All of this goes to show that what counts as an explanation is a pragmatic issue. The success conditions for explanations come from the audience of the explanation, and the factors at play include the audience's background knowledge and the contrast class they have in mind.

So far, I've presented several different approaches to explanations (DN, IS, SR, Counterfactuals, and some pragmatic concerns). Before we finally move to a discussion of Go-playing AIs, it is worth drawing the connection between theories of explanation and XAI. In 2020, Phillips et al put forward four principles of XAI.

1. Explanation: this principle states that an AI system must supply evidence, support; or reasoning for each decision made by the system.
2. Meaningful: this principle states that the explanation provided by the AI system must be understandable by, and meaningful to, its users. As different groups of users may have different necessities and experiences, the explanation provided by the AI system must be fine-tuned to meet the various characteristics and needs of each group.
3. Accuracy: this principle states that the explanation provided by the AI system must reflect accurately the system's processes.
4. Knowledge limits: this principle states that AI systems must identify cases that they were not designed to operate in and, therefore, their answers may not be reliable.[8)]

8) This is Angelov *et al* 2021's paraphrased version of the principles laid out in Phillips *et al* 2020.

In the subsequent sections of this paper, I will focus on how the different theories of explanation, as applied to Go-playing AIs, succeed or fail at living up to these four principles.

## IV. DN, IS, and AlphaGo

I will start the discussion of XAI and Go by considering how the DN/IS model might work for explanations of AI-chosen moves in a game of Go. Remember that the DN model involves inferences from premises about lawlike regularities and empirical facts to the phenomenon or event to be explained. We will start with the conclusion of these inferences. Consider again move 37 in AlphaGo's game against Lee Sedol. This is exactly the kind of move for which we desire explanations. In our DN inference, we will take "AlphaGo plays P10 on move 37" as the conclusion—it is the explanandum for which we will seek an explanans.

We now need to construct the premises of such an explanation. First, the initial conditions. These are simple enough to lay out. We need to know the board position and how that is provided to AlphaGo as an input. The initial conditions may be different for different AIs, but in general they will involve the current board position or the current position plus the history of the game. Here, for example, are the inputs given to AlphaGo Zero, one of the successors of AlphaGo:

> The input to the neural network is a 19 × 19 × 17 image stack comprising 17 binary feature planes. 8 feature planes Xt consist of binary values indicating

> the presence of the current player's stones (Xi t = 1 if intersection i contains a stone of the player's colour at time-step t; 0 if the intersection is empty, contains an opponent stone, or if t < 0). A further 8 feature planes, Yt , represent the corresponding features for the opponent's stones. The final feature plane, C, represents the colour to play, and has a constant value of either 1 if black is to play or 0 if white is to play. These planes are concatenated together to give input features st = [Xt , Yt , Xt-1, Yt-1, ..., Xt-7, Yt-7, C]. History features Xt, Yt are necessary because Go is not fully observable solely from the current stones, as repetitions are forbidden; similarly, the colour feature C is necessary because the komi is not observable. (Silver et al 2017, p. 27)

Things are much more complicated when we move to the lawlike premise. To make a deductive inference we will need an AI model that has no place for randomness. When it is presented with the same inputs it provides the same outputs. In the context of a competition, most AIs are not deterministic, and so some randomness will play a role. For example, under time constraints, engines may not be able to get to the depth necessary to play the same move every time. Because the numbers of visits to particular positions in a tree are not guaranteed to be the same on each run of the inputs, we cannot form a deductively valid argument from the initial conditions and the functioning of the AI to the actually played move.

Of course, the ways computers perform randomness is not actually random. The internal states of computers and the random numbers they generate are for our purposes deterministic. So, if we include the architecture of the model, the settings for all of the parameters in the model, and the internal states of the computer that are relevant for generating random numbers we

can get the lawlike regularities necessary to create a deductively valid inference from the inputs to the output.

Once we have our deductive inference, we have our explanation of why the AI chose the move that it chose. But is it really an acceptable explanation of the move? Knowing the setup and the lawlike properties of the AI would be sufficient for us to determine the move it will play (given a large enough amount of time). So, in some sense we would understand why the AI made its move. This is not, however, what we mean by XAI. To see why, consider again the case of Sally who has had her loan application denied on the basis of an AI decision. She requests some explanation for why she was rejected. The bank gives her a list of the details from her loan application and maybe a report from a credit bureau. The bank also gives her all of the details of the AI model that was used, including the value of every parameter in the model. Would Sally be satisfied with the explanation? Would she understand why she was rejected? The answer to both questions is "no." The inner workings of neural networks and machine learning algorithms are still areas that require expertise that far exceeds the knowledge of the general public. Not only does the putative explanation in this case exceed the capabilities of the person whose loan is rejected—it also exceeds the knowledge of those who programmed or use the AI. After all, humans cannot keep track of all the parameters in the kinds of AI models in use in loan application adjudication or in Go playing. This is just the problem of XAI. If the weights of the parameters of the model were sufficient explanations of AI decisions, then there would be no need for making explainable AIs. Of course, the programmers can print out the weights of all of the parameters in a model, but that doesn't suffice as an explanation. The same, of course, holds in the domain of Go-playing AIs. Knowing the weights of the parameters does give us an

explanation of a Go move, but it doesn't suffice for an average Go player, an expert, or the programmers. A DN explanation of an AI as complicated as AlphaGo will certainly fail to meet the 2nd principle of XAI.

An AI that allows for randomness could potentially be explained in terms of an IS inference. The difference between DN and IS is that IS allows for statistical inferences. So, we can allow for differences in the outputs of the model when fed the same inputs. This provides little help in explaining the moves of AI Go players. After all, if the lawlike premise still includes all of the weights of the parameters, it will still be too complicated to count as an adequate explanation. It doesn't matter that we accommodate the variability in the output of the model.

There may, however, be some more room for explanation when using IS. Consider the construction of AlphaGo Fan, AlphaGo Lee, and AlphaGo Master. All three of them have two networks, a policy network and a value network.[9][10] The policy networks were trained on thousands of games played by top players on the Kiseido Go Server. The task of the policy networks was to learn to predict the probability of a human playing a move in a given position. Suppose instead of training a policy network on games played by humans, we train one on games played by AIs. The goal of this new policy network would be to predict the probability of any given move in a position by an AI.

The new policy network can then be used to explain the moves of an AI using a statistical inference. The initial conditions, once more, are the inputs

9) See Silver *et al* 2017, p. 21-22 for a breakdown of the differences between the versions of AlphaGo.

10) Note that Woosuk Park discusses the removal of the separate policy network in the move from AlphaGo Master to AlphaGo Zero. He considers whether this removal worsens the explainability of AlphaGo Zero's decisions. See Park 2022.

to the AI. The lawlike premise is now the predicted probability by the new policy network of given moves in similar positions to the ones on the board. We can then say that the AI was likely to pick the actual move given the board position because that is the kind of propensity that the AI has.

Note that we do not even need a sophisticated policy network to get inferences like this. We can see certain policies popping up in the AI that diverge from our human policies. For example, the AI has a habit of playing early 3-3 invasions. So, when an AI plays an early 3-3 invasion, we can explain this as a habit of the AIs. For an analogous situation, imagine someone who has built up an implicit bias towards things on their left. When offered two equally good pieces of cake, they choose the one on the left. When presented with two equally pleasant roads down which to walk, they choose the one on the left. When the person comes upon a choice in the future between an object on their left and another on their right. We can explain their having chosen the one on the left by means of their history of a left-leaning bias. Such an explanation doesn't get to the root cause of their bias, but it could suffice in some situations. The same can be said of explaining some AI-played Go moves. The AI has built up some bias and we can explain its actions by means of that bias. Such an explanation doesn't explain how the AI came to that bias, and so it fails to meet the first principle of XAI, but it does succeed at meeting principle 2. That is to say, an explanation like this would be understandable to most people. Unfortunately, this kind of explanation fails the 1st principle. Merely knowing the propensities of the AI does not tell us its justifications for its decisions.

We get a better explanation in this case than in the DN case. The average Go player can understand that the AI has picked up certain habits and that its choice in a given situation can, at times, be explained by these habits. Un-

fortunately, we cannot explain every move of the AI with surface level propensities. For some, we will need the much more complicated policy network trained on AI played games. In addition, the explanations we get in these cases are not really the kinds of explanations we are looking for. They tell us what kinds of moves the AI generally favors, but they fail to tell us why the AI likes those moves. They are only very surface level explanations of the moves of the AI. We may be able to do better with one of the other theories of explanation.

## V. The Statistical Relevance of Go Moves

In this section, I consider both the SR theory and counterfactuals since both function similarly in the context of Go-playing AIs. Starting with the SR theory, we must consider how the conditional probability of the AI's playing the move it actually plays changes given different board positions. The focus here is on the board positions because they are the variables that can change between multiple runnings of the AI. Consider Figure 3:

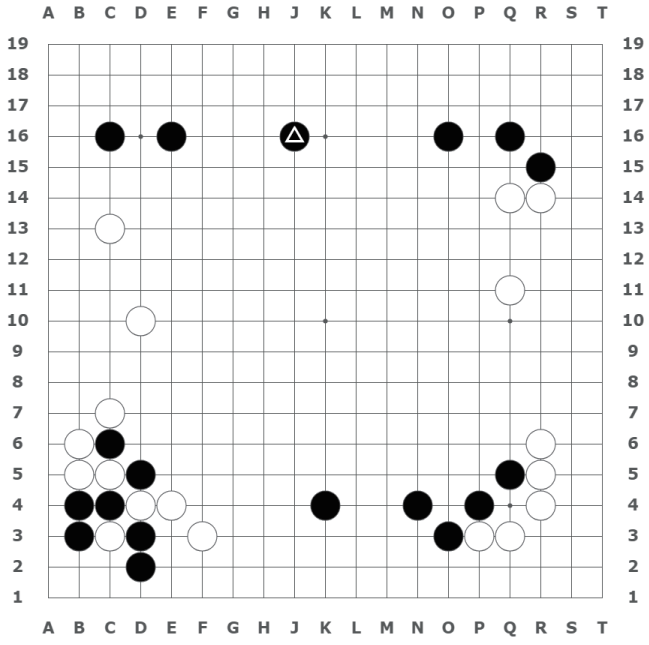


**Figure 3: An altered version of the 2nd game between AlphaGo and Lee Sedol**

The stone that was originally on J17 has been moved to J16. We can now consider how much different the AI's decisions are than they actually were. In this new position, KataGo gives roughly similar evaluations to the various candidate moves to the evaluations given when the stone was at J17. So, the probability that it will play one of those moves is mostly unchanged. This tells us that the stone's being at J17 is less relevant to the explanation.

This process can be repeated for each stone in the actual board position and the various places that stone could have been instead. By doing this we can determine which stones are most statistically relevant to the AI's decision. The explanation of the AI's move, then, is just the list of statistically

relevant stones. That is to say, the AI made the move that it made because of the arrangement of a particular subset of the stones on the board.

Applying the counterfactual approach to a Go-playing AI, we get a similar kind of explanation to the one provided by the SR analysis. We need to make a list of counterfactuals that would result in the AI playing a different move than it actually did. In other words, we make a list of changes to the inputs to the AI that result in a different output. We can then combine that list to know exactly which stones on the board are important for the AI's decision. The stones that are most important for getting the move that was actually played are the ones that, when changed in a counterfactual scenario, result in a different AI-played move. Note that this will give us a list similar to the one given by the SR analysis.

These methods of explaining AI decisions are present in the attempts at explaining the decisions of image recognition AIs. When an image classifier tells the user that an image contains a cardinal instead of a robin, it can be asked to highlight which pixels played the largest role in determining its decision. This method has allowed programmers to find interesting loopholes discovered by the AI. For example, Ribeiro et al (2016) trained an image classifier to identify wolves as opposed to huskies in images. They purposely trained the network on biased data where all of the images containing wolves also contained snow. Their goal was to test how humans interpreted the outputs of the biased AI. As part of the process, they have the AI explain how it comes to its verdict on a given picture. This is accomplished by identifying the pixels in the image that were most important for the AI's decision. Of course, the pixels showing snow ended up being the AI's justification.

We are getting closer to an acceptable XAI for Go, but SR and the counterfactual approach aren't yet sufficient. An expert Go player may be able to

glean something from data about which stones played the largest role in the AI's decision. For example, if a particular stone or subset of stones was relevant to the AI's move, then an expert may be able to couch that knowledge within their broader understanding of the game to learn something about the AI's decision. The stones on a Go board can bear complex relationships to one another. Moving one stone over can have drastic effects on the direction of the game. This means that the SR and counterfactual methods will often run into the problem that just about every stone is relevant. As such, if we were confused about why the AI took the current board position and output its actual move, then we will be just as confused knowing that every stone was relevant to that decision. This problem is particularly thorny when we consider that the goal is to explain why the AI makes one move over an ostensibly equal move elsewhere on the board. A small change in the inputs has a high chance of changing the output when the potential outputs are very close together.

In addition, the complexity of such an explanation would put it out of the reach of people who are not already Go experts. This reveals a dissimilarity between an XAI that evaluates loan applications and one that plays Go. A list of relevant counterfactuals regarding loan decisions are understandable to the layman. One can easily grasp the idea that they made too little income to qualify for a loan or that they have too many late payments in their credit history. A layman in Go, on the other hand, cannot understand why changing one stone on the opposite side of the board could have such drastic effects on the AI's decision.

To sum up this section, the SR theory and the counterfactual approach both meet the 1st principle, since they are fully grounded within the relationships between the inputs of the neural network and its output. They both

fail on giving meaningful information, and so they both fail to meet the 2nd principle. All things considered; they still do better than the DN approach in this respect.

## VI. Reading the Future

Before continuing to the pragmatic aspects of explanation, I will take a short detour to discuss the kinds of explanations that Go-playing AIs currently provide to players. At any point in a game, the player can ask the AI about its expected probability for either of the players winning. Players can also see data on the other moves that the AI considered and how many times it visited those parts of the tree. The players can see just how much the AI prefers one move over another. Most importantly, players can have the AI play out a sequence of moves from the current position.

All of these points of data count as partial explanations for why the AI made its move. We know that the AI made its move because it thought that move would increase its probability of winning. In particular, we know just how much more the AI thinks that the actual move improves its situation over alternative moves. We also can see how the AI thinks the game might proceed from the move it just played. Many top players learn from the AI by playing out sequences like this. They attempt to understand why the AI made a particular move by allowing the AI to play out the sequence until the proper consequences of the AI's move become clearer. Perhaps the AI played a move in order to set up a future attack. One way to find this out is to allow the AI to play out the sequence it saw until we get to that attack.

Is this a sufficient explanation for the AI's move? Perhaps in some cases

this may be sufficient. When, for example, the payoff of the AI's move is only a handful of moves away. Unfortunately, we cannot use this method to explain every move an AI makes. Playing out individual sequences, even if they are the AI-determined optimal sequences, will give us too small of a view of the tree that the AI is searching. We may know one potential payoff of the AI's move, but not why this particular payoff is better than what the AI could receive from a different move. We have here the particularly tricky problem that I set up in section 2. Why did AlphaGo play at P10 instead of D13? We can play out sequences from both positions, but we will not be able to see why the many sequences that can result from P10 were preferable to those that result from D13 or other nearly equivalent moves. Once we have multiple board positions that are some number of moves after the AI's chosen move, we still must rely on the AI to output for us an evaluation of the position. All we have done is delay the inevitable request for clarification from the AI, and if we can't explain the AI's original decision, we won't be in a much better position trying to explain its evaluation of a slightly progressed board position. The kind of explanation described in this section would fail to meet principle 1 of the 4 principles of XAI. In many cases it doesn't give much justification for the AI's move.

## VII. Useful Explanations

One recurring problem with the theories of explanation I have discussed so far is that they do not provide explanations that are accessible to non-experts. In the case of some of the explanations, they aren't even accessible to experts. Some in the literature on explanations, like Peter Achinstein (1983), have made the move to treating explanations as pragmatic entities. Expla-

nations have purposes and audiences. What counts as an explanation of an event is determined by the knowledge and desires of the audience of the explanation.

Consider Sally, the woman whose loan application was rejected. She desires to know what she can change about her life to eventually be approved for the loan. She also wants to know whether the decision was made for potentially illegal reasons. Her background knowledge may not include any information about machine learning algorithms or the structures of neural networks. So, an attempt to explain to her the decision in terms of the weights of the parameters in the model would fail. It neither gives her actionable information nor any understanding at all.

The SR and counterfactual explanations of AI-played Go moves could be sufficient to give experts some understanding of the reasons why the AI makes some of its moves. Note that the understanding they get is not just an understanding of the habits of the AI, like they would get from the IS theory. Instead, they may be able to get knowledge of the underlying principles of Go that the AI has discovered. These approaches, however, fail to give non-experts a genuine understanding of why the AI makes its moves.

Is it even possible to give an explanation that is sufficient to give non-experts genuine understanding of AI-played moves? There is skepticism in the literature on XAI about the possibility of giving explanations of AI decisions. The source of this skepticism comes from human inability to explain our own decisions. Human decisions are heavily influenced by many factors that we cannot isolate and explain. For example, teachers grading student written essays may try to fit their decisions into a rubric, but a lot of the process is done based on the hard to articulate feelings of the teacher. Jocelyn Maclure puts this complaint thusly, “Those who seek to deflate the explainability problem argue that we should not be excessively troubled by the lack

of transparency of automated decision-making because humans are equally opaque when they think and judge" (Maclure 2021).

Is the situation any better in Go? Well, let's consider how Go players explain their moves. Sometimes the explanation is easy enough. Playing the vital point that saves or kills a group or playing a simple sente move during the endgame, are moves that humans can easily explain. But moves in the midgame that are out in the open board can be significantly harder to explain. As Egri-Nagy and Törmänen put it:

> On the one hand, we do not know exactly how we play the game. It is difficult to verbalize our Go knowledge. Explanations for a move often get replaced by an 'it felt right' statement. (Egri-Nagy and Törmänen 2020, p. 4)

This kind of explanation is even worse than the kinds of explanations that current AIs give. At least KataGo can explain its decisions in terms of a probability estimate for the winner of the game. A human who explains their move by saying "it just felt right" cannot even provide an estimate of how much their move has changed their probability of winning.

Our inability to explain our Go moves is clear from the abundance of Go proverbs. We have proverbs like "Don't throw an egg at a wall" which informs us that we ought not play a weak stone near our opponent's strength. Proverbs are by their nature vague and situational. There are times when one ought to go against the advice of a proverb, but they stand as good rules of thumb. It is impossible to condense complex and important pieces of Go knowledge into phrases that can be properly understood by beginners, and so we explain Go moves with pithy little sayings.

So, we already have an explainability problem in Go. Unlike explaining

the decisions regarding loan applications, explaining Go moves is particularly difficult for humans. We develop intuitions that we often cannot express in words. For a similar example, consider the work of a chicken sexer. It is incredibly difficult to identify the sex of a chick, but some humans are able to quickly identify the sex and sort the chicks. Robert Brandom puts it this way,

> Industrial chicken-sexers can, I am told, reliably sort hatchlings into males and females by inspecting them, without having the least idea how they do it. With enough training, they just catch on. [...] At least in this way of telling the story, they are reliable noninferential reporters of male and female chicks, even though they know nothing about how they can do it, and so are quite unable to offer reasons (concerning how it looks or, a fortiori, smells) for believing a particular chick to be male. (Brandom 1998)

They are clearly identifying some features of the chick, but it is difficult for them to put into words what they are noticing.

What lesson can we draw from the pragmatic concerns about explanation? We should be skeptical about the possibility of making XAI for functions that are particularly complicated. In the case of Sally's loan application, the number of variables and the relevant changes to those variables are limited enough that a list of counterfactuals could be created that would give Sally enough understanding to make informed decisions in the future. A position in the middle of a Go game has too many parameters to keep track of and too many ways that those parameters could be changed for us to make a list of counterfactuals that would give a player enough understanding to make future decisions.

Albert Einstein supposedly said, “If you can’t explain it to a six-year-old, then you don’t understand it yourself.” Taking this quote at face value undermines its message. There are plenty of things that experts cannot explain to six-year-olds. A lot of background knowledge is needed to understand quantum mechanics. It is possible to explain some of the topic to a six-year-old, but they will only gain surface level understanding. The same is true of Go. To be able to understand the limited explanations that Go-playing AIs give, one must already have a sufficient amount of background knowledge. There isn’t much more an AI can do to explain its moves to a general audience than show the change in its expected territory and probability of winning. For experts, on the other hand, some of the methods of explaining AI-played Go moves described above may be sufficient to gain a limited understanding. For example, knowing which stones were the most important ones for the AI’s decision could shed some light on why the AI made its move, as opposed to a seemingly equal move. The understanding we can gain in this way will, unfortunately, be incomplete.

The function describing optimal Go play from a given position is so complicated that there is some reason to be pessimistic that an XAI for Go can be created. If humans could understand the function, we wouldn’t need to explain Go-playing AIs in the first place. Again, there is a substantial disanalogy between Go-playing AIs and loan adjudicating AIs. Humans can understand who would be a good debtor and who would not. The point of the AI in these cases is to cut down on human labor and make less biased decisions. The loan adjudicating AI isn’t meant to do something humans cannot do. Go-playing AIs, on the other hand, are designed to be superhuman in their abilities. It is no wonder they cannot explain their justifications to us—they are just that much better than us.

# References


Achinstein, Peter. *The nature of explanation*. Oxford University Press, USA, 1983.

Angelov, Plamen P., Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. "Explainable artificial intelligence: an analytical review." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, no. 5 (2021): e1424.

Arun, Kumar, Garg Ishan, and Kaur Sanmeet. "Loan approval prediction based on machine learning approach." *IOSR J. Comput. Eng* 18, no. 3 (2016): 18-21.

Baron, Sam. "Explainable AI and Causal Understanding: Counterfactual Approaches Considered." *Minds and Machines* (2023): 1-31.

Brandom, Robert B. "Insights and blindspots of reliabilism." *The Monist* 81, no. 3 (1998): 371-392.

Hempel, Carl G. *Aspects of scientific explanation*. Vol. 1. New York: Free Press, 1965.

Hempel, Carl G., and Paul Oppenheim. "Studies in the Logic of Explanation." *Philosophy of science* 15, no. 2 (1948): 135-175.

Ghasemi, Mehdi, Daniel Anvari, Mahshid Atapour, J. Stephen Wormith, Keira C. Stockdale, and Raymond J. Spiteri. "The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism." *Criminal Justice and Behavior* 48, no. 4 (2021): 518-538.

Lewis, David. "Causation." *The journal of philosophy* 70, no. 17 (1973): 556-567.

Maclure, Jocelyn. "AI, explainability and public reason: the argument from

the limitations of the human mind." *Minds and Machines* 31, no. 3 (2021): 421-438.

Nieva, Richard. "Cigna Sued Over Algorithm Allegedly Used To Deny Coverage To Hundreds Of Thousands Of Patients." Forbes. July 24th, 2023. https://www.forbes.com/sites/richardnieva/2023/07/24/cigna-sued-over-algorithm-allegedly-used-to-deny-coverage-to-hundreds-of-thousands-of-patients/ (accessed September 9th, 2023).

Park, Woosuk. "How to Make AlphaGo's Children Explainable." *Philosophies* 7, no. 3 (2022): 55.

Phillips, P. Jonathon, Carina A. Hahn, Peter C. Fontana, David A. Broniatowski, and Mark A. Przybocki. "Four principles of explainable artificial intelligence." *Gaithersburg, Maryland* 18 (2020).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* pp. 1135-1144. 2016.

Salmon, Wesley C. *Statistical explanation and statistical relevance.* Vol. 69. University of Pittsburgh Pre, 1971.Scientific explanation and the causal structure of the world. Princeton University Press, 1984.

Selbst, Andrew, and Julia Powles. ""Meaningful information" and the right to explanation." In *conference on fairness, accountability and transparency,* pp. 48-48. PMLR, 2018.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert et al. "Mastering the game of go without human knowledge." Nature 550, no. 7676 (2017): 354-359.

Received: 11, Oct, 2023

Revised: 26, Oct, 2023

Accepted: 02, Nov, 2023