philosophies

MDPI

Article

How to Make AlphaGo's Children Explainable

Woosuk Park

Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea; woosukpark@kaist.ac.kr

Abstract: Under the rubric of understanding the problem of explainability of AI in terms of abductive cognition, I propose to review the lessons from AlphaGo and her more powerful successors. As AI players in Baduk (Go, Weiqi) have arrived at superhuman level, there seems to be no hope for understanding the secret of their breathtakingly brilliant moves. Without making AI players explainable in some ways, both human and AI players would be less-than omniscient, if not ignorant, epistemic agents. Are we bound to have less explainable AI Baduk players as they make further progress? I shall show that the resolution of this apparent paradox depends on how we understand the crucial distinction between abduction and inference to the best explanation (IBE). Some further philosophical issues arising from explainable AI will also be discussed in connection with this distinction.

Keywords: abduction; AlphaGo; Baduk (Weiqi; Go); explainable AI; inference to the best explanation (IBE)

check for updates

Citation: Park, W. How to Make AlphaGo's Children Explainable. *Philosophies* **2022**, *7*, 55. https://doi.org/10.3390/philosophies7030055

Academic Editor: Marcin J. Schroeder

Received: 23 April 2022
Accepted: 18 May 2022
Published: 24 May 2022

1. Introduction

We are living in the age of AI. If DeepBlue was a mere astral signal, AlphaGo was more than a symbolic announcement of the beginning of a new era. Now, as AlphaFold 2 takes another huge step towards the goal of AGI, AI is changing all aspects of our lives. How should we understand this phenomenon and its far-reaching philosophical implications? While machine learning is known to be the key to enhancing AI algorithms, recent advances in cognitive neuroscience have been another crucial factor. Nevertheless, much is still unknown about how these advancements in AI are possible. How are we to fathom super-intelligent machine minds in order to live well together with them in the future? Above all, the explainability or interpretability of AI is arguably one of the most heatedly discussed issues of our time. Turning now to abductive cognition seems to be a natural strategy to answer these questions. As it has elements of both intuition and inference, we may be more hopeful to bridge the gap between humans and machines. Pioneered by C.S. Peirce in the late 19th century, abduction has been studied extensively in logic, philosophy of science, cognitive and computer science, AI, law, and semiotics during the 20th century, and several notable monographs on abduction have been published in the first two decades of the 21st century, uncovering the logical form and various patterns of abduction. Could abduction not be both the final fortress for human creativity and the desirable link between humans and AI?

Under the rubric of understanding the problem of explainability of AI in terms of abductive cognition, I propose to review the lessons from AlphaGo and her more powerful successors. As AI players in Baduk (Go, Weiqi) have arrived at superhuman level, there seems to be no hope for understanding the secret of their breathtakingly brilliant moves. Without making AI players explainable in some ways, both human and AI players would be less-than omniscient, if not ignorant, epistemic agents. Are we bound to have less explainable AI Baduk players as they make further progress? I shall show that the resolution of this apparent paradox depends on how we understand the crucial distinction between abduction and inference to the best explanation (IBE). Some further philosophical issues arising from explainable AI will also be discussed in connection with this distinction.

The preliminary discussion in Section 2 of an explainable Baduk player is not merely for clarifying the notion of explainability of AI in general; it is needed for understanding what exactly we want in our pursuit of explainable AI Baduk players. In Section 3, by briefly reviewing the evolution of AlphaGo to her more powerful descendants, I will pin down the combination of the separate policy and value networks into a single neural network as the most important change in the evolution of AlphaGo to its children. I shall introduce a thought experiment in which one suspects whether DeepMind intentionally has made the AI Baduk agent less explainable. If this suspicion is not without any ground, we seem to be in a paradoxical situation: By making AlphaGo less explainable, DeepMind made explainability of AI the focal issue of our time once and for all. Such a suspicion stems from a hypothesis contrasting the role and function of the policy network and value network of AlphaGo in terms of the distinction between abduction and inference to the best explanation (IBE). In order to examine this hypothesis critically, the distinction between abduction and IBE will be briefly discussed in Section 4. Ultimately, I shall resolve the apparent paradox by presenting a positive interpretation of DeepMind's seeming obliteration of abduction and IBE. Finally, in Section 5, I shall discuss briefly some further philosophical issues naturally arising from the central question as to whether we are bound to have less explainable AI Baduk players as they become more powerful.

2. What Is an Explainable Baduk (Go, Weiqi) Player?

Recently, one can witness the surge of interest in the explainability of AI (e.g., [1–14]). Among the many intricately related advances that have contributed to this, the monumental success of AlphaGo of DeepMind looms large (Silver et al. [15–17]). Even though it is still merely a dream to develop general purpose AI, AlphaGo is the symbol for the age of AI. As everyone knows, DeepMind has not only generalized AlphaGo to AlphaZero, but also moved on to more ambitious projects like AlphaFold [17–19]. This gambit seems rather hasty, for there are still a lot of lessons of various sorts we can learn from AlphaGo and its superior descendants. After all, is it not AlphaGo that made the question of the explainability of AI the focal issue of our time?

As the title "The Game Is Not over Yet—Go in the Post-AlphaGo Era" indicates, Egri-Nagy and Törmänen aptly summarize our current situation described above in their recent paper [20]. The centrality of the issue of explainable AI is, above all, masterly pinned down:

> The AI's principal variation (the most advantageous sequence of play) can give information about a better way of playing, but how much a human can infer from it depends on the player. *An AI cannot explain why it plays its moves beyond giving an estimated winning probability and possibly a score estimation, but humans cannot generate these numbers on their own*. The AI's output is generally the most useful for a human if the AI's idea was originally conceivable for the human, but depends on an important key move or realization later in the sequence to truly work. More often, however, an AI chooses its moves by its superior whole-board judgement, which is difficult to explain in human terms. Ref. [20] (Section 4.2) (Emphasis added)

The last sentence in the quoted passage makes it explicit that the moves of AI Baduk players are not explainable in human terms. More importantly, the italicized sentence above seems to question the explainability of AI even in its own terms. Of course, it does not explicitly claim that the moves of AI currently not explainable "even in its own terms". At least, however, it seems to insinuate that very idea.

Whether such an impression in the readers' part was intended or not, it seems meaningful to examine the claim that AI is currently not explainable "even in its own terms". In order to appreciate and evaluate this claim, it may be helpful to know why and how Egri-Nagy and Törmänen distinguish between the concepts of "ultra-weakly solved", "weakly solved", and "strongly solved". They need the distinction in order to correct people's mistaken belief that the game of Baduk has been solved. According to them, such a belief is wrong, for "creating a superhuman Go playing entity is not the same as mathematically solving the game":

> Ultra-weakly solved: The result of perfect play is known, but not the strategy. For example Hex, where a simple mathematical proof shows that the first player wins, but we do not know how to achieve that result.
>
> Weakly solved: Perfect result and strategy are known from the starting position. Checkers is solved this way; therefore, there are still positions for which we do not know the perfect result.
>
> Strongly solved: For all legal board positions, we know the perfect result and we can demonstrate a sequences of moves leading to that. Ref. [20] (Section 7.1)

Now, we can see that they are using three (or possibly four) pronged criterion for drawing the distinction: (1) whether the perfect result, i.e., the result of perfect play, is known, (2) whether the strategy is known, (3) whether they are known at the starting position, and (4) whether they are known at all legal board positions. It might be their ultimate hidden purpose of all this to raise some rhetorical questions: for example, if some simple game like Checkers or Hex are solved merely in a weak or ultra-weak sense, how could there be any chance for sophisticated games like chess or Baduk to be solved in a strong sense?

Indeed, Egri-Nagy and Törmänen go on to claim the following about Baduk:

> The combinatorial complexity of the game prohibits us from having answers to these questions with a brute-force method. How far are the current Go engines from perfect play? There are some indirect ways to measure that distance, e.g., the added value of the tree search to the raw output of the policy network [21]. The current AIs are certainly not at perfect play yet, as omniscient neural networks would have only three distinct output win rate probabilities of 0%, 50%, and 100% for all board positions. Talking in terms of probabilities other than these three values is admitting ignorance. Ref. [20] (Section 7.1)

Chess might have been solved at least weakly or ultra-weakly. It is arguably the case that we do have a brute-force method in chess. Further, if so, many people would say that automatically we are equipped with a winning strategy in chess. (See [2,18,22,23]) However, it is by no means possible to claim that Baduk has been solved weakly or ultra-weakly. There is no brute-force method in Baduk, nor is there any winning strategy in Baduk.

Roughly speaking, then Egri-Nagy and Törmänen are claiming that, regardless of the huge difference between human and AI Baduk players after AlphaGo, both are on a par in that they are still ignorant, less than omniscient epistemic agents.

3. Who Is Afraid of Making AlphaGo Explainable? An Apparent Paradox

Now, I will briefly review the evolution of AlphaGo to her more powerful descendants. AlphaGo Zero is the culmination of the evolution of AlphaGo, i.e., in its history from AlphaGo Fan, AlphaGo Lee, AlphaGo Master to AlphaGo Zero. It must have a unique place in broader contexts, say in the more inclusive history, of the Alpha projects of the DeepMind including AlphaFold as well as Alpha Zero, but what exactly was its secret of success or the crucial factor that made the success possible? The AlphaGo team of the DeepMind assesses the innovative aspects of AlphaGo Zero as follows:

> Our program, AlphaGo Zero, differs from AlphaGo Fan and AlphaGo Lee 12 in several important aspects. First and foremost, it is trained solely by self-play reinforcement learning, starting from random play, without any supervision or use of human data. Second, it only uses the black and white stones from the board as input features. Third, it uses a single neural network, rather than separate policy and value networks. Finally, it uses a simpler tree search that relies upon this single neural network to evaluate positions and sample moves, without performing any Monte-Carlo rollouts. To achieve these results, we introduce a

> new reinforcement learning algorithm that incorporates look ahead search inside the training loop, resulting in rapid improvement and precise and stable learning. Further technical differences in the search algorithm, training procedure and network architecture are described in Methods. Ref. [16] (p. 354)[1]

For my present purpose, i.e., understanding the problem of the explainable AI Baduk player in connection with abductive cognition, the most pertinent factor seems to be the third one: "Third, it uses a single neural network, rather than separate policy and value networks". Roughly put, I think, AlphaGo with the separate policy and value networks is dazzlingly contrasted with AlphaGo Zero which combines them in a single network. We seem to have already secured a prima facie promising working hypothesis that while the former can be interpreted in terms of abductive cognition, the latter seems to block that possibility.

Silver [16] tries to contrast AlphaGo Zero with the three versions of AlphaGo: (1) AlphaGo Fan, (2) AlphaGo Lee, and (3) AlphaGo Master. The names of these versions were taken from AlphaGo's opponents, Fan Hui (October, 2015), Lee Sedol (March, 2016), and the top human players (January, 2017)[2]. What we need to note from the comparison of AlphaGo Zero with the three versions of AlphaGo in [16] is that a truly revolutionary change was made in the transition from the latter to the former:

> AlphaGo Zero is the program described in this paper. It learns from self-play reinforcement learning, starting from random initial weights, without using rollouts, with no human supervision, and using only the raw board history as input features. It uses just a single machine in the Google Cloud with 4 TPUs (AlphaGo Zero could also be distributed but we chose to use the simplest possible search algorithm). Ref. [16] (p. 360)

> However, we should not conclude that the crucial step was taken when the last version of AlphaGo, i.e., AlphaGo Master, was transformed into AlphaGo Zero. For, according to [16], AlphaGo Master "uses the same neural network architecture, reinforcement learning algorithm, and MCTS algorithm as described in" it (ibid.). Of course, it was immediately pointed out that "it uses the same handcrafted features and rollouts as AlphaGo Lee 12 and training was initialised by supervised learning from human data" (ibid.). What was happening in AlphaGo Master, about which DeepMind has never published a separate paper? Although there is no explicit mention made, the big difference between AlphaGo Master and its earlier versions, i.e., AlphaGo Fan and AlphaGo Lee, must be that the former no longer has policy and value networks as the latter did. [16] is so anxious to emphasize that AlphaGo Zero discovered a remarkable level of Go knowledge during its self-play training process. This included fundamental elements of human Go knowledge, and also non-standard strategies beyond the scope of traditional Go knowledge. Ref. [16] (p. 357)

After having briefly reviewed how AlphaGo evolved into her more powerful descendants, it seems instrumental to introduce a thought experiment in which one suspects whether DeepMind intentionally made the AI Baduk agent less and less explainable. If this suspicion is not without any ground, we seem to be in a paradoxical situation:

> (An Apparent Paradox) By making AlphaGo less explainable, DeepMind made the explainability of AI the focal issue of our time once and for all.[3]

This imaginary suspicion cannot be merely motivated by any appeal of conspiracy theories. According to Medianovskyi and Pietarinen:

> XAI has set itself an ambitious goal of making autonomous AI systems succeed in responding to requests for explanations of its own states, behaviours and outputs. The need for XAI has arisen from perceiving current ML as opaque or even representing solely behaviouristic black-box learning models that are in

> some sense incomprehensible and ill-motivated in their actions and intentions. Ref. [5] (p. 2)

As is clear from the quote, explainable AI is being pursued by everyone in machine learning owing to its own black box characteristic. If so, DeepMind's intentional decision to make AlphaGo less explainable appears self-refuting.[4]

Since this would-be paradoxical result is by no means a genuine logico-mathematical or semantic paradox, we need to reflect on what exactly is paradoxical here. The only thing we can be certain of is that such a doubt stems from a hypothesis contrasting the role and function of the Policy Network and Value Network of AlphaGo in terms of the distinction between abduction and inference to the best explanation (IBE).[5]

4. How to Resolve the Apparent Paradox

In order to examine this hypothesis critically, after elaborating why we might have reason to be suspicious about DeepMind's crucial decision to abolish the separation of policy and value networks, the distinction between abduction and IBE will be briefly discussed in this section. Ultimately, I shall resolve the apparent paradox by presenting a positive interpretation of DeepMind's obliteration of abduction and IBE.

4.1. The Apparent Paradox in the Suspicious Mind

As was already pointed out, the suspicion in our thought experiment of whether DeepMind intentionally made the AI Baduk agent less explainable stems primarily from a hypothesis contrasting the role and function of the policy and value networks of AlphaGo in terms of the distinction between abduction and inference to the best explanation (IBE), but it may also need some other small implicit assumptions. For example, I might have assumed that the problem of explainability came to loom large only after the appearance of Alpha Zero, if not AlphaGo Zero or AlphaGo Master, though it may not be so important or necessary for my argument. Also, I might have assumed that AlphaGo is more explainable than her children, including AlphGo Zero, which is, in fact, never proven. For argument's sake, let me focus on the main hypothesis, and how it is led to the apparent paradox.

In the suspicious mind, probably some reasoning like the following is at work. (Let us ignore that there might be some redundancies involved. Also, do not take seriously the order of presentation of all these propositions. It would be good enough to be sympathetic with the suspicious mind in the thought experiment and his or her shock and disappointment when DeepMind's announcement of giving up the separation of policy and value networks in its AlphaGo Zero paper, i.e., [16].):

1. AlphaGo had two separate neural networks, i.e., Policy network and Value network. (Fact)
2. The functioning of the policy network is like abduction. (My assumption)
3. The functioning of the value network is like IBE. (My assumption)
4. AlphaGo Master, AlphaGo Zero and AlphaGo's other children have only one neural network by integrating the policy and value networks. (Fact)
5. Insofar as there was a division of labor between the policy and value networks, it was in principle possible to make sense of AlphaGo's strategic moves or decision making processes. (My assumption seemingly shared by DeepMind)
6. Therefore, the policy network is functioning as an abducer, while the value network is doing IBE. [from lines 2 and 3]
7. Since there is no longer such a division of labor in AlphaGo Zero and AlphaGo's other children, it becomes a harder task to understand their workings. (Corollary of 5)
8. In other words, AlphaGo and AlphGo's other children are less explainable than AlphaGo. [from lines 5, 6, and 7]
9. It was DeepMind's deliberate decision to give up the separation of policy and value networks. (Fact)
10. AlphaGo's children far outweigh AlphaGo in strength, and are thereby less explainable. (My assumption)

Therefore, we have reason to accuse DeepMind of intentionally making AlphaGo and her children less explainable.

It is important whether 5, 6, and 7 can be safely used in our reasoning as indubitable premises. Of course, readers might be recommended to take a supercritical attitude toward my assumptions 2, 3, and 10. It is not difficult to justify 5, 6, and 7 from DeepMind's own reports about the roles and functions of policy and value networks. Above all, at least in Silver et al. [15], the efficiency of the division of labor between policy and value networks was strongly emphasized:

> "compensating by selecting those positions more intelligently, using the policy network, and evaluating them more precisely, using the value network—*an approach that is perhaps closer to how humans play*" Ref. [15] (p. 489) (Emphases added)

It is interesting to note, by the way, that DeepMind was making this point together with the fact that "AlphaGo evaluated thousand times fewer positions in her game against Fan Hui than Deep Blue did in its match against Kasparov" (Ibid.). As Park et al. [25] (p. 134) indicated, the italicized part in the quote above is most interesting, for this suggests a very useful perspective for understanding the roles and functions of policy and value networks. This perspective counts "what is more intelligent" as "what is closer to how humans play" without any argument. Whether such an assumption is justified or not, that must be presenting a very important point of reference in discussing the similarities and differences between human decision making and AlphaGo's decision making. Further, it cannot be a coincidence that the quoted passage correlates "intelligibility" with the policy network, and "precise evaluation" with the value network.[6]

4.2. Abduction and IBE

Now, in order to deepen our understanding of what is going on, it might be useful to review briefly the recent debates regarding identity or non-identity of abduction and IBE. For this purpose, citing the apt summary in Cabrera's recent survey article on IBE seems most useful. As he reports:

> It is common in contemporary discussions of IBE to link abduction to IBE in some way, either to directly equate IBE with abduction, or to trace IBE back to Peirce, or at least to use "IBE" and "abduction" interchangeably. (e.g., [26–31])
>
> While Peirce's notion of abduction is often cited as the intellectual forbearer of IBE, this view has been rejected by a great number of commentators. (e.g., [32–37])[7]

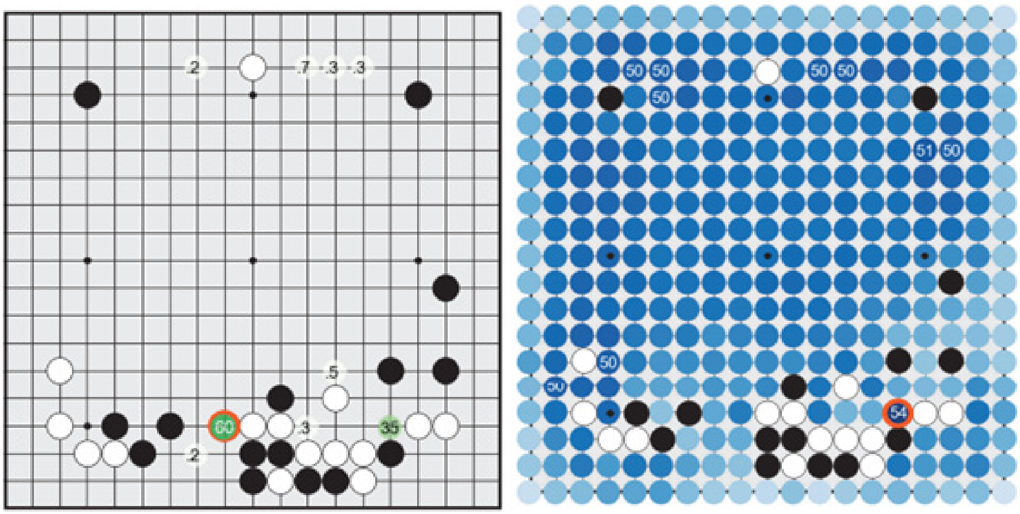
It would be nice if we provide rigorous definitions of abduction and IBE. However, we have failed to secure the core meaning of abduction and IBE. After all, we are still unclear about the differences between abduction and induction. Certainly, in terms of the differences between abduction and the well-known hypothetico-deductive method, and the differences between abduction as inference and abduction as intuition, the situation has improved much in the last quatercentury. Unlike the 20th century, in which the focal issue was to identify the logical form of abduction, we have come to appreciate all the different patterns or types of abduction. Further, thanks to these developments, our interest in abduction as reasoning has expanded to the interest in abductive cognition, which is found not only in human enterprises, but also in non-human animals. The recent controversy on the identity or non-identity of abduction and IBE should be understood against such a background history. In fact, I myself have discussed this issue several times, for I count it as pertaining to two of the most urgent current agenda in the study of abduction:

> (1) Is classifying abduction compatible with the search for the logical form of abduction, and (2) could there be any non-explanatory abduction? If abduction is just IBE, the problem of the logical form of abduction is nothing other than that of the logical form of IBE. If abduction is IBE, there would be no more mind-boggling for non-explanatory abduction: How could there be any "Non-explanatory IBE"? (Ref. [38] (p. 27); See also [39–42])

For my present purpose, it is neither necessary nor useful to settle this issue here, though I have found Gerhard Minnameier's attempts to contrast abduction with IBE in terms of their functions somewhat conclusive: "Thus, while abduction marks the process of generating theories—or, more generally, concepts—IBE concerns their evaluation"([26] (p. 75); see also [43–48]).[8] As we saw above, DeepMind also would not have any qualms with such a basic contrast of abduction and IBE.

One thing that is evident is that the parallelism of the abduction-policy network and IBE-value network can be imposed in interpreting the games of AlphaGos older than AlphaGo Master. Let me use a game between Fan Hui and AlphaGo Fan, which was masterfully discussed in [15]. The intriguing points are well represented in Figure 1a,b and Figure 2. Park et al. [25] commented on all this as follows:

> What is most shocking is that there is a dazzling contrast between the policy network and the value network. In the policy network shown in Figure 1a, the focus is on the move that scored 60 and the move scored 35. For, all the other moves scored less than 1. AlphaGo's judgment that humans prefer the former move to the latter move seems correct. It could be the case that most Baduk players, including the advanced ones, would play the former move without serious consideration. The move not only guarantees ample territory but also promises to secure sente. For, Black can capture two White stones by ladder, unless White responds to Black's move that scored 60. Even if one considers the invading move in the right bottom corner that scored 35, it would rarely be executed, since it is not so attractive. For, as shown in Figure 2 [sic], though it is a quite nice move destroying White's territory in the right bottom corner, it is not a fatal move threatening White's group. There is even a worry due the uncertainty involved in case White counterattacks by thrusting a wedging move, which was in fact the choice Fan Hui made in the actual game. Now, we can see that in the value network shown in Figure 1b [sic] the invading move in the right bottom corner got the highest score 54, and there are many other moves that scored 50, while the hane move in the bottom side got extremely low evaluation even failing to get serious consideration. Ref. [25] (pp. 134–136)



(**a**) (**b**)

Figure 1. (**a**) Policy Network (Figure 5d in [15]). (**b**) Value Network (Figure 5a in [15]). Reprinted/adapted with permission from Ref. [15]. Copyright year: 2016, copyright owner's name: DeepMind".

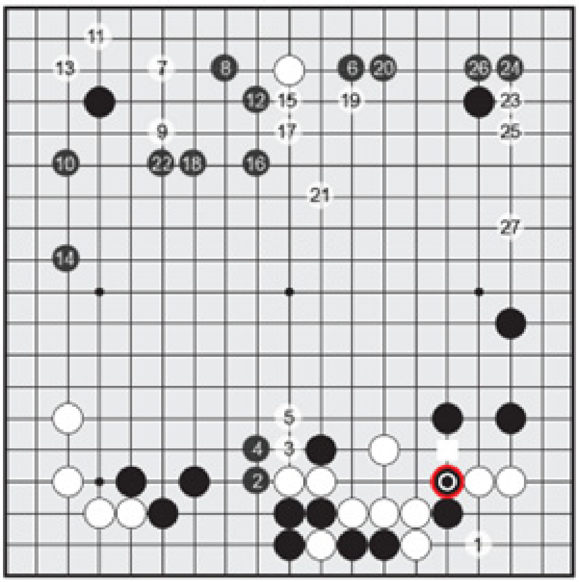


Figure 2. A variation reference to Figure 1 (Figure 5f in [15]). Reprinted/adapted with permission from Ref. [15]. Copyright year: 2016, copyright owner's name: DeepMind".

One possible source of confusion and misunderstanding lies in that I have been arguing with an implicit assumption that IBE is not so much an abduction as an induction. I do not believe that I am idiosyncratic here with the terminology of IBE or induction. I cite Thagard's recent identification of IBE as induction as evidence:

> A narrow use of the term "induction" covers only generalization from some to all, but the broader use covers any inference that differs from deduction in introducing uncertainty. There are many such kinds of induction ranging from analogy to statistical inference, but one of the most common goes by the name "inference to the best explanation". Refs. [27,30,49,50] (p. 4)

To those who count abduction as IBE, nevertheless, my assumption could be extremely troublesome. By the same token, another assumption that what the value network does is nothing but IBE could be unacceptable. So, it seems necessary to make it explicit how I understand the relationship of IBE with abduction and induction. For this purpose, it seems useful to quote from Gangle's recent discussion of the basic inferential mechanism of machine learning systems:

> Arguably, the typical tasks for which machine learning systems consisting of backpropagation algorithms are trained fit the schema of inductive reasoning rather more closely than that of abductive reasoning. Training an A.I. machine learner to distinguish images of cats from those of dogs, for instance, seems to be a cognitive task more closely aligned with generalization than with explanation. Nonetheless, any such implementation of an algorithmic process to fulfill such a task may be understood as being abductive in principle to the degree that the trained network is intended to function successfully with regard to new data that is sufficiently dissimilar to its original training data. The trained network as a whole may in this respect be understood as a type of abductive hypothesis with respect to the successful fulfillment of relevantly similar tasks. Of course, the network itself does not understand itself in this way, but external trainers and collaborators might very well see things in such a light. Ref. [51] (p. 10)

I fully agree with Gangle's characterization of backpropagation algorithms as more inductive than abductive reasoning. Also, by assuming that we are talking about the workings of the value network, I might accept his claim that machine learning's distinguishing between cats from dogs is more with generalization than with explanation. Further, I can also accept his understanding that "the trained network as a whole may . . . be understood as a type of abductive hypothesis". What is important for me is that Gangle's understanding of machine learning is concordant with my understanding that, in both cases, together with induction, abduction and IBE can have proper places in a single neural network.

4.3. Criticism of the Implicit Assumptions of the Apparent Paradox

What is wrong in the ungrounded accusation of DeepMind for making AlphaGo and her stronger children less and less explainable by not separating policy and value networks? The most troublesome aspect of the reasoning of the suspicious mind in the thought experiment reconstructed might be that he or she might be begging the question by assuming that there is no room for abduction in AlphaGo Zero and other children of AlphaGo without a separate policy network. It is one thing that the policy network and value network can be correlated with abduction and IBE respectively in their major functions, but it is quite another that the resulting sole neural network left is simply a value network. There could still be abductions in the working of the sole neural network in AlphaGo Master, AlphaGo Zero, and Alpha Zero. We do not have any proof that, in the value network in AlphaGos older than AlphaGo Master, there cannot be abductions, even if IBE rather than abduction is its primary role and function.

Now we can understand why the suspicious mind in the thought experiment might be so puzzled and disappointed by the merging of the policy network and value network into one neural network. He or she assumed, rightly, that the old value network of any earlier AlphaGos must persist after the merge. But he or she assumed wrongly that there would be no remnant of the policy network after the merge. It is true that there is no separate neural network, i.e., policy network, whose primary function is to do abduction, but it is not so much abolished without any trace as integrated into value network. In other words, however weakened, what had been done by the policy network is still there to be done by the sole neural network of AlphaGo's children.

At this stage we might argue that, even though we need to distinguish sharply between abduction and IBE, imposing the distinction by the separation of the policy and value network is neither necessary nor justified without begging the question. In that vein, all implicit assumptions involved in the accusation of DeepMind for inviting the alleged paradox could be criticized as an ungrounded human-centered prejudice against non-human intellectual agents.

5. Further Reflections on Abduction and IBE in the AlphaGo Experiment

Now we can see that we cannot blame DeepMind for making AlphaGo less explainable simply because it has given up a separate policy network. Even if there is a slight loss in the accuracy of predicting moves (see below Section 5.1), as long as the elements of abduction and IBE are still there in the single neural network of AlphaGo's children, the accusation that DeepMind intentionally made AlphaGo less explainable is by no means justified. So, in a sense, we have resolved the apparent paradox about which we were worried so much. Then, what would be the lesson of all this?

5.1. Abduction and IBE in the Single Neural Network of AlphaGo's Children

What is going on within the sole neural network in AlphaGo Master, AlphaGo Zero, and Alpha Zero? If I am right, both abduction and IBE must be there. As for the abduction in that neural network, we may probe the question as to their source. Here are some of the possibilities:

(1) All of them are from the surviving elements of the previous policy network.
(2) Some of them are from the remnants of the policy network, and others are from the ever existing value network component.
(3) All of them are from the value network component.

We should not simply explain away any of these possibilities. Nevertheless, (1) seems not only highly unlikely but also pointless. If the policy network survives without its name, integrating policy and values networks into one would not be a significant renovation at all. It would be much more convincing if the role and function of the policy network is mostly transferred to the value network. But, then, we might be assuming that there were some redundancies or possible conflicts between the policy network and value network in AlphaGo. Be that as it may, we seem to have secured very good motivations to look into the collaboration of policy and value networks in AlphaGo. Also, we may want to have a sort of affidavit comparing AlphaGo with policy and value networks and AlphaGo Master, AlphaGo Zero, and Alpha Zero without such separate neural networks in terms of all the roles, functions, and efficiencies.

Both possibilities (2) and (3) are quite interesting. What is striking at first blush is that, in both cases, we were blind to the possibility of the existence of abduction in the workings of the value network. In (2), there might be abductions in the value network component, which is not coming from the policy network component. If (3) is the case, the previous policy network was almost redundant, and as a result there was nothing to lose in simply deleting it. Though hard to believe, again that would be a very interesting possibility full of lessons.

Once we begin to reflect upon these possibilities seriously, one extremely important report in Silver et al. [16] becomes all important. In a section titled as "Empirical training of AlphaGo Zero training" we read:

> To separate the contributions of architecture and algorithm, we compared the performance of the neural network architecture in AlphaGo Zero with the previous neural network architecture used in AlphaGo Lee . . . . Four neural networks were created, using either separate policy and value networks, as in AlphaGo Lee, or combined policy and value networks, as in AlphaGo Zero; and using either the convolutional network architecture from AlphaGo Lee or the residual network architecture from AlphaGo Zero. Each network was trained to minimise the same loss function . . . using a fixed data-set of self-play games generated by AlphaGo Zero after 72 h of self-play training. Using a residual network was more accurate, achieved lower error, and improved performance in AlphaGo by over 600 Elo. *Combining policy and value together into a single network slightly reduced the move prediction accuracy, but reduced the value error and boosted playing performance in AlphaGo by around another 600 Elo.* This is partly due to improved computational efficiency, but more importantly the dual objective regularises the network to a common representation that supports multiple use cases. Ref. [16] (pp. 356–357) (Emphases added)

From this dazzlingly informative passage, we are captivated by the italicized sentence. What could this mean for our previous discussion of the imaginary suspicion of whether DeepMind intentionally made AlphaGo less explainable by combining policy and value networks into a single network? The reduction of the move prediction accuracy is here described as "slightly reduced", as if there is nothing remarkable for any future research. Even though 600 Elo improvement is truly breathtaking, we should not ignore the fact that, however slight, combining policy and value networks into a single network yields the reduction of the move prediction accuracy.[9] In other words, it yields the drawback in the abductive capabilities of the AI Baduk player, thereby making it less explainable. That means, though refuted as rather ill-conceived and based on groundless prejudices in the previous section, the imaginary suspicion as to whether DeepMind made AlphaGo less explainable might still have a grain of truth.[10]

It is important whether DeepMind intentionally made AlphaGo and her children less explainable, for, if such a suspicion is supported by direct or indirect circumstantial evidence, we may try to recover their characteristics in the past. However slight, we seem to have some such evidence now. Reinstating policy and value networks as separate neural networks should be considered seriously insofar as we truly want explainable AI. If so, what kind of research should we pursue with our enhanced understanding of the functioning of the sole neural network of AlphaGo and her children?

5.2. Abduction and IBE from the Perspective of Human/Computer Interaction in Baduk

At this stage, it seems appropriate to resume my project of comparing human and AI Baduk players inspired by AlphaGo. So far I have published three articles belonging to this project [25,39,40]. However, with [16], which announced integrating policy and value networks into one neural network, my project lost the impetus to be continued. For, if AlphaGo and her children make indubitably and rapidly solid progress by making them less and less explainable, what is it for to compare human and AI Baduk players in all sorts of their behaviors? However, as a result of our discussion of abduction and IBE, time ripens to renew my project. Let us briefly review what I have found in my previous studies on human/computer interaction in Baduk.

In [39] I proposed to view each move in a game of Baduk as presenting an enthymematic argument. It was largely inspired by Paglieri and Woods [56,57], in which they suggested parsimony rather than charity as the driving force of enthymematic argumentation. Based on Fan Hui's detailed commentary of the games between AlphaGo Lee and Lee Sedol [58], I showed how to interpret them as enthymematic interactions between the players. In other words, AlphaGo won the match against Lee Sedol, thanks to her superior capabilities of executing enthymematic interactions. While AlphaGo's moves can be consistently interpreted as taking parsimony view of enthymemes, Lee Sedol seems to be fluctuating between parsimony and charity views of enthymeme [3] (p. 1145).

In any entire game of Baduk, as we encounter a very long chain of moves, such an understanding of Baduk as enthymeme interaction would have far-reaching implications to logic of enthymemes as well as to Baduk. Such a long chain of reasoning must be a combination of deductive, inductive, and abductive arguments. As Charles S. Peirce already realized clearly, these three different logical inferences are intricately intertwined in human reasoning. Even though there seems to be great advance quite recently (see, e.g., [43,44]), our understanding of the abduction-deduction-induction circle is not yet far superior to that of Peirce. Now, if AlphaGo has such a great ability in handling enthymemes, it is highly likely that it has been programmed to perform deductive, inductive, and abductive inferences at appropriate stages. But, of course, it would be even better if AlphaGo has acquired such a great mastery of handling enthymemes by her unsupervised self-learning. Further, in that case, AlphaGo may have the ability to exploit the combined force of all these three different types of reasoning [39] (p. 1164). If any Al Baduk player like AlphaGo Master, AlphGo Zero or Alpha Zero uses both abduction and IBE in her sole neural network, we do have actual instances of a logical agent exploiting all three different types of inferences, i.e., deduction, induction, and abduction. In other words, AI Baduk players may provide us with better information for understanding the deduction-induction-abduction cycle than human reasoners.

In retrospect, our discussion of abduction and IBE in AlphaGo may be instrumental for the future studies on enthymemes as well as on the relationship between abduction and IBE. Above all, if we learn something from AlphaGo's children's single neural network's handling of the subtle relationship between abduction and IBE, we might develop some new notions like enthymematic abduction in such a way that we can improve our understanding of deduction-induction-abduction cycle.

Park [40], which analyzes the concept of strategy in Baduk, was another attempt to compare human and AI Baduk players. Contrary to what one might believe, and to our dismay, we seldom find serious attempts to capture the essence of strategic reasoning even in more recent trends in game theory, such as evolutionary or epistemic game theory. In spite of the recent trends of rather active collaboration of epistemic logicians and game theorists, such as game logic or strategy logic, there is an unbridgeable gap between the concept of strategy in game theory and in real games. As an antidote to this unfortunate state of the study of strategic reasoning, I tried to capture the essence of the concept of strategy*, i.e., good strategy, in Baduk, according to which (1) it is not necessarily the case that a strategy is found in any game, (2) there has to be an intriguing interaction between a strategy and tactics, (3) it is inconsistency-robust.

What is notable is that the many interesting features of the games between AlphaGo and Lee Sedol should be fruitfully analyzed in terms of my notion of strategy in Baduk. So, I even expressed my hope as follows:

> One evidently nice outcome is that we secured a novel question: When is a strategy in games? By asking "When?" rather than "What?", we can give a fair hearing to the players in the formation, evaluation, and revision of the strategies.
>
> . . .
>
> One of the most urgent question to raise must be this: Does AlphaGo have any strategy?
>
> . . .
>
> How are we to explain such a rapid and unbelievable success of AlphaGo? What was the secret of success of Google DeepMind in developing AlphaGo? Of course, we need to examine carefully which component of AlphaGo was the crucial factor: Monte-carlo method, reinforcement learning, deep learning, machine learning, or what? But one thing evident is that even the DeepMind does not know exactly how AlphaGo achieved all the victories. If so, we have to ask whether AlphaGo has any strategy in playing games of Baduk. In answering this question, what I discuss here can be a point of departure. Ref. [40] (p. 1169)

In fact, I tackled the problem of whether AlphaGo has a strategy in a paper co-authored with a cognitive scientist, a neuroscientist, and an artificial intelligence researcher [25]. Of particular interest in the paper was the problem of understanding grand scale sacrificial strategies so frequently found in the self-play games of AlphaGo Master. I will not go into the details here, but will point out that in principle there seems to be no difficulty at all in applying the notion of strategy* developed in Park [40] to AlphaGo Master's games. The toughest part of applying the notion of strategy* to AlphaGo Master's games was the problem of determining when such a large scale sacrificial strategy began and ended. As one might imagine, our previous discussion of abduction and IBE in AlphaGo's single neural network must have far-reaching implications to this problem. For, while there must be an abductive element in conceiving the large scale sacrificial strategy*, the details of executing it must also involve elements of IBE. Evidently, large scale sacrificial strategy must involve initial sacrifices and clear cut calculi of gains and losses involved, which invite us to the even harder general problem of trade in Baduk.

In passing, we may note that Silver et al. [16] mentioned AlphaGo Zero's novel strategies in a few places:

> AlphaGo Zero discovered a remarkable level of Go knowledge during its self-play training process. This included fundamental elements of human Go knowledge, and also non-standard strategies beyond the scope of traditional Go knowledge. Ref. [16] (p. 357)
>
> Humankind has accumulated Go knowledge from millions of games played over thousands of years, collectively distilled into patterns, proverbs and books. In the space of a few days, starting tabula rasa, AlphaGo Zero was able to rediscover much of this Go knowledge, as well as novel strategies that provide new insights into the oldest of games. Ref. [16] (p. 358)

Though interesting and encouraging, these remarks are at best a cryptic hints rather than presenting a refined notion of strategy. On the other hand, as we hinted at in the last paragraph, we can be hopeful to flesh out our concept of strategy* in AI Baduk players by furthering our study of the relationship between abduction and IBE in their single neural network.

5.3. More Philosophical Issues from Abduction and IBE in Human/Computer Interaction

One might think that what I claim here is still merely a dream or at best a programmatic sketch. I am willing to concede that such a criticism is absolutely right, but I do want to indicate that by taking human/computer interaction more seriously, i.e., more philosophically, my dream or programmatic sketch might look much more realistic.

First of all, we should have treated AI Baduk players, i.e., AlphaGo and her children, more seriously as genuine individuals. As confessed in Section 4.1 above, I was begging the question by assuming that only by supplying abductive elements through the policy network could AlphaGo, AlphaGo's children, or any AI Baduk players do abductive reasoning. DeepMind has demonstrated that I was wrong by Alpha Go and Alpha Zero's marvelous successes. The power of unsupervised deep machine learning turns out to be beyond human understanding.

All Baduk players, by now, accept it as a matter of fact that AlphaGo Zero outweighs the strongest human player by more than 1000 Elo. That means there can't be even games between human and AI Baduk players, but only handicap games. Or, we might describe what is going on by invoking a sort of reverse Turing test by AI Baduk players. We might be now in a situation in which we should pass the test so that we can be judged as minimally rational, so to speak. For a long time, the foremost issue for computer Go was developing a computer Go program that plays Baduk as humans do. But, the table is turned, and now every human Baduk player is trying hard to imitate AlphaGo's play. If so, how about DeepMind? Have creators of AlphaGo and her children treated them as genuine individuals, as rational Baduk players? To a certain extent, the answer must be "Yes", for, at least, they differentiated between AlphaGo Fan, AlphaGo Leen, AlphaGo Master, AlphaGo Zero, and Alpha Zero. However, in a sense, they could have treated them more seriously as genuine individuals. In principle, they could have been much more discriminating in making small changes to the mechanism of AlphaGo and her children.[11] But, of course, we can anticipate how this stream of thought will be led to the arduous task of handling the ontological problems such as identity through change or personal identity. A huge field of exciting interdisciplinary research for computer scientists, AI researchers, neuroscientists, behavioral economists, game theorists, psychologists and philosophers of mind is wide open, thanks to human/computer interaction in Baduk.

Another issue to which we could apply the results achieved on paper seems to be how to sharpen our conception of explainability itself. We do have an urgent need for such a sharpening not only in the discussion of explainability of AI but also in the discussion of IBE. Above all, we may note that our discussion in Section 1 already indicates the need to distinguish between the different concepts of explainability, starting with the distinction between explainability-as-such and explinbility-in-use. So far explainability has been treated implicitly as having only the former meaning. In this regard, the situation seems to be similar in the case of the concept of solvability. Unlike the solvability or explainability-as-such, explainability-in-use might be a triadic relation presupposing who is explaining certain things to whom: x explains y to z. In Baduk, for example, AlphaGo may explain "why she played f-8 at her 156th move" to a human opponent by using winning-rate graphs. The situation in which an AI Baduk player Leela explains her sacrificial strategy in the middle game to another AI player Fine Art is certainly imaginable. In principle, it seems not only possible but also necessary to mark the explainer and explainee, so to speak. As we can probe questions as to what explanations should be like, even though the original problem was IBE (see [31]; see also Cabrera [37], and the works cited there), our pursuit of explainability of AI Baduk players may lead us even to the deep problem of the metaphysics of explanation (in Popper's sense).

Still another issue could be found in the so-called self-play games of AlphaGo (or any other AI players). Since, in these self-play games, an AI player is playing against an opponent, who has exactly the same background common knowledge, we could have a wonderful opportunity to deepen our understanding of the problem of rationality in formal epistemology or game logic. Even though it is not easy to formulate the issue clearly and rigorously, I tend to believe that it may be connected to the issues I touched upon under "proto-logic", which concern the comparative studies of non-human animals, humans, artificial intelligences, and possibly angels. As we are already dealing with the problem of comparing human and AI Baduk players, we are bound to face with the problem of other minds ([42,59]; see also [60]). Which is more difficult: understanding a superhuman AI Baduk player's moves or making sense of a novice's moves? Which is tougher to explain: Einstein's mind or a nameless worm's movement?

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicble.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author is indebted to the editors and the three anonymous reviewers of *Philosophies* for their extremely helpful criticisms.

Conflicts of Interest: The author declares no conflict of interest.

Notes

[1] See also Pumperla and Ferguson's interesting remarks on the great success of AlphaGo Zero: "To us, the most astonishing thing about AlphaGo Zero is how it does more with less. Inmany ways, AGZ is much simpler than the original AlphaGo. No more handcraftedfeature planes. No more human game records. No more Monte Carlo rollouts. Insteadof two neural networks and three training processes, AlphaGo Zero used one neural network and one training process." [24] (p. 290)

[2] The following excerpts should be enough for following the discussion of DeepMind's characterization of AlphaZero contrasted with AlphaGo programs: "1. AlphaGo Fan is the previously published program 12 that played against Fan Hui in October 2015. . . . 2. AlphaGo Lee is the program that defeated Lee Sedol 4–1 in March, 2016. It was previously unpublished but is similar in most regards to AlphaGo Fan 12. However, we highlight several key differences to facilitate a fair comparison. First, the value network was trained from the outcomes of fast games of self-play by AlphaGo, rather than games of self-play by the policy network; this procedure was iterated several times—an initial step towards the tabula rasa algorithm presented in this paper. Second, the policy and value networks were larger than those described in the original paper . . . 3. AlphaGo Master is the program that defeated top human players by 60–0 in January, 2017 34. It was previously unpublished but uses the same neural network architecture, reinforcement learning algorithm, and MCTS algorithm as described in this paper. However, it uses the same handcrafted features and rollouts as AlphaGo Lee 12 and training was initialised by supervised learning from human data. 4. AlphaGo Zero . . . ". [16] (p. 360). (The point 4 is cited in the main text.)

[3] It is my presumption that DeepMind's demonstration of the power consists of a dramatic series of events: (1) AlphaGo Fan's winning against a professional Baduk player, i.e., Fan Hui, (2) AlphaGo Lee's victory against Lee Sedol, (3) AlphaGo Master's perfect winning against 60 top level professional Baduk players on internet, and (4) AlphaGo Zero's winning against Ke Jie, the current world champion. Not any one of these events but the entire series of them made the explainability of AI the focal issue of our time once and for all.

[4] Whether the imaginary suspicion is justified or not, we have more than enough reasons to be interested in philosophical issues embedded in XAI. For example, according to Medianovskyi and Pietarinen, "[t]he theoretical issues affecting XAI thus reflect long-standing issues in the philosophy of language, science and communication". [5] (p. 3)

[5] One anonymous reviewer made a penetrating criticism against the argument sketched in Section 3. Right now, it is simply beyond my ability to respond to the criticism. So, please gently allow me to report it: "It seems that there is an equivocation: the argument sketched in Section 3 suggests that the strong integration of policy and value networks in post-AlphaGo systems made those systems less explainable (which seems quite plausible) and that because that design choice was made by DeepMind, the loss of explainability is "intentional"; but in the same section, such "intentionality" is also conceived as if it were _essentially_ a decision about explainability. If the loss of explainability were simply an unintended consequence of this design choice (isn't this indeed plausible? if not, why not?), then is the second sense of "intentionality" warranted? At the very least, this question should probably be addressed explicitly."

[6] My assumptions 2 and 3 will continuously be discussed in the following, and my assumption 10 will be discussed in broader context in Section 5.

[7] One anoymous reviewer rightly pointed out the need to define abduction and IBE. In order not to prejudge the issue, however, I would ask the readers to compare Cabrera's presentation of IBE as "four-step argument schema" with the usual shema for Peircean abduction. [37]

[8] Minnameier's opinion is widely shared. Cabrera's list could be expanded by including, e.g., 45, 46, 47, and even 48. (See [38] (pp. 29–34.))

[9] We may note that, inasmuch as prediction is a standard issue in philosophy of science, we have very good reasons to expect fruitful results from the study of prediction in human/computer games of Baduk. Magnani's inquiries on anticipation as a kind

of abduction in human and machine cognition could be one of the rare precedents in that direction of research [52–55]. In that sense, the abrupt disappearance of the policy network in AlphaGo's mechanism is unfortunate.

[10] According to Medianovskyi and Pietarinen [5], induction has been the paradigm of machine learning (ML): "Instead of broadening the theoretical base of the types of reasoning, ML has stuck to its guns of induction and founded even the otherwise promising newcomer of self-supervised learning on just the same modalities of inductive reasoning" [5] (pp. 1–2). So, it might even be possible to strengthen the suspicion gainst DeepMind's decision to combine the policy and value networks into a single neural network.

[11] What I have in mind could be exemplified well by using the different versions of AlphaGo and their human opponent. Lee Sedol was confronting AlphaGo Lee rather than AlphaGo Fan, and Ke Jie was confronting AlphaGo Zero rather than AlphaGo Lee. It could have been nicer if those human players could have been fully informed of how much progress their AI opponents had achieved.

References

1. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
2. Doran, D.; Schulz, S.; Besold, T.R. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv* **2017**, arXiv:1710.00794.
3. Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. 2017. Available online: https://arxiv.org/abs/1702.08608 (accessed on 19 May 2022).
4. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]
5. Medianovskyi, K.; Pietarinen, A.-V. On Explainable AI and Abductive Inference. *Philosophies* **2022**, *7*, 35. [CrossRef]
6. Schubbach, A. Judging machines: Philosophical aspects of deep learning. *Synthese* **2019**, *198*, 1807–1827. [CrossRef]
7. Zednik, C. Will machine learning yield machine intelligence? In Proceedings of the 3rd Conference on Philosophy and Theory of Artificial Intelligence, Leeds, UK, 4–5 November 2017; Springer: Cham, Switzerland; pp. 225–227.
8. Zednik, C. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philos. Technol.* **2021**, *34*, 265–288. [CrossRef]
9. Hoffman, R.R.; Clancey, W.J.; Mueller, S.T. Explaining AI as an Exploratory Process: The Peircean Abduction Model. *arXiv* **2020**, arXiv:2009.14795v2. [cs.AI].
10. Burrell, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data Soc.* **2016**, *3*, 205395171562251. [CrossRef]
11. Cappelen, H.; Dever, J. *Making AI Intelligible: Philosophical Foundations*; Oxford University Press: Oxford, UK, 2021.
12. Heuillet, A.; Couthouis, F.; Díaz-Rodríguez, N. Explainability in deep reinforcement learning. *Knowl.-Based Syst.* **2021**, *214*, 106685. [CrossRef]
13. Montavon, G.; Samek, W.; Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Processing* **2018**, *73*, 1–15. [CrossRef]
14. Ras, G.; van Gerven, M.; Haselager, P. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 19–36.
15. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
16. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef]
17. Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **2018**, *362*, 1140–1144. [CrossRef] [PubMed]
18. McGrath, T.; Kapishnikov, A.; Tomašev, N.; Pearce, A.; Hassabis, D.; Kim, B.; Paquet, U.; Kramnik, V. Acquisition of Chess Knowledge in AlphaZero. *arXiv* **2021**, arXiv:2111.09259v1. [cs.AI].
19. Thornton, J.M.; Laskowski, R.A.; Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* **2021**, *27*, 1666–1671. [CrossRef] [PubMed]
20. Egri-Nagy, A.; Törmänen, A. The Game Is Not over Yet—Go in the Post-AlphaGo Era. *Philosophies* **2020**, *5*, 37. [CrossRef]
21. Egri-Nagy, A.; Törmänen, A. Derived metrics for the game of Go—Intrinsic network strength assessment and cheat-detection. In Proceedings of the Eighth International Symposium on Computing and Networking (CANDAR), Naha, Japan, 24–27 November 2020.
22. Kasparov, G. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*; John Murray: London, UK, 2017.
23. Hsu, F.-H. *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*; Princeton University Press: Princeton, NJ, USA, 2002.
24. Pumperla, M.; Ferguson, K. *Deep Learning and the Game of Go*; Manning: Shelter Island, NY, USA, 2019.
25. Park, W.; Kim, S.; Kim, G.; Kim, J. AlphaGo's Decision Making. *J. Appl. Log.—IFCoLog J. Log. Appl.* **2019**, *6*, 105–155.
26. Harman, G. The inference to the best explanation. *Philos. Rev.* **1965**, *74*, 88–95. [CrossRef]

27. Thagard, P. The best explanation: Criteria for theory choice. *J. Philos.* **1978**, *75*, 76–92. [CrossRef]
28. Lycan, W.G. *Judgement and Justification*; Cambridge University Press: Cambridge, UK, 1988.
29. Barnes, E. Inference to the loveliest explanation. *Synthese* **1995**, *103*, 251–277. [CrossRef]
30. Lipton, P. *Inference to the Best Explanation*, 2nd ed.; Routledge: New York, NY, USA, 2004.
31. Douven, I. *"Abduction", The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Spring: Berlin/Heidelberg, Germany, 2021. Available online: https://plato.stanford.edu/entries/abduction/ (accessed on 17 May 2022).
32. Minnameier, G. Peirce-suit of truth—Why inference to the best explanation and abduction ought not to be confused. *Erkenntnis* **2004**, *60*, 75–105. [CrossRef]
33. McKaughan, D. From Ugly Duckling to Swan: C. S. Peirce, Abduction, and the Pursuit of Scientific Theories. *Trans. Charles S. Peirce Soc. Q. J. Am. Philos.* **2018**, *4*, 446–468.
34. Campos, D. On the Distinction between Peirce's Abduction and Lipton's Inference to the Best Explanation. *Synthese* **2009**, *180*, 419–442. [CrossRef]
35. McAuliffe, W. How did Abduction Get Confused with Inference to the Best Explanation? *Trans. Charles S. Peirce Soc.* **2015**, *51*, 300–319. [CrossRef]
36. Park, W. On Classifying Abduction. *J. Appl. Log.* **2015**, *13*, 215–238. [CrossRef]
37. Cabrera, F. Inference to the Best Explanation—An Overview. In *Handbook of Abductive Cognition*; Magnani, L., Ed.; Springer: Dordrecht, The Netherlands, 2022. Available online: http://philsci-archive.pitt.edu/20363/1/Inference%2Bto%2Bthe%2BBest%2BExplanation%2B-%2BAn%2BOverview%2BPenultimate%20Draft.pdf (accessed on 17 March 2022).
38. Park, W. *Abduction in Contect: The Conjectural Dynamics of Scientific Reasoning*; Springer: Berlin, Germany, 2017.
39. Park, W. Enthymematic Interaction in Baduk. In *Logical Foundations of Strategic Reasoning, Special Issue of Journal of Applied Logics—IFCoLog Journal of Logics and their Applications*; Park, W., Woods, J., Eds.; College Publications: London, UK, 2018; Volume 5, pp. 1145–1167.
40. Park, W. When Is a Strategy in Games? In *Logical Foundations of Strategic Reasoning, Special Issue of Journal of Applied Logics—IFCoLog Journal of Logics and their Applications*; Park, W., Woods, J., Eds.; College Publications: London, UK, 2018; Volume 5, pp. 1169–1203.
41. Park, W. On Abducing the Axioms of Mathematics. In *Abduction in Cognition and Action: Logical Reasoning, Scientific Inquiry, and Social Practice, Sapere*; Shook, J., Paavola, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2021; pp. 161–175.
42. Park, W. What Proto-logic Could not be. *Axiomathes* **2021**. Available online: https://doi.org/10.1007/s10516-021-09582-3XXX (accessed on 17 March 2022). [CrossRef]
43. Minnameier, G. Abduction, Selection, and Selective Abduction. In *Model-Based Reasoning in Science and Technology. Models and Inferences: Logical, Epistemological, and Cognitive Issues, Sapere*; Magnani, L., Casadio, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 309–318.
44. Minnameier, G. Forms of abduction and an inferential taxonomy. In *Springer Handbook of Model-Based Reasoning*; Magnani, L., Bertolotti, T., Eds.; Springer: Berlin, Germany, 2016; pp. 175–195.
45. Magnani, L. *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Ewasoning*; Springer: Berlin/Heidelberg, Germany, 2009.
46. Woods, J. *Errors of Reasoning: Naturalizing the Logic of Inference*; College Publications: London, UK, 2013.
47. Mackonis, A. Inference to the Best Explanation, Coherence and Other Explanatory Virtues. *Synthese* **2013**, *190*, 975–995. [CrossRef]
48. Hintikka, J. What is abduction? The fundamental problem of contemporary epistemology. *Trans. Charles S. Peirce Soc.* **1998**, *34*, 503–533.
49. Harman, G. *Thought*; Princeton University Press: Princeton, NJ, USA, 1973.
50. Thagard, P. Naturalizing Logic: How Knowledge of Mechanisms Enhances Inductive Inference. *Philosophies* **2021**, *6*, 52. [CrossRef]
51. Gangle, R. Backpropagation of Spirit: Hegelian Recollection and Human-A.I. Abductive Communities. *Philosophies* **2022**, *7*, 36. [CrossRef]
52. Magnani, L. Playing with anticipations as abductions. Strategic reasoning in an eco-cognitive perspective. In *Logical Foundations of Strategic Reasoning, Special Issue of Journal of Applied Logic—IfColog Journal of Logics and their Applications*; Park, W., Woods, J., Eds.; College Publications: London, UK, 2018; Volume 5, pp. 1061–1092.
53. Magnani, L. AlphaGo, Locked Strategies, and Eco-Cognitive Openness. *Philosophies* **2019**, *4*, 8. [CrossRef]
54. Magnani, L. Anticipations as Abductions in Human and Machine Cognition: Deep Learning: Locked and Unlocked Capacities. *Postmod. Open.* **2020**, *11*, 230–247. [CrossRef]
55. Magnani, L. Human Abductive Cognition Vindicated: Computational Locked Strategies, Dissipative Brains, and Eco-Cognitive Openness. *Philosophies* **2022**, *7*, 15. [CrossRef]
56. Paglieri, F.; Woods, J. Enthymematic parsimony. *Synthese* **2011**, *178*, 461–501. [CrossRef]
57. Paglieri, F.; Woods, J. Enthymemes: From reconstruction to understanding. *Argumentation* **2011**, *25*, 127–139. [CrossRef]
58. Hui, F. Commentary on DeepMind Challenge Match between Lee Sedol and AlphaGo. 2016. Available online: https://deepmind.com/research/alphago/ (accessed on 25 December 2017).
59. Carruthers, P. *Human and Animal Minds: The Consciousness Questions Laid to Rest*; Oxford University Press: Oxford, UK, 2019; Introduction.
60. Park, W. How to Learn Abduction from Animals? From Avicenna to Magnani. In *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues*; Magnani, L., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 207–220.