Journal of AI Humanities Vol.16 (April 2024)
DOI https://doi.org/10.46397/JAIH.16.4

Can AI become an Expert?*

Hyeongyun Kim**

Abstract

With the rapid development of artificial intelligence (AI), understanding its capabilities and limitations has become significant for mitigating unfounded anxiety and unwarranted optimism. As part of this endeavor, this study delves into the following question: Can AI become an expert? More precisely, should society confer the authority of experts on AI even if its decision-making process is highly opaque? Throughout the investigation, I aim to identify certain *normative* challenges in elevating current AI to a level comparable to that of human experts. First, I will narrow the scope by proposing the definition of an expert. Along the way, three normative components of experts —trust, explainability, and responsibility—will be presented. Subsequently, I will suggest why AI cannot become a trustee, successfully transmit knowledge, or take responsibility. Specifically, the arguments focus on how these factors regulate expert judgments, which are not made in isolation but within complex social connections and spontaneous dialogue. Finally, I will defend the plausibility of the presented criteria in response to a potential objection, the claim that some machine learning-based algorithms, such as AlphaGo, have already been recognized as experts.

Key words : AI Ethics, Applied Ethics, Expert, Trust, Explainability, Responsibility, Transparency

---

* I am grateful for the invaluable feedback from Prof. Jovana Davidovic, as well as from three anonymous referees for the *Journal of AI Humanities*.

** Ph.D. Student, Department of Philosophy, The University of Iowa

〈목 차〉

1. Introduction

Public discourse on artificial intelligence (AI) often oscillates between anxiety and optimism. On the one hand, the emergence of new technologies has historically engendered concerns about potential job displacement. A classic example is the Luddite movement, where workers destroyed weaving machines that menaced their livelihoods. In a similar vein, AI is now regarded as a latent threat, provoking reactions akin to repugnance and Luddite behavior among workers and consumers (Youn & Jin, 2021, p. 4). The anxiety has been bolstered by the prediction that AI will not only perform tasks that were once exclusive to humans, but will also ultimately replace their roles with superior efficiency, accuracy, and objectivity. According to one comprehensive survey of machine learning researchers, "AI will outperform humans in many activities in the next ten years, such as translating languages (by 2024), writing high-school essays (by 2026), […] and working as a surgeon (by 2053)" (Grace, Salvatier, Dafoe, Zhang, & Evans, 2018, p. 729).

On the other hand, advancements in AI also foster optimism about progress. Many people envision a future where AI autonomously takes on

the role of an expert, equipped with extensive knowledge in specific areas to advise humans. For instance, AlphaGo, a machine learning-based algorithm that has mastered the game of Go and has been advising even professional Go players to improve their strategic skills, is often cited as evidence supporting such a vision.

However, the question of whether AI can be considered an expert goes far beyond worries about job prospects and a blueprint for the promising future. Imagine a scenario where the public opinion described above becomes so prevalent that society is now considering whether to confer the authority of experts on AI. The endorsement of such a status is not trivial; for it would ensure the legitimate use of highly *opaque* AI[1] in domains where *ethically problematic* concerns (about, say, killer robots, predictive policing algorithms, and AI for medical diagnostics) are intertwined. Ross (2022), for instance, suggests conferring the quasi-intellectual authority of experts on opaque AI by appealing to the similarity between the (allegedly opaque) decision-making processes of both AI and human experts.

> I have suggested we take our successful social practice of deferring to specialized experts as a guide for developing an epistemically and ethically sound method for utilizing opaque AI. To this end, we will need to examine when (i.e., under what conditions) it is epistemically and ethically responsible to defer to experts rather than relying on one's own reasoning. We also need to know what features make an individual a genuine expert, how, as a society, we determine that some individual is an expert, and what methods we use for deciding how to act when multiple experts

1) Johnson (2021) succinctly explains the general notion of opaque AI as follows: "often machine learning programs instantiate so-called "black box" algorithms, i.e., those where it is difficult (and arguably impossible) for human observers to describe the rationale for a particular outcome" (p. 9942).

disagree in their decisions. (p. 6)

I argue that this consideration must take into account whether AI currently satisfies the following *normative* criteria: *trust, explainability, and responsibility*. Such factors are deemed normative in that they hold both epistemic and ethical significance. For instance, the issues of trust and explainability raise the question of whether it is epistemically justified to accept the judgment of an expert AI even when its decision-making processes are inscrutable to humans. Furthermore, it is widely known that the increased autonomy and complexity of AI algorithms reduce the developer's control over them, giving rise to concerns about "responsibility gaps" (Matthias, 2004). When it comes to the role of experts, the situation becomes even more tricky; for experts assume several *moral* responsibilities, such as blameworthiness, intellectual honesty, and expert communication.[2)]

Throughout the investigation, I aim to identify certain *normative* challenges in elevating current AI to a level comparable to that of human experts. Accordingly, I will outline the appropriate roles and limitations of AI in decision-making processes within the expert community. The paper is divided into three parts: first, I will narrow the scope by proposing the definition of an expert. Along the way, three normative components of experts–trust, explainability, and responsibility–will be presented. Subsequently, I will suggest why AI cannot become a trustee, successfully transmit knowledge, or take responsibility. Specifically, the arguments focus on how these factors regulate

2) I borrow this concept from Desmond (see Desmond, 2021 and 2024). Although expert communication seems to overlap with explainability at first glance, it embraces an ethical dimension where "a fundamental dilemma between prioritizing actionability and prioritizing scientific transparency" occurs (Desmond, 2021, p. 24). Thus, by adopting this notion, I am more likely to focus on experts' *moral* responsibility.

expert judgments, which are not made in isolation but within complex social interactions and spontaneous dialogue. Finally, I will defend the plausibility of the presented criteria in response to a potential objection, the claim that some machine learning-based algorithms, such as AlphaGo, have already been recognized as experts.

2. What is an Expert?

I shall commence with the following question: What is the definition of an expert? One common approach is to define an expert within the context of the association with *specialized* knowledge that stems from the rapid development of society. According to this perspective, experts are individuals characterized by possessing "comprehensive and authoritative knowledge in a particular area not possessed by most people" (Caley et al., 2014, p. 232). Correspondingly, expertise or expert knowledge could be defined as "substantive information on a particular topic that is not widely known by others" (Martin et al., 2012, p. 30).

There are, of course, countless experts and bodies of expert knowledge satisfying these conditions. Given our main purpose, however, it would be advantageous to narrow the scope to *intellectual* experts. This is because our concerns regarding AI as an expert primarily revolve around ethically significant contexts that require complex reasoning and a rational decision-making process. According to Goldman (2001), an intellectual expert can be characterized by having "a superior quantity or level of knowledge in some domain and an ability to generate new knowledge in answer to questions within the domain" (p. 91). Then one might formulate a

tentative definition of an expert as follows: A person, S, is an expert iff:

(i) S possesses substantive information on a particular topic x, i.e., substantial propositional knowledge of a domain x;
(ii) S's knowledge is not possessed by most people;
(iii) S's knowledge is comprehensive and authoritative.

However, this definition is not a full-fledged description capturing the nature of experts and requires further analysis. First, as contained in Goldman's explanation above, expertise is not merely confined to propositional knowledge. In addition to substantive expertise (knowledge of a domain), experts are required to demonstrate certain abilities: normative and adaptive expertise. While normative expertise refers to the ability to communicate judgments with clarity and accuracy, adaptive expertise indicates the ability to adapt to new circumstances (Martin et al., 2012; Caley et al., 2014). This is the reason experts are not regarded as mere information *gatherers*. Knowledge of a domain is often transformed and developed by the process of transmission as well as its application. Each ability, as it were, rotates around propositional knowledge as its axis. In this regard, experts encompass the notion of a possessor, transmitter, and developer of specialized knowledge.

Second, it is worth noting that both experts and expertise are defined in terms of their relationship with *laypeople*. Here, the term 'laypeople' refers to individuals who lack both knowledge in a specific field and the ability to address issues related to that domain. Generally speaking, laypeople acquire knowledge or solve problems by conferring intellectual authority to experts and relying on their judgments. As Hardwick (1985) points out, "appeals to the authority of experts often provide justification for claims to

know, as well as grounding rational belief" (p. 336). The main point is that a layperson is justified in believing experts' judgments even *without* understanding the underlying rationale (Hardwick, 1985, p. 339). In fact, we trust experts and their judgments precisely because we are unable to fully comprehend or evaluate them; it is simply far-fetched to become an expert in every domain due to limiting factors like time, cost, or intellectual ability (Goldman, 2001, p. 89). Trust thus constitutes one of the normative criteria required for experts.

Third, an expert's authority is maintained insofar as they are a member of the expert community. Unlike door-to-door salesmen, experts do not visit people to prove their authority firsthand. Instead, they demonstrate their expertise by attaining institutional certification and participating in their professional community. Therefore, even if a layperson's trust in an expert is necessarily blind (Goldman, 2001, p. 86), explainability is still required for effective communication among experts. Within the expert community, each member not only cultivates knowledge within their specialization through adaptive expertise but also should be subject to the verification of the ownership of that knowledge by utilizing normative expertise. Additionally, although explainability plays a pivotal role in the expert-to-expert interactions, it also assigns certain tasks and duties to experts in the novice-to-expert relationship (I will address this issue in Chapter 3 with more details).

Based on the analysis presented, I shall now revise the previous definition of an expert. A person, S, is an expert iff:

(i) S possesses substantial propositional knowledge of a domain x and has ability to transmit and develop it;

(ii) S's knowledge is largely inaccessible to layperson, who trust S's judgment without understanding the underlying rationale;
(iii) S is a member of the expert community of domain x, which makes S's knowledge authoritative.

The refined definition clearly emphasizes the necessity of trust and explainability as normative criteria for experts. Furthermore, it is essential to include responsibility as another normative criterion stemming from this definition. Take, for instance, the second condition stating that to become an expert, one must serve as a trustee. As argued by Ryan (2020), "The trustee needs to be able to understand and act on what is entrusted to them and be held *responsible* for those actions" (p. 2761). Arguably, responsibility is extensively involved in both the *preparation* and the *outcomes* of expert judgments; experts, on the one hand, are responsible for providing clear instructions, preventing potential misunderstandings, and maintaining intellectual honesty in their decision-making process. In terms of outcomes, on the other hand, experts shoulder various responsibilities such as accepting blameworthiness and mitigating the repercussions of professional misconduct.

It is also worth noting that responsibility is not solely attributed to individual experts but extends to the entire expert community. The expert community shoulders the responsibility of sanctioning individuals who have committed wrongdoing. Moreover, in cases where such misconduct is structural or widespread, they may also become a direct target of blame. This collective responsibility constitutes a significant ethical dimension within the expert group, as well as among individual experts.

3. Can AI become an Expert?

Now, I will critically examine the capacities of AI in serving as a trustee, successfully transmitting knowledge, and taking responsibility, in turn.

3.1. AI and Trust

The fact that our trust in experts can offset our ignorance of their actual decision-making process is often used as a rationale to justify the use of machine learning-based opaque AI. As mentioned earlier, Ross (2022) espouses this view as follows:

> Laypeople believe the judgments of specialized experts because they trust those experts—not because they understand their reasoning—and they trust those experts because their social framework includes institutions whose role it is to verify the legitimacy of specialized experts. [···] we routinely employ opaque processes in ethically significant domains. And I will argue that there is no special reason to embrace (2) [the use of an opaque process is ethically impermissible] in the case of AI while rejecting it in the case of human experts. (p. 6)

I will argue that this suggestion overlooks two significant differences between the reasoning of human experts and that of AI. First, unlike human experts whose reasoning can be partially fathomed by surmising their mental phenomena, we *cannot* access the AI's chain of reasoning behind its actions in the least, due to the lack of a shared mental model (Danks, 2016; Roff & Danks, 2018). This raises serious concerns about forming trust, both between experts and laypeople, and among experts

within the expert community. Second, for that very reason, AI also lacks a sufficient social framework to verify the legitimacy of specialized experts.

Let me start with the first point. The question of whether AI can effectively serve as a trustee is both complicated and controversial. Nevertheless, the most common starting point for the discussion is to make a clear distinction between trust and reliability (Kirkpatrick, Hahn, & Haufler, 2017; Roff & Danks, 2018; Ryan, 2020; Ross, 2022). Reliability is often characterized by the extent to which the trustee fulfills the trustor's expectations for their actions. Trust, however, encompasses more than just predictability. For instance, Kirkpatrick, Hahn, & Haufler (2017) illustrate the difference between a trustee with good intentions and one who is *merely reliable* by delineating the following scenario:

> [...] we may be lost while driving to our favorite bookshop. While stopped at a red light, we see a man stepping into a taxi and overhear him tell the driver that his destination is the very same bookshop that we too are trying to locate. We follow the driver, *relying* on her to successfully reach our destination. Now imagine the same situation, but after the passenger calls out his destination, we roll down the window and tell the driver that we too are headed to the same bookshop, and we will follow her. The driver happily assents. [...] In the first case, we rely on the driver because we believe that she is bound and constrained by her role as a taxi driver; she is fulfilling a professional obligation, and we know that she will predictably come through for us. By contrast, in the second case, we rely on the driver not only because it is her professional obligation, but also because she is disposed to us as individuals; the driver recognizes our vulnerability and shows us goodwill in her acknowledgment of our dependency on her in the given situation. (p. 145)

As the taxi driver case plausibly suggests, trust *prima facie* presupposes spontaneous dialogue and interaction between a trustor and a trustee; in this regard, trust is essentially interpersonal, requiring moral agency, intention, and consciousness (Kirkpatrick, Hahn, & Haufler, 2017, p. 149).

However, this explanation seems too broad to pinpoint the trust between an expert and a layperson. Let's consider a scenario where an expert *mechanically* handles tasks driven not by benevolent intentions but solely by wages. Such an expert, of course, can become a trustee in relation to laypeople. Nevertheless, we cannot simply say that this expert is a trustee in the sense that they are merely reliable like a calculator; further explanation is needed.

I propose that what characterizes the trust between an expert and a layperson is its distinctive subject: *autonomy*. A reliable machine is, as I put it, one that performs its tasks accurately in accordance with our expectation. In contrast, what we seek from experts is not predetermined action but the specific judgment they *form*. Unlike reliable machines, our trust in experts is primarily oriented toward their *ability* to make accurate judgments. Hence, we would not feel betrayed even if experts were to make judgments and take actions that are completely different from our expectations for problem-solving.

Our trust in experts is thus better understood within a more fine-grained classification: on the one hand, we trust experts' substantial expertise, with beliefs that their knowledge is true and well-justified within the expert community. On the other hand, we value experts' adaptive expertise, i.e., their ability to cope with unforeseen situations. We regard them as autonomous rational beings in this respect. The question then arises: how can we foster this kind of trust that pertains to adaptive expertise and autonomy?

To elucidate this, Roff & Danks (2018) introduce the notion of a mental model of the world shared by the trustor and the trustee. In this model, the trustor does not have to stipulate or observe any predetermined actions. Instead, such a model enables the trustor to know "roughly *what* the trustee will do, and also *why* she pursues that course of action" (p. 7). Indeed, we do not have any difficulty understanding other people's beliefs, desires, and intentions underlying their actions, at least in *a very rough sense* (pp. 6-7). Danks (2016) further emphasizes the significance of this common ground as follows:

> [...] attempts to interpret an autonomous technology in terms of human-like beliefs and desires can go spectacularly awry. When a human driver sees a ball in the road, most of us automatically slow down significantly, to avoid hitting a child who might be chasing after it. If we are riding in an autonomous car and see a ball roll into the street, we expect the car to recognize it, and to be prepared to stop for running children. The car might, however, see only an obstacle to be avoided. If it swerves without slowing, the humans on board might be alarmed – and a kid might be in danger.
>
> Our inferences about the "beliefs" and "desires" of a self-driving car will almost surely be erroneous in important ways, precisely because the car doesn't have any human-like beliefs or desires. We cannot develop interpersonal trust in a self-driving car simply by watching it drive, as we will not correctly infer the whys behind its actions.

In a similar vein, it is well-known that large language models (LLMs) such as ChatGPT sometimes generate content that is filled with lies and deception. However, "they are unconstrained by any concern regarding truth or falsity. Indeed, they have no conception of truth or falsity precisely

because they have no mental model of the world" (Herzfeld, 2023, p. 669). So the fact that AI *might* generate deceptive answers is relatively trivial because human experts also sometimes lie. Rather, the crux of AI deception is that we are unable to know, even in a rough sense, *when AI will lie and why it does so*; in contrast, human experts' deception reveals their intentions, desires and beliefs.

It is therefore questionable how the legitimacy of AI's judgment can be verified within a social framework. Regarding expert-to-novice trust, it might be tempting to conclude that there is no significant difference between the reasoning of human experts and that of AI, because the judgments of both are mostly arcane to laypeople. However, AI does not even have any mental models to share, nor does it have any promising alternatives that allow us to roughly know its intentions. And reliability alone is not enough to accomplish this task due to AI's *autonomy*. Danks concisely encapsulates this dilemma as follows: "ironically, the very feature that makes self-driving cars valuable – their flexible, autonomous decision-making across diverse situations – is exactly what makes it hard to trust them" (Danks, 2016).

This raises a serious concern about forming expert-to-expert trust as well. Arguably, the social network that constitutes trust in experts includes the review and monitoring process of an expert's reasoning by others. A common example would be the peer review process of research papers. In this sense, the decision-making process of experts remains transparent, which is one of the key reasons why laypeople trust them.

In contrast, such a regulatory role of social networks is lacking in opaque AI. Even experts working in the exact same field can only rely on indicators such as accuracy to evaluate the judgments produced by AI's reasoning. As a corollary, the decision-making process of AI remains

shrouded in mist; what we can verify would be nothing but its reliability. Such AI would, *de facto*, make judgments without belonging to the expert community.

3.2. AI and Explainability

The significance of explainability is already included in my accounts of trust. In a nutshell, explainability is crucial in expert-to-expert interactions due to its regulatory function, confirming the ownership of experts' knowledge and abilities. Taking a step further, I will argue that explainability also holds importance in *novice-to-expert* relationships.

As Goldman (2001) points out, the distinction between expertise and novicehood is not like an all-or-nothing relationship: "There are, of course, degrees of both expertise and novicehood. Some novices might not be so much less knowledgeable than some experts. Moreover, a novice might in principle be able to turn themselves into an expert, by improving his epistemic position" (p. 89). Hence, when one describes an expert as a transmitter of knowledge, it does not mean that a layperson can only be notified of the final decision. Laypeople, depending on their proximity to expertise, will comprehend the reasoning of an expert to a greater or lesser extent. In this respect, the role of a successful knowledge transmitter includes enhancing the epistemic position of these potential experts.

Indeed, an expert is often regarded as having both the ability and obligation to explain their knowledge to a layperson at a suitable level. In principle, there can be numerous ways to describe a given situation. For instance, we can explain either material implication in everyday language to students who are not familiar with symbolic logic, or in more technical

terms to students at an intermediate level by drawing truth tables for P ⊃ Q .≡. ~P ∨ Q.

Admittedly, one might argue that there are numerous experts who do not effectively transmit their knowledge and skills to the general public, despite their outstanding performance in their fields.[3] Hence, it might be argued that making accessible and varied explanations may not be necessarily required for experts but would be more suitable for critics or commentators.

My response to these objections is twofold. First, the *fact* that many experts are not proficient in explaining their knowledge does not undermine the *normative principle* that experts *ought to* explain their knowledge to a layperson at a suitable level. Rather, this intuitively plausible principle shows the fundamental difference between AI and human experts: according to the "Ought implies Can" (OIC) thesis, it implies that experts *can* explain their knowledge to a layperson at an appropriate level—which is always false when applying to opaque AI.

Second, such a fact can easily be explained given that there are three kinds of expertise: substantive, normative and adaptive. These represent different proficiencies, all of which are necessary to become an expert. In assessing English proficiency, for instance, the strengths in the four sections —speaking, listening, writing, and reading may vary among students. However, this does not mean that any of these four elements might be unnecessary for demonstrating expertise in English.

Therefore, critics or commentators, especially in the realm of intellectual

3) This objection was raised by one of the anonymous referees for the *Journal of AI Humanities*. The referee also suggests that explainability is required for—and is a virtue of—critics or commentators, rather than experts.

experts, are also regarded as experts in a broader sense. The distinction between experts and critics or commentators lies in the areas in which they are relatively more well-versed. And even if experts are superior to critics across all aspects, it is worth mentioning that the expert-to-novice relationship is a matter of *degrees*. This perspective makes critics or commentators viewed as an expert to novice and novice to experts. They, so to speak, stand at a halfway point on the expert-novice scale.

3.3. AI and Responsibility

It is challenging to determine who must take moral responsibility when AI has made a fatal mistake. This challenge is evident in cases such as misfires on civilians by autonomous weapons systems (AWS) or medical accidents caused by AI applications in health care. As Sparrow (2007) puts it, AWS can be compared to weapons of mass destruction or anti-personnel mines, in the sense that "when they are used, no one is taking responsibility for the decision about who does and does not get killed. [...] The use of these weapons [...] demonstrates a profound disrespect for the value of an individual human life" (p. 68). Should AI be given an authority of an expert, we face an analogous challenge. It is quite clear that every stakeholder would be reluctant to take responsibility for the repercussion of AI's fatal errors. Simultaneously, it would be just absurd to condemn AI and shut it down.

One viable option is to appeal to the principle of the *transitivity* of responsibility: if A has committed an act for which responsibility should be assumed, and B's actions constitute the cause of that act, then B should also bear responsibility. In this regard, it is often said that responsibility has

a distributed character (Coeckelbergh, 2020, p. 2056). I have already illustrated this with an example where responsibility is attributed not only to an individual expert but also, at times, to the entire expert community. From this perspective, one might hold that the issues concerning AI and responsibility are reducible to matters of either AI developers or AI users.

However, van de Poel & Sand (2021) point out that there are roughly two kinds of responsibility: backward-looking and forward-looking. According to their classification, the former encompasses blameworthiness, accountability, and liability, while the latter pertains to obligations and virtues (p. 4773). Arguably, both responsibilities—as obligations and virtues—are indispensable for characterizing the protective and productive aspects of expert actions and judgments, as well as the specific character traits of experts (pp. 4773-4774). It can also be taken for granted that these responsibilities are assigned neither to AI developers nor to AI users. Intellectual honesty is a classic example that shows even substantial expertise is not only relevant to facts but also becomes entangled with responsibility-as-virtue. Desmond's discussion of expert communication is another interesting example that illustrates the ethical dilemma associated with responsibility-as-obligation (Desmond, 2021; Desmond, 2024). According to Desmond, weather announcements, albeit seemingly innocuous, suddenly involve ethical dilemmas and moral responsibilities as follows:

> should an announcer mistakenly forecast sunshine, they may face annoyance, but typically not *moral* indignation. However, what if a deadly hurricane materializes instead of sunshine? This is a very different situation, and the moral dimension of weather announcements then quickly materializes. In fact, meteorologists are only too aware of this sudden *moral* responsibility, and interestingly, something of an ethics of expert communication has arisen

> spontaneously among meteorologists. Some have promoted the default strategy to err on the side of caution and emphasize the *possibility* rather than the *probability* of the worst¯case scenario, just so that the general population will make the requisite preparations. However, others have pointed out that too many false alarms can lead to forecasting communities to lose credibility and trust and desensitize the public to future weather warnings. Even though, unsurprisingly, no universal rule has been found, the lesson for us is that meteorologists must morally deliberate on how to frame weather forecasts once the stakes are sufficiently high. (Desmond, 2024, p. 39)

The point is that normative expertise is not just about conveying facts to the public in a value-neutral manner. Desmond (2024) characterizes this as a tension between actionability and scientific transparency. For instance, consider a scenario where an expert tries to warn about a virus with a 1 % fatality rate. It's not enough for the expert to simply convey this fact because the public would react totally differently to warnings about a virus that could lead to millions of deaths, even though this is just another description of the very same virus (p. 37). Expert advice, in this sense, is formed through spontaneous dialogue between experts and the public, which often leads to an ethical dilemma.

> The fundamental dilemma of expert communication, in effect, describes how the scientist must choose how much to anticipate the interests and goals of the intended audience. If there is "too little" anticipation, the scientist is not providing useful and focused expert advice. If there is "too much" anticipation, and the scientist is deciding to an excessive extent on what the audience should or should not hear, then this becomes manipulative. (p. 42)

To sum up: responsibility is not only involved in the matter of blameworthiness but directly affects the expert judgments and advice. In contrast, we cannot expect AI to be bound by such regulation.

4. AlphaGo's Expertise

In light of the discussion thus far, AI is unqualified to be considered as an expert; AI lacks normative elements such as trust, explainability, and responsibility, which exert a significant influence on the decision-making process of experts.

However, an objection could be raised against the three criteria I have presented, suggesting that they are just too stringent. AlphaGo can be a strong candidate for serving as a counterexample. First, it certainly has substantive expertise, demonstrated by its ability to play the game of Go at a level much higher than any human. Silver et al. (2017) dramatically describes AlphaGo's achievement as follows:

> 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. (p. 358)

Moreover, AlphaGo appears to possess both normative and adaptive expertise. On the one hand, it not only gains knowledge but also develops its competence through gameplay. On the other hand, many human Go players have already been learning this game from AI based on AlphaGo's

strategy.

My initial response to this objection is that AlphaGo is not a cognitive or intellectual expert despite its appearance. In a nutshell, what AlphaGo possesses is *not knowing-that but knowing-how*, akin to "violinists, billiards players, and textile designers" (Goldman, 2001, p. 91; see also Lewis, 1988, p. 288). It is a specific skill that cannot be reduced to propositional knowledge.

It should be noted here that my arguments steer clear of addressing a metaphysical issue, a question of whether AlphaGo can (or, has) propositional knowledge. Instead, I will argue that, *at least epistemically*, we can only perceive the knowledge possessed by AI as a certain kind of knowing-how insofar as it is opaque AI, as in the case of AlphaGo.

Consider, for instance, AlphaGo's normative expertise. It is widely known that AlphaGo's teaching approach is significantly different from traditional methods. Instead of explaining why a particular move is good in a given situation, AlphaGo only provides the possibility of winning for each potential move. In this respect, compared to human experts, its normative expertise is unsatisfactory on both the expert-expert and expert-novice dimensions. Of course, AlphaGo's knowledge can be *interpreted* by proficient human professional players. However, this does not mean that AlphaGo can systematically teach human novices, nor does it mean human experts can effectively interact with it. Due to a lack of social networks, it neither belongs to an expert community nor possesses trust (in contrast to mere reliability).

The conceptual distinction between plausibility and probability may shed light on the crux of the issue with AlphaGo's approach. According to Brennan-Marquez (2017), the former is about explanatory power, while the

latter is about predictive likelihood (pp. 1258-1259). More importantly, he articulates the nature of inference based on explanatory power, which cannot be reducible to mere predictive likelihood, as follows:

> All observed facts invite many possible inferences as to what brought the facts about. For Inference A to be plausible, it must provide an explanation of observed facts that meshes with an observer's understanding of the world. Moreover, whether Inference A is more plausible than Inference B (or vice versa) depends on which inference supplies the better explanation: which inference is simpler, consistent with a greater share of facts, and more compatible with "background beliefs." Inference A is relatively plausible if, in comparison to other inferences, it is worth entertaining. (Brennan-Marquez, 2017, pp. 1258-1260)

Evaluating *plausible* inference requires several qualitative standards (simplicity, applicability, and coherence) and a complicated network composed of facts and beliefs. This kind of inference is often superior to explanations based solely on predictive likelihood, when we deal with either cases that involve unlikely but tailored explanations (e.g., diagnosis of a rare disease) or likely but untailored predictions (e.g., the Court's reasoning in individual cases, which must respect the presumption of innocence) (Brennan-Marquez, 2017, pp. 1260-1267). Within this framework, we can say that what AlphaGo provides is not plausibility but probability.

And this raises further concerns about AI and responsibility. When it comes to AlphaGo, we may say that its primary task and accompanying responsibility are relatively straightforward and marginal (winning a game with specific rules). On the contrary, the tasks performed by AWS or medical AI are not only complex but also entangled in serious ethical

issues. In these areas, the primacy of plausible accounts is often stressed; accordingly, the potential risks of AI are too significant to be ignored.

5. Conclusion

This paper argues that AI cannot function as a trustee, a successful transmitter of knowledge, or a subject taking responsibility. This is because the decision-making process of experts is established within social networks, rather than solely achieved by individual experts without any regulation. Here, we can recognize the importance of *transparency*. Contrary to appearances, the expert does not neglect transparency in their decision-making process by appealing to their epistemic authority. Instead, such a process is formed through intangible factors within complex social connections and spontaneous dialogue. In this central respect, experts judgments are still brought to light. Consequently, the role of AI within expert community should be somewhat limited unless the transparency of AI has been secured.

References

Brennan-Marquez, K. (2017). Plausible Cause: Explanatory Standards in the Age of Powerful Machines. *Vanderbilt Law Review*, 70, 1249-1301.

Caley, M. J., O' Leary, R. A., Fisher, R., Low-Choy, S., Johnson, S., & Mengersen, K. (2014). What is an expert? A systems perspective on expertise. *Ecology and Evolution*, 4(3), 231-242.

Coeckelbergh, M. (2020). Artifcial Intelligence, Responsibility Attribution, and a Relational Justifcation of Explainability. *Sci Eng Ethics*, 26, 2051-2068.

Danks D. (2016). Finding Trust and Understanding in Autonomous Technologies. *The Conversation*, December 30, 2016. Accessed April 13, 2024. https://theconversation.com/finding-trust-and-understanding-in-autonomous-technologies-70245

Desmond, H. (2021). Expert Communication and the Self Defeating Codes of Scientific Ethics, *The American Journal of Bioethics*, 21:1, 24-26.

Desmond, H. (2024). The ethics of expert communication. *Bioethics*, 38, 33-43.

Goldman A. I. (2001). Which Ones Should You Trust? *Philosophy and Phenomenological Research*, 63(1), 85-110.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When will ai exceed human performance? evidence from AI experts. *Journal of Artificial Intelligence Research*, 62, 729-754.

Hardwig, J. (1985). Epistemic Dependence. *The Journal of Philosophy*, 82(7), 335-349.

Herzfeld, N. (2023). Is Your Computer Lying? AI and Deception. *SOPHIA*, 62, 665-678.

Johnson, G.M. (2021). Algorithmic bias: on the implicit biases of social technology. *Synthese*, 198, 9941-9961.

Kirkpatrick, J., Hahn, E. N., & Haufler, A. J. (2017). Trust and Human−Robot Interactions. In: P. Lin, K. Abney, and R. Jenkins (Eds). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York, NY: Oxford University Press, 142-156.

Lewis, D. K. (1988). What Experience Teaches. *Proceedings of the Russellian*

Society, 13, 29-57. Reprinted in: *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press, 1999.

Martin, T. G., Burgman, M.A., Fidler, F., Kuhnert, P. M., Low-Choy, S., McBride, M., & Mengersen, K. (2012). Eliciting Expert Knowledge in Conservation Science. *Conservation Biology*, 26(1), 29-38.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6, 175-183.

Roff, H. M. & Danks, D. (2018). Trust but Verify: The Difficulty of Trusting Autonomous Weapons Systems, *Journal of Military Ethics*, 17(1), 2-20.

Ross, A. (2022). AI and the expert; a blueprint for the ethical use of opaque AI. *AI & Soc*, 1-12. https://doi.org/10.1007/s00146-022-01564-2

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Sci Eng Ethics*, 26, 2749-2767.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550, 354-359.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, 24, 62-77.

van de Poel, I. & Sand, M. (2021). Varieties of responsibility: two problems of responsible innovation. *Synthese*, 198, 4769–4787.

Youn, S & Jin, S. V. (2021). “In A.I. we trust?” The effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging “feeling economy.” *Computers in Human Behavior*, 119, 1-13, https://doi.org/10.1016/j.chb.2021.106721

Received: 9 March. 2024, Evaluated: 20 March. 2024~12 April. 2024, Accepted: 23 April. 2024