Original paper UDC: 004.8:7.01(045)
doi: 10.21464/sp38203
Received: 8 March 2023

Hyeongjoo Kim
Chung-Ang University, Humanities Research Institute, 84 Heukseok-ro, Dongjak-gu, KR–06974 Seoul
godwithhj@cau.ac.kr

Unexplainable Explainable AI

Abstract

This paper critically investigates the explainable artificial intelligence (XAI) project. I analyze the word "explain" in XAI and the theory of explanation and identify the discrepancy between the meaning of the explanation claimed to be necessary and that which is actually presented. After summarizing the history of AI related to explainability, I argue that American philosophy in the 1900s operated in the background of said history. I then extract the meaning of explanation in view of XAI, to elucidate the relationship among AI, logic, and the theory of explanation. In so doing, I aim to reveal DARPA's surreptitious definitional retreat in terms of its contents and formal fallacy of sophisma figurae dictionis*, drawing from Kant's paralogism. I conclude that this intentional fallacy preexists the XAI project and that presumptuous use of reason, which Kant criticizes, is underlying.*

Keywords

XAI, AI, explanation, paralogism, *sophisma figurae dictionis*, Immanuel Kant

1. Introduction

The main goal of this paper is to provide a critical perspective on the XAI Project that is being carried out, for example, at the initiative of the U.S.'s Defense Advanced Research Projects Agency (DARPA).[1] In particular, I delve into the word "explain" to (i) determine the rationale for DARPA's emphasis on "explainability", (ii) identify the purpose of explaining XAI, which DARPA claims is a new AI technology, and (iii) highlight the discrepancy between the meaning of the "explanation" claimed to be necessary and that which is actually presented. To this end, an overview of the history of AI related to its explainability is provided, and the relationship between logical positivism as its underlying mechanism and pragmatism as the mainstream American philosophy is examined. In this process, the method of putting forth the theory of explanation—the main research field of the related philosophy—is investigated, and the three-tiered relationship of AI, logics, and the theory of explanation are conclusively elucidated. This process of comparison and analysis aims at revealing DARPA's surreptitious definitional retreat through a conceptual explanation of its content and, in terms of form, the fallacy of *sophisma figurae dictionis*—with a comparison drawn from an example of Kant's paralogism.

1
DARPA – "Defense Advanced Research Projects Agency" – is a U.S. military research agency created in 1958 in response to the USSR's launching of the first Sputnik satellite in 1957. Many DARPA-funded projects have non-military purposes, such as computer networking and information technology.

This work is licensed under CC BY 4.0.

2. The Philosophical Logic Behind the History of XAI

The history of AI is the 70-year history of computing, which is tantamount to the history of AI philosophy, which began based on the groundwork of philosophy of mathematics and logic. Throughout the history of AI, the topic of "explainability" has emerged intermittently in the fields of AI-related philosophy and research, triggering heated discussions before fading away once again. This chapter presents a deductive classification of AI into rule-based, machine-learning, and deep-learning systems within the background framework of symbolism and connectionism, followed by an analysis of their respective philosophical characteristics based on examples, with its scope limited to philosophical logic.

As is the case with all historical interpretations, archived fact-finding information on the inception of AI history, even if there is a general consensus about the information, may not be acceptable for all applications from all perspectives. Although the term "artificial intelligence" was coined at the Dartmouth conference in 1956, the precursor of related research dates back to the 1940s. It is a well-known fact that the attempt at designing an information-processing system in line with the human brain structure was undertaken by McCulloch and Pitts in 1943. It is around that time that the development of an artificial neural network of the connectionism lineage represented by deep learning began. Orthodoxically, AI has been considered to have developed separately into two camps of symbolism centered around imitating the functions of the human brain and connectionism constituted around the schematics of the operating principles of the human brain.[2]

However, symbolism and connectionism only serve as vague ideologies that contribute to setting the direction of the development and application of AI technology; thus, symbolism and connectionism cannot serve as the basis for classifying concrete methodologies of AI technology. Therefore, this study defines respective characteristics of the two abovementioned "-isms" as rules and learning and dichotomizes AI into rule-based and learning-based AI. After the deep-learning technique developed by Hinton in 2006 put an end to the 3rd "Ice Age" of AI, the terms "deep learning" and "AI" seem to be used synonymously. Given this situation, we obtain the following three-tiered classification system: Rule-based AI, Machine-learning AI, and Deep-Learning AI.

The characterization and categorization of these three AI types and their relational structures will be discussed later. For now, the idea of "explainability" must be considered to lay the groundwork for further discussion.

In the development history of AI, the term "explainable" first appeared in the phrase "explainable decision-making system" , which was designed based on the "rule-based conditional probability approximation" in 1975. Since then, it has mainly been used in the field of expert systems based on knowledge-based data-symbol processing systems. Later, an expert system based on probability theory was revitalized, and the relevant "explanation" has been continuously developed – mostly in connection with machine learning showcased by explanation-based learning (EBL). However, the "learning" in this context is different from the current deep learning that requires a large amount of supplied or autogenerated training data. This type of learning, rather, seeks to derive a generalized explanation from a small number of specific learning data, and pursues Explanation-Based Generalization (EBG). As examined above, AI,

which uses the possibility of an "explanation" as the technological driving force, spans between rule-based AI and machine-learning AI. Nevertheless, the situation has changed, and related discussions are underway in the connectionism AI lineage, particularly in the research field of deep learning. If the explainability in XAI originally meant to express the strengths of AI technology, it has now turned into a slogan that AI should be explainable. Simply put, explainability, which is currently lacking, has become a necessary condition toward improvement. To determine why, we need to examine the operating principles of the representative models of rule-based, machine-learning, and deep-learning AI. However, this must be preceded by a brief overview of the history of the effect of philosophy, or more specifically mathematical logic, on laying the foundation for AI.

Newell, a cognitive psychologist and the father of symbolic AI, stated that "AI researchers consider philosophy more immediately relevant to their work than they do psychology" (Nutter 1987: 284). This is a meaningful statement. In my point of view, psychology and philosophy imply connectionism and symbolism, respectively. Although Newell was a cognitive psychologist, his prioritization of philosophy over psychology had much to do with the pioneering role he played in the research on the symbolism lineage, such as the development of the typical rule-based program "Soar". That being said, how has philosophy contributed to AI concretely?

In the early and mid-twentieth century, philosophy of language, with a focus on propositional logic, was prevalent in tandem with the emergent skepticism about the metaphysical idealism that was established basically in Europe – above all in Germany. In this context, there was a strong philosophical trend called "logical positivism", initiated by Frege, who put forth anti-psychologism, and Russell, Wittgenstein, and Carnap, who were Frege's successors and the three giants of 20th-century philosophy (Glymour, Ford & Hayes 1995: 17). They distinguished between perceptually meaningful and meaningless worlds as the keynote of "meaning precedes truth", similar to Kant's distinguishing between recognizable and unrecognizable domains. Above all, they excluded ethics and metaphysics, which had been regarded as an essential domain of philosophy for over 2,000 years, from the realm of science, dismissing metaphysics as a cognitively meaningless pseudoscience and ethics as a mere emotivism, not pertaining to the realm of cognition that can discuss truth and falsehood. Thus, they argued that any object can be meaningful only when verbally expressible and verifiable (Carnap 1931).[3] A strong emphasis was also laid on the necessity of the pre-existence of the formulation rules governing the well-formed formula of the language system, and such formulation rules were constituted de facto under the name "artificial language". This idea became "the philosophical foundation that has governed the thinking of the

2
Schematizing the functions and mechanisms of the brain is rather like mimicking rigid human intelligence because it is a human understanding based on scientific instrumentalism and functionalism. For the purpose of this paper, the scope of discussion is limited to this rigid human understanding targeted in the artificial intelligence engineering, putting aside the history and background of profound human understanding made in other fields of study. Symbolism and connectionism fall under this scope of discussion.

3
In this respect, Carnap emphasized the meaninglessness of metaphysical propositions, stating: "Saying anything about something 'beautiful' or 'good' stands in fact for nothing." (Carnap 1931: 236).

founders of AI" (Lee 1993: 75), and "the influence of logical positivism was decisive in their work of creating an artificial programming language" (Lee 1993: 75). On the historical side as well, Pitts and Simon[4] were students of Carnap, the originator of logical positivism. Glymour mentioned that Carnap influenced both camps of AI through his two students (Glymour 1992: 367). Carnap also taught another student, Hempel – the most prominent proponent of the "explanation debate" to be covered intensively in this article. This theory of explanation was the underlying theory of the Dendral/Meta-Dendral programs, which are the basic programs for the expert system that dominated an epoch of AI (Glymour, Ford & Hayes 1995).

In fact, Pitts actively uses in his study the artificial language put forward by Carnap.[5] The present paper examines the characteristics of artificial language put forward by Carnap and determines how it influenced rule-based AI language using simple examples. According to Carnap (1931), the concept of *Seiende* (being) and *Nicht-Seiende* (not-being) has played a crucial role in the history of metaphysics ever since its inception in ancient Greek philosophy; however, the conclusion as below "being and not-being coexist" cannot exist in a logically valid language and are thus meaningless (Carnap 1931: 214–219, 234). Nevertheless, putting aside the quest for the meaning of this language, these expressions paradoxically coexist in a meaningful world. To scrutinize how being and not-being coexist in the world of artificial language, Parmenides' theory of being can be invoked:

1) What does not exist is not-being.
2) Not-being is nothing.
3) Nothing does not exist.
4) Therefore, all things exist.

Let being, not being, and not-being be denoted by S, N, and –S, they are then translated as follows in an artificial language system using quantifiers:

′1) $(\forall x)(\sim(x(\in S)\rightarrow(x\in\text{-}S)$
′2) $(\forall x)(\sim(x(\in\text{-}S)\rightarrow(x\in N)$
′3) $\sim(\exists x)(x\in N)$
′4) $(\forall x)(\in\text{-}S)$

This argument can be proven through a *reductio ad absurdum*. Briefly put, if it is assumed that all things do not exist by denying ′4), from which a contradictory conclusion "something that is nothing must exist and not exist concurrently" $((x\in N)\wedge\sim(x\in N))$ is drawn, then ′4) must be true.

The next argument is Aristotle's theory of not-being refuting Parmenides' argument.

1) What does not exist is not-being.
2) Not-being is nothing.

These statements can be expressed symbolically:

′1) $(\forall x)(\in S)$
′2) $(\forall x)((x(\in\text{-}S)\rightarrow(x\in S)$

This argument can be also proven through the following *reductio ad absurdum*:

′3) $(\exists x)\sim((x(\in\text{-}S)\rightarrow(x\in S)$
′4) $(\exists x)((x(\in\text{-}S)\wedge\sim(x\in S))$

´5) (a∈-S)∧~(a∈S)
´6) (a∈S)
´7) ~(a∈S)
´8) (a∈S)∧~(a∈S)
´9) ´3)→´8)
´10) ´2)[6]

As examined above, being and not-being coexist in the world of artificial language devoid of objective reference of natural language. Such contradictions can be resolved because an artificial language takes the rule fitness complying with the mathematically based rules of logic as the sole measure of inferential reasoning without regard to objects and contexts.

Rule-based AI is essentially based on this idea.[7] The Attempto Project Group of the Department of Computer Engineering at Zurich University developed the Research and Development in Advanced Communications in Europe (RACE) Program, which verifies the validity of the inference directed by natural language, as shown in the above examples. Parmenides' theory of being and Aristotle's theory of not-being were entered into this program, which yielded the following proof:

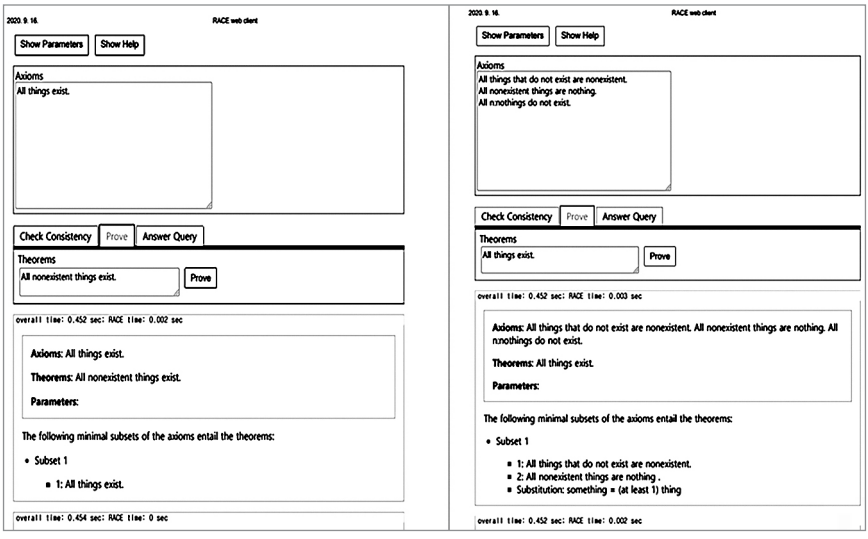

Figure 1: The programming was done with the help of Professor Heeryon Cho at Chung-Ang University, to whom I am grateful.

4
Pitts authored *A Logical Calculus of the Idea Immanent in Nervous Activity*, which is regarded as the precursor of AI, and Simon developed the General Problem Solver, which has been credited as the conceptual basis for the term "problem-solving" and is frequently mentioned in the context of the *raison d'être* of AI.

5
It may also be worth tracing the influence of Carnap on Pitts, but this is beyond the scope of this study.

6
Lee (1993: 77) explains that these two arguments were presented and interpreted by Weingartner in 1974 at a seminar at University Salzburg.

7
For example, LISP, an AI language developed by McCarty that is still in use, is based on the predicate-expression method of formal logic and the logical achievements through the use of conditional expressions (if, then) and connective/declarative expressions (and, or). In particular, the self-reference- and

What is noteworthy in this proof is that it pinpoints and explains the part of the premise to which the conclusion pertains. This program accommodates the fact that the truth value of the conclusion is inferred from the truth value of the premise in a deductive argument, i.e., the analytical characteristic that the conclusion is implied in the premise. This suggests that the mathematical quest of the human intrinsic-reasoning system has been extended to AI language through artificial language. This construct arises from the philosophical attitude of confidence toward the reasoning system inherent to human beings in that truth is preceded by semiotic logic inherent in the AI language and the logical sense underlying the semiotic logic. There is no reason to request an explanation for a rule-based AI program because explainability is the intrinsic principle of this program.

By contrast, in the field of machine learning, research on the learning aspect has mainly been carried out in conjunction with inductive logic. As examined previously, the concept of an "explainable decision-making system" arose from the "rule-based conditional probability approximation" and developed into research on the decision-making process. Although it is admittedly an important research topic to determine whether the "rule-based conditional probability approximation" is a concrete method, this study is focused on the words "rule-based" and "probability". These terms are closely associated with deductive and inductive logic, respectively. In view of this, the explainability will likely emerge during the transition from deductive logic to inductive logic, i.e., from the intersection of rule- and probability-based machine learning. The conditional probability expressed by the formulation $P(a|b)=P(a\wedge b)/P(b)$ simply refers to the probability of event *a*'s occurring when event *b* has already occurred under the assumption that events *a* and *b* influence each other. In other words, the formula expresses the probability of event *a*'s occurring only if event *b* has occurred or will occur.

A flagship machine-learning algorithm that takes account of the advantages and drawbacks of this idea is the "Naive Bayesian Classifier", which is primarily used for spam e-mail classification. The well-known Bayesian theorem is expressed by $P(a|b) = (P(b|a)*P(a))/P(b)$. Simply put, this means that the probability of an event's occurring is proportional to the probability before it is given. Here again, the meaning and method of calculation differ depending on whether events *a* and *b* affect each other. However, it is impossible to clearly determine whether events *a* and *b* affect each other and to estimate the initial value for *a*. Regardless, the formula cannot be established without entering an initial value and judgment over the inter-event relationship, which leaves us no choice but to rely on a subjective decision. This is why Bayesian theory is classified as subjectivism and criticized by positivistic scientism, which regards the objectivity of the natural world as a barometer of theoretical validity. By contrast, the first reason for the naive Bayesian classifier's being naive is that each event is considered an independent event initially as a naive strategy to avert the criticism of subjectivism. This hinders an accurate understanding of the situation, but it has the advantage of becoming increasingly accurate with an increase in the number of antecedents and related data, although the formula becomes more complicated. However, it cannot be completely free from the criticism of subjectivism because data selection does not have an objective basis but is still in the hands of the user, nor can the user determine the logical reason for the dataset design. What can be explained is merely the problem-solving ability, i.e., the performance, of the classifier's

use of the dataset selected by the user. This is the second reason why the Bayesian classifier is naive.

This leads us back to the key aspect of this article, i.e., the meaning of explanation. Suppose that the formula for deriving the result to be predicted using the naive Bayesian classifier is complex and that there are 100 conditional words for a refined classification of spam e-mail. In addition, suppose that certain e-mails have been classified as spam. If the conditions are this complex, it is difficult to understand intuitively the process by which the result was obtained. If the number of data is extremely large or the method of obtaining such data is unclear, it becomes more difficult to explain the process, although not impossible. No matter how complex the formula or how large the number of data is, the process becomes explainable in due course if the required efforts are made. Apart from this, however, given that the result was obtained through a probabilistic process, if the "explainability" here means the correspondence of the resulting value with the actual value, the naive Bayesian classifier cannot be assured of full explainability. This can be considered tantamount to the limitations of inductive logic with respect to deductive logic, given that the truth value of the result depends entirely on the truth value of the premise (data) in deductive, unlike inductive, logic. The fact that the data setting depends on the user's own selection and computational convenience is also outside of the scope of explainability in the strictest sense of the word. In view of this, machine learning AI is considered to have a lower explainability than rule-based AI in the sense that machine-learning AI has lower conformity of the resulting value with the actual value. On the other hand, if the domain of the explanation request is limited to the mechanical process to the exclusion of the designer's intention and the user's interpretation around the naive Bayesian classifier, the explainability is not impaired.

To summarize the discussion thus far, explainability is not mentioned in rule-based AI but in machine-learning AI owing to the difference in the degree of explainability. Both are explainable in that the process of drawing conclusions is traceable. Nevertheless, the concept of explainability has emerged in the field of early-phase machine learning for two reasons. First, when the meaning of an explanation is strictly defined and the target of its application is expanded, the unexplained parts surface. For example, the naive Bayesian classifier regards each event as separate, not for any logical reason but for the efficiency of obtaining the resulting value. An explanatory gap still exists if the entire sequence of processes must be explained, based on the motives of applying AI to the principle of the operation. Second, even if the part to be explained is narrowed down to the part after the assumption, i.e., excluding the reason for the assumption from the explanation, and only the operation process within the AI model is to be explained, the process is relatively complex and takes significant effort to explain.[8]

recursion-problem solutions proposed by Russell and Tarski are reflected in LISP's advantage of free implementation of meta-language.

8

Regarding the limitations and hopes of the explainability of the connectionism, Clark said, as early as 1990, that: "The methodology of connectionist explanation is perfectly geared to the avoidance of ad-hoc organizing principles and sentential, linguistic bias. There remain important and unresolved questions as connectionism may provide. But […] techniques are already being developed and will no doubt become well-understood." (Clark 1990: 304). In 2020, three decades later, the problem of explainability would sink deeper into the mire. From this, it can be inferred that the dilemma of AI's explainability began around the time that AI of the connectionism

→

3. Explanation, Problem-Solving, and Explainable Deep Learning

As mentioned above, the concept of explanation has rarely been mentioned in relation to rule-based AI, but it was used as a self-compliment in the prime of machine-learning AI. However, in the era of deep learning, explainability is required, sought, or ethically imposed. Given that an imposition arises from the absence of an attribute, DARPA's XAI Program emphasizes the inexplicability of deep-learning AI. Nevertheless, DARPA raises the claim about explainability as a type of justifiable request while seeking to achieve technical feasibility. That being said, how should we understand the word "explanation"?

Admittedly, "(the term) explanation is highly ambiguous" (Thagard 1993: 44), and this ambiguity has made an explanation a crucial topic of the 20th-century scientific philosophy. G. H. Von Wright divides the tradition of scientific explanation into the Aristotelian tradition and the Galilean tradition (Wright 1971: 2). These traditions are referred to as (finalistic-) teleological and (mechanistic-) causal explanation, respectively. On the other hand, scientific explanation gradually considered teleological explanation based on the power inherent in the cause or the intention of the actor as unscientific, and either excluded it from the discussion or tried to reduce it to a causal explanation. It is here that the concept of causation stands out as the main concept to explain "explanation". In the sense that the cause already has the power to produce the effect, teleological explanation implies causal necessity. On the other hand, in the philosophy of science, based on the Humean tradition of taking a skeptical stance on the causal necessity of the objective world, the theories that consider the regularity as the core of causality were developed. It can be said in the broad sense to follow the Galilean tradition. C. Hempel's deductive-nomological model is a representative example of this. The model is based on a structure consisting of two major parts (*explanans* and *explanandum*). In Hempel's words, "[e]xplanans falls into two subclasses; one of these contains certain sentences, which state specific antecedent conditions; the other is a set of sentences, which represent general laws" (Hempel & Oppenheim 1948: 247). From these two subclasses, the *explanandum* is deduced. Hempel schematized this structure as follows:

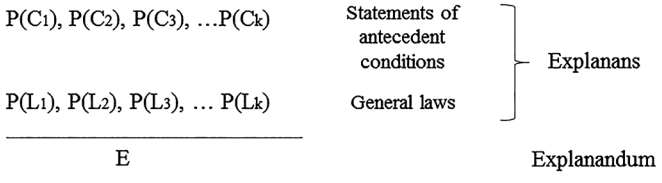

Figure 2

This is a deductive-nomological (D-N) explanation in the sense that C_1, C_2, C_3 ... C_k are each captured by their corresponding laws, L_1, L_2, L_3 ... L_k, respectively, to form logical clusters, and the *explanandum* can be logically inferred from the *explanans*. According to Hempel, an essential feature of explanations is their factivity (Páez 2019: 445). Later, realizing that the limitation (Salmon 1971: 29) of the D-N model is that deductive inference is only possible if specific conditions are satisfied, Hempel designed an inductive-statistical (I-S) model. The I-S model can be understood as an expanded D-N mode in that it shares the latter's predictive relevance, which is why the D-N

model can be considered an explanatory model, and underlying principle of these two models is that particular facts are subsumed under the laws (Lee 2010: 21–22).

Against the strong reliance on the sufficient conditions implied in Hempel's D-N model, Salmon presented the following counterexample:

> "John Jones avoided becoming pregnant during the past year, for he has taken his wife's birth control pills regularly, and every man who regularly takes birth control pills avoids pregnancy (Salmon 1971)." (Salmon 1971: 34)

Although John Jones' regular use of birth control pills is the sufficient condition for not getting pregnant, his failure to get pregnant was not due to his regular use of such pills (Salmon 1970: 61). When we apply this to Hempel's D-N model, a law can be considered a sufficient condition for an event to occur, but it does not mean that the occurrence of that event (*explanandum*) was caused by that law (*explanans*). To put it succinctly, Salmon raises the problem of explanatory relevance between general law and relevant event as the explanatory gap of the D-N model. As an alternative, Salmon proposes a Statistic-Relevance (S-R) model that relies on conditional probability. Let the general public be denoted by A, the use of birth control pills by B, the avoidance of pregnancy by C, and men by D. Men's becoming pregnant or not has no relevance regarding the use of birth control pills. If this fact is overlooked, we can conclude that the probability of getting pregnant is higher when birth control pills are not taken than when they are.[9] However, when the relevance of a man's using birth control pills is considered, it should be borne in mind that the probability of getting pregnant by a person who takes birth control pills is not the same when this person is a man.[10] Thus, Salmon's explanatory model is based on a probabilistic model.

Fraassen attributes the limitations of these theories to the "lack of contextual factors" and presents a pragmatic theory of explanation as an alternative. According to Fraassen, an explanation is neither a proposition or argument nor an enumeration of propositions, but answers to why-questions about "some topics in formal pragmatics (which deals with context-dependence) and in the logic of the questions" (Fraassen 1980: 134). The explanatory power of these questions depends on the topics of the questions and the relevance relation between the question and its context. The core characteristic of this relevance relation is that it tends to be extremely specific, is based on individual desires and interests, and is dependent upon the circumstances (Fraassen 1980: 156).

This theory is criticized by essentialists, such as Hempel and Salmon, who argue that pragmatic explanations are only "pseudo-explanation" and "explanation in appearance". In Hempel's words, "[i]t is neither necessary nor sufficient for the scientific adequacy of an explanation that it should reduce the explanandum to ideas with which we are already familiar" (Hempel 1965: 433). This theory has also built its own domain and is still developing. However, if its target area of explanation is inquiry in social sciences, such as our living world, it leaves room for criticism of Hempelians, despite the high plausibility and utility of Fraassen's theory, given that the ideology of

lineage was improving its problem-solving capacity.

9
This statement can be expressed by the formula $P(B/A\&\neg C) < P(B/A\&C)$.

10
This statement can be expressed by the formula $P(B/A\&C) \neq P(B/A\&C\&D)$.

science aiming at an objective description of natural phenomena underlies explanation. Because the goal of this article is not to understand the contextual understanding of the impact of AI on our lives but to reveal the explanatory potential of the product of science and technology called "AI", the author agrees with this criticism within the scope of this article.

The discussion thus far has verified the following facts: As the D-N model develops into a pragmatic explanatory model, the conditions for consideration of such explanation become sophisticated, but the explanatory power weakens. In particular, if an explanation is focused on the principles of AI technology, a pragmatic explanation does not fit the definition of an "explanation" in the strictest sense of the word.

An important interim conclusion must now be drawn. The pattern of change in the theory of the explanatory is similar to that of the previously examined development of AI. We stated that rule-based AI based on deductive argumentation has developed into learning-based AI based on statistics and inductive argumentation while maintaining its own characteristics. This resembles the pattern of change in the explanatory model just examined. Therefore, we can understand the principle of rule-based AI from the perspective of the D-N model, and the principle of machine-learning AI from the perspective of the S-R model; the method of explanation presented by the RACE Program, taken as an example of a typical rule-based program, involved embedding the conclusion in the premise. In other words, in verifying the validity of a specific argument, it is important to catch which part of the conclusion—presented as the basis for the argument—is embedded in which premise, as suggested by the "explanation" of the RACE Program.[11] Here, the conclusion embedded in the premise is, in principle, identical to the law's seizure of the condition described in the D-N model. This principle, which governs both the RACE Program and D-N model, is the principle of deduction. It was also noted that the naive Bayesian classifier is a typical machine-learning AI. It is a well-known fact that Salmon's theory was inspired by Bayesianism. Following in the footsteps of Reichenbach, who rated himself as "the greatest empiricist of the 20th century" (Salmon 1977: 3), Salmon seeks to design the preconditions based on neutral knowledge, excluding all elements of subjectivism, which is the main criticism of Bayesianism, and is hence classified as an objective Bayesian (Galavotti 2022). The design of the S-R model can also be understood along these lines. One of the reasons for the Naive Bayesian Classification's being "naive" is also its pursuit of conditional objectivity. As examined previously, a conditional probability is calculated by reconfiguring two independent events into two related events. The fact that the naive Bayesian classification presupposes that two events are independent means that it operates on the probability calculation from a neutral standpoint. In this sense, the naive Bayesian classifier is based on objective Bayesianism. Briefly put, the commonality between naive Bayesian classification and the S-R model is objective probabilism. Conclusively, rule-based AI and statistics-based learning AI can be considered XAI in that they seek to reveal the (causal) relationship between events, whether a strict or loose definition of explanation is applied.

Against this background, we will examine the explanatory potential of deep learning. Figure 3 shows a schematic of the explanatory differentiation of XAI as presented by DARPA.

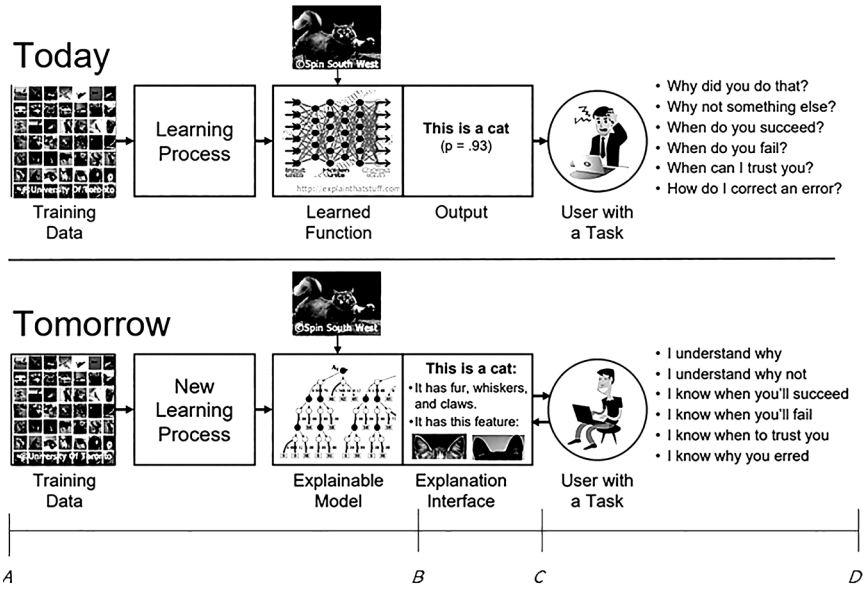

Figure 3. Image taken from:
David Gunning, "DARPA/I2O Program Update", November 2017, p. 5.

To work out the difference in XAI with respect to the current deep-learning technology represented by this schematic, it is necessary to understand how it works, at least at an elementary level. In Figure 4, an example is given to explain the principle underlying the learned function in Figure 3.

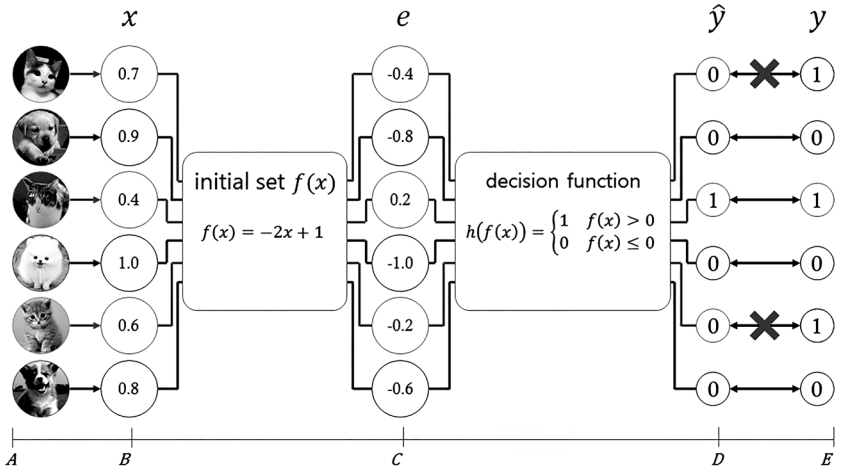

Figure 4: From the lecture materials of
Jaesung Lee of Chung-Ang University.[12]

11
The condition of an explanation in a stricter sense can be met by determining the principle underlying the explanation in natural language on the interface. This process will be described in a follow-up study because it is beyond the scope and purpose of this article. However, given the impossibility of the explanation in this interface to deviate from clear logical rules in principle, it can be assumed that the algorithmic operation behind it will also be transparent.

12
I would like to express my special thanks to Prof. Lee. The explanation in the paragraph below for this figure is from (Kim 2022: 140).

This cat-recognition process model is widely used as an example. To explain this briefly, the algorithm numerically quantifies the input images and substitutes them into multi-layered functions until the final output is expressed as zero or 1 to check the match or mismatch. Although omitted in the image of Figure 4, the error range in the D–E section is reflected in the B–C section, and this iterative process lasts until the error range reaches the minimum value. The Backpropagation algorithm governs the entire process. Because it constantly changes the mapping function f(x), it is understood herein as a meta-function. In Figure 4, x denotes the initial value, e denotes the output value of f(x), $\hat{y}$ denotes the output value of the decision function (0 or 1), and y denotes the true value of the target image. Because it is beyond the purpose of this article, instead of providing a more detailed explanation regarding deep learning, three key points are mentioned.

First, the value of y is given with the image prior to the learning process. The cat image is given with the number 1, as in the first picture, which means that the image matches the cat, or it is given with the number 0, as in the second picture, which means that the image does not match the cat. Second, we need to understand why the value of x is given – specifically, why the numerical value of the first pattern was set to 0.7. Does this mean that the match rate is 70% when the pattern is decomposed into pixels and checked against a specific prototype? Like this question, we try to determine how to associate 0.7 with the first pattern; however, this effort is pointless, for the answer to this question is simply that the highest match rate was obtained when the pattern was set to a mere 0.7. Based on error backpropagation, the value of each pattern is adjusted. Third, there is an extremely large number of hidden layers in the A–B section of the deep-learning algorithm that are actually used. How can XAI overcome these three hurdles?

Considering the theoretical understanding of "explanation" in light of the AI-related context of its use, the explanatory task of deep learning can be clearly described as follows: 1) the internal mathematical structure in which a numerical value is assigned to each pattern; 2) the traceability of the changing trend of the mapping function and the presentation of a mathematical analysis model; and 3) the explanation of the causality in the relationship between the input pattern and each hidden layer. Assuming that care should always be taken to understand the concepts used by many research groups on the same topic as being geared toward consistency, even though the identical meaning cannot always be maintained, the goal of XAI in the past 70 years of AI research history should be the dismantlement of the black box autogenerated by deep learning. However, the direction of explanation shown in the flowchart in Figure 4 proves the contrary. As can be easily seen, DARPA's XAI does not even attempt to discern the structures of the deep layers of deep learning; rather, it keeps this structure intact and attaches explanatory labels to the input data. Specifically, the information passing through the units in each layer is forced to pass them in a form recognizable by humans, and to try to output the feline characteristics after the termination of the learning session, as shown in Figure 3, with the values that have passed the units. It has a completely different goal from providing an explanation regarding the mechanism by which data are segmented at a certain layer and combined with other data. The black box remains the same, but the data passing through it are combined to represent the black box's judgment to the user facing the interface. This is similar to identifying endangered animals living in the wild and determining

years later how they had grown in size. We can identify an animal by the tag we attached to it and see how its physical state has changed, but we have no knowledge regarding where it ate, what it did, and what activities are responsible for its present shape. Thus, while the A–B section in Figure 3 needs to be explained, an explanation is attempted in the B–D section.[13] The addition of explanatory functions to deep learning through the process steps, such as adding explanatory labels, visualization, and writing explanatory notes, can be compared to covering a black box with a patch of black cloth.

However, from DARPA's statement that DARPA is considering a trade-off between problem-solving capacity and explainability or even mentioning an internal conflict relationship, it can be inferred that they are aware of this situation to a certain extent (Gunning *et al*. 2019). To cite DARPA, among learning-based AI technologies, the "decision tree" technology has the highest explainability and the lowest problem-solving ability, while deep learning has the best performance but the lowest explainability. This is consistent with the position that this paper has held. Apart from these facts, there are reasons for the significantly degraded credibility of this project when viewed in the framework of the explanatory theory discussed above: DARPA uses the word "explanation" in multiple senses for various reasons. First, the DARPA's explanation relies on the designer's domain-dependent knowledge, and that explanation is requested by many different groups. This in turn means that the explanation of XAI is ultimately evaluated by the user's subjective judgment, presumably depending on the clarity and usefulness of the explanations presented in the C–D section. As examined above, the explanation of XAI does not follow any predefined guidelines of scope, degree, direction, or object. DARPA even notes, referring to the internally planned and implemented XAI research, that "[r]eliable and consistent measurement of the effects of explanations is still an open research question" (Gunning *et al*. 2019). The Hempelians can criticize the "explanation" of XAI in the same context as the criticism they made of the pragmatic theory of explanation. If knowledge of designers and data producers regarding the algorithm constitutes the background theory of the explanation, and the wish of AI users determines the success of the explanation, the purpose of epistemic explanation is once again dissolved in the sea of context. The concept of "explainability" was deliberately released during the history of AI development, and a similar concept, also called "explainability", appeared as well. What is behind this phenomenon?

As has been explained, the philosophy of logical positivism set the stage for the birth of AI. On a related note, in examining the landscape of American philosophy around 1950, it becomes it clear that pragmatism was also a large, mainstream philosophy during the same period. In the 1920s, when Dewey consolidated his position as the leader of pragmatism, logical positivism

13
However, in his article "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)", Páez argues the opposite, namely that "explanation" as terminology in the field of the philosophy of science cannot be applied to the discussion about the artificial intelligence, meaning that the XAI project should instead aim toward pragmatic explanation rather than attempting to explain causal factivity (Páez, 2019). Bringing Páez's claim into the context of this article, the explaining area of an XAI should be section C–D in Figure 3. However, I argue that since artificial intelligence is essentially a machine based on scientific technology, a more rigorous concept of scientific explanation, or at least a perspective based thereon, should be applied to AI research.

appeared on the stage of American philosophy. The exchange between these two camps became increasingly active, and a philosophical common denominator was established. Despite that the decisive difference between the two owes to pragmatist advocacy of abduction as a third form of reasoning, pragmatism and logical positivism have a common denominator as scientifically oriented philosophies. At a more concrete level, they have in common the belief that the true source of knowledge is experience, having inherited this position from the British empiricists, as well as the tenet that "philosophy is a method rather than theory" (Nekrašas 2001: 41). It is also noteworthy that by the time AI was born in the 1950s, the merging of the two philosophical trends was underway.

"Problem-solving" is never a mission when it comes to defining AI (Rich 1987: 10).[14] *Artificial Intelligence: A Modern Approach*, which is the standard AI textbook, defines AI as a "problem-solving agent" (Russell & Norvig 2010: 64). In other words, AI is a tool that solves problems that require intelligence. Pragmatists saw "problems" in events confronted by humans and regarded the process of solving them as a living process. Therefore, "problem-solving" itself was the philosophy of pragmatism. This philosophy significantly contributed to defining the roles of AI (Kieras 1987: 113–115; Holyoak 1987: 115–118). A flagship example is Dewey's discourse of problem-solving, which emphasizes the concept of inquiry as implemented in the General Problem-Solving program. It is also noteworthy that abduction, which was developed by Peirce, the founder of pragmatism, as a problem-solving approach, is gaining traction in AI prediction programs. Pointing out that our lives have many aspects that cannot be covered by deductive and inductive reasoning, Peirce argued that formal arguments such as deductive reasoning have a firm theoretical legitimacy but a weak impact in the process of solving our questions and forming beliefs. Thus, he developed abduction, which takes into account the importance of setting up hypotheses and having a good pre-understanding of the world as the third category of reasoning. The syllogistic form he suggested is as follows (Peirce 1958: 1958):

A surprising fact, C, is observed.
However, if hypothesis H is true, C will be a matter of course.
Hence, there is reason to suspect that hypothesis H is true.

From the viewpoint of deductive reasoning, the above reasoning is committing the fallacy of affirming the consequent. However, Peirce, who was of course aware of this, argued that this reflects the actual problem-solving cognitive process in real life. From the stance of deductive reasoning, it is an error to infer that it rained because the land was wet, but this is an acceptable inference in our world of experience. The problem-solving mind that can give a suitable answer to questions at hand, instead of seeking validity in light of the deductive principle, is in fact a long tradition in the history of science. For example, Newton used calculus and assumed gravity to explain with plausibility the motion of planets in the solar system. Newton's purpose was to give a phenomenal account of an accurate calculation of the motions of the planets of the solar system, but not to determine the cause of gravity. He urges readers of the *Philosophiæ Naturalis Principia Mathematica* not to inquire after the cause of gravity. Substituting this reasoning process into the above syllogistic form yields the following:

The movements of the planets of the solar system were observed as in A.
If there is gravity, a movement like A can be explained well.

Therefore, there is gravity.

The discovery of germanium by Mendeleev, the creator of the periodic table, is similar to the logic of abductive reasoning, i.e., logic of the hypothesis. According to the composition of the periodic table, there must be an element with similar properties and an atomic mass of approximately 70 next to silicon, which has an atomic mass of 28. This assumption cannot be proven by deduction. However, the presence of such an atom adds to the completeness of the periodic table. Eventually, this puzzle of the periodic table was solved by the discovery of germanium, which has an atomic mass of 72.63. Peirce compares the ability to make a hypothesis to the ability to find the minor premise in a syllogistic task. Inferring the minor premise based on the premise and conclusion is similar to the deep-learning methodology of optimizing a function by giving input and output values. Peirce called this ability an inexplicable "miracle" which he saw as a kind of instinct of reason.

However, not all positions in science are geared toward solving problems. Semantic realism, which is a type of scientific outlook, supports the coherence theory of truth, while semantic anti-realism replaces the realistic concepts of truth, such as a guaranteed argument and limit of inquiry, with epistemological concepts. Semantic realism is divided into scientific realism, which believes that all scientific statements have a truth value, and scientific instrumentalism, which regards science only as a tool of scientific inquiry, deferring the allocation of truth values to scientific laws and theories. Scientific realism also includes methodological realism, which regards truth as an important purpose of scientific inquiry, and methodological non-realism, which replaces truth with methodological substitutes, such as a successful prediction, empirical relevance, and problem-solving ability.[15] AI research, whose main focus is on problem-solving rather than a theoretical quest for truth, may have generally evolved on the basis of methodological non-realism and scientific instrumentalism, albeit to varying extents. The quest for truth decreases with increasing importance attached to the drawing of practical results for scientific phenomena. It is natural that, as more importance becomes attached to problem-solving, explanatory power diminishes proportionally. Requiring only problem-solving is, strictly speaking, tantamount to disregarding the need for explainability. As mentioned previously, DARPA seems to be aware of this dilemma.

4. The Presumptuousness of Linguistical Reason in the Light of Kant's Paralogisms

As examined in the foregoing chapters, the current "explanation" of explainable AI (XAI) is similar in appearance to the "explanation" of explainable AI that was prevalent 30 years ago, albeit the meaning has changed. As confirmed while examining the theory of explanation, the latter tried to retain

14
"Artificial intelligence programs are designed to solve problems."

15
"Scientific realists in turn include methodological realists who take truth (usually together with information or systematic power) to be an important aim of scientific inquiry and methodological non-realists who replace truth as an aim of science by some methodological surrogate (e.g., successful prediction, empirical adequacy, problem-solving ability)." (Niiniluoto 1986: 258)

the essence of explanation in terms of scientific causal explanation, whereas the former – while attempting to assume the same appearance as the latter – does not. In the world of science, using notions and terms to serve one's own purpose has been common throughout history. In fact, Kant considered in his *Critique of Pure Reason* that such an arbitrary use of concepts causes serious confusion in the area of metaphysics; to be more precise, this stems from the already existing confusion of metaphysics. The chapter "Paralogisms" in Kant's *CPR* is particularly focused on critiquing the arbitrary use of concepts. In this chapter, while analyzing the first syllogism in the first (A) edition of the *CPR*, I will examine the essential aspect of the nature of human reason that causes metaphysical transcendence. Through this analogy, I intend to reveal the duality of the concept of "explanation" of XAI and how the human desire underlying this duality causes linguistic presumptuousness.

The chapter "Paralogisms" is the most modified chapter in the *CPR*. Therefore, although it would be beneficial to clarify the structure and content of both the first and second editions of "Paralogisms" and the reason why Kant wrote this chapter anew, such a discussion will be omitted as it would exceed the scope of the undertaking in this paper. However, despite the superficial difference in content, Kant's intention in both editions of that chapter was to criticize the metaphysics of his time – specifically rational psychology – based on the common factor that "*I think* is […] the sole text of rational psychology" (A343/B401).[16] In view of this, I will take and analyze the first example of paralogisms in the first edition.

This ([AP]) is:

[APO] That the representation of which is the absolute subject of our judgments, and hence cannot be used as the determination of another thing, is **substance**.

[APU] I, as a thinking being, am the absolute subject of all my possible judgments, and this representation of Myself cannot be used as the predicate of any other thing.

[APS] Thus, I as a thinking being (soul), am **substance.**

Let us first simplify this argument. According to Ameriks, all paralogisms in the first and second editions have the following basic form:

"Whatever is X, is Y.
I am X.

Therefore, I am Y."[17]

This can be reconstructed as the following:

For all X: if X is M, then X is Y
The I is M

Also: the I is Y.

Let us keep this basic form in mind and return to [AP]. While the basic form above appears to be very simple, [AP] seems relatively complicated. In fact, [APO] and [APU] each consist of two clauses. We analyse [APO] as follows, by interpreting the expression "and hence" as a signal word of an explanation:

[APO_1] That representation of which is the absolute subject of our judgments is substance.
[APO_2] This absolute Subject cannot be used as the determination of another thing.

In the same way, [APU] can be analysed as follows:

[APU_1] I, as a thinking being, am the absolute subject of all my possible judgments.

[APU_2] This representation of Myself cannot be used as the predicate of any other thing.

According to Kant, the term "subject" is defined in such a way that "it is distinguished from mere predicates and determinations of things" and "cannot be used as the determination of another thing" (A 349). From this it can be concluded that [APO_2] and [APU_2] do not add any new information at this point but merely present a confirmatory explanation. Therefore, we will bracket it to simplify the syllogism. Then, as a first step, [AP] is simplified as follows:

[APO_1] That representation of which is the absolute subject of our judgments is substance.
[APU_1] I, as a thinking being, am the absolute subject of all my possible judgments.

[APS] Thus, I as a thinking being (soul), am **substance.**

The major premise of a syllogism is understood as "a general rule" (AA. IX: 120), and the minor as "subsumption of the condition under this rule" (Höffe 2011: 228; cf. AA. IX: 120). In fact, Kant writes in [APO1], "our judgments", but in [APU1], "my […] judgments". In this sense, the major premise [APO] is expressed with a universal quantified proposition.

For all X: if X is an absolute subject, then X is a substance.
The "I" is an absolute Subject.

Also: The "I" is a substance.

This is equivalent to the following argument [AP]**:

[APO_1]* All absolute subjects are substance.
[APU_1]* The "I" is an absolute subject.

[APS]* Therefore: The "I" is a substance.

Now the discussion has reached an important point. If the concept of "subject" is understood here merely in the logical sense, the conclusion "the I is a substance" must be understood in such a way that the "I" may here be regarded as a substance solely in the sense that the "I" is a logical "vehicle" (B399) that "accompanies all my representations" (B131). However, if the concept of the subject is understood as a real being to which an intuitive concept is applied, this must be criticized from a Kantian perspective. It may be recalled that to understand the "the I" as anything more than a logical vehicle that the subject must employ to make sense of the subject's representations is to assume a metaphysical position of the sort that Kant intended to criticize; Kant understands the "I" in the A- "paralogisms" merely as a logical function. It is noteworthy that this also applies to the B "paralogisms". The parallel passage in the B "paralogisms" is as follows: "that the I that I think can always be considered as subject, and as something that does not depend on thinking merely as a predicate, must be valid – this is an apodictic and even

16
All page numbers refer to the pagination of the Academy Edition.

17
"1. Whatever is X (i.e., 'cannot be employed as determination' or 'can never be regarded as the concurrence of several things,' or 'is conscious of the numerical identity of itself at different times,' or 'can only be inferred as a cause') is Y (i.e., substance, or simple, or person, or 'in merely doubtful relation' to us). 2. I ('as a thinking being' or 'the soul', or, in the fourth paralogism, 'outer appearance') am X. 3. Therefore, I am Y (substance, simple, person; in the fourth paralogism the conclusion of is that 'outer appearance is merely doubtful,' but this can be transposed into a claim about us, that we are in a 'merely possible' epistemic relation to what is outside)." (Ameriks 1998: 374).

an identical proposition" (B407). Kant says clearly that the statement that the "I" "can be considered" as a subject is apodictically true.

As mentioned above, the "I" is identified with the pure logical subject in the major premise; the "I" as a "constant logical subject" (A350) accompanies all of my representations. In this sense, this "I" can be described as the "vehicle of all concepts whatever" (A341/B399) and as "the standing and lasting I" (A123). This sentence, i.e., the major premise [APO1]*, interpreted in this way, summarizes Kant's position. According to rational psychology, however, the "I" of subordinate premise [APU1]*, in contrast to that of [APO1]*, is seen as an empirical, i.e., "real", subject, for it is to be viewed from the perspective of a categorical, determinable intuition. From this, it seems that an "allegedly new insight" (A350) can be derived, namely that the "I" represents "a standing and abiding perception" (A350). This is where the error of rational psychology begins. Firstly, this "I" is not identical with that of the major premise; secondly, the "I" understood in this way is not to be understood as a subject but as an object; and thirdly, it even deviates from the Kantian view of the objectified "I", because the "I" as an object is by no means categorically determinable and intuitive but rather indeterminable.

Kant explicitly warns against subsuming this empirically understood "I" of [APU1]* under the transcendental concept of substance, which leads to the conclusion's [APS]* being misunderstood. In other words, if the rational psychological proposition [APU1]* is subsumed under the Kantian proposition [APO1]*, this leads to the illusional inference that the "I" is an empirical substance. Kant's decisive critique thus refers to the fact that [APS]* is understood empirically. More concretely, Kant's focus is on the concept of "empirical substance". In this, there is a contradiction, because the term "substance" already contains the transcendentality as its essential property, which is indeed the basis of experience (cf. A182/B224–A189/B232) (of "empirical knowledge"; B218), but itself is nothing empirical. If we nevertheless insist on this notion of "empirical substance", then it points to "something chimerical" (A315/B371), namely something that is at once empirical and transcendental. The condition of this "empirical substance" is not the transcendental one for the possibility of experience but rather the real one, showing "which is always-perceptually-present" (Bennett 1974: 76). If the "I" is in turn associated with this empirical substance, the concept of the "enduring" (A349) arises; in other words, the "I" of "everlasting duration" (A351). In this way, rational psychology can finally assert "immortality" (A345/B403) with [APS]*.

Based on the discussion presented above, the argument that a rational psychologist is likely to make can be constructed as follows:

If a being is a substance, then it is immortal.
The thinking "I" is a substance.

Therefore, the thinking "I" is immortal.

From a Kantian point of view, there is a significant gap of thought between the minor premise and the conclusion. For Kant, the thinking "I" is substance in the sense that all of its thoughts are inherent in it (cf. A80/B106; Bennett 1974: 77); it is something that underlies all ideas, that is, substance. Therefore, it is a purely logical and epistemological, but by no means ontological, concept. Likewise, from a Kantian perspective, rational psychology does not make a clear distinction between "transcendental" and "empirical". Accordingly, it mixes these two predicates, which cannot coexist simultaneously, in the

concept of the "I". In other words, the absence of the term "transcendental" causes the "transcendental illusion" (A297/B353) in rational psychology, i.e., the assumption of an immortality of the human soul. Kant sharply criticizes the fact that rational psychologists have not noticed that deceptive difference between the transcendental and the empirical "I" and that they therefore believe the syllogism [AP] to be true, from which they finally derive the "immortality" (A345/B403) of the soul believed to be allowed to derive.

5. Conclusion

Kant, who consistently insisted on the modest use of reason through self-criticism, pinpointed and criticized that rational psychologists were unable to control metaphysical desires and threatened the right course of science by allowing the speculative world to penetrate the world of experience. DARPA's desire to add explainability to a top-notch problem-solving capacity that disregards the understanding of the process seems to have inherited the desire of the metaphysicist to add materiality to the concept of self as a premise for experience. With a comparison of Kant's paralogism, DARPA's surreptitious definitional retreat is revealed through a conceptual explanation of its of content and, in terms of form, the fallacy of *sophisma figurae dictionis*.

In every decisive phase of AI development, the corresponding AI technology has been accompanied by criticism. Searle responded to the attempt to attach semantic value to rule-based AI based on a syntactic design using the Chinese Room argument. Dreyfus criticized the overheated expectations on AI, which revived with glory an expert system mounted with knowledge-acquisition technology, by putting forward the everyday-language definitions of expert concepts permeated with practical wisdom (phronesis). My criticism of the XAI project is aimed at its attempt to pragmatically disassemble the essence of explanation, behind which attempt is lurking the presumptuous use of language by lazy reasoning. As mentioned, Kant had already given this warning.

Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A 6A 3A 01078538).

Bibliography

Ameriks, K. (1998): "The paralogisms of pure reason in the first edition", in: Immanuel Kant, *Kritik der reinen Vernunft*, G. Mohr, M. Willaschek (eds.), Akademie Verlag, Berlin, pp. 371–389.

Bennett, J. (1974): *Kant's Dialectic*, Cambridge University Press, Cambridge.

Carnap, R. (1931): "Überwindung der Metaphysik durch logische Analyse der Sprache", *Erkenntnis* 2 (1931), pp. 219–241.

Clark, A. (1990): "Connectionism, Competence, and Explanation", in: M. A. Boden (ed.), *The Philosophy of Artificial Intelligence*, pp. 281–308, Oxford University Press, Oxford.

Fraassen, B. v. (1980): *The Scientific Image*, Oxford University Press, New York.

Galavotti, M. C. (2022): "Wesley Salmon". *Stanford Encyclopedia of Philosophy* (2022). Available at: https://plato.stanford.edu/entries/wesley-salmon/ (accessed on 15 December 2023).

Glymour, C.; Ford, K. M.; Hayes, P. J. (1995): "The prehistory of android epistemology", in: K. M. Ford, C. Glymour & P. J. Hayes (eds.), *Android Epistemology*, Cambridge, MIT Press, pp. 3–21.

Glymour, C. (1992): *Android epistemology. Computation, artificial intelligence, and the philosophy of science*, Cambridge, The MIT Press.

Gunning, D. *et al*. (2019): "XAI – Explainable artificial intelligence", *Science Robotics* 4 (2019) 37, pp. 1–2, doi: https://doi.org/10.1126/scirobotics.aay7120.

Hempel, C.; Oppenheim, P. (1948): "Studies in the logic of explanation", *Philosophy of Science* 15 (1948) 2, pp. 135–175.

Hempel, C. (1965): *Aspects of scientific explanation*, The Free Press, New York.

Höffe, O. (2011): *Kants Kritik der reinen Vernunft. Die Grundlegung der modernen Philosophie*, C. H. Beck, München.

Holyoak, D. (1987): "Cognitive Psychology", in: S. C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, vol. 1, pp. 115–120, New York *et al*., John Wiley & Sons.

James, W. (1978): *Pragmatism and the Meaning of Truth*, Harvard University Press, Cambridge (MA) – London.

Kant, I. (1998): *Critique of Pure Reason*, transl. P. Guyer – A. Wood, Cambridge University Press, New York.

Kim, H. J. (2022): "Tracing the Origins of Artificial Intelligence", in: H. J. Kim, D. Schönecker (eds.), *Kant and Artificial Intelligence*, De Gruyter, Berlin – Boston, pp. 129–143.

Lee, C. S. (1993): *Philosophy of artificial intelligence*, Korea University Press, Seoul.

Lee, J. (2010): *Explanation and Its Place in Metaphysical and Scientific Inquiries*, Dissertation, Indiana University Bloomington.

Nekrašas, E. (2001): "Pragmatism and positivism", *Problemos* 59 (2001) 2, pp. 41–52, doi: https://doi.org/10.15388/Problemos.2001.59.6830.

Niiniluoto, I. (1986): "Theories, approximations, and idealizations", *Studies in Logic and the Foundations of Mathematics* 114 (1986), pp. 255–289, doi: https://doi.org/10.1016/S0049-237X(09)70696-2.

Nutter, J. (1987): "Epistemology", in: S. C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, vol. 1, pp. 280–287, New York *et al*., John Wiley & Sons.

Russell, S. J.; Norvig, P. (2010): *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River *et al*.

Páez, A. (2019): "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)", *Minds and Machines* 29 (2019), pp. 441–459.

Peirce, C. (1958): *The collected papers of Charles Sanders Peirce reproducing Vols. I-VI* ed. Charles Hartshorne and Paul Weiss, Vols. VII–VIII ed. Arthur W. Burks, Cambridge, Harvard University Press.

Rich, E. (1987): "Artificial Intelligence", in: S. C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, vol. 1, pp. 9–16, New York *et al*., John Wiley & Sons.

Salmon, W. (1977): *Hans Reichenbach. Logical empiricist*, Springer. Heidelberg.

Salmon, W. (1971): "Statistical explanation", in: W. C. Salmon, R. C. Jeffrey, J. G. Greeno, *Statistical explanation and statistical relevance*, University of Pittsburgh Press, Pittsburgh, pp. 29–88.

Kieras, D. (1987): "Cognitive Modelling", in: S. C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, vol. 1, pp. 111–115, New York *et al*., John Wiley & Sons.

Thagard, P. (1993): *Computational Philosophy of Science*, MIT Press, Cambridge (US).

Wright, G. H. v. (1971): *Explanation and Understanding*, Routledge & Kegan Paul, London.

Hyeongjoo Kim

Neobjašnjiv objašnjiv AI

Sažetak

Ovaj rad kritički istražuje projekt objašnjive umjetne inteligencije (XAI). Analiziram riječ »objasniti« u XAI-ju i teoriji objašnjenja i identificiram neslaganje između značenja za »objašnjenje« za koje se tvrdi da je potrebno i onoga što je stvarno predočeno. Nakon sažetka povijesti umjetne inteligencije vezane uz objašnjivost, tvrdim da je američka filozofija 1900-ih djelovala u pozadini navedene povijesti. Zatim izdvajam značenje objašnjenja s obzirom na XAI, da bih razjasnio odnos među umjetnom inteligencijom, logikom i teorijom objašnjenja. Čineći to, nastojim otkriti DARPA-ino prikriveno definicijsko povlačenje u smislu njegovog sadržaja i formalne pogreške sophisma figurae dictionis*, polazeći od Kantova paralogizma. Zaključujem da ova namjerna pogreška postoji prije projekta XAI-a i da mu je u podlozi preuzetna uporaba uma koju Kant kritizira.*

Ključne riječi

XAI, AI, objašnjenje, paralogizam, *sophisma figurae dictionis*, Immanuel Kant

Hyeongjoo Kim

Unerklärbare erklärbare KI

Zusammenfassung

In dieser Arbeit wird das Projekt der erklärbaren künstlichen Intelligenz (XAI) kritisch untersucht. Das Wort „erklären" in der XAI und Erklärungstheorie wird analysiert, und die Unstimmigkeit zwischen der Bedeutung des Wortes „erklären", die angeblich notwendig ist, und dessen, was wirklich vorgestellt wird, wird identifiziert. Nach einer kurzgefassten Geschichte der künstlichen Intelligenz mit Bezug auf die Erklärbarkeit, behaupte ich, dass die amerikanische Philosophie der 1900er Jahren im Hintergrund der besagten Geschichte wirkte. Danach ziehe ich die Bedeutung der Erklärung im Hinblick auf die XAI heraus, um das Verhältnis zwischen der künstlichen Intelligenz, Logik und Erklärungstheorie zu erläutern. Auf diese Weise versuche ich DARPAs heimlichen Definitionsrückzug im Sinne ihres Inhaltes und des formalen Fehlschlusses sophisma figurae dictionis*, von Kants Paralogismus ausgehend, zu entdecken. Ich schlussfolgere daraus, dass dieser absichtliche Fehlschluss vor dem XAI-Projekt existiert und dass ihm der von Kant kritisierte vermessene Gebrauch der Vernunft zugrunde liegt.*

Schlüsselwörter

XAI, KI, Erklärung, Paralogismus, *sophisma figurae dictionis*, Immanuel Kant

Hyeongjoo Kim

L'IA explicable inexplicable

Résumé

Le présent document examine de manière critique le projet d'intelligence artificielle explicative (XAI). Le mot « expliquer » est analysé dans le cadre du XAI et de la théorie de l'explication en déterminant l'écart entre la signification de l'explication prétendument nécessaire et celle qui est réellement présentée. Après avoir résumé l'histoire de l'IA liée à l'explicabilité, j'affirme que la philosophie américaine des années 1900 a opéré en arrière-plan de ladite histoire. J'en dégage ensuite la signification de l'explication dans le contexte du XAI afin d'élucider la relation entre l'IA, la logique et la théorie de l'explication. Ce faisant, j'entends révéler le retrait définitionnel dissimulé opéré par DARPA en termes de contenu et l'erreur formelle du sophisma figurae dictionis*, tirée du paralogisme de Kant. Je conclus que cette erreur intentionnelle est préexistante au projet XAI et que l'utilisation présomptueuse de la raison, critiquée par Kant, y est sous-jacente.*

Mots-clés

XAI, IA, explication, paralogisme, *sophisma figurae dictionis*, Emmanuel Kant