Hyeongjoo Kim

4 Tracing the Origins of Artificial Intelligence: A Kantian Response to McCarthy’s Call for Philosophical Help

Abstract: Computer scientist John McCarthy has been tremendously influential in our understanding of what Artificial Intelligence really is. I shall argue that, from a Kantian point of view, the underlying theoretical framework of McCarthy’s position – which I summarize as the claim that AI as a technical entity is an imitation of the computational ability of human intelligence for problem solving in the empirical physical world – can be understood as transcendental realism. McCarthy dispels the distinction between phenomena and things-in-themselves by fundamentally blocking the reflective ability of intelligence, that is, the ability to intuit oneself as a synthetic act. The reflective ability of intelligence, i.e. the self-consciousness contained in the “cogito, ergo sum”, is the barometer, as it were, for distinguishing transcendental idealism from transcendental realism; at the same time, it is the watershed that separates Kant from McCarthy. In a philosophical paper, McCarthy called for the help of philosophers to define AI; Kant, I shall argue, can offer such help.

1 Introduction

The term “artificial intelligence” (AI) has become a global buzzword. AI chatbots on smartphones speak to us, today’s children are called digital natives, and AI companion robots serve at senior centers. However, what exactly is AI? Is it a robot, a program, or a kind of intelligence? While AI is a commonly used term, people seem to use it in reference to different concepts. In other words, the meanings encapsulated by this term are diverse.

In the 1780s, Kant took issue with the ambiguity of the concept of “metaphysics.” At that time, the term “metaphysics” was a hot buzzword among philosophers as well; just as AI, metaphysics had become a “battlefield of endless controversies” (A VIII).¹ Kant introduced the synthetic-analytic distinction

¹ For citations, see References.

Hyeongjoo Kim, Chung-Ang University

Open Access. © 2022 Hyeongjoo Kim, published by De Gruyter. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. https://doi.org/10.1515/9783110706611-004
(of subject-predicament judgments) as a methodological means for concept clarification. The emphasis on this distinction is still valid in the quest for the meaning of artificial intelligence. A synthetic proposition’s predicate concept is not contained in its subject concept and is therefore knowable only through experience, whereas an analytic proposition’s predicate concept is contained in its subject concept and thus is knowable through concept analysis (A6/B10 ff). I will expand this synthetic-analytic distinction to include a distinction between concepts referring to empirical objects and concepts relating to speculative and philosophical ideas; I will then apply the latter distinction to the semantic layers of the concept of AI. While this article aims to trace the original meaning of AI, the scope of discussion will be limited to the works by McCarthy, who boldly coined this oxymoronic term as early as 1956. More specifically, the paper will focus on the definition and explanation of “artificial intelligence” as discussed in his philosophical writings (McCarthy 1987, 1995, 2007).

On the assumption that “philosophers have not really come to an agreement [concerning the definition of intelligence: HJ. K.] in 2500 years” (McCarthy 2007, p. 5), McCarthy considered it hardly possible to develop a “solid definition of artificial intelligence” (McCarthy 2007, p. 2). Nonetheless, he discussed the question of “Why Artificial Intelligence Needs Philosophy” (McCarthy & Hayes 1969, p. 2) from an epistemological perspective, defined artificial intelligence, and attempted a philosophical explanation of it. This attempt set a milestone for the foundation of future AI research, as expressed in his call for support: “Philosopher! Help!” (McCarthy 1995, p. 5)²

This article is my response to McCarthy’s call from a Kantian viewpoint. In doing so, I pursue two goals: one is related to the definition of artificial intelligence and the other to the epistemological background of AI. First, I will provide an analysis of McCarthy’s definition of artificial intelligence. To give a brief preview of the results of this analysis, there are, I submit, three interrelated semantic layers in McCarthy’s concept of AI: engineering AI, philosophical AI, and literary AI. Engineering AI is the object of direct reference as a technical entity; philosophical AI is a conceptual premise of engineering AI; and literary AI is the ideology of engineering AI development. In other words, I define literary AI as AI geared toward the goal of AI engineering that is outside the realm of current technology, such as strong AI or super AI, which is mainly

² Kant is not, of course, what McCarthy has in mind here. In this context, he mentions that he received the help of linguistic philosophers such as Searle, Austin, and Grice. From his mention of Carnap, Quine, and Putnam, it can also be assumed that what he had in mind is the philosophy of logical positivism, which became the philosophical basis for the development of artificial intelligence in the early and mid-20th century.
discussed in the post-humanist world based on the fundamental skepticism about the distinction between humans and machines. Based on this categorization, what McCarthy ultimately wanted to assert in his definition of AI is that *artificial intelligence* (as a technical entity) is a model for the human computational ability for problem-solving in the empirical physical world. Second, by analyzing its epistemological background, I argue that it corresponds to Kantian transcendental realism. Furthermore, I will demonstrate that both a Kantian and McCarthy’s understanding of intelligence have a common denominator: they both define intelligence as problem-solving computational ability in the world of empirical realism. However, Kant demands another necessary condition for intelligence: the self-consciousness represented by the “I think.” While this demand functions as the core evidence of the Kantian worldview of transcendental idealism, it was neglected by McCarthy.

### 2 McCarthy’s Concept of Intelligence

McCarthy’s definition of artificial intelligence is densely formulated in his article “What Is Artificial Intelligence?” (McCarthy 2007). It can be summarized as follows:

1. Artificial intelligence is the science and engineering of intelligent machines, especially intelligent computer programs. 2. Artificial Intelligence is related to the similar task of using computers to understand human intelligence. However, the concept of *intelligence* is not completely independent from the concept of human intelligence yet. 3. It is certain that intelligence is the computational part of the ability to achieve goals in the world. 4. Varying kinds and degrees of intelligence occur in people, animals, and machines.³

Now, the AI referred to in [1] is a technical entity as an *a posteriori* designation of an empirical embodiment existing in reality (be it a robot with a body or a program), not a term analytically derived from the concepts of *artificial* and *intelligence*. [2], [3], and [4], however, deal with the original meaning of intelligence, the traditional subject of epistemology. In this regard, McCarthy uses the term AI, though unintentionally, in two senses: AI as a technical entity and AI as a type of ‘intelligence’. On this premise, while [2] points out the dependence of the concept of *intelligence* on *human intelligence*, [4] talks about the ubiquity of the concept of intelligence. Two seemingly contradictory (yet compatible) arguments

---

coexist in these two statements.\(^4\) [Claim 1] is derived from [4], and [Claim 2] is derived from [2]:

[Claim 1]: *Intelligence* is a concept that encompasses *AI* and *human intelligence*.
[Claim 2]: *Artificial intelligence* is dependent on *human intelligence*.\(^5\)

In [Claim 1], intelligence is a necessary condition both of AI and of human intelligence. Furthermore, AI has an intelligence like human intelligence; accordingly, it can be regarded as *artificial* intelligence, that is, a kind of intelligence. AI and human intelligence have one thing in common: intelligence. They are distinct from each other only by what (or who) possesses this intelligence. In this sense, the term *animal intelligence* is also possible. In other words, intelligence does not originate from the human being and therefore is not a human-specific element. In fact, McCarthy emphasizes the expression “our notion of general intelligence” (McCarthy & Hayes 1969, p. 3) as an ability to perceive “commonsense knowledge” (McCarthy 1987, p. 1030) of the world and uses it as a key concept to guide the discussion on the definition of intelligence.\(^6\) According to McCarthy, “an entity is intelligent if it has the intellectual world of mathematics, an understanding of its own goals, or other mental processes [including the human mind: HJ. K.]” (McCarthy & Hayes 1969, p. 4). Furthermore, he

---

\(^4\) However, the coexistence of two claims described contrastively in the above summary does not mean that they are incompatible. This aspect will become clear in the discussion that follows.

\(^5\) In this regard, McCarthy (1995) says, “Human level artificial intelligence requires equipping a computer program with a philosophy, particularly epistemological attitudes” (p. 1). And I point out that the basic idea for the distinction between this two claims was drawn from Kim (2016).

\(^6\) This is the conceptual basis of artificial general intelligence (AGI) currently under discussion. However, it is not what McCarthy intends to express through this concept. AGI, which is considered the next-generation AI or a term synonymous with strong artificial intelligence (SAI) as the ultimate goal of AI development, means the ability of an intelligent agent to perform any intellectual task that a human being can. AGI belongs to the category of synthetic-empirical concepts according to the classification I previously made, i.e., empirical reality. In his article “Generality in Artificial Intelligence,” McCarthy still regards the purpose of AI as reasoning and problem-solving ability (McCarthy 1987, p. 1032). Here, “generality” refers to the possibility that AI abilities (problem-solving and computation) can have human abilities and his emphasis on its imperativeness. This is ultimately linked to the implementation of one of the rule-based AI programs, “General Problem Solver” (McCarthy 1987). In the course of these discussions, McCarthy does not clearly distinguish between the synthetic and analytical concepts of AI. To develop my argument, I pay attention to the expression “our notion” in the phrase “our notion of general intelligence.” McCarthy himself initially used the concept of generality in the sense of *human-like* rather than in the sense of transcending that of humans and machines. Therefore, Claims 1 and 2 are connected, which will be addressed later.
claims that “the physical world exists and already contains some intelligent machines called people” (McCarthy & Hayes 1969, p. 5). Considering his comments, [Claim 1] can be rephrased as follows:

[Claim 1*]: The intelligence of AI is general intelligence.

But what is the artificiality of AI? The dictionary meaning of “artificial” is: “made or produced to copy something natural” (https://dictionary.cambridge.org/). “Something natural”, however, does not refer to anything other than intelligence itself. AI as an object of philosophical analysis, different from the synthetic sense of AI referring to a discipline of science or a concrete entity such as robots as defined in [1], is imitative intelligence. More explicitly, AI as part of any machine or program is artificialized intelligence of (human) intelligence, that is, an intelligence modeled after (human) intelligence.7 Re-examining [Claim 2], AI should be considered as an intelligence in the sense that AI shares some of the faculties of human intelligence. Thus, at first glance there seems to be a difference between [Claim 1*] and [Claim 2]. This difference can be expressed by the thesis of the independence of intelligence in general from human intelligence; it is a matter of determining whether endowing AI with the status of intelligence is admitting that (general) intelligence encompasses AI and human intelligence, logically preceding them, or that it is similar to human intelligence.

On a related note, McCarthy’s statement mentions the first premise of general intelligence: “The physical world exists and already contains some intelligent machines called people” (McCarthy & Hayes 1969, p. 6). This premise serves as the rationale supporting the principle behind this interpretation: humans are intelligent machines, these machines already exist in the world, and homo sapiens is only one of them. This is consistent with [4], “Varying kinds and degrees of intelligence occur in people, many animals, and some machines.” Against this background, combining [Claim 1] and [Claim 2] allows for the understanding that both machines and humans have intelligence; however, since humans have had a higher form of intelligence so far, machines are modeled after human intelligence.

Now, let us turn to [Claim 3]: “It is certain that intelligence is the computational part of the ability to achieve goals in the world.” This statement makes it clear that McCarthy limits the scope of the definition of intelligence to computational power. Regarding the purpose of using this power, he says:

7 The Turing Test, the first experiment related to the feasibility question of artificial intelligence which has served as a direct model for McCarthy’s concept of AI, is based on the concept of intelligence as imitation (McCarthy, 2007).
Work on artificial intelligence, especially general intelligence, will be improved by a clearer idea of what intelligence is (…). We have to say that a machine is intelligent if it solves certain classes of problems requiring intelligence in humans.

(McCarthy & Hayes 1969, p. 4)

According to McCarthy, the purpose of AI (as a technological entity) is problem-solving, which requires human intelligence. That is, problem-solving ability is what endows AI with a type of intelligence – machine intelligence – with a status similar to that of animals and humans. Thus, McCarthy’s statement [3*], “Intelligence is an ability to solve problems through computation,” helps us to understand the link between the seemingly contradictory [Claim 1*] and [Claim 2] (again, [Claim 1*]: The intelligence of AI is general intelligence; [Claim 2]: AI is dependent on human intelligence. With [3*] as an underlying statement, the two claims can be formally synthesized as follows:

[Claim 3]: All intelligence (general intelligence) is a computational ability to solve problems. Therefore, AI and human intelligence have computational abilities.

However, [Claim 3] does not explain general intelligence and dependence on human intelligence. Therefore, we must revisit the three interrelated layers mentioned in the introduction: engineering AI, philosophical AI, and literary AI. I have already pointed out that the first two concepts are implied in [1] and [2, 3, 4], respectively. From a scientific viewpoint, literary AI is a regulative idea, to borrow the term from Kant.

If the “general intelligence” implied in [Claim 3] directly refers to such literary AI, how should we understand it? If intelligence means general intelligence that precedes any concrete form of intelligence possessed by a specific entity, general intelligence logically precedes AI or human intelligence, and, accordingly, what bestows the status of intelligence on AI would not be human intelligence but intelligence itself. The attributes and limits of general intelligence constitute an uncharted terrain for us. Against this background, strong AI, super AI, or any intelligence that goes beyond such limits and attributes is free from the restraints of discussion. However, although this argument can serve the purpose of AI engineers, including McCarthy, as an ideology for AI development as an engineering entity, it cannot be its premise. As stated in [3], McCarthy clearly limits the scope of intelligence to computational power. Even if the general ability he advocates is intelligence as a “general representation of the

---

8 McCarthy’s argument that the purpose of AI is problem-solving is still valid. For example, Artificial Intelligence: A Modern Approach, which is the standard of AI textbooks, defines AI as a “problem-solving agent” (2010, p. 64).
world” (McCarthy & Hayes 1969, p. 1), that is, intelligence with common sense and human-like flexibility (McCarthy & Hayes 1969, p. 3), it merely means upgraded reasoning and computational ability, far from strong AI with intuition and self-consciousness. To reiterate, the essence of AI as established by McCarthy is the ability to solve problems in a “specific situation” (McCarthy & Hayes, 1969, p. 4) in the external physical world based on logical reasoning. As such, we can concretize his claim as follows:

[Claim 3*]: AI as a technical entity is an imitation of the computational ability of human intelligence for problem solving in the empirical physical world.

3 Kant’s Concept of Intelligence and McCarthy’s Transcendental Realism

As McCarthy states, philosophical discussions about intelligence have been around for 2,500 years. It is true that human interest in intelligence dates to ancient Greece. The Greeks persistently inquired into what can be known (noetos; intelligible) by the Geist (nous) and its cognition or function (noesis) (see Plato 2008, p. 235 f.). Parmenides was the first philosopher who put the Geist over physical or sensible abilities, and this line of thought led to Plato, who subdivided the concept of Geist and extended it into discussions about the relationship between the Idea and its cognition, as well as the activities and divisions of Geist’s abilities. This conception of intelligence was represented by the concept of “intellectus or intelligentia” (Copleston, 1993, p. 191) which is the etymological background of the concept of intelligence we are discussing now.

9 In this sense, Russell and Norvig (2010) attempted to organize the definition of existing AI (with four frameworks: thinking humanly, thinking rationally, acting humanly, and acting rationally) through a combination of two separate categories: ‘ability to act and think’, and ‘similarity with humans and rationality’. Thinking humanly relates to understanding and developing programs in relation to human cognitive abilities, specifically the relationship between sensory organs, the ability to think, and knowledge. Thinking rationally (with the subtitle “Approach based on the ‘law of thought’”) is based on human (formal) logical thinking. The definition of AI based on acting humanly is no different from the keynote of the Turing Test. Acting rationally is an extended concept that includes thinking rationally and aims to solve a given problem under specific circumstances based on the ability to reason (pp. 2–5).

10 This is related to Kant’s transcendental realism, which will be addressed later. Transcendental realism is the theory that space-time exists independently of our sensitivities; therefore, representations that exist in this space-time are considered things-in-themselves.
Against this linguistic tradition, Kant limits intelligence to the cognitive faculty of cognizing empirical objects along with the faculty of cognizing the self. That is, Kant boldly discards the role of human intelligence as the faculty that intuits God, and simultaneously asserts an agnosticism of metaphysical objects beyond the empirical world. Kant describes his worldview as *transcendental idealism*, and the worldview he criticizes as *transcendental realism*. This leads us to the following passage:

[K][K1] I understand by the transcendental idealism of all appearances the doctrine that they are all together to be regarded as mere representations and not as things in themselves, and accordingly that space and time are only sensible forms of our intuition, but not determinations given for themselves or conditions of objects as things in themselves.

[K2] To this idealism is opposed transcendental realism, which regards space and time as something given in themselves (independent of our sensibility). The transcendental realist therefore represents outer appearances (if their reality is conceded) as things in themselves, which would exist independently of us and our sensibility. (. . .) [K3] The transcendental idealist, on the contrary, can be an empirical realist. (. . .) [K4] He can concede the existence of matter without going beyond mere self-consciousness and assuming something more than the certainty of the representation in me, hence the cogito, ergo sum. (A369)

As can be seen from [K1], transcendental idealism clearly distinguishes between the thing-in-itself and the phenomenon and limits the area that is cognized by human intelligence to the world of phenomena. The cause of this distinction and limitation is the consciousness-immanence of space and time. According to it, space-time is not an entity that exists independently of human intelligence but is a form of human intelligence that enables cognition. Hence, according to Kant’s transcendental idealism, “we have to do only with our representations” (A190/B235) that we self-create, intuited through space-time.

From the Kantian viewpoint, McCarthy’s statement [3]12 matches the transcendental realism designated by Kant as an object of criticism in [K2]. The epistemological premise of [Claim 3] is as follows:

---

11 In this linguistic context, Kant distinguishes two types: archetype intelligence (*intellectus archetypes*) and ectype intelligence (*intellectus ectypus*). Archetype intelligence indirectly expresses divine intelligence. Divine intelligence implies intuition as part of it. In other words, it means intelligence without distinction between intelligence (*Verstand*) and intuition (*Anschauung*), that is, intuitive intelligence (*intuitiver Verstand*). For God, cognition is knowledge. “God cognizes the object as it is” (AA XXVIII, 606). Therefore, divine intuition is a logical premise about the existence of the thing-in-itself. On the other hand, human cognitive ability, which is ectype intelligence for divine intelligence, is limited.

12 [3] reads: “It is certain that ‘Intelligence’ is the computational part of the ability to achieve goals in the world.”
The physical world already containing intelligent machines called people exists. Information about this world is obtainable through the senses and is expressible internally. Our common-sense view of the world is almost right, and that is our scientific view. The right way to think about the general problems of metaphysics and epistemology is not to attempt to clear one’s own mind of all knowledge and start with “Cogito, ergo sum” and build up from there. (McCarthy & Hayes, 1969, p. 6)

According to [M1] and [M2], the understanding of the world in which problems to be solved are assigned to AI, that is, the world underlying the situation requiring a specific output value, posits a scientific worldview (in [M3]) in which the existence of an external physical object is taken for granted as common sense (in [M1]). [M2] justifies direct knowledge of the physical world, that is, the world external to consciousness, by claiming that the knowledge of the physical world obtained through our sense organs can become the immanent knowledge of the cognitive subject. This can be reformulated as:

\[ M^* \] The existence of a physical world that contains humans is real. Therefore, as Kant says, the physical world exists outside human cognition. However, empirical knowledge of the physical world obtained through sense organs is immanent. This is our common scientific sense. Therefore, this is correct.

Applying \[ M^* \] to \[ K \] yields transcendental realism, as described in \[ K2 \]:

\[ K2 \] This idealism is opposed to transcendental realism, which regards space and time as something given in themselves (independent of our sensibility). The transcendental realist therefore represents outer appearances (if their reality is conceded) as things in themselves, which would exist independently of us and our sensibility.

Acknowledging that time and space are things-in-themselves acknowledges the reality of physical objects that occupy time and space, regardless of the perceiver’s abilities. More specifically, transcendental realism affirms the existence of things-in-themselves, as opposed to transcendental idealism. In \[ M^* \], the physical world, which exists outside the perceiver and whose existence is justified beyond doubt, has the same contextual meaning as that of things-in-themselves \([K2]\). This confirms transcendental realism as the theoretical framework for \[ Claim 3^* \], “AI as a technical entity is an imitation of the computational ability of human intelligence for problem solving in the empirical physical world.” Thus, when viewed from the Kantian perspective, McCarthy is a transcendental realist \([K2]\).

On the other hand, from the statement in \([K3]\) that a transcendental idealist can be an empirical realist and thus concede the existence of external things, Kant argues that transcendental idealism and empirical realism are compatible (as seen in A369 above), even though the point to emphasize through the two concepts is different. While
it can be inferred that there is a common denominator in these seemingly contradictory positions. As suggested by the term itself, transcendental idealism means that anything outside the realm of human intelligence is transcendental, that is, only ideal. Against this background, the realm of things-in-themselves and the realm of phenomena are divided, with the field of the activity of human intelligence limited to the latter. However, from the viewpoint of the phenomenal world, only empirically perceived external things are real. Transcendental idealism and empirical realism differ in that the former defines the realm of human cognition in a negative way, that is, by *reductio ad absurdum*, and the latter in a

transcendental idealism focuses on the separation between thing-in-itself and phenomenon due to the consciousness-immanence of the space-time form, in empirical realism, the perception of external things derives the reality of things in space directly from reality, and based on this, the correspondence between representations and external objects, that is, the objective validity of representation is discussed. This has resulted in a variety of interpretations and controversies. At the time of republishing the *Critique of Pure Reason*, Jacobi assessed that Kant’s explanation of materialism had failed and Kantian idealism, or transcendental idealism, was likely to be a non-idealism (*Nicht-Idealismus*), since it took on an object corresponding to representation, in *David Hume über den Glauben oder Idealismus und Realismus* (1787). “It is still controversial whether empirical realism connected to transcendental idealism can be justified as a kind of realism” (see Heidemann 2015, p. 1894, translated by the author). Putnam regards Kant’s empirical realism as intrinsic realism, while Allison protects Kant’s idealism from the overemphasis of the realist stance (see Heidemann 2015, p. 1894). It is another philosophical issue to establish a philosophical stance on transcendental idealism. Transcendental idealism can also be interpreted as “idealism, phenomenology, and realism” (see Edmundts 2015, p. 1109). For example, Strawson regards transcendental idealism as a phenomenological idealism (Strawson 1966, p. 246), whereas Allison argues that there is a clear difference between the two concepts (see Allison 2004, p. 41). Whether transcendental idealism can be regarded as a phenomenology or phenomenological idealism was already triggered by Garbe and Feder when Kant was alive has been a debate between “two worlds or two perspectives,” as introduced and attended by many researchers, including Van Cleve.

Despite this conceptual difference and many controversies arising out of it, this paper focuses on the statement that “a transcendental idealist can be an empirical realist” since “the certainty of representation in me, that is, ‘Cogito, ergo sum,’ the certainty of self-consciousness” (A369) can be the basis for both theories; and it considers that “transcendental idealism is a realism only possible when one pays attention to the subjective conditions of our cognition” (Edmundts 2015, p. 1108). The meaning of empirical realism is nothing more than the theory that time and space and our representations in them are meaningful only within the limits of our experience (Kim 2016, p. 25). It is well identified that the primary commonality between the two must be the consciousness-immanence of space and time, if we temporarily reserve the emphasis itself based on Kant’s plan for the difference in expression, the interpretation by expression, criticism of skepticism, and the objective validity of cognition. In addition, reserving various secondary interpretations that can be derived from transcendental idealism, I will just focus on this point to proceed with the discussion by contrasting transcendental idealism (empirical realism) and transcendental realism centered on self-consciousness.
positive way. By virtue of these rules, Kant presents the task of finding an answer to the question of how limited human intelligence works as the ultimate task of his epistemology. In other words, the goal of his transcendental philosophy is to justify how synthetic a priori propositions are possible (B73). Kant responds to this task with the following statements:

*In this way* synthetic a priori judgements are possible, if we relate the formal conditions of a priori intuition, the synthesis of imagination and the necessary unity of the synthesis in a transcendental apperception, to a possible empirical knowledge in general.

\[A158/B197\]

[T]he unity of apperception in relation to the synthesis of the imagination (Einbildungskraft) is the understanding (. . .) In the understanding there are therefore pure a priori cognitions that contain the necessary unity of pure synthesis of the imagination in regard to all possible appearances.

\[A119\]

The imagination is the ability to derive a schema that is “intellectual on the one hand and sensible on the other” (A138/B177), and the unity of apperception is “the formal unity of consciousness in the synthesis of the manifold of the representations” (A105). In summary, for Kant, the necessary conditions for legitimate knowledge are the synthesis of receptive sensibility and spontaneous intellect, and the awareness that the subject of this synthesis and the destination of the resultant representation is the self. What is important from our standpoint, i.e., the perspective of empirical realism, is that this process of experience must relate to possible empirical cognition in general. What matters here is that this possible empirical cognition in general signifies “possible appearances” for Kant. Bearing this in mind, let us take a brief look at the working principle of AI.

First, I will briefly describe the core contents of deep learning technology, which has gained significant interest among the currently available AI techniques. Although deep learning is technically different from rule-based AI techniques such as list processing (LISP), which is an AI language designed by McCarthy, they share the same goal, that is, problem solving. Deep learning is an AI technique that processes inputted information based on the vast amount of data already provided to produce a desired result as accurately as possible. For example, if we draw the number “3” in handwriting, the deep learning technique will print a clearly recognizable number “3” based on the numerous data already obtained. Handwriting “3” is compared with other similar handwriting “3” data already entered; then, common features are extracted as a group of related things. With this process of repeated feature extraction, handwriting “3” is recognized as a definite character, that is, digitized character “3”. Handwriting “3” is stored as data to be used for the next input information. Language and face recognition, which are a
tagging function of Facebook, work by a similar mechanism. When a picture of a human face is entered into a computer and deep learning is activated, a series of processes lead to the final judgment that “this is a human face.”

Let us consider the widely used cat-recognition process model as an example.

In this model,\(^\text{14}\) the algorithm quantifies the input images and substitutes them into multilayered functions until the final output is expressed as 0 or 1 to determine a match or mismatch. Although omitted in the figure above, the error range in the D-E section is reflected in section B–C, and this iterative process lasts until the error range reaches the minimum value. The back-propagation algorithm is the algorithm governing the entire process. This is understood herein as a meta-function because it continuously changes the mapping function. In the figure above, \(x\) 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. In summary, computers with deep learning can conceptualize information. Computers can take handwriting “3” as a digitalized number “3” and analyze data with patterns in the pictures of human faces to define the concept of “human face”.

To return to Kant, he explains that the key to human intelligence is simply the conceptualization of sensible information. The object of the external becomes the object of the internal, and in the process, synthesis by affinity (Affinität) occurs. Then, the synthesized diversity is unified into one representation, and when it is defined by concepts, it is stated in a form of judgment. If we see a set of red dots in the space of a circle, these manifold (mannigfaltige) dots are

\(^{14}\) The lecture materials of Prof. Jaesung Lee of Chung-Ang University. Special thanks to him.
synthesized by the principle of affinity in the circle; then, by the understanding of the synthesized representation, we derive the judgment that “this representation is an apple.” When other similar judgments are given, by reason, we can make abstract judgments such as “An apple is a fruit.” Given a picture of “3,” the principle of AI, which conceptualizes it as a digitalized number “3” and derives new judgments through reasoning based on the learned (determined) data, is similar to the principle of human intelligence just discussed. This similarity can be expressed in McCarthy’s terms, “problem-solving ability using intelligence,” or in Kant’s terms, “faculty of combination of understanding.” If the world of objects of “problem solving” (McCarthy) and “empirical cognition” (Kant) is limited to the phenomenal world, McCarthy can also be classified as an empirical realist in the sense that the experience of the existence of an object can be obtained by artificial intelligence through its synthesis ability.

As we have pointed out, however, McCarthy is also a transcendental realist, and the core of Kant’s criticism of transcendental realism is the lack of the distinction between things-in-themselves and phenomena; behind the rationale for this criticism lies the concept of “intelligence” as defined by transcendental idealism. With this in mind, let us compare Kant’s [K4] and McCarthy’s [M4].

[K4] One can concede the existence of matter without going beyond mere self-consciousness and assuming something more than the certainty of the representation in me, hence the cogito, ergo sum.

[M4] The right way to think about the general problems of metaphysics and epistemology is not to attempt to clear one’s own mind of all knowledge and start with “cogito, ergo sum” and build up from there.

In the Critique of Pure Reason, “cogito” (I think) often means self-consciousness, and [K4] matches this concept. In this context, self-consciousness is a definite representation of “I think (therefore, I am)”. For this reason, “without going out of his consciousness” can be matched with “not to attempt to clear one’s own mind of all knowledge,” and “without assuming more than the certainty of the representation within me (‘I think, therefore I am’) with “start with ‘cogito, ergo sum’ and build up from there.” Thus, [K4] is simplified as follows:

[K4*] An empirical realist acknowledges the existence of matter without going out of the cogito of self-consciousness.

[K4*] has two semantic layers: (1) the cognition realm of self-consciousness, and (2) “I think must be able to accompany all my presentations” (B130). Based on these two thoughts, [K4] can be reformulated as follows:

15 As a prime example, see B130 f., B413. For further information, see Kim (2016).
Our knowledge starts with our self-consciousness of “I think, therefore I am” and ends there.

In the same vein, [M4] can be simplified as follows:

[M4*] Our knowledge is not an epistemological quest for self-knowledge and therefore does not start with “I think, therefore I am.”

A direct comparison of [K4*] and [M4**] shows that the decisive difference between the two, that is, between empirical realism and transcendental realism, lies in the answer to whether the self-consciousness of “I think” is acknowledged. According to Kant, this self-consciousness constitutes human “intelligence.” In the “deduction” of the Critique of Pure Reason (§ 25), which deals with the problem of self-consciousness as the main subject, Kant says the following about intelligence:

( . . . ) through which [intuition of the manifold in me: HJ. K.] I determined this thought [I think myself: HJ. K.]; and I exist as an intelligence that is merely conscious of its faculty for combination ( . . . ) [T]his spontaneity is the reason I call myself an intelligence.

(B158 footnote)

In this passage, intelligence is defined first and foremost as the faculty of combination. As examined above, this holds true for both Kant and McCarthy. However, Kant adds the faculty to be aware of this faculty as an additional essential attribute of intelligence. For Kant, the human being is a subject constantly conscious of the fact that the representations and thoughts that the subject is thinking certainly belong to the subject. This is where Kant and McCarthy part.

References


