‘인공지능’과 ‘인간지능’ 개념에 대한 철학적 분석 시도*
-맥카시와 칸트의 지능개념을 중심으로-

김형주**

주제분류 형이상학, 인식론, 인공지능, 독일철학

주 요 어 인공지능, 인간지능, 칸트, 맥카시, 자기의식, 선험적 관념론

요 약 문

본 논문은 ‘인공지능’이라는 개념을 처음 사용한 공학자인 존 맥카시의 ‘인공지능’에 대한 정의, 설명을 분석하고 이로부터 도출된 결론을 칸트의 철학을 도구삼아 규정된 ‘인간지능’과 비교하여 양자 간의 유사점과 차이점을 밝히고 그 차이점의 원인을 찾는 것을 목적으로 한다. 이를 위하여 첫째 ‘인공지능’과 ‘지능’ 그리고 칸트의 철학을 통해 조망된 ‘인간지능’간의 개념적 연관성을 규명한다. 둘째 인공지능과 관련한 맥카시의 언급과 인간지능에 대한 칸트의 설명을 비교하여 인공지능과 인간지능의 유사점이 ‘지능을 사용한 문제해결능력, 달리 표현하면 주어진 문제에 대한 판단력이라 논증한다. 셋째 칸트의 선험적 관념론과 선험적 실재론의 구분을 살펴보고 양자가 논의되는 인식론적 전제가 차이가 있음을 규명하여, 양자의 차이점이 자기의식의 소유여부라는 사실을 논증한다. 그러나 사태에 대한 판단은 자기의식 밖으로 나갈 수 없다는 선험적 관념론의 기본입장

* 본 논문은 지난 2016년 6월 3일 “인공지능과 인간의 존엄성”이라는 주제로 중앙대학교 부설 중앙철학연구소와 연구모임 **未名**이 공동으로 개최한 하계학술발표회에서 “인간지능과 인공지능 -매카시와 칸트의 ‘지능’개념을 중심으로”라는 제목으로 발표한 원고를 수정, 보완한 것이다. 본 지면을 빌어 생산적인 질문과 토론을 하여주신 분들과 유익한 지적과 조언을 하여주신 익명의 심사자들께 감사의 마음을 표한다.

** 동서울대학교

을 통하여 역설적으로 인공지능 역시 우리와 마찬가지로 자기의식을 가질 수 있다는 개연성을 열어 놓을 수 있음을 역설한다. 이를 통하여 칸트의 선험적 관념론이 탈인간중심적 인공지능연구의 개념적 뒷받침이 될 수 있다고 주장한다.

1. 들어가며

인공지능 알파고(AlphaGo)가 이세돌 9단을 이겼다. 이 승리에 대해 이세돌 9단은 ‘이세돌이 패배한 것이지 인간이 패배한 것은 아니’라고 답하였다. 많은 사람들은 이 답변에 그가 펼친 인공지능과의 대결, 그리고 그에 대한 결과 이상으로 주목하였다. 이 발언은 이세돌이 자기 자신을 단지 여러 유능한 바둑기사 중 한 명으로 묘사하여 비록 자기 자신은 졌지만, 자기가 세계최고의 바둑기사가 아니기 때문에 자신이 졌다는 사실이 모든 다른 바둑기사들의 패배를 뜻하는 것은 아니라는 것으로 해석되었다. 이 인터뷰로 인해 이세돌은 바둑실력에 걸 맞는 겸손한 인격을 갖춘 것으로 많은 사람들로부터 칭찬을 받았다.

그러나 우리는 이 발언을 조금 다른 각도에서 해석해 볼 수도 있다. 바둑기사 이세돌, 구체적으로 말해, 바둑을 잘 두는 이세돌의 능력은 알파고에게 못 미쳤지만, 이것이 이세돌이라는 인간이 가진 모든 능력이 인공지능 알파고에게 미치지 못하는 것을 의미하는 것은 아니다. 다시 말해 우리는 이 발언을 바둑을 두는데 필요한 인간의 여러 가지 능력들이 지금의 인공지능 로봇들에게 미치지 못하는 것은 사실이지만, 이 사실이 곧 인간의 지성이 가진 모든 능력들을 인공지능 로봇이 가지고 있고, 그 능력들 또한 인간지성보다 월등한 것은 아니라는 것을 뜻한다고 해석할 수 도 있다. 인공지능의 어떤 능력이 인간의 어떤 능력보다 앞선다는 것은 양자 사이에 유사점이 있다는 것을 뜻한다. 한편 인공지능이 모든 면에서 인간의 능력을 앞서지는 않는다는 사실은 앞에서 언급했듯이 인간이 가지고 있는 어떠한 능력은 아직 인공지능이 구비하고 있지 않다는 것을 의미한다. (물론 그 역도 성립할 수 있다.)

본 논문은 이 유사점과 차이점에 주목한다. 그러나 유사점과 차이점을

논증함에 있어 인공지능이라는 기계가 가지는 기능적 측면을 주제적으로 다루지는 않겠다. 다음 장에서 기술하겠지만, '인공지능' 의 기능은 매우 빠르게 발전하고 있을 뿐더러, 이 개념이 가지는 외연은 매우 넓기 때문이다. 본 논문은 '인공지능'이라는 말이 처음 등장하였을 당시, 이 개념을 사용하였던 사람인 존 매카시(John McCarthy)의 언급들을 분석하여 그 개념에 대한 구체적인 이해를 시도한다. 그리고 이를 '인간지능'과 비교한다. '인간지능'이라는 말 역시 '인공지능' 못지않게 다양한 의미로 해석이 가능하기에, 이 의미의 해석의 폭을 규정하기 위하여 지능 개념에 대한 칸트의 이해를 살펴본다. 이를 통하여 '인간지능'에 '인공지능'과는 차별화되는 부분이 있을 수 있는지, 있다면 어떤 것일 수 있는지, 이 차이가 가지는 철학적 의미는 무엇인지 고찰하는 것이 본 논문의 목적이다.

2. 인공지능과 지능의 함수관계

'인공지능'이라는 개념은 1956년 존 매카시와 마빈 민스키(Marvin Minsky), 그리고 그들의 동료들에 의해 개최된 다트머스(Dartmouth) 컨퍼런스에서 처음 공론화되었다.[1] 이후 이 개념은 시간이 지나고, 과학과 기술이 발전함에 따라 여러 영역에서 다양한 의미로 사용되고 있다. 그렇기 때문에 '인공지능'을 딱히 무엇이라고 일의적으로 정의하는 것은 매우 어렵다. 보다 정확히 말하자면 이는 불가능에 가깝다. 왜냐하면 인공지능은 일종의 선재적(**先在的**) 개념이라기보다는 기술의 발전에 따른 프로그램 혹은 로봇, 그리고 넓게는 이를 다루는 학문의 분야를 추후적으로 가리키는 개념이기 때문이다. 다시 말해 인공지능이라는 개념에 그것

1) '인공지능'이라는 말이 1956년 다트머스 컨퍼런스에서 처음 사용된 것은 주지의 사실이나 1950년 발표된 튜링(Turing)의 논문 "Computing Machinery and Intelligence"에서 이와 관련된 중요한 논의들은 이미 등장하였다.

이 지칭하는 특정한 대상을 대입시키는 방식으로 이를 이해하거나 그 의미를 규정할 수 없다. 급속히 변화되고 발전하고 있는 기술과 이 기술이 적용된 프로그램, 사물 그리고 이들을 포괄하는 학문의 분야들의 공통적인 특징을 기반으로 이들을 아울러 '인공지능'이라는 말을 통하여 논의하고 있는 것이 우리의 실정이다. 요컨대 인공지능은 합의적 개념이다.[2] 다만 우리가 '인공(artificial)'-'지능'이라는 개념자체를 접하였을 때, 그것의 용례와 실제 관련 학문분야에서 어떻게 사용되고 있는지에 대한 지식과는 별도로 그 자체만으로 알 수 있는 한 가지 사실은, 그것이 인공적(artificial)인 한 어떠한 것에 대한 모방이라는 것이다. 우리의 논의는 여기서 시작된다. 모방의 대상이 되는 어떤 것이란 무엇인가? 이에 대한 가장 손쉬운 답변은 '인간의 지능'일 것이다. 만약 지능이라는 개념이 본래 인간을 비롯한 정신적 존재와 관련하여서만 사용되었던 역사적 배경을 고려한다면, 간단히 '지능'이라 답할 수 있다. 풀어 설명하자면 어떠한 기계, 프로그램이 갖추고 있는 인공지능은 (인간)지능에 대한 인공지능, 다시 말해 (인간)지능을 모형으로 만든 것이다.[3] 그러나 이에 대한 보다 섬세하고 정확한 답변을 내리는 것이 쉽지 않다는 사실은 앞으로의 논의를 통해 드러날 것이다. 이러한 배경에서 우선 '인공지능'이라는 말이 처음 사용되었을 당시 부여된 의미에 대한 고찰을 통해 오랜 시간동안 다양한 형태로 발전한 '인공지능'이라는 체계가 가지는 여러 복합적인 요소들을 제외한, 그것이 가지는 순수한 의미를 파악하는 것을 시도해 보겠다.

초기 인공지능연구의 기틀을 다지고 지금의 연구에도 많은 영향을 미

2) 이와 관련하여 이초식은 "인공지능의 정의 문제는 일종의 규약이기 때문에 자의적으로 규정해도 무방한 것으로 여겨지기도 한다"(이초식, 1993 10쪽)고 말한다.

3) 실제로 본 논문이 다루게 될 맥카시의 인공지능 개념의 직접적인 모델인 인공지능의 가능성물음과 관련한 최초의 실험인 튜링검사(Turing-Test)는 "흉내 내기로서의 지능개념에 기초하고 있다."(이상욱, 2009, 53쪽, MacCarthy, 2007 참조)

친 맥카시는 2007년 “인공지능이란 무엇인가?”라는 제목으로 발표된 인터뷰형식의 보고서에서 인공지능을 “지능적인 기계, 특별히 지능적인 컴퓨터 프로그램을 만드는 과학 혹은 기술”(MacCarthy, 2007)이라고 말한다. 인터뷰의 내용을 요약하면 다음과 같다.

> 인공지능이란 지능을 가진 컴퓨터 프로그램이나 기계를 만드는 기술, 과학이다. 이는 인간의 지능을 이해하는 컴퓨터의 사용과 같은 사안에 관계한다. 그러나 ‘지능’이라는 개념은 아직까지는 ‘인간 지능’이라는 개념으로부터 전적으로 독립되어서는 확실한 정의를 가지를 수 없다. 다만 확실한 것은 ‘지능’은 기본적으로 이 세계에 존재하는 특정한 목적을 성취하기 위한 계산적(computational) 능력의 일부라는 것이다. 이 능력은 동물, 기계, 사람에게 다양한 형태와 정도로 갖추어져 있다. 한편, 인공지능의 목적은 인간지능 만큼의 능력을 갖추는 것이다. 그러나 2007년 현재의 관점에서 이 목적의 달성은 회의적으로 보인다.[4]

맥카시는 인공지능을 위에서 언급했듯이 ‘지능적 기계, 지능적 프로그램을 만드는 과학 혹은 기술’이라고 간략하게 정의하지만 이 정의의 배후에는 지능, 인공지능, 인간지능 세 개념 사이에 놓인 복잡한 긴장관계가 놓여 있다. 이 관계를 구성하는 개별항이 무엇인지, 각 항들의 관계를 어떻게 정리할 수 있는지를 보여주는 것이 본 논문의 중요한 목표중의 하나인 맥카시의 언급에 입각한 고전적 인공지능 개념의 분석을 위한 첫 걸음이다. 위의 요약에는 일견 최소한 다음의 두 주장이 담겨있는 것처럼 보인다.[5]

---

4) MacCarthy, 2007 참조.

5) 그러나 위의 요약문에 상이하게 기술된 두 주장이 공존한다고 하는 것이 이 두 주장이 양립되지 않는다는 것을 의미하지 않는다. 이는 이어지는 논의에서 밝혀질 것이다.

[주장1] '지능'은 '인공지능'과 '인간지능'을 아우르는 상위개념이다.[6]

[주장2] '인공지능'이라는 개념은 '인간지능'이라는 개념으로부터 비독립적이다. '인간지능'은 '인공지능'의 롤-모델이고, 따라서 '인간지능'이 없었더라면 '인공지능'이라는 개념은 애초부터 존재할 수 없었을 것이다.[7]

만약에 우리가 [주장1]에 주목하여 인공지능의 의미를 명확히 하고자 한다면, 지능은 '인공지능', '인간지능'의 필요조건이며, '인공지능'은 '인간지능'과 마찬가지로 지능을 가지고 있으므로 '인공*지능*', 즉 하나의 지능으로 간주될 수 있다고 이해할 수 있다. 요컨대 인공지능이 지능일 수 있는 이유는 인간지능이 지능일 수 있는 이유와 마찬가지로 *그것*이 지능이기 때문이다. 실제로 그는 위의 글을 발표하기 훨씬 이전인 1969년에 "인공지능의 관점에서 본 몇 가지 철학적 문제들(Some philosophical problems form the standpoint of artificial intelligence)"에서 "일반지능(our notion of general intelligence)"(McCarthy & Hayes, 1969, 3쪽)이라는 표현을 강조하면서 이를 지능의 정의와 관련한 논의를 이끌어가는 핵심개념으로 사용한다. 그에 따르면 "한 집합체가 수학의 지성적(intellectual) 세계, *그것*의 고유한 목적들의 이해, 그렇지 않으면 [인간의 정신을 포함한: 역자 *추가*]*다른 정신적 절차*의 지성적 세계를 포함하는 세계의 적절한 모델들 가지고 있다면, 그 집합체는 지능적"(McCarthy & Hayes, 1969, 4쪽)이다. 심지어 그는 "물리적 세계는 존재하고 이미 이

6) 실제로 그는 "인공지능의 관점에서 본 몇 가지 철학적 문제들"에서 인공지능에 대비되는 개념으로 '자연지능(natural intelligence)'이라는 개념을 사용하는데 이는 위에서 언급한 그의 논문의 맥락에서나 지금 우리의 논의의 맥락에서 '인간의 지능'을 의미하는 것으로 볼 수 있다. (McCarthy & Hayes, 1969, 4쪽)

7) 이와 관련하여 그는 "인공지능과 철학의 관계"(What has AI in Commo with Philosophy")에서 "인간지능 수준의 인공지능은 특정한 철학적 태도, 특별히 인식론적 태도를 갖춘 컴퓨터 프로그램의 구비를 요구한다"(MacCarthy, 1995 1쪽)고도 말한다.

*세계 안에는 인간이라 불리는 지능적 기계*들이 존재한다"(McCarthy & Hayes, 1969, 5쪽)고도 이야기한다. 이와 같은 그의 언급들을 주목해 본다면 [주장1]은 다음과 같이 구체화될 수 있다.

[주장1*] 인공지능의 지능은 일반지능이다.[8)]

한편, [주장2]에 주목한다면 우리는 '인공지능'을 '인간지능', 구체적으로 말하자면 *그것*의 계산적 능력을 원형으로 만들어진 모형적(ectypus) 개념으로 받아들여야 한다. [주장2]의 핵심은 인공지능은 인간지능의 일부 능력을 공유하고 있다는 의미에서 지능으로 간주되어야 한다는 것이다. 이상의 분석에서 [주장1*]과 [주장2]사이의 차이점이 두드러진다. 인공지능에게 지능의 지위를 부여하는 것이 인공지능과 인간지능을 포괄하면서 이 양자에 논리적으로 선행하는 '(일반)지능'인가 아니면 '인간지능과의 유사성'인가?

일단 해석의 선의의 원칙(principle of charity)에 따라 두 주장이 서로 모순되지 않는다고 생각해 보기로 한다. 위에서 살펴보았듯이 맥카시는 지능이 인간, 기계, 심지어 동물에게 다양한 형태와 등급으로 나타난다고 주장한다. 이 주장에 근거하여 [주장1]과 [주장2]를 연결시켜 본다면, 기계와 인간 모두 지능을 가지고 있지만 현재까지는 인간의 지능이 더 높기 때문에, 기계는 인간의 지능을 모델로 삼고 있다고 이해할 수 있다. 이 해석을 검토하기 위해 계속해서 맥카시가 "인공지능의 관점에서 본 몇 가지 철학적 문제들"에서 지능과 인공지능의 관계에 대해 설명한 부분을 살펴보자.

"인공지능, 구체적으로 말해 일반지능에 대한 논의는 지능이 무엇

8) 이상욱은 같은 맥락에서 일반지능을 "일반화된 지능"(이상욱, 1990, 64쪽)이라고 표현한다.

> 인지에 대한 보다 명확한 이해를 통해 발전될 수 있다. (...) 만약 기계가 인간의 지능을 필요로 하는 특정한 문제들을 해결한다면, 우리는 그 기계가 지성적이라고 말해야만 한다."(McCarthy & Hayes, 1969, 4쪽)

위의 인용은 맥카시가 *인공지능*의 본질을 *인간의 지능*이 요구되는 사안에 대한 문제해결능력으로 보고 있다는 사실을 명확히 보여준다.[9] 그는 인공지능이 인간지능, 동물의 지능과 마찬가지로 지능 그 자체(Intelligence it self)로 인해 지능으로 불릴 수 있다는 주장을 정당화하기 위하여 인간지능이 소유하고 있는 것과 같은 문제해결능력의 소유여부를 끌어들인다. 즉, 지능적 존재 그 자체로서의 기계를 정당화하기 위하여 전통적인 의미에서의 지능적 존재, 다시 말해 지능의 소유에 대해 의심의 여지가 없는 존재인 인간과의 유사성을 정당성의 척도로 삼는다. 이러한 입장을 견지한다면 [주장2]는 [주장1]로부터 구체화 내지는 추론된 것이라 할 수 있고, 우리가 다루고 있는 요약문의 주장은 결국 [주장2]로 수렴된다. 그리고 이때의 일반지능은 사람이 이미 가지고 있는 지능이 이제는 기계도 소유하고 있다는 의미에서 양자가 공유하고 있는 문제해결능력 이상을 의미하지 않는다. 그렇다면 일반지능으로서의 인공지능은 일상적 의미에서 인간의 지능이 가지는 능력 중의 일부에 대한 모사에 불과하다. 그러나 위에서 언급된 '일반지능'이 현재 진행되고 있는 일공지능 연구에서 사용되고 있는 의미로, 구체적으로 말하자면 "인간지능에 구속받지 않(이상욱, 2009, 65쪽)"는 지능을 지칭하는 의미를 담고 있다면 문제는 달라진다. 왜냐하면 '일반지능'이라는 개념이 지능이 딱히 누구의 지능, 어떠한 지능으로 구체화되기 이전을 지칭하는 '지능 일반'을

9) 이와 유사한 주장을 찾는 것은 어렵지 않다. 예를 들면 마르(Marr)는 그의 논문 "인공지능을 보는 하나의 관점"(Artificial Intelligence: A Personal View)에서 "인공지능의 목적은 (...) 정보처리과정과 관련한 해결 가능한 문제들을 규정하고 이를 해결하는 것"(Marr, 1990, 133쪽)이라고 말한다.

의미한다면, 이는 앞서 언급했듯이 '인공지능', '인간지능'보다 논리적으로 선행하는 것일 테고, 그렇기 때문에 인공지능에 지능의 지위를 부여하는 것은 인간지능이 아니라 지능 그 자체일 것이기 때문이다. 만약 그렇다면 맥카시는 '일반지능'이라는 개념을 통해 인공지능에게 인간지능의 제한된 모사품이라는 멍에를 벗겨주는 동시에 인공지능에게 주어진 문제에 대한 자기주도적인 해결력 이상의 능력을 부여하고자 했을 것이다.[10]

문제는 '지능'을 어떻게 이해할 것인가에 놓여 있다. 실제로 맥카시는 지능 개념으로부터 파생된 본질적인 문제들에 대해 "철학자들은 2500년 동안 합의를 보지 못했다"(McCarthy & Hayes, 1969, 5쪽)고 말한다. 이는 맥카시가 '지능', 구체적으로 말해 철학자들의 주된 탐구대상이었던 인간지능에 대한 논의가 지난 2500년 동안 지속되어왔다는 사실을 잘 알고 있다는 사실을 나타낸다. 또한 이는 인간지능에 대한 탐구는 매우 복잡하고, 그렇기 때문에 이를 단 몇 마디로 명확히 정의한다는 것은 불가능하다는 사실도 말해준다. 인간의 지능에 대한 철학적 해석의 다양성에 대해서는 다음 장에서 다루기로 하고, 본 장에서는 우선 맥카시의 인공지능 개념은, 그가 아무리 기계에 독립적인 지능체의 지위를 부여하고자 하였더라도, 그 스스로도 고백했듯이 사실은 인간지능 개념에 빚을 지고 있다는 사실을 주지하고자 한다. 실제로 맥카시의 지능개념은 비판을 받는다. 현재 미국을 비롯한 세계 각국의 많은 공과 대학에서 교과서로 사

10) 그러나 실상 맥카시는 일반지능 개념을 강한 인공지능 개념으로 사용하지는 않았다. 'Generality in Artificial Intelligence'에서 그는 일반지능의 목적을 여전히 추론과 이에 따른 문제 해결능력으로 보고 있다.(MacCarthy, 1987, 1032쪽 참조) 여기서 '일반성'은 인공지능의 (문제 해결, 계산) 능력이 인간의 능력에 미칠 수 있는 가능성, 그래야만 한다는 당위성에 대한 강조를 표현한다. 본문에서 요약문에 대한 순수한 분석을 통해 위와 같이 논의를 확장시킨 까닭은 첫째, 주목하고 있는 텍스트에 대한 해석의 가능성을 선입견을 배재한 채 검토한다는 본 논문의 목적에 충실하기 위해서 둘째, 이를 통해 우리가 다루고 있는 맥카시의 언급에서 추후 언급될 현대 인공지능 연구의 두 가지 흐름에 대한 논의의 가능성의 단초를 찾아보기 위해서이다.

용되고 있는 "인공지능(artificial intelligence)"에서 저자 러셀(Russel)과 노르빅(Norvig)은 맥카시의 "인공지능이란 개념보다 계산적 합리성(computational rationality)이 훨씬 더 정교하고 [그 사용에 있어서: 역자 추가] 덜 위험하다"[11]고 말한다. 필자가 생각하기에 맥카시 스스로도 이 비판, 좀 더 정확히 말해 그의 인공지능 개념의 본질에 대한 핵심적 지적에 수긍할 것 같다. 왜냐하면 위에서 언급했듯이 맥카시 역시 인공지능의 핵심을 논리적 추론(reasoning)을 기반으로 물리적 외부세계[12] 내에 주어진 "특정한 상황"(McCarthy & Hayes, 1969, 4쪽)에 있어서의 문제해결능력이라 여기기 때문이다.[13] 이를 바탕으로 우리는 [주장2]를 다음과 같이 구체화할 수 있다.

---

11) Russel & Norvig, 2010, 17쪽.

12) 이는 앞으로 다루게 될 칸트의 세계관 분류 중 "선험적 실재론(Transzendentaler Realismus)"에 관계한다. 선험적 실재론은 시공간은 우리의 감성과는 독립적으로 그 자체 실재하는 것이므로 이 시공간 안에 존재하는 표상(Representation)들이 물자체일 수밖에 없다는 이론이다. (A369 참조; 칸트의 저작 중 "순수이성비판"은 관례에 따라 초판은 A, 재판을 B로 표시한 후, 학술원판(Akademie-Ausgabe) 쪽수를 표시한다. 이외의 칸트 원저는 학술원판 권수와 쪽수를 표시한다.) 모든 인용은 기존의 번역서를 참고한 필자의 번역이다.

13) McCarthy, 2007 참조. 이러한 입장에서 위 책의 저자 러셀과 노르빅은 '행위와 사고', '인간과의 유사성과 합리성'이라는 두 개의 개별 범주의 조합을 통해 '인간과 같이 생각하기(Thinking humanly), 합리적으로 생각하기(Thinking rationally), 인간과 같이 행위하기(Acting humanly), 합리적으로 행위하기(Acting rationally)' 등 총 네 가지 틀로써 산재되어 있던 기존의 인공지능의 정의에 대한 정리를 시도한다. '인간과 같이 생각하기(Thinking humanly)는 인간의 인지적 능력들, 구체적으로 말해 감각기관과 사고력, 지식들 간의 관계에 유비하여 프로그램을 이해하고 개발, 발전시키는 것과 관계한다. '"사고의 법칙"에 의거한 접근'이라는 부제가 붙은 '합리적으로 생각하기'는 인간의 논리적 사고, 아리스토텔레스로부터 발전한 형식 논리학에 근거한다. '인간과 같이 행위'하는 것에 근거한 인공지능의 정의는 튜링에 의해 제안된 소위 튜링테스트의 기조에 다름 아니다. 인공적 행위자, 즉 컴퓨터의 (언어)행위의 수행여부가 인공지능의 핵심이다. '합리적으로 행위하기'는 '합리적으로 생각하기'를 근간으로 이를 포함하는 확장된 개념으로 추론능력을 기본으로 하여 특정한 상황 하에서 주어진 문제를 해결하는 것을 목적으로 한다. (Russell & Norvig, 2010, 2-5쪽 참조.)

[주장2*] 인공지능은 물리적 경험세계 내의 문제해결을 위한 인간의 계산적 합리성의 모형이다.[14]

이 두 주장들은 인공지능 연구의 커다란 두 줄기 흐름과도 연결된다. [주장2*]에서의 인공지능은 넓은 의미에서 고전적 인공지능으로 분류된다. 고전적 인공지능은 일종의 "흉내내기 지능"(이상욱, 2009, 65쪽)으로서 인간지능의 계산적 합리성을 롤-모델로 한다. 한편 [주장1*]의 '일반지능'으로서의 '지능'은 원칙적으로 굳이 "인간지능에 구속받지 않(이상욱, 2009, 65쪽)"는 지능이다. 바꾸어 말하면 [주장1*]은 인공지능 연구의 롤-모델이 굳이 인간일 이유가 없다는 사실을 나타낸다. 보덴(Boden)의 구분에 따르면 [주장2*]은 고전적 인공지능 연구의 입장이고 [주장1]은 현재 활발히 진행되고 있는 미래지향적인 연구인 "지능 일반에 관련한 과학으로서의 인공지능"(Boden, 1990, 1쪽)의 입장을 대변한다.[15]

이상에서 우리는 "인공지능이란 무엇인가?"이라는 인터뷰에 실린 인공지능에 대한 맥카시의 정의에 대한 분석으로부터, 인공지능에 대한 그의 설명과 정의의 뒷면에는 두 가지 상이한 주장이 공존하고 있다는 사실을 도출하였다. 그리고 "인공지능의 관점에서 본 몇 가지 철학적 문제들"에 기술된 '지능'에 대한 그의 언급을 통해 맥카시는 [주장1*]을 주장하고 있지만 실제로는 [주장2*]의 입장을 취하고 있다는 사실을 밝혔다.

---

14) 이와 관련하여 이상욱은 "인공지능의 한계와 일반화된 지능의 가능성"이란 논문에서 "인공지능 연구자들은 인간의 다양한 지적 능력 중에서 상대적으로 인공적 구현이 손쉬운 영역은 데카르트가 인간이성의 핵심으로 보았던 *논리적, 합리적 추론능력*"(이상욱, 2009; 강조 필자)이라고 보았다고 말한다.

15) 이는 지능연구에 대한 인간중심주의, 탈인간중심주의와도 연관이 된다. 만약 우리가 [주장2*]에 따른다면 인간중심주의의 입장을 견지한다고 할 수 있고, 이와는 달리 [주장1*]에 따른다면 탈인간중심주의의 입장을 취하는 것이라고 할 수 있다.(이상욱, 2009, 63쪽 참조) 이에 대한 논의는 본 논문의 마지막에 이어질 것이다.

다음 장으로의 논의를 이어가기에 앞서 인간지능과 인공지능의 유사점과 차이점에 대한 규명이 본 논문의 우선과제로 제시되었다는 사실을 잠시 환기하기로 한다. 메카시의 지능 개념이 사실상 "인간의 지능을 필요로 하는"(McCarthy & Hayes, 1969, 4쪽) 특정한 문제를 해결하기 위한 계산적 합리성을 의미한다는 것은 이것이 인공지능과 인간지능 사이의 유사점이라는 사실을 말해준다. 그리고 그가 인공지능의 원본 모델로서 계산적 합리성을 지목하는 배경에 물리적 경험세계, 각주 13에서 언급했듯이 선험적 실재론적 세계가 놓여있다는 사실을 다시 한 번 주지하고자 한다.

3. 칸트와 '인간지능'

3.1 칸트의 지능개념

맥카시의 언명을 통해서도 확인하였듯이 지성 혹은 지능에 대한 철학적 논의의 역사는 2500년 전부터 지금까지 지속되어 왔다. 맥카시의 말을 문자 그대로 받아드리면, 지성에 대한 인간의 관심은 고대 그리스 시대로 거슬러 올라간다. 그들은 정신(nous)과 그것의 "인식 작용 혹은 기능[인] noesis[지성]"(플라톤, 1997, 436쪽 각주)에 "의해서 알 수 있는(noetos; *intelligible*) 것"(거드리, 1994, 64쪽, 강조 필자)에 대해 집요하게 물었다. 파르메니데스는 정신을 감각의 우위에 둔 최초의 철학자였으며, 이러한 사조는 플라톤에게로 이어져서 이데아와 이에 대한 인식의 관계, 이를 설명하는 과정에서 파생하는 정신의 능력들의 활동과 구분에 대한 논의로 세분화되고 확장되었다.[16] 이러한 지성 개념은 중세에 이르

16) 이와 관련한 플라톤의 논의는 그의 저서 전역에 걸쳐서 다양한 형태로 등장하지만 필자는 특별히 국가 6권에 집약된 논의를 참조하였다(플라톤, 1997, 433-445

러 신을 알 수 있는 유일한 통로로서 “intellectus 혹은 intelligentia”(코플스톤, 1988, 314쪽)이라는 개념으로 표현되면서 지금 우리가 논의하고 있는 지성 개념의 어원적 모습을 띄게 되었다. ‘intelligentia’가 intellectus를 필요조건으로 하여 실천적 영역에서의 지성의 활동과 그 능력까지도 포괄하는 개념이라면 intellectus는 단어의 의미상 인식능력 자체를 의미한다고 할 수 있다.[17]

이러한 언어적 배경에서 칸트는 intellectus를 두 가지로, 즉 원형 지성(intellectus archetypus)과 모형 지성(intellectus ectypus)으로 구분한다.[18] 원형 지성(intellectus archetypus)은 우회적으로 신적 지성을 표현한다. 신적 지성은 그 안에 직관을 함축한다. 다시 말해서 지성(Verstand)과 직관(Anschauung)의 구분이 없는 지성, 즉 직관적 지성(intuitiven Verstand)을 뜻한다.[19] 신에게 있어서 인식은 곧 지식이다. “신은 대상을 그 자체로 인식한다”(XXVIII 606). 이러한 의미에서 신적 직관은 물자체의 존재에 대한 논리적 전제라 할 수 있다. 반면 신적 지성에 대한 모형적 지성(intellectus ectypus)을 뜻하는 인간의 인식능력은 제한적이다. 그렇기 때문에 직관과 사고력으로서의 지성(Verstand)은 분리되어 있고 대상은 직접적으로 포착되지 않는다. 이러한 지성은 그것이 현상만을 인식할 수

쪽 참조)

17) 온라인 독일어 어원 사전(http://www.dwds.de) 참조.

18) XXVIII 606, V 408 참조. 본 논문에서는 칸트의 지능 개념을 intellectus가 가지는 의미를 중심으로 이론철학의 범위 내에서만 다루겠다. 그러나 intelligentia에 주목한다면 ‘인간-지-능’을 ‘지성계의 존재자, 즉 예지자로서의 인간의 능력’(Die Fähigkeit des Menschen als das Wesen in der intelligibelen Welt)으로 이해하여 인간지능에 대한 논의가 실천철학의 영역에까지 확장될 수 도 있다.

19) ‘intuitiv’와 ‘anschaulich’는 칸트의 철학에서는 둘 다 ‘직관적’으로 번역될 수 있으나, 의미는 완전히 다르다. 전자는 감성적 직관(Anschaung)으로부터 독립적인 인식을 지칭하거나, 또는 어떠한 개념적 매개 없이 직접적으로 주어지는 지식을 표현한다. 반면 후자는 감성적 직관을 통해서 구상될 수 있는 인상을 나타낼 때 사용된다. (II 407 참조) 이러한 의미에서 직관적 지성(intuitiver Verstand)은 초감성적 직관이다.

있는 이상, 감성에 의존적, 달리 말해 감성에 얽매여 있다. 감성과 지성의 선험적 종합을 통하여 대상은 인식되고 지식은 산출된다. 인식의 과정이 직접적이지 않다는 점, 일련의 인식과정, 사고과정의 진행이 불가결하다는 점에서 이러한 인식은 논변적(diskursiv)이다.

3.2 선험적 관념론과 선험적 실재론 그리고 자기의식

위에서 살펴본 바와 같이 지성개념에 대한 고대와 중세의 전통적 입장은 칸트가 이를 두 가지 유형으로 명확히 나누어 고찰하는 태도에 이론적 배경이 되었다.[20] 절대적 일자의 정신을 분유 받은 존재, 완벽한 정신적 존재로서의 신의 형상을 따라 창조된 존재로서의 인간이해는 고대로부터 중세에까지 이르는 인간이해의 전통이었으며 칸트 역시 이를 수용하고 있다. 이러한 전통은 모형적 지성(intellectus ectypus)의 특징을 설명하는 선험적 관념론(transzendentaler Idealismus)을 태동시키는 모태가 되었다. 그는 선험적 관념론, 그리고 이와 대별되고 쌍을 이루는 세계관인 선험적 실재론을 다음과 같이 정리한다.

> "나는 모든 현상들의 선험적 관념론을 이 현상들을 물자체가 아닌 단지 표상으로만 간주하는 교설로 이해한다. 이 교설에 따르면 시간과 공간은 단지 우리의 직관의 감성적 형식들일 뿐 그 자체 주어진 규정 혹은 물자체로서의 대상의 조건이 아니다. 이 관념론에 대립하는 것이 선험적 실재론이다. 선험적 실재론자는 시간과 공간을 (우리의 감성과는 무관하게) 그 자체 주어져 있는 것으로 본다. 그렇기 때문에 선험적 실재론은 외적 현상을 (만약 우리가 그것의 현실성을 인정한다면) 물자체로 여긴다. 이 외적 현상들은 우리와 우리의 감성에 독립적으로 존재하며, 그렇기 때문에 순수 지성개념들(Verstandesbegriffe)에 따라서도 우리의 외부에 있는 것으로 간주된다. (...)

20) Kern, 2015, 1173, 1174쪽 참조.

> 선험적 관념론자는 이와 반대로 경험적 실재론자가 될 수 있다. 이는 소위 이원론자이다. 즉 그는 자기의식의 밖으로 나가는 일이 없이, 그리고 내 안에 있는 표상의 확실성, 즉 '나는 생각한다 그러므로 존재한다' 이상의 것을 상정함이 없이 물질의 현존을 인정한다."(A369 이하)

위에서 알 수 있듯이 선험적 관념론은 물자체와 현상을 명확히 구분하며, 모형적 지성으로 표현되는 인간지성의 인식영역을 단지 현상계로 한계 짓는다. 이러한 구분, 한계 짓기의 원인은 시공간 개념의 의식 내재성이다. 시공간은 그 자체 인간 지성의 외부에 독립적으로 존재하는 실체가 아니고 인식을 가능하게 하는 인간지능의 형식에 불과하다. 그렇기 때문에 칸트의 선험적 관념론에 따르면 우리는 단지 시공간을 통해 받아들인 "표상과만 관계할 수 있을 뿐"(A190/B235)이다. *선험적 관념론에서의 인식 주체는 자기안의 표상, 자기의식에 수반된 지식에 대해서는 "데카르트적 명증성"*(Henrich, 1976, S. 86)*을 확신할 수 있지만, 다른 인식 주체의 인식의 과정, 인식의 소유에 대해서는 함구(緘口)한다.*[21] '나는 생각한다'라고 하는 주체의 주체됨을 보증하는 자발적 사고력은 주어진 지각들(Perzeptionen)의 기저에서 이들을 종합적으로 통일하여 이 지각들이 바로 나의 지각들임을 의식하게 하는 근원적 능력으로서 통각(Apperzeption)으로도 표현된다. 그리고 순수 통각 그 자체는 바로 자기의식(Selbstbewusstsein)에 다름 아니다. 선험철학의 주체는 그렇기 때문에 자기 의식적 주체, 설명하자면, 지금 그리고 과거에 자기에게 주어진 지식들, 감정들, 판단들이 무엇이었는지를 끊임없이 *의식 할 수 있는, 의식 하여야만 하는* 주체이다. 다른 말로 표현하자면 지금 내가 떠올리고 있는 표상들, 생각들이 확실히 나에게 귀속되어(zugeschrieben) 있다는 사실을 끊임없이 의식하고 있는 주체이다.

---

21) 맹주만, 2006, 19쪽 참조.

한편, 맥카시가 말하는 인공지능은 다음의 인식론적 전제들 위에서 논의된다.

> 1. "인간이라 불리는 지능적 기계를 이미 함유하고 있는 물리적 세계는 존재한다."
> 2. "세계에 대한 우리의 상식적 관점은 거의 옳으며 그것이 바로 우리의 과학적 관점이다."
> 3. "형이상학과 인식론의 일반적인 문제들에 대해 올바르게 생각하는 방법은 이에 대한 모든 지식에 대한 우리 스스로의 마음을 [모든 경험적 요소로부터: 필자 추가] 순수하게 하는 것에 대한 시도가 아니다. 이는 'Cogito ergo sum'으로부터 시작할 수 없을 뿐더러 이로부터 견고해 질 수 도 없다."(McCarthy & Hayes, 1969, 6쪽)

요컨대 인공지능에게 해결해야할 문제가 주어지는 세계, 다시 말해 특정한 출력 값이 요구되는 상황의 배경이 되는 세계에 대한 이해는 (1)외부의 물리적 대상의 현존에 대해 아무런 의심도 하지 않는 것을 (2)상식으로 하는 자연과학적 세계관을 전제한다. (3)이러한 세계관은 'Cogito ergo sum', 즉 지능으로서의 자기 자신에 대한 의식, 짧게 말해 자기의식과는 무관하게 정당화된다. 이 세 가지 특징으로 비추어 볼 때, 인공지능의 세계는 자연과학적 실재론을 배경으로 하며, 이는 다시 위에서 살펴본 칸트의 인식-존재론적 세계 모델 중 선험적 실재론으로 설명이 가능하다. 선험적 실재론을 풀어 설명하자면 선험적 직관형식, 즉 시공간이 우리 지성과 관계하는 인식능력으로서의 형식이 아니라 그 자체 실재한다는 세계관을 나타내는 이론이다. 다시 말해 선험적 실재론은 시간과 공간을 우리의 감성의 형식이 아닌, 그 자체로 우리의 외부에 실재하는 것으로 간주한다. 그렇기 때문에 사물의 현존이 귀속되는 곳은 자기의식이 아니라 우리의 의식과는 상관없이 실재하는 시공간이다. 이러한 의미에서 칸트는 '경험적 실재론자 노릇을 하게 되는 선험적 관념론자는 "'나

는 생각한다 그러므로 존재한다' 이상의 것을 상정함이 없이 물질의 현존을 인정한다"고 말한 것'이고, 위의 3.에서 맥카시는 '형이상학과 인식론이 'Cogito ergo sum'으로부터 시작할 수 없'다고 말한 것이다.[22] 이렇

22) 칸트는, 위의 인용 A369에서 본 바와 같이, 선험적 관념론과 경험적 실재론을 양립가능한 것으로 제시하고 있으나, 두 개념을 통하여 각각 강조하고 싶은 내용은 상이하다. 선험적 관념론을 통해서는 시공간 형식의 의식 내재성에 기인한 물자체와 현상간의 분리에 주안점을 두고자 했다면, 경험적 실재론을 통해서는 외적 지각이 현실성으로부터 공간 안의 사물의 현실성을 직접적으로 도출하고 이를 근거로 표상과 외적 대상의 일치, 다시 말해 표상의 객관적 타당성을 이야기하고자 한다. 이러한 사실은 당시부터 지금까지 다양한 해석과 이에 따른 논쟁의 빌미를 제공하고 있다. 『순수이성비판』 재판 발표 당시 야코비(Jacobi)는 『David Hume über den Glauben oder Idealismus und Realismus』(1787)에서 유물론에 대한 칸트의 설명은 실패하였다고 평가하며, 칸트적 관념론, 즉 선험적 관념론은 표상에 상응하는 대상을 취하기 때문에, 사실상 반관념론(Nicht-Idealismus)가 되고자 한다고 말한다. "선험적 관념론과 연결된 경험적 실재론이 실재론의 한 종류로 정당하게 간주될 수 있는지는 지금도 논쟁거리이다"(Heidemann, 2015, 1894쪽). 퍼트남(Putnam)은 칸트의 경험적 실재론을 내재적 실재론으로 간주하는데 비해, 앨리슨(Allison)은 실재론적 입장의 과잉강조로부터 칸트적 관념론을 보호한다(Heidemann, 2015, 1894쪽 참조).

이렇듯 선험적 관념론에 대한 철학적 입장을 정위하는 것 역시 칸트의 철학이 배태하고 있는 또 하나의 철학적 문제다. 선험적 관념론은 "관념론으로, 현상학주의로, 그리고 실재론"(Edmundts, 2015, 1109쪽)으로도 해석될 수 있다. 예를 들어 스트로슨(Strawson)은 선험적 관념론을 현상학적 관념론으로 치부한 반면, (Strawson, 1966, 246쪽 참조) 앨리슨(Allison)은 양자 사이에는 극명한 차이가 있다고 말한다. (Allison, 2004, 41쪽 참조) 선험적 관념론을 현상학주의 혹은 현상학적 관념론으로 간주할 수 있는지는 칸트가 살아있을 당시 이미 가르베(C. Garbe)와 페더(G. Feder)등에 의해 촉발되어 지금까지 반 클레베(Van Cleve)등 많은 연구가들이 소개하고 참여하는 '두 세계 혹은 두 관점'논쟁과 연결된다. 이 연구가들의 이름과 논쟁의 내용은 잘 알려져 있기 때문에 별도의 언급은 하지 않겠다.

본 논문은 위에서 간략하게 소개된 선험적 관념론을 바라보는 개념상의 차이, 이로부터 불거진 많은 논쟁사를 배후에 두고, 이를 촉발한 칸트의 직접적인 설명이 등장한 구절인 A369, '내 안의 표상의 확실성인 Cogito ergo sum, 즉 자기의식의 확실성'을 기반으로 하는 '선험적 관념론자는 경험적 실재론자가 될 수 있다'는 칸트의 언급을 근거로 "선험적 관념론은 오로지 우리 인식의 주관적 조건에 주목할 때만 가능한 실재론"(Edmundts, 2015, 1108쪽)으로 간주한다. 경험적

듯 선험적 실재론에서 자기의식은 선험적 관념론에서처럼 더 이상 인식의 '근본명제', '최상원칙', '근원적 능력'으로서의 지위를 갖지 않는다.[23] *이러한 이유에서 필자는 자기의식이야말로 선험적 관념론과 선험적 실재론을 가르는 분수령이라고 주장한다.*

양자의 차이점에 대해서는 잠시 후 이어서 논의하기로 하고, 선험적 관념론이 가지는 내용을 살펴보고 이를 토대로 양자의 유사점에 대해 검토하고자 한다. 칸트의 철학, 특히 이론철학에서 '선험적(transzendental)'이라는 개념은 독특한 지위를 갖는다. 이 개념은 '초월적(transzendent)'에 어원을 두고 있지만 전혀 다른 의미를 지닌다. 지금 논의의 맥락에서 말하자면, '초월적'은 'intellectus archetypus'에, '선험적'은 'intellectus ectypus'에 상응한다. 다시 말해 초월적은 초월적 지성을 가진 존재자, 즉 신과 같은 존재자들의 존재방식을 표현하는 말인 반면, '선험적'은 규제적 지성(Intellektus)[24]을 가진 존재자, 즉 불완전한 인식주체인 인간의 의식구조에 대한 분석에 관계한다. 이러한 의미에서 '선험철학(Transzendental-Philosophie)'은 인간지성이 아프리오리(a priori)하게 가지

실재론의 의미는 시간과 공간, 이에 담지되어 있는 우리의 표상들은 우리의 경험의 한계 안에서만 의미를 갖는다는 이론에 다름 아니다.(Kim, 2016, 25쪽 참조) 표현상의 차이, 표현에 따른 해석의 차이, 회의주의에 대한 비판과 인식의 객관적 타당성 확보에 대한 칸트의 기획에 근거한 강조점 자체를 잠시 유보한다면 양자의 일차적 공통분모는 시공간의 의식내재성일 수밖에 없다는 사실은 잘 드러난다. 그리고 필자는 선험적 관념론으로부터 도출될 수 있는 수많은 이차 해석의 문제를 접어두고 바로 이점에만 주목하여 자기의식의 소유여부를 중심으로 선험적 관념론(경험적 실재론)과 선험적 실재론을 대비시키면서 논의를 전개하고자 한다.

23) 인간지능의 세계인 선험적 관념론과 인공지능의 세계인 선험적 실재론의 차이점에 관해서는 양자의 유사점을 비교한 후 이어 논의하겠다.

24) 앞으로의 논의에서 우리말 '지성'을 두 가지 의미로 사용된다. 첫째는 앞에서는 지능으로 번역하였던 'intelligence'의 문맥을 고려한 다른 번역어이다. 이는 인간의 지적 능력 일반을 뜻하는 의미에서 인간지능과 같은 의미를 갖는다. 둘째, 지성(Verstand)은 『순수이성비판』의 연역장의 주요개념으로서 범주개념들의 발원지를 뜻한다.

고 있는 인식능력들의 체계에 대한 학문이다.[25] 이 체계는 크게 세 부분, 즉 감성론, 분석론, 변증론으로 구성되어 있는데, 이들은 각각 인간의 상이한 인식능력인 감성, 지성(Verstand), 이성의 속성과 능력에 대해 다룬다. 각각의 인식능력에 대한 상세한 논의는 본고의 목적상 다루지 않기로 한다. 그러나 앞으로의 논의를 위해 이들의 특성을 요약하면, 감성은 외감의 대상을 내감의 대상으로 표상양태의 변화를 통하여 받아들이는 수용성을, 지성은 아프리오리한 범주개념들에 따라 주어진 표상들을 종합하는 순수 자발성을 의미한다. 이성은 칸트의 철학에서 많은 의미와 용례로 사용되지만, 이를 선험철학의 체계와의 연관에서 이론이성으로 국한하여, 그것이 가지는 가장 핵심적인 기능을 꼽자면 추론능력이라 할 수 있다. 그러나 상술한 각각의 능력들, 즉 수용성, 자발성, 추론능력이 명확히 구분되어 오로지 각각 하나의 인식능력에만 귀속하는 것은 아니다. 근원적 종합능력으로서의 지성의 종합능력은 감성이 대상들을 받아들임과 동시에 작동한다. 그렇지 않으면 시간의 계기성은 무의미한 말이 될 것이다. 한편 지성 역시 이성과 마찬가지로 추론적 능력으로도 표현된다. 다만 지성은 칸트에 따르면 직접추론, 다시 말해 소전제가 필요 없이 하나의 전제로부터 곧장 결론에 이르는 추론능력인 반면, 이성은 삼단논법을 가능하게 하는 간접적 추론능력이다.[26] 이렇듯 인간의 인식능력들은 개념적으로는 구분되지만 작용적 측면에서는 긴밀하게 연관되어 있다. 확실하고 생산적인 지식, 즉 아프리오리한 종합적 지식은 상상력을 매개로 한 감성과 지성의 선험적 종합을 통해서만 가능하고, 이렇게 산출된 지식에 이성은 질서와 통일을 부여한다. 쉽게 말해서 감성, 지성, 이성은 지식의 산출을 위해 필연적으로 연결되어 협력한다.

이상에서 우리는 우리가 인간지능을 인간의 지적 능력으로 해석한다면, 그것이 칸트의 지성에 대한 설명 틀에 입각하여 어떻게 이해될 수 있는

25) B25 참조.
26) B360 참조.

지 살펴보았다. 이를 바탕으로 선험적 관념론에서의 인식주체인 인간지능과 선험적 실재론에서의 인식 주체인 인공지능과 비교해보기로 한다. 친숙한 설명을 위해 우선 알파고가 여전히 고전주의 인공지능 연구의 입장에서 논의될 수 있고, 이러한 의미에서 선험적 실재론의 인식주체로 여겨질 수 있다고 가정해 보자. 알파고의 핵심 기술은 인간의 두뇌의 신경망 네트워크에서 영감을 얻어 개발된 뉴럴 네트워크(Neural Network Model)의 최신 모델인 딥 러닝(Deep Learning)이다.[27] 이러한 이유에서 딥 러닝의 뿌리는 연결주의(Connectionism) 인공지능이라고 할 수 있다.[28] 딥 러닝은 이미 주어진 방대한 데이터를 바탕으로 입력된 정보를 처리하여 원하는 결과를 가능한 한 정확하게 도출하게 하는 소프트웨어이다. 예를 들어 우리가 손 글씨로 3이라는 숫자를 그리면, 딥 러닝기법을 장착한 컴퓨터는 이미 확보하고 있던 수많은 데이터를 기반으로 확실한 숫자 3을 출력한다. 손 글씨 3은 이미 입력되어 있던 이와 유사한 다른 자료들과의 비교를 통해 "상관이 있는 것을 한 묶음으로 해서 [공통의: 필자 추가] 특징"(마쓰오 유타카, 2015, 164쪽)이 추출된다. 이러한 추상화의 과정이 몇 번이고 되풀이 되면 손 글씨 3은 확실한 글씨 3으로 인식된다. 그리고 방금 입력된 손 글씨 3은 다시 데이터로 저장되어 다음 입력될 정보에 활용된다. 그림판에 그린 한글인식, 나아가 페이스북의 태그 기능인 얼굴인식 기능이 이와 유사한 원리로 작동된다. 컴퓨터에 사람의 얼굴을 입력시키고 딥러닝을 작동시키면 일련의 과정을 거쳐서 '이것은 사람의 얼굴이다'라는 판단이 도출된다. 실제로 구글(Google)의 Image Auto Caption은 특정한 이미지가 입력이 되면 저장된 데이터를

27) 정상근, 2015, 11쪽 참조; 도안구, 2015, 7쪽 참조.

28) Boden, 1990, 2쪽 참조. 연결주의 인공지능은 지금 우리의 논의의 대상인 튜링, 맥카시의 고전주의 인공지능과 "방법론적으로는 정반대이지만"(Boden, 1990, 2쪽), 연결주의 인공지능의 핵심아이디어에 착안한 딥 러닝 역시 인간지능을 모델로 하고 있다는 의미에서 본 논의의 선상에 포함시킬 수 있다.

통해 이미지를 분석하고 분석한 이미지를 Language Generating RNN 프로그램을 통해 자연언어로 출력한다.[29] 딥 러닝의 작동원리에 대한 보다 상세한 설명은 본 논문의 목적과 지면상의 이유로 생략하기로 한다. 다만 이상의 약술의 핵심은 딥 러닝을 장착한 컴퓨터는 정보를 개념화할 수 있다는 것이다. 숫자 3의 형태의 특성을 가진 정보를 3으로 규정할 수 있으며 사람의 얼굴의 형태의 특성들을 가진 데이터를 받아, 이를 사람의 얼굴이라는 개념으로 규정한다.

우리는 이상에서 약술한 인공지능의 작용원리와 이에 앞서 살펴본 본 인간지능의 인식원리와의 유사점을 발견할 수 있다. 칸트가 설명한 인간지능의 핵심은 간단히 말해 감성의 개념화이다. 외감의 대상이 내감의 대상이 되고 그 과정에서 동종성(Affinität)에 의한 종합이 일어난다. 그리고 종합된 다양들은 하나의 표상으로 통일되며 이 표상이 개념들에 의해서 규정되면 판단 형식으로 언명된다. 만약 우리가 원형이라는 공간 안에 붉은 점들의 집합을 보게 되면, 이 다양한(mannigfaltig) 점들은 원형이라는 구획 안에서 동종성의 원리에 의해서 종합이 되고, 종합된 표상이 지성의 범주를 거치면 '이 표상은 사과이다'라는 판단이 도출된다. 이와 유사한 다른 판단들이 주어지면 이성은 추론을 하며 '사과는 과일이다'등과 같은 추상적인 판단을 할 수 도 있다. 3을 그린 그림이 주어졌을 때, 이를 3으로 개념화하고, 학습(판단)된 자료들을 바탕으로 추론하여 새로운 판단을 도출하는 인공지능은 큰 맥락에서 방금 살펴본 인간지능의 원리와 유사하다. 이 유사점은 위에서 언급한 '지능을 사용한 문제해결능력, 즉 판단력'이라는 말로 대변된다.

그러나 우리는 이전의 논의에서 맥카시의 인공지능이 문제를 해결하여

29) 예를 들어 재래시장 사진이 입력되면 '사람들이 재래시장에서 물건을 사고 있다', '야채시장에는 많은 야채들이 있다'와 같은 문장이 출력된다. http://techcrunch.com/2014/11/18/new-google-research-project-can-auto-caption-complex-images/ 참조.

하나의 판단을 도출하는 세계와 칸트의 인간지능이 활동하는 세계는 다르다는 사실을 확인하였다. 이 두 세계관의 차이점이 다름 아닌 외적 대상의 객관적 타당성확보에 대한 자기의식의 개입여부였다. 이와 관련하여 칸트는 다음과 같이 말한다.

> "'나는 생각한다'는 모든 나의 표상을 동반할 수 있어야만 한다. 왜냐하면 그렇지 않으면 내 안에 표상된 어떤 것은 전혀 생각되지 않은 것일 수 있게 되기 때문이다. 이는 표상이 불가능하다는 것을 뜻하거나 이 표상이 최소한 나에게는 아무것도 아니라는 것을 뜻한다."(B130)

위의 단락으로부터 우리는 두 가지 중요한 사실을 알 수 있다. 첫째, '나는 생각한다' 즉 자기의식은 선험적 관념론의 인식주체인 인간지능에 있어서 판단의 근본원리이다. 둘째, '나는 생각한다'에 수반된 표상의 객관성은 오로지 나의 판단에만 적용된다. 이를테면 선험적 관념론적 주체는 '너는 생각한다'는 모든 너의 표상을 동반할 수 있어야만 한다'고 말할 수 없다. 이 주체는 '나는 '너는 생각한다'는 모든 너의 표상을 동반할 수 있어야만 한다'고 생각한다'라고만 말할 수 있다.[30] 이러한 의미에서 선험적 관념론은 "자기의식의 밖으로 나가는 일이 없다"(A369). 반면 선험적 실재론은 물자체를 자기의식 밖으로 해방시킨다. 즉 인식주체의 대상인식 여부와 상관없이 대상은 존재한다. 자기의식은 대상의 객관성에 아무런 역할을 하지 않는다.

정리하자면, 맥카시의 인공지능과 칸트의 인간지능은 양자 모두 '지능을 사용한 문제해결능력, 즉 판단력'을 갖는다는 점에서 유사하지만, 이 문제가 놓여있는 세계관은 완전히 다르다. 그리고 이 세계관의 차이는 각 세계에 존재하는 대상에 대한 인식의 객관적 타당성 확보를 위한 인

30) Rosefeldt, 2000, 23쪽 참조.

식주체의 자기의식의 소유 여부에 달려있다.

4. 나가며

알파고가 이세돌을 이겼다. 만약 알파고가 여전히 위에서 우리가 논의한 선험적 실재론적 세계에서 행위하는 고전주의 인공지능이라면, 다시 말해 약한 인공지능이라면, 그(것)의 자기의식은 없어야만 한다. 만약 그렇다면, 그(것)은 '바둑두기'를 수행하는 동안 최적의 답을 찾도록 프로그래밍되어 있기 때문에, 바둑을 두는 주체가 바로 자기 자신인지, 대국에서 이겼을 때 이긴 주체가 바로 자기 자신인지 의식하는 기능은 애초에 갖추고 있지 않다. (현재 우리가 이러한 기능을 개발할 수 있는 기술력을 갖추고 있는지, 언젠가는 그러한 기술이 가능한지 역시 미지수이다.) 그(것)은 그저 자기 외부의 세계에서 자기에게 던져진 문제에 대해 최적의 답을 찾을 뿐이다. 만약 인간지능 이세돌이 선험적 관념론적 인식-행위주체라면, 그는 바둑을 두는 내내, 바둑을 두면서 생각을 하고 수를 계산하는 주체가 자기 자신임을 의식한다. 이 자기의식은 그가 의도적으로 의식하여 주어지는 것이 아닌, 그가 바둑을 두는 내내 하는 경험들의 기반에서 이들을 한데 묶어주고 통일성을 부여하는 의식이다. 그는 그가 패배한 후에도 다른 어떤 것이 아닌 바로 자기 자신이 패배한 사실을 즉각적으로 의식한다. 그렇다면 이세돌만 패배의 주체가 바로 자기 자신임을 의식하는가? 알파고는 승리의 주체가 자기 자신임을 의식하지 못하는 것이 확실한가? 우리가 선험적 관념론과 선험적 실재론을 제 3자의 입장에서 구분하는 메타적 입장에 서있다면 위의 질문에 대해 아무런 고민 없이 '그렇다'라고 답할 수 있다. 그러나 우리가 칸트의 입장을 충실히 따라서 인간지능으로서의 우리를 선험적 관념론의 인식주체로 간주한다면

이 질문에 대해 확답할 수 없다. 나 자신이 나에 대한 의식을 가지고 있음은 최소한 나에게 있어서는 명증적이지만 이 명증성을 다른 사고주체에게 적용하는 것은 허용되지 않기 때문이다.[31]

자기의 모든 인식과 경험이 '나는 생각한다'라는 자기의식의 확실성위에서만 수립된다는 사실은 인식에 대한 우리의 확신을 우리의 경험 세계 안에 가두어 놓는다. 선험적 실재론에 대한 이러한 해석은 칸트가 인식론적 회의주의자 혹은 유아론자라는 비판을 받게 하는 근거가 될 수 있다. 그러나 칸트는 선험적 관념론을 주장하면서 자기의 경험에 대한 확실성을 주장하였을 뿐이지 외부세계의 현존을 부정하지 않았다. 다만 그는 외부세계를 나에게 있어 명증적인(assertorisch) 것이 아닌 개연적인(problematisch) 것으로 보았을 뿐이다. 보다 정확히 말하자면, 칸트가 선험적 실재론을 비판하면서 선험적 관념론을 주장한 이유는 외부세계의 현존을 부정하기 위한 것이 아니라 자기경험의 확실성을 주장하기 위한 것이다. 인식의 경계설정은 일면으로는 한계설정이지만, 다른 한 면에서 보면 새로운 세계의 가능성에 대한 확보이다. 이러한 사실을 염두에 두고 논의의 시작에서 제시한 멕카시의 인터뷰 글에 담긴 두 주장을 상기하여 보자.

> [주장1*] 인공지능의 지능은 일반지능이다.
> [주장2*] 인공지능은 물리적 경험세계 내의 문제해결을 위한 인간의 계산적 합리성의 모형이다.

선험적 관념론은 [주장2*]을 아무런 주저함 없이 받아들이지 않는다. 인간지능인 '내'가 지능을 가졌기 때문에 인공지능인 '너'도 반드시 나와 같은 방식의 지능을 가져야하는 것은 아니다. 내가 말할 수 있는 것은 알파고가 지금 나와 같이 바둑을 두는 것으로 보아 나와 같은 지능을 가

31) Rosefeldt, 2000, 23쪽.

졌을 것이라고 '나는 생각한다'는 것이다. 재차 언급하지만 인식의 경계 설정은 일면으로는 한계설정이지만, 다른 한 면에서 보면 새로운 세계의 가능성에 대한 확보이다. 이러한 의미에서 [주장2*]을 흔쾌히 받아들이는 것에 대한 주저는 탈인간중심적 관점의 인공지능 연구의 핵심이념을 대변하는 [주장1*]의 가능성을 열어 놓는다.

칸트는 근대적 의미의 계몽의 완성자라 평가받는다.[32] 이는 데카르트적 주체, 인간중심주의의 완성자를 의미하기도 한다. 그러나 역설적으로 칸트의 선험적 관념론은 우리를 지성이 인간의 고유한 속성이라는 신념으로부터 해방시킬 수 있는 가능성 또한 갖는다. 근대적 자아 개념의 확립에 결정적인 역할을 하였던 칸트의 철학은 역설적으로 새로운 지능의 출현의 가능성에 대한 이론적 뒷받침이 될 수 도 있다. 알파고는 어쩌면 우리와 마찬가지로 자기의식이 있을지도, 혹은 우리와 전혀 다른 방식의 지능이 있을지도 모른다.

32) Wundt, 1964, 1쪽 참조.

참고문헌

국내문헌

거드리, (1994),『희랍철학입문』, 박종현 옮김, 종로서적.

도안구, (2015),「인공지능의 혁신 딥러닝」,『철도저널 18(6)』, 6-9쪽.

마쓰오 유다까, (2015),『인공지능과 딥러닝』, 박기원 옮김, 동아엠엔비.

맹주만, (2006),「칸트의 이성비판과 포스트모던 칸트」,『이성과 비판의 철학』, 9-43쪽.

정상근, (2015),「인공지능과 심층학습의 발전사」,『정보과학회지 33(10)』, 10-13쪽.

이상욱, (2009),「인공지능의 한계와 일반화된 지능의 가능성: 포스트 휴머니즘적 맥락」,『과학철학 12』, 49-69쪽.

이초식, (1993),『인공지능의 철학』, 고려대학교 출판부.

코플스톤, (1988),『중세철학사』, 박영도 옮김, 서광사.

플라톤, (1997),『국가』, 박종현 옮김, 서광사.

외국문헌

Allison, H. (2004). *Kant's Transcendental Idealism*. New Haven/London: Yale University Press.

Boden, M.(Hrsg.) (1990). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

Cramer, K. (1990). Über Kants Satz: Das: Ich denke, muß alle meine Vorstellungen begleiten können. In *Theorie der Subjektivität / Hrsg*. von Konrad Cramer (S. 167-202). Frankfurt am Main: Suhrkamp.

Marr, D. (1990). Artificial Intelligence: A Persnal View, In *The Philosophy of Artificial Intelligence / Hrsg. von Margaret A. Boden* (S. 133-146). Oxford: Oxford University Press.

Edmundts, D. (2015). Idealismus, Transzendentalter, In *Kant Lexikon* (*Online*). Berlin/New York: Walter de Gruyter.

Heidemann, D. (2015). Realismus, Transzendentaler/Empirisher. In *Kant Lexikon* (*Online*). Berlin/New York: Walter de Gruyter.

Henrich, D. (1976). *Identität und Objektivität*. Heidelberg: Carl Winter.

MacCarthy, J. (1995). "What has AI in Common with Philosophy? (Online)."

MacCarthy, J. (2007). "What is artificial Intelligence? (Online)."

MacCarthy, J, & Hayes, P. (1969). "Some philosophical Problems from the Standpoint of artificial Intelligence (Online)."

MacCarthy, J. (1987). "Generality of artificial Intelligence" In *Communications of the ACM Volume 30 Issue 12* (S.1030-1035). New York: ACM.

Rohlf, M. (2010). *Stanford Encyclopedia of Philosophy* (*Online*). Von http://plato.stanford.edu/entries/kant/#TraIde abgerufen

Rosefeldt, T. (2000). Das logische Ich. Berlin/Wien: Philo.

Russell. S, & Norvig, P. (2010). "Artificial Intelligence: A Modern Approch." Boston: Prentice Hall.

Strawson, P. (1966). *The Bounds of Sense*. London: Methuen & Co. Ltd.

Wundt, M. (1964). *Die deutsche Schulphilosophie im Zeitalter der Aufklärung*. Olms.

Kant, I. (1990 ff.). *Kants gesammelte Schriften (Sog. Akademie-Ausgabe)*. Berlin/New York: Walter de Gruyter.

Kim, H. (2016). *Zur Empirizität des „Ich denke" in Kants Kritik der reinen Vernunft*. Univ. Diss. Siegen. Digitaler Zugriff unter: http://dokumentix.ub.uni-siegen.de/opus/volltexte/2016/999/index.html

Kern, A. (2015). intellectus archetypus/ectypus. In *Kant Lexikon (Online)*. Berlin/New York: Walter de Gruyter.

Artificial Intelligence and Human Intelligence

-With Emphasis on Intelligence-Concept in MacCarthy and Kant

Kim, Hyeongjoo (Dongseoul Univ.)

The aim of this article is to analyze John MacCarthy's concept of "artificial intelligence" and compare it with Kant's concept of "human intelligence." Thereby I will clarify their commonalities and differences. In order to do this, I will, at first, attempt to establish a conceptional relationship between "artificial intelligence," "intelligence," and "human intelligence" (Kant). Secondly, I will argue that MacCarthy's artificial intelligence and Kant's human intelligence have resemblance when they are considered as a problem-solving faculty, i.e. an ability to make judgements about problems given under certain conditions. Thirdly, I will show that these two concepts have essential differences because Kant and MacCarthy adopt different epistemological frameworks. Kant's distinction between transcendental idealism and transcendental realism is laid out, and this shows that the difference between two concepts depends on whether one possesses self-consciousness or not. However, the basic idea of transcendental idealism, that our judgement can be valid only when it is related with our self-consciousness, suggests a paradox that it is still possible for "artificial intelligence" to acquire self-consciousness. Consequently, Kant's transcendental idealism can provide theoretical justification for any research on post-human artificial intelligence.

Key words: Artificial Intelligence, Human Intelligence, Kant, MacCarthy, Self-Consciousness, Transcendental Idealism

김형주 E-mail: godwithhj@hanmail.net

투 고 일	2016년 07월 25일
심 사 일	2016년 08월 01일
게재확정	2016년 08월 10일