과학철학 22권 2호 (2019) pp. 53-77

초지능이 실현될 것인가:
보스트롬의 정의를 기준으로[†‡]

고 인 석[*]

이 논문은 초지능의 실현 전망, 즉 인간의 지능을 능가하는 기계지능이 등장할 것인지 그 개연성을 평가한다. 논자는 초지능을 “사실상 모든 관심 영역에서 인간의 인지능력을 뚜렷이 상회하는 지능”으로 규정하는 보스트롬의 정의를 활용하면서 인공지능이 그러한 수준에 이르렀는지를 평가할 수 있는 수학적 구도를 제안한다. 그것은 인지능력의 다양한 요소들이 그 공간의 차원들에 대응하는 다차원의 추상적 공간에서 인간이나 특수한 기계지능 같은 지능 체계가 점유하는 부피를 평가하고 비교하는 것이다. 보스트롬이 말한 의미의 초지능이 실현된다면 그것은 이와 같은 다차원의 인지능력 공간에서 인간 수준의 지능이 점하는 부피를 고스란히 진부분집합으로 포함하는 부피로 나타날 것이다. 그러나 그처럼 다양한 인지능력의 차원들을 모두 포괄하는 기계지능을 개발하는 것은 비현실적 과제이고, 무엇보다도 인간의 지능이 계속 새로운 차원의 능력을 개발해낸다는 사실을 고려할 때 사실상 실현불가능한 과제다. 이로써 이 논문의 물음은 해결되었지만, 인간에게 없는 지능의 차원을 가진 인공지능이 등장할 수 있다는, 모종의 현실적 위험을 함축하는 새로운 문제가 제기된다. 지능이 작동하는 것처럼 보이는 현상이 있다고 해서 반드시 지능을 가진 존재, 곧 지능적 주체가 있다고 볼 이유는 없다. 그러나 진정한 지능적

† 인하대학교 교내연구비의 지원을 받아 작성된 논문임.

‡ 이 논문의 이번 버전들은 2018년 10월 중앙대에서 <지능의 진화와 인간환경의 확장>을 주제로 열린 제2회 인공지능인문학 전국학술대회에서, 그리고 2019년 6월 대만 국립양명대에서 열린 Workshop on Interdisciplinary Philosophy에서 발표되었다. 또 이 논문에 정서된 아이디어의 일부는 고인석 (2018a, 2018b)에서 개진된 바 있다.

* 인하대학교 철학과, insok@inha.ac.kr.

주체의 존재 여부와 무관하게 우리는 인공지능 기술이 구현하는 날로 새로운 지능 현상을 우리의 제어가능성 범위 안에 두는 일에 유념하지 않으면 안 된다.

【주요어】 초지능, 인간 수준의 기계지능, 인지능력의 공간, 지능적 존재, 제어가능성

1. 초지능의 개념과 실현 전망의 관계

인간의 지능을 뛰어넘은 기계지능, 즉 '초지능(superintelligence)'이 출현할 것인가? 만일 출현한다면 언제쯤, 어떤 방식으로 출현할 것인가? 초지능이 우리가 사는 현실 세계의 일부가 된다면 우리의 삶은 어떻게 변할 것인가? 인공지능 기술의 빠르고 광범위한 발달과 인공지능의 효용에 대한 사회적 관심의 증대가 맞물리는 가운데 이러한 물음들이 학계와 세간의 공통된 관심으로 부상하고 있다.[1)]

이런 물음들에 대한 유의미한 토론이 성립하기 위해서는 먼저 "초지능이 무엇인가?" 하는 물음에 대한 답이 제시되어야 한다. 다시 말해, 초지능의 개념에 대한 규정이 필요하다. 초지능에 대한 가능한 정의는 여럿일 수 있다. 그러한 복수의 정의 가운데 어느 것이 다른 것보다 초지능의 개념을 더 적절히 포착하는지 토론하게 될 수도 있다. 그러나 어쨌든, 초지능에 관한 예측을 제시하거나 나아가 기술 개발의 방향에 대한 지지나 반대 같은 모종의 태도를 표명하려는 자는 먼저 자신이 논의의 대상으로 삼는 초지능이 어떤 것인지 명시하지 않으면

1) 닉 보스트롬은『슈퍼인텔리전스: 경로, 위험, 전략』의 첫머리에서 다음과 같이 말한다. "미래의 어느 날 우리가 인간의 일반 지능을 능가하는 기계 두뇌를 만들게 된다면, 이 새로운 슈퍼인텔리전스(superintelligence, 초지능)는 매우 강력한 존재가 될 것이다. 그리고 마치 지금의 고릴라들의 운명이 그들 스스로가 아니라 우리 인간에게 달린 것처럼, 인류의 운명도 기계 초지능의 행동에 의존하게 될 것이다"(보스트롬 2017, p. 11).

안 된다. 그리고 이런 예측에 대한 평가나 태도 표명에 대한 찬반을 포함하는 토론은 "초지능을 이러이러한 것이라고 규정한다면"이라는 조건절의 제약 하에서 진행된다.[2)]

우리의 경험에 이미 친숙한 대상이거나 적어도 그 개념에 부합하는 뚜렷한 사례가 단 몇 개라도 존재하는 경우라면 사정이 다르겠지만, 초지능의 경우는 아직 적절한 의미의 사례(instance)가 존재하지 않는 경우다.[3)] 이 경우 우리는 아직 현실에 존재하지 않지만 원칙적으로 존재하게 될 수 있는 '초지능'을 상상을 통해 형상화하고 그렇게 상정된 속성들을 논의의 대상으로 삼을 수밖에 없다. 이것이 초지능에 대한 논의의 특수한 상황이다.

개략적으로 말하자면, 초지능은 앞에서 언급한 것처럼 인간의 지능을 뛰어넘는 기계지능(machine intelligence)[4)]을 뜻한다. 그러나 기계지능이 인간의 지능을 뛰어넘는다는 것이 정확히 어떤 경우를 가리키는 말인지 따질 필요가 있다. 그것은 인공물이 구현하는 지능이 특정한 기능의 차원에서 인간의 지적 능력을 능가한다는 것인가? 만일 초지능

---

2) 이것은, 뒤집어 말하면, 초지능의 개념을 제시하지 않고 진행되는 초지능 관련 전망이나 토론은 그 유의미성을 담보할 수 없다는 뜻이기도 하다. 이 논문은 보스트롬의 정의를 토대로 한 논의이다. 보스트롬의 정의는 합리적이지만, 가능한 유일의 정의는 아니다. 그러나 이 논문의 취지는 초지능을 규정하는 가능한 정의들의 스펙트럼을 두루 살피는 작업이 아니라 보스트롬의 정의가 지시하는 초지능의 전망을 평가하는 데 있다(이 논문을 심사한 두 심사위원의 논평 덕분에 이 점을 더 분명히 할 필요를 인식하게 되었다).

3) 만일 그와 같은 적절한 사례가 존재한다면, 서두에 거론된 세 물음 가운데 앞의 둘은 물음으로서의 지위를 상실할 것이다.

4) '기계지능' 대신 '인공지능'이라고 쓸 수도 있겠지만, 학술적인 의미에서 후자는 연구영역 혹은 연구활동을 지칭하는 개념이라는 점을 고려하여 그러한 애매성의 위험이 없는 전자의 표현을 사용한다. 어떤 이는 '합성지능(synthetic intelligence)'이라는 명칭을 사용하기도 한다. 애매성의 위험이 없는 곳에서는 이 글도 기계지능 또는 합성지능을 가리키는 표현으로 '인공지능'을 사용할 것이다.

의 개념을 그렇게 규정한다면, 초지능이 과연 출현할 것인지, 언제 그렇게 될 것인지는 의미 있는 물음으로 성립하지 않게 된다. 그런 의미의 초지능은 이미 오래 전에 개발되어 우리 주변에 산재하고 있기 때문이다. 5278×737이나 365.24÷27.5 같은 연산을 하도록 우리에게 문제를 제시하는 동시에 계산기에 앞의 식을 입력하면 우리가 숫자 두어 개를 적기도 전에 전자계산기는 셈의 답을 낸다. 다량의 자료를 기억했다가 재생하는 일이나 복잡한 자료를 특정한 차원에서 정리하고 비교하는 일 따위는 정확성이나 속도 어느 면에서 보더라도 사람보다 컴퓨터가 윗길이다. 따라서 특정 기능의 차원에서 인간의 능력을 능가한다는 것은 초지능을 규정하는 속성으로 적절하지 않다.

그렇다면 초지능은 '특정한 기능의 차원'이 아니라 '모든 기능, 혹은 모든 지적 능력의 차원'을 기준으로 규정되어야 할 것처럼 보인다. 이것은 닉 보스트롬이 '초지능'을 주제로 한 저서 『슈퍼인텔리전스: 경로, 위험, 전략』에서 제시한 초지능의 개념[5]과 부합하며, 이 논문은 3절 이하에서 보스트롬의 정의를 기준으로 초지능의 전망을 논할 것이다. 기계의 지능이 그것을 만드는 인간들의 지능을 능가하게 될 것인가? 이 논문은 이 물음을 따진다.

문법적 관점에서 이 물음에 대한 가능한 답은 "그렇다", "아니다", "알 수 없다" 세 가지다. 세 번째 답은 저 물음이 미래에 관한 물음이고 미래가 우리처럼 유한한 시공간적 존재의 손아귀에 주어져 있지 않다는 사실을 고려할 때 실패할 위험이 거의 없는 안전한 답이다. 그러나 그것은 그 '알 수 없음'의 세부 양태와 근거를 명시하는 논의가 없는 한 거기서 출발하여 유의미한 결론들로 나아가는 사유의 전개를 더 이상 기대하기 힘든 무기력한 답변이다. 이런 점을 고려하면서 우리는 "그럴 것이다. 왜냐 하면…"과 "아닐 것이다. 왜냐 하면…"의 두 가지 대답을 저울질할 필요가 있다. 이와 같은 평가는 기술의 발전 양상에

5) "사실상 모든 관심 영역에서 인간의 인지능력을 뚜렷이 상회하는 지능"(Ibid., p. 53). 원문은 "any intelligence that greatly exceeds the cognitive performance of humans in virtually all domains of interest"이다. 원문의 'greatly'를 고려하여 번역본의 표현을 보완하였다.

대한 예견처럼 보이지만, 이하의 논의 과정에서 드러나겠듯이 단순히 현상에 대한 예견이 아니다. 이 물음은 인공지능이라는 기술과 인간의 관계에 대한 철학적 성찰을 요구하며, 인공지능이라는 중요한 기술 영역을 우리가 어떻게 인식해야 할 것인가 하는 관점의 문제를 제기한다.

보스트롬은 초지능이 언제쯤 개발될 것인지에 관해서는 전문가들의 다양한 견해를 인용한다. 이 물음에 관한 전문가들의 견해는 일률적이지 않다.[6] 그런 지능이 아예 실현되지 않는다고 보거나 아니면 실현되더라도 아주 먼 미래에나 가능하리라고 보는 이들도 있다. 그러나 인간 수준의 지능(human-level machine intelligence)이 개발되는 시점에 대한 전문가들의 예견은 대략 2050년 전후 이삼십 년 사이에 분포한다.[7] 그는 "미래의 어느 날 우리가 인간의 일반 지능을 능가하는 기계 두뇌를 만들게 된다면"이라고 가정법을 써서 말했지만, 그의 저서는 그러한 지능이 실현될 것이라는 저자의 확고한 믿음을 드러낸다. 이제 이런 판단 혹은 예측의 타당성을 검토해보자.

2. 알파고가 이세돌보다 바둑을 잘 둔다?

이 일을 위해 먼저 따져보아야 할 것은 '우월함'(superiority)이라는 개

6) 도밍고스(Pedro Domingos)는 그의 저서 도밍고스 (2016)에서 역연역법(inverse deduction), 역전파(backpropagation), 유전자 프로그래밍(genetic programming), 베이즈 추론(Bayesian inference), 서포트 벡터 머신(support vector machine) 등 다섯 갈래의 머신러닝이 하나로 수렴됨으로써 '마스터 알고리즘'이 탄생할 수 있을 것이라고 전망한다. 컴퓨터공학자인 저자 도밍고스가 말하는 마스터 알고리즘은 인간 수준의 지능을 초월한다는 지위를 핵심 속성으로 내세우지 않는다는 점에서 보스트롬의 초지능과 차이가 있지만, 인공지능 기술이 지향하는 정점을 가리키는 개념이라는 점에서 유사하다.

7) 보스트롬 (2017), pp. 47-52.

념이다. 인간 수준의 지능을 뛰어넘는 지능을 가진 기계가 등장할 것인가 하는 물음이 상이한 지능 간에 성립하는 우월함의 관계를 함축하고 있기 때문이다. 우월함이란 어떤 것인가?

2016년 봄, 알파고는 이세돌에게 바둑을 이겼다. 적어도 수많은 뉴스와 논평이 그렇게 보도했고, 우리도 실제로 그런 식으로 평가한다. 그런데 실제로 알파고가 이세돌보다 우월한 능력을 지녔나? 누가 그렇게 묻는다면, 당연히 우리는 먼저 이렇게 되물어야 할 것이다. "**어떤 능력에서** 우월한지를 묻는 것이지요?" 우월하다는 것은 특정한 능력 또는 속성의 비교를 전제하는 관계이기 때문이다. 상대방은 아마도 "그야 물론 바둑을 두는 능력에서지요"라고 대꾸할 것이다. 그런데 우리는 이 지점에서 이세돌과 알파고 중 어느 편의 바둑 두는 능력이 우월한지 평가하기 위해 '바둑을 두는 능력'이 구체적으로 어떤 능력인지 돌아볼 필요가 있다. 그리고 이를 위해서는 바둑을 둔다는 것이 어떤 것인지(what it is to play Go) 살펴야 한다.

바둑이 '수담(手談)'이나 '난가(爛柯)'라는 별칭으로도 불린다는 점을 상기해보자. 그런 별칭에는 바둑이 바둑을 두는 사람 간의 소통과 교류라는 의미, 그리고 시간이 가는 줄도 모르게 만드는 특유한 종류의 즐거움 또는 여유 같은 의미가 담겨 있다. 종국에 가로세로 19줄씩의 바둑판 위에 만들어진 집들의 수를 세어 승부를 가리는 일은 중요하지만, 그런 계산이 바둑의 전부라고 보는 것은 편중되고 제한된 관점이다. 그런데 알파고는 바둑판을 사이에 두고 이세돌과 마주 앉아 있지도 않았고, 스스로 바둑돌을 놓지도 못했다. 물론 알파고는 그것이 도대체 누구와 무엇을 하는지도 의식하지 못했다. 중립적인 관점에서 본다면, 2016년 봄의 5번기 대국에서 알파고가 한 일은 '이세돌을 상대로 바둑을 둔 것'이라기보다 '어느 곳에 다음 수를 두면 승리의 확률이 극대화되는지를 계산하여 그 결과를 아자황에게 전달한 것'이라고 평가하는 것이 적절할 것이다.[8] 알파고라는 인공지능 프로그램 덕에

8) 이후 유명해진 이 바둑시합에서 이세돌 9단이 상대한 인공지능 프로그램 알파고에 관해서는 하사비스와 실버, 그리고 아자황 등이 공저하여 2016년 1월 네이처(Nature)지에 게재한 논문 Silver et al. (2016)을 참고하라.

하사비스(Demis Hassabis)가 이끄는 팀이 바둑이라는 특별한 게임의 형식을 매개로 인공지능 기술의 놀라운 잠재력을 보여줄 수 있었던 것은 부인할 바 없는 사실이지만[9], "알파고가 바둑을 두었다"는 진술은 아주 너그러운 비유와 생략의 맥락 속에서만 참된 진술로 용인될 수 있다.

어떤 이는 "인공지능이 그 깊고 어렵다는 바둑도 인간보다 더 잘 둔다고 인정하면 될 일을 교묘한 이야기로 회피하고 있는 것이 아니냐!"라고 핀잔하듯 따질 것이다. 그러나 이것은 인간의 자존심을 지키기 위한 변명과는 거리가 멀다. 이러한 논의를 통해 먼저 지적하고 싶은 것은 인공지능에 관한 사회적 담론에 일정한 의인화(anthropomorphizing)의 정조가 스며들고 있다는 사실이다. 인공지능이나 지능형 시스템에 관해 이야기할 때, 자주 우리는 인간과 다르면서도 인간의 마음에 상응하는 어떤 것을 가진 "낯선 주체"가 어느 새 우리 앞에 서 있다고 가정하고 있는 듯하다.[10] 그러나 전자지능(electronic intelligence)으로 작동하면서 우리와 **상호**작용하는 가운데 그 자신—또는 그것 자체—의 생각이나 감정을 우리와 **주고받는** 인공물이나 인공시스템은 아직 현실에 없다. 무엇인가가 나와 **서로의 생각이나 감정을 주고받는다**면, 그것은 적어도 하나의 존속하는 주체(enduring subject)이어야 한다. 예컨대 내 노트북컴퓨터가 만일 그러한 주체의 지위를 갖는다고 인정하려면 우리는 먼저 노트북컴퓨터에 관하여 철학의 주요 문제 가운데 하나로 꼽히는 인격동일성의 문제에 상응하는 문제를 해결했어야 한다. 그러나 그것은 희망적으로 보더라도 요원한 과제다.

어떤 전문가들은 "(여러 가지 난관에도 불구하고) 그것은 충분히 가능한 일"이라는 논조로 대꾸한다. 그러나 논자가 보기에 그들이 말하는 '가능성'은 '불가능하다는 것이 증명되지 않았다'는 것 이상의 무엇

9) 이미 널리 알려진 것처럼, 알파고가 보여준 힘의 배경은 심층학습(deep learning)과 강화학습(reinforcement learning)이라는 두 가지 기계학습의 경로를 성공적으로 활용한 데 있다.

10) 인공지능이나 인공지능로봇의 존재론적 지위에 관한 일반적이면서도 다각적인 논의는 이중원 외 (2018)을 참고하라.

이 아니다. 이런 종류의 가능성에 관한 이야기는 문학과 예술의 소중한 재료가 되지만, 사회의 현안을 다루는 바탕으로는 적합하지 않다. 인공지능 기술의 발달을 전망하고 검토하는 일은 오늘의 시점에서 기업, 시민사회, 국가, 글로벌 공동체 등 다양한 수준의 공동체의 미래를 좌우할 만큼 중요한 과제다. 그런데 이런 고찰과 그것에 근거한 사회적 차원의 관리(social management of technology)가 문학적 상상력을 작동원리로 삼게 된다면 위험한 일이다.

정대현은 인공지능의 발달과 그로 인한 도전이 현실의 영역 너머에 펼쳐진 가능성의 영역을 다루는 인문학 고유의 상상력에 외면해서는 안 될 과제들을 제시하고 있다고 말했다.[11] 그의 이러한 평가에 공감한다. 그런 인문학적 상상력은 자연과 사회의 현실을 넘어 최대한 자유롭게 작동할 필요가 있다. 그것이 인문학 고유의 가치를 극대화하는 길인 동시에 인문학의 사회적 기여를 증진하는 길이다. 그러나 그렇게 펼쳐진 상상력의 산물들은 현실을 해석하고 현실의 문제를 다루기 위한 생각의 토대를 풍부하게 만드는 요소인 반면, 그 자체 사회적 실천의 준거로 삼을 수 있는 것이 아니다.

인문학적 상상력의 산물들은 다양성의 빛을 발하고, 우리는 그것들에 통일성이나 수렴의 양상을 요구할 이유가 없다. 반면에 사회적 실천의 문제들은 비록 언제라도 더 나은 다음 해법에 자리를 내어줄 잠정적인 것일지라도 공유된 해법을 요청한다. 우리는 항상 그 시점에 동원할 수 있는 지식과 지혜를 모아 공유할 만한 최선의 해법을 찾고, 현실에 적용한다.

3. 지능의 우월함에 관한 수학적 해석

보스트롬은 초지능을 “사실상 모든 관심 영역에서 인간의 인지능력을 뚜렷이 상회하는 지능”[12]으로 정의한다. 그런데 이 정의는 그가 초지

11) 이는 2018년 8월 대만 국립정치대에서 열린 제4회 동아시아철학대회(CCPEA 2018)에서 정대현과 논자가 나눈 대화에서 따온 내용이다.

능을 바라보는 중요한 관점을 노정한다. 그리고 논자 역시 이 관점에 공감한다. 그것은 인공지능에 대한 평가가 우리의 관심과 결부되어 있다는 생각의 관점이다.[13]

아래에서 좀 더 상세히 서술할 인지능력 비교의 수학적 구도에서 보자면 이것은 N차원 공간에 분포하는 영역들 간의 비교로 번역할 수 있다. 보스트롬의 정의를 이러한 비교의 언어로 번역하자면, 초지능은 이 공간에서 인간의 인지능력에 상응하는 N차원의 부피를 고스란히 포함하는 N차원의 더 큰 부피에 대응된다. 그리고 이 공간을 결정하는 차원들은 우리의 관심과 결부된 것들이다. 우리의 관심과 무관한 그 밖의 차원은 고려되지 않는다. 다시 말해, 그런 차원은 방금 말한 공간의 구성에 관여하지 않는다.

5는 3보다 크고, 면적 10제곱미터의 침실은 10평 넓이의 거실보다 좁다. 대한민국의 국토 면적은 방글라데시보다 작고, 2016년 국내 총생산(GDP)은 방글라데시보다 크다. 우리는 이처럼 대한민국의 국토 면적과 방글라데시의 국토 면적을, 또 두 나라의 국내 총생산을 비교할 수 있지만, 대한민국의 국토 면적과 방글라데시의 국내 총생산을 비교하여 우열을 가릴 수는 없다.

우리는 한 변이 10센티미터인 정사각형과 지름이 10센티미터인 원 가운데 어느 편이 더 큰지 비교하여 전자가 더 크다고 판정할 수 있다. 그렇다면 한 변이 10센티미터인 정사각형과 지름 10센티미터의 공(球)은 어떤가? 둘의 관계는 대한민국의 면적과 방글라데시의 국내 총생산 간의 관계와는 달리 둘 다 미터로 환산된 공간적 크기라는 점에서 공통의 요소를 지녔지만, 그럼에도 불구하고 둘 중 어느 편이 더 큰지 말하는 일은 간단치 않다. 만일 전자와 같은 정사각형 모양의 구멍을 뚫어 놓았다면 후자의 공이 그 구멍을 통과할 수 있겠지만, 그렇다고 전자가 후자보다 큰 것은 아니다. 오히려 전자를 적절히 휘거나 접음으로써 후자의 공간에 포함시켜 넣을 수 있다는 점에서 후자가 전

12) 앞의 각주 6) 참조.

13) 이러한 점에서 보스트롬의 정의는 그가 이 부분에서 참고하고 있는 렉(S. Legg)의 정의나 굿(I. J. Good)의 정의와도 차별된다.

자보다 더 크다.[14]

초지능의 후보로 간주될 만한 몇 가지 인공지능 체계가 개발되었다고 가정해 보자. 이것들과 인간의 지능을 비교하여 우열을 가리는 작업은 어떤 방식으로 이루어질 수 있을까? 이세돌이 알파고를 상대로 바둑 시합을 한 것은 이러한 테스트의 한 형태였다고 볼 수 있다. 이세돌이 그 시합에서 4 대 1로 패했지만 우리는 알파고를 초지능이라고 평가하지 않는다. 뿐만 아니라 그런 평가는 이세돌과 겨룬 AlphaGo Lee보다 한참 더 강한 역량을 지녔다는 AlphaGo Zero에 대해서도 똑같이 적용된다.[15] 알파고가 바둑 시합에서 승리의 확률을 높이는 수를 찾는 데는 세상 어느 인간보다도 유능하지만, 할 줄 아는 것이 그것 하나뿐이기 때문이다.

그렇다면 무엇 무엇을 인간보다 더 잘 해야 인간의 지능을 능가했다고 할 수 있을까? 앞에서 언급한 보스트롬의 기준은 이 물음에 관하여 "사실상 모든 관심의 영역"이라고 답한다. 다시 말하자면, 우리가 의미 있다고 생각하는 모든 영역이 지능을 저울질하는 비교의 차원들을 형성한다는 것이다.

체스나 바둑을 잘 두는 능력이 이에 해당할까? 최근 인공지능 기술 발전의 역사를 고려하면 우리는 이 물음들에 당연히 그렇다고 답하게 될 것이다. 그러나 체스나 바둑에서 판세를 읽고 이기는 수를 찾는 일이 인간이 삶을 영위하는 데 필수불가결하거나 핵심적인 역량이냐고 물으면 대답은 불분명해진다. 체스, 그리고 이어서 바둑이 인공지능 연구의 도전 과제가 된 것은 그것들이 제한된 조건 속에서 문제 해결(problem solving under restricted conditions)의 모색이라는 인지 연구와 인공지능 연구 공통의 관심을 구체화할 수 있는 모형이 될 수 있었

14) 보자기나 종이를 구기거나 뭉쳐서 공 안에 넣는다고 생각해보면 되겠다. 보자기나 종이는 실제로 상당한 두께를 가졌다는 점에서 2차원 대상이 아니라 3차원 대상이기 때문에 공 안에 구겨 넣을 수 있는 종이가 한정되겠지만, 두께가 0인 진정한 2차원 대상이라면 그 크기가 어떻든 상관없이 예컨대 테니스공 크기의 3차원 부피 안에 구겨 넣을 수 있다.

15) AlphaGo Zero에 관해서는 Silver et al. (2017)을 참조하라.

기 때문이라고 보는 것이 적절하다. 그것들은 말하자면 실세계에서 발생하는 문제 해결의 상황을 특정의 제한된 방식으로 간략화한 모형으로 간주된다. 이렇게 해석하면 체스나 바둑을 잘 두는 능력이 왜 "관심의 영역"에 포함될 만한지 수긍할 수 있을 것이다.

그렇다면 브릿지나 포커 같은 카드 게임을 잘 하는 능력도 마찬가지일까? 체스와 바둑을 인정한 터에 그것들을 아니라고 내치려면 그럴 듯한 구실이 필요할 텐데, 그런 구실은 얼른 눈에 띄지 않는다. 게다가 그런 게임마다 게임에 참여하는 사람이 발휘하게 되는 미세하면서도 고유한 지적 능력이 있을 법도 하다. 그러나 그렇게 온갖 종류의 게임을 잘 하는 능력까지 초지능이 포섭해야 할 "관심의 영역"에 포함시키자니 초지능을 만드는 사람들에게 한편으로는 자질구레하니 잡다하고 다른 한편으로는 그런 이유로 도무지 해내기 어려울 일을 의뢰하고 있는 것 같은 생각이 든다.

어려움은 이 수준에 그치지 않는다. 관심의 대상이 될 만한 인지능력의 영역이 카드 게임이나 보드 게임을 하는 능력에 국한되지 않을 것이기 때문이다. 조금 더 생각해보자. 시를 잘 짓는 능력, 또 같은 이야기라도 한층 더 그럴 듯하고 재미나게 늘어놓는 능력은 어떠한가? 또 오페라 아리아를 멋지게 부르는 능력이나 신체 각 부분의 동작을 조율하여 감탄을 자아낼 만큼 아름답게 춤을 추는 능력은 어떤가? 이런 능력은 신체의 부분들과 결부된 능력이고, 그렇기 때문에 인간 같은 성대와 흉부, 그리고 유연한 팔다리를 가지지 못한 인공물에 기대하기 힘든 능력이다. 그러나 그런 이유 때문에 그런 능력들을 초지능이 뛰어넘어야 할 인지능력의 목록에서 제외하는 것은 일종의 본말전도로, 옳지 않다.

실제로 논자는 인공지능에 대한 지금처럼 고조된 관심 속에서 어느새 우리 스스로 '인간의 인지능력'의 범위를 기계지능이 흉내 낼 수 있는 능력의 영역으로 한정하게 될 개연성이 있다고 본다. 그러한 개연성은 결과적으로 인간 인지의 풍부하고 미세한 영역들 가운데 상당한 만큼이 무관심의 그림자 속으로 밀려들어가 버릴 위험을 함축한다. 이것은 인간이 지닌 정신문화의 영역을 스스로 축소하고 빈곤하게 만

드는 일일 뿐만 아니라 인공지능 같은 새로운 기술의 수용과 활용이 필연적으로 수반할 까닭이 없는 종류의 불행한 진행이다. 따라서 우리는 이런 위험을 인간이라는 종의 집합적 관점에서 경계하고 메타적 사고를 통해 조절할 필요가 있다.

이 지점에서 확인하게 되는 것은 보스트롬의 정의가 말하는 "사실상 모든 관심의 영역"이 무궁무진하다는 사실이다. 이를 인지능력의 공간에 관한 서술로 환산하여 말하면, 이 공간을 구성하는 차원의 수가 무궁무진하다는 것이다. 이러한 사정은 설령 수백 가지 카드 게임의 종류를 몇 가지로 크게 분류하고 각각의 분류군에 속한 게임과 관련된 인지능력의 차원들을 추출하는 방식으로 인지능력의 공간을 단순화하는 경우에도 달라지지 않는다.

제약은 두 방향에서 온다. 하나는 두 종류의 카드 게임이 거의 비슷한 방식으로 작동하더라도 세밀한 부분에서 상이한 규칙이 적용되는 경우, 양자가 플레이어들에게 동일한 인지능력을 요구한다고 판정하기 어렵다는 점이다.[16] 이런 어려움은 예를 들어 카드 게임과 관련된 인지능력에 대하여 그것의 차원들을 분별하는 일의 정당한 원리가 무엇인가 하는 물음을 끌어들인다. 다른 하나는 설령 게임들을 적절한 군으로 묶고 각 묶음에 해당하는 인지능력의 차원들을 확인하는 데까지 인정했다 하더라도 새로운 인지능력의 차원이 새로운 게임에 업혀 얼마든지 등장할 수 있다는 점이다.[17] 결과적으로, 카드 게임과 결부된 인지능력의 공간만 해도 확정하기 어려울 뿐만 아니라 점차로 증가하는 수의 차원들을 가진 공간이라는 사실이 다시 한 번 확인된다.

지능을 주된 연구주제로 다루어 온 심리학에서는 인간의 지능을 적절한 수의 요소들로 분석하는 작업이 수행되었다. 지능에 관하여 제시

16) 예컨대 동일한 종류의 카드게임에서도 조커를 포함시켜 진행하는지, 또 그렇게 포함시킨 조커에 어떤 기능을 부여하는지에 따라 게임 전개의 양상이 달라진다. 조커에 부여된 특별한 기능(또는 약점)의 성격에 따라 카드게임에 참여하는 인지 주체가 그런 조커가 없었던 때와는 다른, 새로운 인지능력을 발휘하도록 유도되는 것은 자연스러운 일일 것이다.

17) 바로 앞의 각주 참조.

된 영향력 있는 심리학 이론으로 평가되는 스턴버그(R. J. Sternberg)의 '인간 지능의 3원론'(triarchic theory of human intelligence)과 가드너(Howard Gardner)의 '다중 지능(multiple intelligences)' 이론이 이에 해당한다. 그러나 이런 지능이론을 활용하더라도 이런 상황은 근본적으로 바뀌지 않는다.[18] 예를 들어, 가드너가 단일한 지능 요소로 분석한 언어 지능은 다시 여러 차원의 하위 지능으로 분석될 수 있다. 예컨대 언어를 이해하는 지능과 언어 표현을 산출하는 지능은 단일한 차원의 능력이라기보다 상이한 지능의 차원으로 분석하는 것이 적절하다. 또 언어 이해만 해도 모국어를 이해하는 것과 외국어를 이해하는 것은 서로 구분되어야 할 상이한 인지능력을 요구하는 활동일 수도 있고, 똑같이 모국어를 이해하는 능력이라고 해도 시적 표현에서 화자가 의도한 바를 간취하는 능력과 복합적인 정보가 담긴 문장의 내용을 빠르고 정확하게 해득하는 능력은 단일한 차원의 지능으로 보기 어려울 것이다.

4. 인지능력 공간의 관점에서 읽은 초지능의 전망

이상의 논의는 이 논문의 물음과 어떻게 연결되는가? 두 지능의 역량을 비교하는 일은 그것들의 역량 범위를 지금 이야기하고 있는 N차원 인지능력 공간에서 두 지능의 부피를 저울질하는 일로 환산할 수 있을 것이다. 알파고와 왓슨 포 온콜로지의 지능을 그렇게 비교하기는 어렵다. 두 지능이 (중첩되는 몇 개의 차원은 있을 수 있겠지만) 서로 다른

18) 스턴버그는 지능을 구성적-분석적(Componential-Analytic) 능력, 경험적-창의적(Experiential-Creativity) 역량, 실천적-맥락적(Practical-Contextual) 능력의 세 '하위 이론'(subtheories)의 영역으로 분석한다. 한편 가드너는 인간의 마음을 구성하는 지능의 요소들을 음악적 지능, 시각적-공간적 지능, 언어 지능, 논리적-수학적 지능, 신체-운동 지능, 인간관계적(interpersonal) 지능, 자기성찰적(intrapersonal) 지능, 자연친화적 지능, 그리고 실존적 지능의 아홉 가지로 분석한다.

인지능력의 차원들을 구현하고 있기 때문이다. 그러나 보스트롬이 말하는 것과 같은 초지능이라면 특정한 목적과 기능에 국한되지 않는 일반지능(general intelligence)의 속성을 지니는 것일 터이니, 인지능력의 다차원 공간에서 그것이 점하는 부피와 인간 지능이 포괄하는 부피를 비교함으로써 두 지능체계 간의 우열을 객관적으로 가릴 수 있을 것이다. 논자는 이러한 비교가 보스트롬이 초지능을 "사실상 모든 관심 영역에서 인간의 인지능력을 뚜렷이 상회하는 지능"이라고 정의할 때 상정한 상황과 다르지 않다고 본다.

그렇다면 앞 절의 논의가 이 지점에서 함축하는 것이 무엇인가? 그것은, 간단히 말하자면, 인간의 인지능력을 적절히 한정된 수의 차원들을 가진 다차원 부피로 규정할 수 없다는 것, 그리고 인간의 인지능력을 특징짓는 새로운 차원들이 생겨난다는 것이다. 이는 문자 그대로의 의미에서 인간의 지능을 초월하는 초지능, 다시 말해 인지능력의 다차원 공간에서 인간의 지능에 상응하는 부피를 고스란히 그 일부로 포함하는 부피를 지닌 인공지능을 개발하는 것이 근본적으로 어려운 과제라는 운명을 예견하게 한다. 앞의 3절에서 논한 것처럼 예컨대 3차원 도형은 제아무리 부피가 큰 경우에도 4차원 도형의 부피를 포함할 수 없다는 사실을 생각해보라. N개 차원의 인지적 능력에서 인간의 수준을 뛰어넘도록 만들어진 인공지능이라고 해도 거기 포함되지 않은 (N+1)번째의 새로운 차원을 가지게 된 인간의 지능을 포섭할 수 없는 것이다. 이것은 이 논문의 물음에 대한 부정의 답을 함축한다.[19]

이 상황에서 초지능의 전망을 긍정적으로 보며 지지하는 사람들은 두 종류의 반론을 구사할 수 있을 것으로 생각된다. 그 중 한 갈래는 별 힘이 없지만, 다른 한 갈래는 유의해서 살필 필요가 있다. 논자가 보기에 힘이 없는 한 갈래는 "나날이 발전하고 있는 자료 기반의 기계학습(data-driven machine learning)을 통해 인공지능이 인간 지능의 모든 차원들을 추출해내는 일이 결국엔 가능해지지 않겠느냐?"는 반론이

19) 뿐만 아니라, 3절의 내용을 고려할 때, 각 차원의 구체적인 크기가 어떻든 상관없이 (N+1)차원의 부피가 N차원의 부피를 포함할 수 있다는 의미에서 후자보다 더 크다는 평가가 도출된다.

다. 이 역시 새로운 기술, 특히 인공지능의 미래와 관련하여 흔히 활용되고 있는 '가능성 논변'의 한 사례에 해당한다. 그리고 원칙적 불가능을 논증하는 일은 어렵지만, 그것이 어째서 사실상 불가능에 해당하는 어려운 일인지는 밝힐 수 있다.[20]

다른 한 가지 반론은 "발달된 인공지능에서 인간의 지능에 존재하지 않는 인지능력의 차원이 발현(emergence)[21]될 수 있지는 않을까?"라는 것이다. 이것은 우리가 따지고 있던 '포함관계'의 승부를 일종의 교착 상태에 이르게 하는 반론이다. 그것은 인공지능이 인간 지능의 모든 차원을 포섭하는 문제를 내버려둔 채 전자가 후자에 없는 새로운 종류, 새로운 차원의 역량을 가지게 될 가능성을 공략한다. 그것은 이 논문이 전개한 비교의 관점에서 인공지능이 인간의 지능을 초월할 수 있다고 주장하는 대신, "인공지능에서도 인간이 이해할 수 없고 그래서 그것을 어떻게 다루어야할지도 알지 못하는 새로운 차원의 지적 역량이 생겨날 수 있다"라고 맞으름장을 놓는다. 이것은 일종의 논점 일탈의 오류라고 평가할 수 있지만, 단순히 그렇게만 평가하고 지나간다면 우리가 인공지능의 미래에 관하여 생각해야 할 중요한 사항을 놓치게 될 위험이 있다.

이러한 반론 앞에서 우리는 인공지능이 왜 생겨났고 점점 더 높은 수준의 인공지능이 개발되는지, 인공지능의 존재의미를 돌이켜볼 필요가 있다. 인공지능은 우리, 곧 인간 사회가 연구개발하고 있는 테크놀로지의 한 영역이다. 그것은 오늘의 시점에서 최대의 자본과 최고 수준의 유능한 인력들을 소모하고 있는 인간 활동의 영역이기도 하다. 막강한 자본력을 가진 기업들이 그렇게 내달리는 이유는 기업이윤의 전망이라는 관점에서 이해하는 것이 적절할 것이다. 그러나 우리는 특정 기업의 관점에서가 아니라 인간이라는 종의 집합적 관점에서 우리 자신이 인류 공동의 자원을 소모하면서 하고 있는 일의 합리성에 관심을 가지지 않으면 안 된다. 이것은 구글이나 아마존 같은 기업들의 투

20) Ko (2017)을 참고하라.

21) 이것은 환원(주의)에 관한 논의에서 오랫동안 다뤄져온 '창발'의 개념에 상응한다.

자일 뿐만 아니라 인류 공동의 투자이기 때문이다. 이러한 사회적 맥락에 관한 논의는 이 논문이 다루는 문제의 가장자리에 걸쳐 있지만, 그렇다고 덜 중요한 것은 결코 아니다. 인공지능 같은 핵심 기술—또는 인공지능이 연루된 수많은 핵심 기술들—은 인류의 중대한 투자이고, 개인의 삶에서도 그렇듯 중요한 투자를 성공으로 귀결시키는 것이 인류의 존속과 번영을 좌우한다.

이처럼 중요한 투자에 참여하는 공동투자자의 일원으로서 우리가 우선적으로 챙겨야 할 것은 이 투자의 최종 목표가 무엇인지를 스스로 분명히 인식하는 일이다. 테크놀로지의 연구개발에 관한 모든 평가와 조정도 그런 목표의 관점에서 이루어질 것이고, 그것이 마땅하다. 그 최종적 목표는 “인간의 지능을 뛰어 넘는 인공지능을 만드는 것”도, “인간이 신이 되는 것”도 아니다. 이미 이천 삼백여 년 전 아리스토텔레스가 간명하게 말한 것처럼, 그리고 세계적인 공학단체들의 윤리강령 제1조가 공통으로 가리키고 있는 것처럼, 그것은 기술을 개발하고 운용하는 우리 자신, 인간의 행복한 삶이다.

기술 산품으로서의 인공지능 역시 오로지 인류의 행복한 삶에 기여하는 한에서 존재의미를 획득할 것이다. 그리고 이런 관점에서 볼 때, 우리 같은 인간들이 이해할 수 없는, 그래서 도무지 조절할 수도 없는 역량을 지닌 인공지능은 기술의 존재의미와 부합하지 않는다. 우리가 이해할 수도 조절할 수도 없는 어떤 것이 우리의 삶을 풍요롭고 행복하게 해줄 것으로 생각하는 것은 합리적 사고가 아니라 일종의 종교적 사고다. “초지능이 너희를 삶의 모든 질고에서 구원하리라”라는 메시지를 내세운 종교가 생겨날 수 있을 것이다. 그것은 여느 사이비종교의 메시지와 다름없는 수준의 메시지이지만, 거기에 현혹되는 사람들도 함께 생겨날 것이다.

5. 인공지능의 존재 지위

2016년 전후로 생산된 유럽연합의회의 문건에 나타난 ‘전자인

격’(electronic person)의 개념과 제안이 인공지능이나 인공지능로봇의 존재 지위에 관해서 상당한 논의를 촉발했다. 어떤 이들은 이것이 인공지능이나 인공지능로봇도 마땅히 인격체로 간주해야 한다는 주장에 대한 신뢰할 만한 수준의 공적 정당화에 해당한다고 여기기도 한다. 그러나 이에 관하여 우리는 적어도 두 가지를 분명히 인식할 필요가 있다.

첫째는 이 문제에 관한 ‘진보진영’이 의지하는 원래의 문건도 전자인격의 위상을 부여할 만한 인공물의 속성에 관해 모호하지만 상당히 높은 수준의 조건[22]을 상정하고 있다는 사실이다. 다시 말해 전자인격의 지위가 원칙적으로 수용된다 하더라도 그 문턱(threshold)의 구체적인 조건에 관한 치열한 토론이 필요할 수밖에 없다. 둘째는 더 중요한데, 그것은 결국 이 문제가 과학적 진위 판단의 문제가 아니라 사회적 관점으로부터의 평가와 결단에 걸린 문제라는 사실이다. 만일 우리 사회가 특정한 속성을 갖춘 인공지능로봇을 그것이 창출한 이윤에 대한 권리를 가지고 그것이 유발한 피해에 대해서는 책임의 의무를 지는 인격체로 인정한다면, 그 이유는 그것이 존재론적 진리에 부합하는 것이어서가 아니라 그렇게 인정하는 것이 사회 전체의 이익에 부합하는 길이기 때문이다.[23]

테크놀로지는 조물주나 자연이 아니라 우리 인간이 스스로를 위해 만들어 온, 그리고 앞으로도 만들어갈 대상이다. 우리는 이런 대상에 대하여 관망자나 분석가의 태도가 아니라 제작자의 관점을 취해야 한다. 허버트 사이먼(Herbert Simon)의 통찰을 응용하여 말하자면, 인공지능이 장착된 로봇시스템 같은 인공물은 인간이 만드는 것이면서도

---

22) “the most sophisticated autonomous robots”(DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL), p. 12). 이후 2017년 2월 발표된 관련 문건 “European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))”도 참조할 것.

23) 유럽의회 내에서 이 문제를 둘러싼 찬반논쟁과 토론이 여전히 진행 중이다.

인간의 내적 환경, 즉 우리의 신체적, 정신적 속성들과 인간 생존의 환경으로서의 세계 사이에서 접속면의 성격을 띤다.[24] 인공물들은 그런 관계 속에서 인간의 생존과 번영을 제약한다. 이 접속면을 사려 깊게 가꾸고 조절하는 능력이 우리, 즉 인간이라는 종의 성쇠를 좌우할 것이다.

인간은 인공지능, 그리고 그것을 활용한 로봇시스템 같은 인공물의 제작자다. 제작자는 무엇을 어떻게 만들 것인지 고심하고, 그 고심의 결론을 자신의 활동을 통해 실현한다. 인간은 그렇게 제작한 산물에 대하여 그것을 그렇게 만든 자로서의 책임을 진다. 또 공학의 다른 산물들도 그렇지만, 특히 기계학습의 과정을 통해 최초의 속성과 다른 속성들을 획득하게 되는 인공지능의 특성을 고려할 때, 그런 인공물을 관리하고 사용하는 이들에게도 응분의 권한과 책임이 있다. 무엇을 어떻게 만들지 뿐만 아니라 만든 것에 어떤 의미, 혹은 지위를 부여할지도 우리가 결정할 사항이다. 이런 결정과 그것에 부수하는 법, 제도, 윤리규범 등은 첨단의 산물을 설계, 제작하는 일부 공학자나 기업의 소관이 아니라 사회적 토론과 합의를 통해 결정하고 실행해야 하는 사안이다.[25] 그러나 결국 사회적 합의에 맡겨야 할 사안이라고 해서 철학이 할 일이 없는 것이 아니다. 오히려 최대한의 합리성을 머금은 합의에 도달할 수 있기 위한 이론적 논구가 필요하다.[26]

---

24) Simon (1996), 1장의 논의 참조. 단, 여기서 사이먼이 말한 접속면(interface)은 인공물과 세계의 접속면이었다. 이 논문에서 논자는 그의 접속면 개념을 원용하면서 '인공물의 설계자-제작자-사용자인 인간과 세계 사이에 놓인 접속면으로서의 인공물'이라는 구도를 설정한다.

25) 앞에서 언급한 사이먼의 관점에서, 이런 법이나 제도 역시 인공물의 범주에 포섭된다.

26) 2018년 2월 고등과학원(KIAS)에서 열린 토론회 <인공지능의 도전, 철학의 응전>에서 인공지능이 도대체 지능인가에 대하여 토론에 참여한 철학자들의 상이한 견해가 충돌했다. 인공지능은 진정한 의미에서 지능조차 아니라는 주장과 지능이 인간을 비롯한 일부 동물에 국한된 것이라는 편협한 생각을 이제는 버려야 한다는 주장이 제시되었다. 토론회에서 이런 견해의 양끝은 각각 이종관과 신상규가 대변하였다. 이종관 (2017), 신상규 (2017)

인공지능이 동물 신체에 귀속된 것이 아니라는 이유만으로 그것이 지능일 수 없다고 판정하는 것은 '인공지능'이라는 개념어가 이미 그것이 '지능'임을 보여준다고 판정하는 것 못지않게 부적절해 보인다. 둘 다 인공지능이 진정한 지능인가, 또는 적어도 장차 지능이 될 수 있는가 하는 물음에 대한 토론을 원천봉쇄한다는 점에서 그러하다. 그러나 우리는 지능이 본래 동물 신체에 귀속된 능력을 가리키는 개념이라는 점을 여전히 유념할 필요가 있다. 지능은 그것을 가진 주체의 존속과 번영의 성패를 좌우한다. 지능은 둘레세계 즉 외부 환경을 이해하는 능력이고, 이러한 지능의 관념은 그 환경의 한가운데에 놓인 주체의 존재를 요청한다. 이 주체는 우리가 지능이라고 부르는 속성을 활용하여 그것 자체 혹은 자신의 관점에서 세계를 읽고 해석한다.

현행의 인공지능 담론에서 우리는 인공지능과 인간의 지능을 비교하면서도 인공지능이 누구의 지능인지, 혹은 무엇의 지능인지 따져 묻지 않고 있는 것처럼 보인다. 그러나 그 물음은 꼭 필요하다. 지능은 무엇인가의, 혹은 누군가의 지능이기 때문이다.[27] 다시 말해, 어디엔가 지능이 현존한다면 거기에는 지능을 가진 존재, 즉 지능적 존재(intelligent being)가 있어야 한다. 그리고 지능은 바로 그 존재의 운명을 좌우한다. 우리가 이제껏 알고 있는 지능은 생명체의 것, 특히 동물의 것이다. 그리고 그것은 뇌를 중심으로 하는 신경체계를 기반으로 구현되는 능력이다. 반면에, 서두에 인용한 보스트롬의 말에서도 드러나듯, 일반적으로 인공지능은 기계에 구현된다.

기계가 보이는 지능적 현상이 진정한 지능의 존재를 의미하는지 우리는 묻고 따질 수 있고, 이 물음 역시 결국엔 개념 사용에 관한 결정의 문제에 귀착할 것이다. 그러나 우리는 그 결정이 충족해야 하는 조건들을 하나씩 찾아냄으로써 그것이 임의적인 결정이 아니라 최대한 객관적인 결정이 되도록 유도할 수 있다. 예컨대, 앞에서 우리는 지능의 존재를 인정하기 위해서는 적어도 그것을 가진, 존속하는 주체가

참조.

27) 인공지능의 몇 가지 정의에 관한 러셀과 노빅의 분석(러셀 외 2016, 1.1 인공지능이란 무엇인가?)을 참조하라.

확보되어야 한다는 것을 확인했다. 자연은 아주 오랜 시간을 들여서 그런 것을 만들어냈다고 생각된다. 진화가 낳은 수많은 종류의 생명체 가운데서도 아주 제한된 일부만이 지능을 가진 주체들이라고 인정된다. 컴퓨터나 로봇이 애초에 그런 주체가 될 수 없도록 만드는 본질적 배제의 이유가 없다 하더라도, 어떤 로봇을 그런 주체로 인정할 수 있기 위해서는 그것이 어떤 분명하고 적극적인 근거에서 자케-드로(Pierre Jaquet-Droz)가 만든 소년[28]이나 스마트폰의 게임 어플리케이션에 등장하는 캐릭터와 달리 존속하는 주체인지에 대한 소명이 있어야 할 것이다.

6. 맺는 말: 제어가능성의 중요성

이 논문은 상이한 기원과 특성을 지닌 지능 간의 우월성을 저울질하는 문제를 다루었다. “인공지능은 그것의 창조자를 능가할 것인가?”라는 이 논문의 물음은 우열의 비교, 다시 말해 어떤 힘겨루기의 구도를 담고 있다. 문득, “우리는 왜 이처럼 우열의 관점에서 사물들을 비교하는 것일까?”하는 메타적 물음이 떠오른다. 어쩌면 그것은 우리가 가진 일종의 동물적 본능에 기인하는 것인지도 모른다. 보스트롬은 ‘슈퍼인텔리전스’를 논하지만, 우리 중 누군가는 “그렇다면 다시 그것을 뛰어넘는 울트라슈퍼인텔리전스는?” 같은 물음을 떠올리며 가슴 두근거리는 기대감과 더불어 우리를 압도하는 힘에 대한 두려움이 뒤섞인 감정을 느끼게 될 지도 모르겠다. 논자는 그것이 더 강한 존재가 약한 존재를 잡아먹고 지배하는 생존경쟁의 역사가 우리에게 남긴 동물적 습성의 유산이라는 설명이 이에 대한 하나의 그럴 듯한 해명이라고 본다.

그런데 이 지점에서 한 가지 분명히 짚어두어야 할 것은, 우월한 존재라고 해서 열등한 존재를 항상 통제할 수 있는 것은 아니라는 사실

28) 18세기의 시계제조업자이면서 뛰어난 기계설계자였던 자케-드로가 1770년 전후로 제작한 자동인형. 스위스 뇌샤텔의 ‘예술과 역사 박물관’(Musée d'Art et d'Histoire de Neuchâtel)에 전시되어 있다.

이다. 『슈퍼인텔리전스』에서 보스트롬의 관심은 초지능으로 인한 파괴적 위험에 있다.[29] 그러나 현실에서 훨씬 더 개연성 높은 위험은 통제가능성의 한계에 내재한다. 그런 까닭에 우리는 현실의 맥락에서 인공지능에 관한 논의의 역량을 통제가능성 혹은 제어가능성(controllability)의 문제에 집중할 필요가 있다. 제어가능성의 열쇠는 어디에 있는가?

인류가 최소한 수천 년 전부터 사용해 온 물레방아와 기중기를 생각해보자. 물레방아는 마을의 어느 아낙이나 장정보다 힘차게 그리고 꾸준히 그 일을 하고 기중기만큼 무거운 돌을 들어 올리는 장사도 없는데도 우리는 물레방아가 곡식을 제분하는 능력에서 인간을 능가했다거나 기중기가 무거운 물건을 들어 올리는 일에서 인간의 능력을 뛰어넘었다고 말하지 않는다. 물레방아나 기중기가 숭배의 대상이 되었던 흔적도 찾기 힘들다. 그런 이유는 무엇일까? 논자의 생각에 그 이유는 두 가지로 분석할 수 있다. 하나는 그것들이 그 자체로 그런 일을 해내는 주체로 이해되는 것이 아니라 자신들이 만든 도구로 인식되었고, 그리하여 그런 것들이 지닌 힘과 하는 일이 인간들 자신이 지닌 힘과 성과의 확장으로 인식되었기 때문이다. 또 한 가지 이유는, 그러한 초인적 작용의 과정이 비록 근대과학의 방식으로는 아니더라도 어떤 합리성의 수준에서 이해되어 있었기 때문이다.[30]

여기서 우리는 인공지능 체계나 인공지능로봇의 경우에 적용할 만한 가르침을 얻는다. 첫째는 새로운 기술이 어떤 방식으로 신기하고 강력한 것이라 할지라도 그것이 우리가 만들어 사용하는 도구라는 사실을 정확히 인식하는 것이 필요하다는 메시지다. 제어가능성과 더 직접 관련된 둘째 가르침은 기술의 작동 원리에 대한 이해가 제어의 선행 조건이 되리라는 것이다. 우리는 이해하지 못하는 것을 제대로 제

29) 논문의 첫머리에 인용된 부분에서 초지능이 실현된 시대에 인간들이 처할 운명을 오늘날 인간의 처분에 달린 고릴라들의 운명에 견주어 묘사했던 것을 상기하라.

30) 이러한 전(前)학문적 이해에 대해서는 'Protophysik'에 관한 로렌첸(Paul Lorenzen)과 야니히(Peter Janich) 등의 논의를 참고할 만하다. Janich (1997) 참조.

어할 수 없기 때문이다. 조금 더 응용하여 말하자면, 우리는 제대로 이해할 수 있는 범위에서 기술을 개발하여 활용해야 한다.

『마스터 알고리즘』의 마지막 부분에서 저자 도밍고스는 다음과 같이 말한다. 그리고 이는 인공지능의 위험과 관련하여 적절한 통찰을 표현하는 것처럼 보인다. "사람들은 컴퓨터가 너무 똑똑해져서 세상을 지배할 거라고 걱정하지만, 실제로 나타난 문제는 컴퓨터가 [아직] 너무 멍청하고 그런 컴퓨터가 이미 세상을 지배하고 있다는 것이다."[31] 아직 너무 멍청한데도 불구하고 그런 컴퓨터가 이미 세상을 지배하고 있는 것처럼 보인다면, 그 이유는 컴퓨터가 가진 인공지능의 힘에 있는 것이 아니라 그것을 다루는 우리의 역량이나 방식에 문제가 있기 때문이다. 이 논문이 논증한 것처럼 지능의 모든 차원에서 인간 지능의 수준을 뛰어넘었다는 의미의 초지능이 불가능하다고 해도, **우리가 적절히 제어하지 못하는 인공지능 기술로 인해 인간이 마치 인공지능의 지배를 받는 열등한 존재인 것 같은 느낌을 주는 집단적 체험을 하게 되는 일**은 충분히 가능하다. 초지능에 관한 토론이 제공하는 중요한 지혜는, 실질적인 문제가 인공지능 그 자체가 얼마나 똑똑하고 강력한가에 있다기보다 우리가 그것을 얼마나 **현명한 방식으로** 만들고 또 다루는가에 있다는 인식일 것이다.

31) 도밍고스 (2016), p. 456.

참고문헌

고인석 (2018a),「기계와의 공생: 철학적 도전과 고민」, 한국연구재단 언론홍보협의회 세미나 <인공지능 시대 인간과 기계의 공생: 철학적 성찰과 해법> 발표.

고인석 (2018b),「미지의 기술을 대하는 태도: 과학기술학이 할 일」, 제4회 과학학 연합학술대회 자료집.

도밍고스, 페드로 (2016)『마스터 알고리즘: 머신러닝은 우리의 미래를 어떻게 바꾸는가』, 강형진 옮김, 비즈니스북스.

러셀, 스튜어트, 피터 노빅 (2016),『인공지능: 현대적 접근방식 1』, 류광 옮김, 제이펍.

보스트롬, 닉 (2017),『슈퍼인텔리전스: 경로, 위험, 전략』, 조성진 옮김, 까치.

신상규 (2017),「인공지능, 새로운 타자의 출현인가?」,『철학과 현실』 112권, pp. 155-178.

이종관 (2017),『포스트휴먼이 온다: 인공지능과 인간의 미래에 대한 철학적 성찰』, 사월의책.

이중원, 박충식, 이영의, 고인석, 천현득, 정재현, 신상규, 목광수, 이상욱 (2018),『인공지능의 존재론(포스트휴먼 시대의 인공지능 철학 01)』, 한울아카데미.

European Parliament 2014-2019 (2016), “DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))”, http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN (검색일 2019. 07. 09.)

European Parliament 2014-2019 (2017), “European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL)”, http://www.europarl.europa.eu/RegData/etudes/STUD/20

16/571379/IPOL_STU(2016)571379_EN.pdf (검색일 2019. 07. 09.)

Gardner, H. (1993), *Multiple Intelligences: The Theory in Practice*, Basic Books.

Janich, P. (1997), *Das Maß der Dinge: Protophysik von Raum, Zeit und Materie*, Suhrkamp.

Ko, I. (2017), “What the Unsupervised Learning Would Teach Us (and What Not)”, Paper read at the 8th Asia-Pacific Conference for the Philosophy of Science, National Chung Cheng University, Taiwan.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel & D. Hassabis (2016), “Mastering the game of Go with deep neural networks and tree search”, *Nature* 529: pp. 484 - 89.

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel & D. Hassabis (2017), “Mastering the Game of Go without Human Knowledge”, *Nature* 550: pp, 354-59.

Simon, H. (1996), *The Sciences of the Artificial* (3rd ed.), MIT Press.

Sternberg, R. J. (1988), *The Triarchic Mind: A New Theory of Human Intelligence*, Viking.

논문 투고일	2019. 07. 09.
심사 완료일	2019. 07. 23.
게재 확정일	2019. 07. 23.

An Estimation of the Realizability of Bostromian Superintelligence

Insok Ko

This paper prospects the future of machine intelligence and examines whether the “superintelligence” that surpasses the human intelligence will be realized. It suggests a mathematical framework for assessment of whether a machine intelligence has reached such a level, on the basis of Nick Bostrom's definition of superintelligence. One can measure and compare the volumes that each intelligent system contains in the multidimensional space, in which every dimension represents a specific element of cognitive ability. Superintelligence corresponds to a volume in that multidimensional space that includes the volume of human-level intelligence as its proper subset. Considering that the human intelligence continues to develop new dimensions of intelligence, it appears to be a practical impossibility, or at least a never-ending task to develop such superintelligence. Hereby it is argued that superintelligence in Bostrom’s sense shall not be realized. But we confront a new problem that implies significant risk. It is that AI can get a novel dimension of intelligence that we humans do not have. It is a real risk, because we will not be able to control properly what we do not understand. Even if there will be no superintelligent artificial agent in proper sense, we should care to keep the outcomes of AI technologies within the boundary of controllability.

Keywords: superintelligence, human-level machine intelligence, space of intelligence, intelligent being, controllability