생명윤리 제21권 제1호(통권 제41호) : 71-85, 2020년 6월
J Korean Bioethics Assoc 21(1): 71-85, June 2020

ISSN 1598-0324

인공지능이 인간 같은 행위자가 될 수 있나?*

최 경 석**

【국문초록】 이 글은 인공지능과 인간 지능의 차이가 무엇인지 논의함으로써 인공지능의 존재론적 지위와 인간의 특징이 무엇인지 밝히고자 한다. 최근 알파고의 등장에서처럼 인공지능은 컴퓨터가 상호 연계되며 빅데이터 기술과 딥 러닝 기술까지 포함하는 형태로 발전하였고 계산의 영역뿐만 아니라 감성의 영역까지 포함하는 것으로 발전하고 있다. 그러나 인공지능은 인간의 감성을 흉내낼 뿐이다. 왜냐하면 감정이나 정서의 영역은 진정성이 중요하게 작동하는 영역으로서 감정을 표출하는 자의 주체성과 총체적 인격성이 전제되어야 하기 때문이다. 또한 인간의 뇌가 수행한 기능은 생존을 위한 것이다. 지능 역시 생존을 위한 문제해결능력으로 정의될 수 있다. 이런 점에서 인공지능은 진정한 지능이라 여길 수 없다. 인공지능이 인간과 같은 진정한 지능을 갖추기 위해서는 인공지능이 자신이 처한 환경에서 생존하려는 문제해결능력의 발휘로서 지능을 사용해야 하기 때문이다. 그러나 인공지능이 생존하려는 욕구를 지니고 있다고 판단하기는 아직 이르다. 게다가 인공지능이 인간과 같은 자의식이라는 고도의 지능을 지니고 있다거나 지닐 것이라고 보기도 어렵다. 인간의 자의식은 생물체로서의 유한성을 인식하고 다른 인간과의 연대와 협력을 통해 자아정체성과 함께 타자성을 인식하는 매우 고차원적인 지능 활동이다. 이와 같이 인간 지능과 인공지능의 차이는 인공지능이 의도를 지닌 행위자로 인식되기 어렵다는 판단에 이르게 하며, 도덕적 행위자로 인식되기는 더욱 어렵다고 본다. 인공지능을 인간처럼 여기는 태도나 기획은 오히려 인간이 인공지능과 어떻게 다른지를 보여주고 있다.

【색인어】 인공지능, 인간 지능, 진정성, 행위자, 정체성, 자의식

* 이 논문은 과학기술정보통신부의 재원으로 한국연구재단 바이오•의료기술개발사업의 지원을 받아 수행된 연구임(NO. 2019M3E5D2A02064496)
** 이화여자대학교 법학전문대학원 교수 및 생명윤리정책 협동과정 교수, choiks@ewha.ac.kr

1. 머리말

지난 2016년 3월, 이세돌과 알파고의 바둑 대결이 있었다.[1] 이 대결에서 필자가 주목하는 부분은 이세돌이 알파고에게 졌다거나 이겼다는 것이 아니다. 오히려 필자가 주목한 것은 인공지능이 명령하는 대로 돌을 두는 대리 기사의 모습이었다. 이 대리 기사가 다른 수를 생각하고 있는 경우, 우리는 어떤 갈등이 발생하는지 생각해 볼 필요가 있다. 만약 이 대리 기사가 알파고와의 대국에서 계속 대리 역할을 했다고 해보자. 그리고 이런 갈등을 겪은 이 대리 기사는 반복적인 학습에 의해 자신의 판단을 불신하는 경향을 형성할 수 있었을 것이다. 그리고 이 갈등이 이제 더 이상 갈등으로 여겨지지 않고 알파고를 신뢰하는 때부터 우리는 인공지능의 명령에 지배되는 현상을 목도하게 될 것이다.

인공지능이 지배하는 사회가 도래할 수 있다고 하는 경고는 결코 미래의 일이 아니라, 현실의 일이라고 볼 수도 있다. 그런데 이런 우려에는 한 가지 매우 중요한 생각을 놓치고 있거나 전제하고 있다. 그것은 바로 우리가 인공지능을 무엇이라고 생각하는지에 대한 우리의 태도이다. 인공지능과 계산기를 비교할 때, 우리는 우리가 인공지능에 대해 취하는 태도가 계산기와는 다르다는 것을 목격한다. 왜냐하면 우리는 계산기가 우리 세계를 지배한다는 경고를 진지하게 받아들이지 않고 있기 때문이다. 계산 능력에서 우리가 전산계산기보다 정확하지 않다는 것 때문에 인간이 무능하다고 느끼거나, 전자계산기가 인간과 같은 계산 능력을 지닌 어떤 존재자라고 여기지는 않는다. 그럼에도 불구하고 인공지능의 경우, 인간이 준 자료 값을 처리하라 명령하고, 복잡한 정보처리 작업을 수행하는 기계라고 인식하지 못하고, 인간과 같은 지능을 지닌 어떤 존재자라고 여기는 경향이 발생하는 이유는 무엇인가? 나아가 인공지능이 사회를 지배할 것이라고 우려하는 이유는 무엇인가?

이러한 우려는 이런 우려를 하는 사람들이 인공지능을 무엇으로 이해하고 있는지, 즉 인공지능의 존재론적 지위를 어떻게 이해하고 있는지와 매우 밀접하게 관련되어 있다. 따라서 "우리가 기계 또는 인공지능과 어떤 관계를 유지할 것인지, 이런 존재자들에 대해 어떤 태도를 취할 것인지는 매우 중요한 철학적, 윤리적 문제이다. 최근에는 인공지능에게 도덕적, 법적 책임을 물을 수 있는지 다루는 논문도 발표되고 있다.[2] 그러나 필자는 이런 탐구가 야기하는 오해에 대해 우려하고 있다.

이세돌이 한 라디오 프로그램의 인터뷰에서 자신의 은퇴와 관련하여 한 대담에서 우리는 인간과 기계의 관계에 대한 이세돌의 견해를 엿볼 수 있다.[3] 이 대담에서 이세돌은 알파고에 대해 혼란스런 양가 감정을 드러내고 있다. 핵심은 자신이 기계인 알파고에 패배했다는 것이고, 더 이상 바둑을 둔다는 것에 의미가 없다고 결정했다는 것이다. 이러한 이유에서 은퇴를 결정한 이세돌의 대응에 대해 서로 상반되는 견해가 있을 수 있다. "그래, 패배를 인정하고 떠나는 것이 더 나을 수 있지", "뭐 그럴 필요까지 있나? 바둑이란 그냥 인간끼리 두는 것으로 하면 돼지." "바둑을 두면서 상대방의 성격도 파악하고, 그런 긴장감 속에서 인간과

1) https://radio.ytn.co.kr/program/?f=2&id=41677&s_mcd=0263&s_hcd=01, YTN radio, [정면인터뷰] "이세돌 9단, 알파고에 대망의 첫 승... 과연 불공정한 경기인가?", 2016.3.14.
2) 이상형, "윤리적 인공지능은 가능한가? : 인공지능의 도덕적, 법적 책임 문제" 법과 정책연구, 16권 4호, 한국법정책학회, 2016, 283-303.
3) https://www.youtube.com/watch?v=QiOEQ4xQSO8, "'쎈돌' 24년 바둑인생, 프로 활동을 마치며(이세돌) | 김어준의 뉴스공장, 2019.11.26.

인간이 만나 승부를 겨루는 것이 바둑이지, 기계가 무슨 그런 성격이나 인격이 있겠어." 등등. 삶의 세계 속으로 점점 침투해 들어오면 인공지능의 존재에 대해 우리들 사이에 상이한 태도가 드러날 것이다. 그리고 이것은 현재 우리가 답해야 하는 문제는 인공지능의 도덕적 책임 여부를 논의하기에 앞서 인공지능의 존재론적 지위가 무엇이냐는 질문이다. 이런 질문은 결국은 인간의 존재론적 지위와 특성에 대한 재조명이 바로 우리가 당면한 과제임을 자각하게 할 것이다.

이 글에서는 이러한 문제의식을 바탕으로 인공지능이 나름 기술적으로 변천하고 있다는 것, 인공지능과 인간 뇌의 차이, 인공지능의 존재론적 지위와 인간의 특징 순으로 논의해 보고자 한다.[4]

2. 인공지능기술의 변천

1990년대 인공지능에 대한 논의는 이미 심리철학에서도 주요한 철학적 쟁점 중 하나였다. 1996년 출판되어 전미를 포함하여 세계적으로 중요한 심리철학 교재 중 하나가 되었던 김재권의 *Philosophy of Mind* 에서는 "Mind as Behavior," "Mind as the Brain," "Mind as a Computer"라는 장이 독립 장으로 구성되어 있다.[5] 당시 심리철학계를 포함하여 철학계의 논쟁의 핵심은 인간이 수행하는 인식적 활동과 유사하거나 동일한 활동을 컴퓨터가 수행하고 있다고 여겨지므로, 인공적으로 지능에 해당하는 활동을 컴퓨터가 수행할 수 있는지 여부, 수행한다면 이런 활동을 인간 지능 활동과 동일하다고 볼 것인가 등등의 문제였다.

따라서 당시 CPU를 뇌에 비유하거나 여러 보조장치를 인간의 감각기관에 비유하거나 이런 주변장치가 합쳐진 컴퓨터를 하나의 개별자로 비유하는 경향이 있었다. 휴머노이드의 발전과 휴머노이드를 또 다른 하나의 존재자처럼 여기는 사고는 이러한 사고에 SF적 요소가 가미된 것으로 볼 수 있다. 외형적으로 인간의 모습을 한 휴머노이드는 이런 사고를 하기에 일반인에게는 매우 친화적인 요소일 수 있다.[6]

그러나 최근 이세돌과 인공지능 알파고와의 대결에서 확인할 수 있듯이, 인공지능은 외양이 휴머노이드의 모습을 해야 하는 것이 아닐 뿐만 아니라, 여러 컴퓨터가 연계되는 형태로 그 기술이 발전하였고, 여기에 빅데이터 기술, 딥 러닝 기술까지 포함시켜 발전하고 있다.

이러한 놀라운 발전은 인공지능의 존재론적 특성이 무엇인가라는 문제가 얼핏 보기에는 좀 더 복잡한 양상으로 전개되게 한다. 소위 "딥 러닝"이란 용어에서 짐작할 수 있듯이, 주로 인간에 귀속되는 속성을 인공지능에도 귀속시키는 표현을 사용하고 있다. 말 그대로 인공지능이 인간처럼 학습도 할 수 있다는 이미지를 부과하고 있다. 그러나 이 학습이 인간의 학습과 얼마나 유사한 것인지는 별도의 논의가 필요할 것이다. 게다가 딥 러닝의 소재나 정보처리 방식의 설계 등도 기본적으로는 인간이 프로그램화하는 것이므로 정말 순수하게 기계나 인공지능이 무엇을 한다는 것인지 의문이다.

그러나 이러한 변화가 지속되고, 설사 더 가속화되면 될수록, 인공지능은 정말 인간과 얼마나 유사한 것인지, 그래서 인공지능이 인간과 같은 존재라는 것인

4) 이 논문의 제목으로 "인공지능이 인간 같은 행위자가 될 수 있나?"라고 물었던 것은 "인공지능"을 마치 인간과 같은 행위자로 전제하고 제기되는 질문들보다 앞서 질문해야 할 것이 인공지능이 인간이기라도 한 것인지, 인간과 유사한 도덕적 행위자이기조차 한 것인지부터 물어야 한다는 점을 강조함으로써 인공지능에 대한 존재론적 질문을 환기하기 위해서이다.

5) Jaegwon Kim, *Philosophy of Mind*, Westview, 1996.

6) 일본 만화, 『공각기동대』를 비롯한 유사한 SF 영화는 인간과 구별되는 또 다른 존재자처럼 로봇을 그려내고 있다.

지, 인공지능 역시 행위의 주체로 볼 수 있다는 것인지, 더 나아가 도덕적 행위자로 볼 수 있는지, 뿐만 아니라 인공지능에 대해 인간은 어떤 관계를 유지하고 어떤 태도를 취해야 한다는 것인지 등등이 점점 더 시급히 진지하게 논의되어야 할 질문이 되었다.

인공지능의 오동작으로 사고가 발생한 경우, 이러한 인공지능 기술을 사용하는 사용자나 인공지능 기술을 제공한 제작자나 설계자가 책임을 지닌 것이 당연함에도 불구하고, 인공지능이 책임을 져야 하는 것인가라는 질문 자체는 이미 인공지능을 어떤 행위자로 전제한 질문이다.[7] 이미 이런 질문에는 인공지능을 인간과 유사한 행위자로 보는 사고가 전제되어 있다.

3. 인공지능과 인간 뇌의 기능

1) 정보처리 영역에서의 인공지능

인공지능은 향후 지속적인 발전을 거듭할 것이다. 정보처리 능력이란 관점에서 볼 때, 컴퓨터는 인간이 수행하는 정보처리 능력보다 훨씬 뛰어나고 효율적인 능력을 보여 주고 있다. 그러나 정보처리 기능 정도로만 이해될 때, 우리는 "인공지능"이라는 매우 다양한 함의를 지닌 용어보다 "고도의 정보처리 능력을 갖춘 컴퓨터"라는 표현을 사용하는 것이 바람직하다. 그러나 이런 긴 어구의 표현을 사용하는 대신 "인공지능"이라는 용어를 사용할 때, 많은 오해와 왜곡을 생산할 가능성이 높다. 그래서 딥 러닝, 빅 데이터 기술과 결합된 인공지능은 인간의 지능처럼 "학습"을 한다거나 그래서 인간의 지능과 유사하다거나 나아가 인공지능이 인간의 지능을 능가한다는 생각이나 상상을 형성하게 한다. 그리고 이런 판단은 인공지능을 지닌 다양한 형태의 개체가 새로운 의사결정능력을 지닌 존재자인 것처럼 생각하게 하는 잘못이 있다.

위와 같은 태도가 바로 인공지능과 대결을 벌이는 바둑이나, 다양한 형태의 게임 등을 대할 때 일부 사람들에게서 드러난다. 이런 기능에 인간적인 성격이 외연적으로 시현되는 경우, 우리는 이 능력을 발휘하는 주체가 기계가 아니라 유사 인간인 것처럼 이해하는 경향도 있다. 예를 들어, 휴머노이드의 외형은 바로 이런 오해를 불러일으키는 데 일조하고 있다. 물론 필자는 이런 태도나 경향성을 매우 의심의 눈으로, 그리고 매우 우려스럽게 바라보고 있다. 우려의 핵심은 인공지능에 대한 두려움이 아니라, 인공지능의 담론을 둘러싼 비유적 표현이 담론을 합리적으로 이끌기보다 우리의 이해를 왜곡하거나 담론을 가능하게 하는 숨은 전제에 대한 논의보다는 자극적인 내용이나 상상에 기반한 비현실적인 주제가 담론의 화두에 놓일 수 있기 때문이다.

알파고 외에 "인공지능 판사"[8], "인공지능 의사"[9]와 같은 용어가 사용되고 있는 것은 아무리 비유적인 표현이라 하더라도, 필자는 지나치게 자극적이라 생각한

7) 이상형은 "도덕적 행위자의 자율적 행위에 책임과 권리를 부여한다."고 말하며, "이런 책임과 권리 또한 역사적으로 이루어져 온 인정투쟁의 결과"라고 주장한다. 따라서 인공지능이 "자신의 행위에 대한 책임과 권리가 인정되기 위해서는 이런 인정 과정이 필수적이며", "지금 우리가 인공지능에게 기대하는 것은 정당한 인정투쟁이 되기 위해 그를 최대한 윤리적으로 만드는 것"이라고까지 주장한다. 이상형, "윤리적 인공지능은 가능한가? "『법과 정책연구』, 16권 4호, 한국법정책학회, 2016, 298-299. 그러나 필자는 인공지능이 이런 인정투쟁을 벌일 수 있는 주체인지부터가 의문이다.

8) http://www.etnews.com/20191121000113 "인공지능 판사와 프레디쿠스", 전자신문, 2019.11.21. 임영익(인텔리콘 메타연구소 대표) 기고.

9) http://health.chosun.com/site/data/html_dir/2019/08/21/2019082100747.html "가천대길병원, 인공지능 의사 '왓슨' 현지화 나선다", 헬스조선, 2019.8.21. 유대형 헬스조선 기자.

다. 이 용어는 "판사의 업무 중 일부 정보처리 업무를 대신한다."는 의미인지, "판사처럼 법률정보를 활용하여 개별 사안에 대한 법적 판단을 내린다."는 의미인지 불분명하다. 인공지능이 판사를 대신할 것이라는 우려는 이미 인공지능을 가치판단이 가능한 행위 주체로 보는 생각을 전제로 한다. "인공지능 의사"라는 용어도 의사 대신 진단하고, 처방하고, 치료하고 수술하는 주체가 인공지능이라는 시각을 형성하게 한다.

우리가 판사, 검사, 변호사의 역할이 무엇인지 다시 되묻는 것이 선행되어야 할 시기에, 그리고 의사의 역할이 무엇인지 다시 되물어야 할 시기에 오히려 인공지능이 고도의 가치판단을 요하는 이런 전문 직역의 임무를 대신할 수 있을 것과 같은 인상을 주는 것은 전문직업성이란 개념에 대한 몰이해를 드러내는 것이고, 직업전문성과 관련된 윤리와 철학을 송두리째 내동댕이치는 격이다. 그리고 이런 용어 사용에 앞서 과연 가치판단과 같은 정성적 판단의 영역이 인공지능을 통해 어떻게 가능하다는 것인지, 정량적 평가가 어떤 알고리즘에 의해 정성적 평가로 전환될 수 있다는 것인지부터 되짚어 보아야 할 것이다.

담론에 사용되는 용어가 일반적인 의미에 부합하지 않아 생기는 오해로서 고인석의 다음 예는 주목할 만하다.

> 이 공장의 모든 밸브 상태를 감시하며 개폐를 제어하는 인공지능시스템 BV2017은 (ISO 13482의 기준에서) 자율성을 지닌 시스템이다.
>
> 자율성을 지닌 존재의 결정과 행위는 존중되어야 하고 외부의 간섭으로부터 보호되어야 한다.
>
> 따라서, BV2017의 모든 작동은 존중되고 (인간 관리자를 포함한) 외부의 간섭으로부터 보호되어야 한다.[10]

위 논증은 고인석이 지적하고 있듯이 애매어의 오류가 개입되어 있다. 왜냐하면 "로봇공학에서 '자율성'(autonomy)이란, [로봇이] 인간의 개입 없이 스스로 임무를 수행하는 역량을 의미"하기 때문이다.[11] 자율성에 대한 이런 정의는 철학이나 윤리학에서 흔히 언급되는 칸트의 자율성 개념과 다르며, 생명의료윤리에서 언급되는 자율성과도 다르기 때문이다.[12]

인공지능 분야뿐만 아니라 첨단 과학 분야가 소개될 때, 비유적 표현으로 보이는 "윤리적 뇌", "이기적 유전자", "확장된 정신"(extended mind)과 같은 용어를 접한다. 필자는 이들 표현을 기술적 표현으로 보기에는 무리가 있고 비유적 표현으로 보아야 한다고 생각한다. 그러나 비유적 표현으로 이해하지 않을 때 이 용어들은 불필요한 담론을 야기하거나 논증되지 않은 전제를 받아들이도록 하는 불편함을 야기한다.[13]

2) 감성의 영역

지금까지는 인공지능에서 수행하는 소위 "지능"이란 부분이 단순히 계산적 지능 활동에 국한되는 것으로 언급했지만, 인공지능의 영역이 이 정도에 머무르

10) 고인석, "인공지능이 자율성을 가진 존재일 수 있는가?" 『철학』, 제133집, 2017, 167면.
11) "D6.2 – Guidelines on Regulating Robotics", p.15. URL=www.robolaw.eu.). 고인석, "인공지능이 자율성을 가진 존재일 수 있는가?" 『철학』, 제133집, 2017, 166면에서 재인용.
12) 철학이나 윤리학에서의 자율성 개념과 생명의료윤리에서의 자율성 개념에 대한 차이는 최경석, "자율성 존중의 원칙 : 정치적 이념과 철학적 이념", 『윤리학』, 제3권 제2호, 한국윤리학회, 2014. 11. 43-64 참조.
13) 필자도 동의하는 바인데, 고인석은 자율 주행 차량에서 '자율'은 사실상 '자동'과 교환 가능한 개념임을 지적한다. 고인석, "인공지능이 자율성을 가진 존재일 수 있는가?" 『철학』, 제133집, 2017, 182면.

지 않고 있다는 것은 주지의 사실이다. 노인 돌보미 인공지능[14]이나 휴머노이드 개발에서 볼 수 있듯이, 로봇 개발자들은 감정이나 정서의 영역까지 로봇의 기능을 확대해 가고 있다.

그러나 과연 로봇이 즉 인공지능이 감정이나 정서를 지니고 있다고 볼 수 있을까? 이미 프로그램화된 반응으로 감정을 표현하는 것임을 우리는 짐작할 수 있다. 진정 "감정"을 인공지능이 갖출 수 있을지에 대해 필자는 매우 부정적인 입장이다. 왜냐하면 감정의 영역에는 유기체의 쾌와 고통, 나아가 진정성이란 것이 매우 중요한 역할을 하기 때문이다.

슬픔, 기쁨, 삶의 애환, 동정심, 정의감 등등. 우리가 느끼는 수많은 다양한 정서가 뇌와 관련이 있다. 그리고 인지적 기능과도 연관된다는 것을 잘 알고 있다. 물론 신경과학이 그 매커니즘을 정확히 밝혀내고 있지는 못하다. 그러나 적어도 우리는 이런 감정이나 정서의 아주 기초적인 기제인 쾌와 고통은 생물체, 그것도 중앙신경체계를 지닌 일부 유기체에서 발생함을 알고 있다. 보다 복잡한 정서가 목격되는 것은 이 유기체가 발생학적으로 더 복잡한 뇌를 지니고 있을 때이다. 적어도 개나 고양이와 같은 수준의 뇌기능을 지닌 동물에게서 감정이나 정서는 발생한다고 여긴다. 그런데 기계가 이런 반응을 보인다면, 그것은 흉내내기이지 기계 그 자체가 외부자극으로부터 쾌나 고통을 느꼈기 때문이라고 여기지는 않는다. 이 직관이 우리에게 시사하는 바는 매우 크다.

우선, 쾌나 고통을 포함한 정서적 반응은 뇌라는 생물학적 기제에서 발생하기 때문이다. 우리가 휴머노이드나 인공지능에 이와 같은 외부 자극을 감지하는 기계를 연결시키고 심지어는 이 자극을 처리한 결과물에 상응하는 반응이나 행동을 보이는 기계장치가 있더라도 과연 이 과정이나 결과물이 쾌나 고통에 해당한다고 보아야 할지는 매우 의문스럽다. 이러한 의문의 핵심에는 쾌나 고통 등의 정서가 그 주체가 지닌 것이라고 볼 수 있는지 여부가 문제되기 때문이다. 즉 진정성에 문제가 있기 때문이다. 다시 말해, 뇌 기관에서 발생한 정서적 반응이 그 뇌를 소유한 주체의 것이라고 여기는 것과 같은 의미부여가 인공지능과 같은 존재자, 필자는 여전히 기계라고 인식하고 있는 존재자에게 가능하냐는 것이다.

우리는 자신에게 보인 타인의 정서가 꾸며진 것이고 단지 그런 척하는 것이라면 매우 불쾌하게 여긴다. 예를 들어, 좋지 않은 일을 겪은 나에게, 타인이 "얼마나 속상하니?" "매우 힘들겠구나, 그래도 힘을 내"라고 하거나, "나도 매우 속상하다"와 같이 공감하는 정서를 보일 때, 이런 정서에 대해 진정성이 결여되었다는 사실을 알게 되면, 매우 실망스럽게 여기고, 인간적인 배신감까지 느끼기도 한다. 우리는 이처럼 인간의 감정에 왜 진정성이란 판단을 부가하는 것일까?

진정성이란 개념은 우선, 발화되거나 표현된 것의 소유자라는 주체성의 개념을 전제로 한다. 그리고 그 반응은 단순한 반응이 아니라 주체의 총체적 반응으로 여겨진다. 다시 말해, 감정이나 정서는 단순히 인지적 영역의 계산적인 기능적 기제가 발휘된 것만이 아니라, 발화하거나 표현하는 주체의 전인적 반응으로 여겨진다는 것이다. 발화나 표현의 내용은 그 주체의 것일 뿐만 아니라, 인격을 대변하는 것으로도 여겨진다.

발화만 놓고 보더라도 우리의 언어적 행위는 자연적 음성 그 자체가 결코 핵심이 아니다. 언어 행위는 그 음성이나 문자를 매체로 하여 그 음성이나 문자에 부여된 의미를 그것이 담아 전달하고자 하는 발화 주체와 그 음성이나 문자가 도달되도록 해야 하는 대상

14) http://isplus.live.joins.com/news/article/article.asp?total_id=23539137&cloc= ""자식보다 낫네"…독거노인 외로움 달래 주는 AI", 일간스포츠, 2019.7.30. 권오용 기자.

자와의 관계 속에서 벌어지는 행위이다. 감정이나 정서를 표현하는 것도 마찬가지이다. 감정이나 정서의 주체가 있고, 이것이 표현되는 경우에는 그 표현이 전달되었으면 하는 대상자가 존재하게 된다.

감정이나 정서로부터 시작된 이상의 논의에서 주목해야 하는 것은 감정과 정서의 진정성이다. 그리고 이 "진정성"이란 개념은 주체성과 총체적 인격성을 전제로 한다. 진정성은 서양 철학에서는 인테그러티(integrity)와 분리되어 이해할 수 없다. 다시 말해 진정성이란 발화의 내용이 그 발화자나 표현자의 정체성과 부합될 때 획득되는 성질이다.

정체성이란 이미 주체성과 총체적 인격성을 담고 있는 개념이다. 주체성이란 개념은 타자를 전제로 한다. 다른 사람의 것이 아니라 내 것이라는 것을 말할 때, '나'라는 주체성은 타인 없이 성립될 수 없다. 총체적 인격성은 삶이라는 역사성을 담고 있다. 나의 삶의 역사가 사상된다면 인격을 논의할 수 없다. 따라서 정체성이란 다른 것과 구별되는 나다움이며, 이 나다움이란 순간적인 삶의 단면이 아니라 통시적으로 축적되어 온 나의 모습이다.

그런데 정체성을 논의하는 위 분석에서 놓치지 말아야 할 것은 자의식이다. 내가 누구인가라는 성격을 구성하게 하는 정체성이란 자기 자신에 대한 이해와 의식이 없다면 성립할 수 없는 개념이다. 어떤 감정을 표출하는 주체가 진정성 있는 감정을 표출하는 것이라면, 이 감정은 표출하는 자의 주체성, 역사적으로 축적된 삶의 특성과 이 특성에 대한 자기 자신의 서사적 자기 의식과 부합하는 것이다.

인공지능이 인간의 감정을 흉내낼 수는 있을 것이다.[15] 그러나 우리는 이 인공지능의 감정 표현을 진정성 있는 감정이라 부를 수 있을까? 이 질문에 대한 긍정적인 답은 앞서 언급했던 매우 복잡한 전제 조건들을 인공지능이 충족시킬 때 가능하다. 그래서 우리는 인공지능은 주체성을 지니고서 자신다움이라는 가치를 추구하는지 답해야 한다. 즉 인공지능이 자의식을 가지고 있는지 답해야 한다. 그러나 필자는 적어도 현재 우리가 상상할 수 있는 인공지능은 그렇지 못하다고 생각한다. 결론적으로 인공지능은 결코 감정을 진정성 있게 표현하는 주체가 될 수 없다.

또한 감정의 표출은 단지 감정의 표출에 그치는 것이 아니라 행위주체자의 다른 특성들과 연계되어 총체적으로 이해되는 측면이 있다. "슬프다"라는 감정은 단지 감정 그 자체만이 아니라, 눈물이라는 생물학적 특성과도 연결될 수 있다. 즉 표현 주체의 총제적인 매커니즘과도 연결되어 있다. 마찬가지로 뇌의 다양한 기능은 사실상 단지 뇌에서 발생하는 국면적인 특성에만 그치지 않고, 뇌를 소유한 행위주체자의 다른 생물학적인 특성과 연관되어 있으며, 의미론적으로는 그 행위주체자의 전인적인 성격과 관계를 맺고 있다.

사실, "지능"이라는 것조차도 흉내내기이다. 이상욱은 "튜링 검사는 지능이 무엇인지에 대한 '정의'로서 제안된 것이 아니라 누구나 인정할 수 있는 인간의 지능을 특정 대상에게 '확정'시킬 수 있는지를 판가름하려는 목적으로 제안된 것"이라고 밝힌다.[16] 그리고 튜링 검사는 "지능 자체를 검사하기보다는 우리가 암묵적으로 지능적이라고 생각하는 인간의 행위를 기계나 다른 지능소유 후보자들이 얼마나 잘 '흉내'내는지를 판별하도록 설계"되었음을 지적한다.[17] 이것이 시사하는 바는 우리가 익숙하게 사용하는 "인공지능"이라는 용어에서의 "지능"조차 인간의 지능이라기보다 인공지능이란 연

15) 필자는 인공지능에 대한 기능주의적 접근은 흉내내기로서의 감정 표출과 진정성을 지닌 감성 표현을 구별하지 못하게 하는 단점을 지닐 수 있다고 본다.
16) 이상욱, "인공지능의 한계와 일반화된 지능의 가능성 : 포스트 휴머니즘적 맥락", 『과학철학』 12, 2009, 7면.
17) 이상욱, "인공지능의 한계와 일반화된 지능의 가능성 : 포스트 휴머니즘적 맥락", 『과학철학』 12, 2009, 8-9면.

구 분야에서 새롭게 정의하는 지능일 수 있다는 점이다. 인간의 지능이 무엇인지 제대로 규명되지 않은 채 인간 지능의 일부를 흉내낸 것을 "인공지능"이라 한다면, "인공지능"이란 용어 역시 앞서 언급한 비유적 표현에 가깝다고 볼 수 있다. 그리고 이것이 비유적 표현이라면 인공지능에 대한 우리의 담론은 혹시 애매어의 오류가 뒤범벅된 담론은 아닌지 성찰할 필요가 있다.

이상욱은 "인간은 오랜 진화의 역사에서 두뇌 구조를 발달"시켜 왔음을 언급하면서도 "인공지능을 갖춘 기계가 인간의 진화과정과 유사한 경험을 축적하지 못할 원리적 이유는 없다"고 말함으로써 매우 원론적인 주장이거나 논리적 가능성 차원의 주장을 하기도 한다.[18] 그러나 과연 지능이 진화의 산물이라면 생물체가 아닌 인공물이 인간의 지능과 같은 지능을 지닐 수 있을지는 대단히 의심스럽다.

3) 인간 뇌와 인간 생존

인간의 뇌가 하는 역할은 단순히 계산하거나 감정을 표출하는 데 그치지 않는다. 인간의 뇌가 수행한 기능은 생존을 위한 것이었다. 이대열은 지능을 다음과 같이 정의한다. "지능은 생명체가 자신의 생존과 번식을 위해 다양한 환경에서 의사결정의 문제를 해결하는 능력으로 정의할 수 있다."[19] 여기서 주목해야 하는 것은 "생명체"라는 용어와 "다양한 환경"이다. 지능은 생명체가 다양한 환경에서 생존과 번식을 위해 탄생하게 된 것이라고 이해할 수 있다. 이러한 지능은 인공지능과 차이가 있다. 이대열은 "인공지능을 진정한 지능이라고 여기지 않는 이유는 그것이 해결해야 하는 문제가 그 자신의 문제가 아니라 인간이 제시한 문제이기 때문"이라고 한다.[20] 인공지능은 인간이 만든 하드웨어에 인간이 만든 소프트웨어를 작동시킴으로써 운용된다. 이대열은 "조만간 인공지능이 인간을 대체하게 될 것이라는 예측은 기우에 지나지 않는다"[21]라고 하면서 그 이유를 다음과 같이 세 가지로 제시한다.

첫째, "인공지능의 문제풀이 능력은 극히 제한적이다."[22] 그리고 "인공지능은 특정 문제의 해결을 목적으로 개발되었기 때문에, 생존과 번식에 관련된 모든 문제를 해결해야 하는 동물의 신경계처럼 다양한 종류의 문제를 해결하지 못한다."[23] 그런데 이대열은 이 한계가 본질적인 것인지 즉 범주적 차이인지에 대해서는 대답을 유보한다. 그는 "이런 기계가 환경이 달라지면서 새로운 문제에 마주쳤을 때 그 문제들을 융통성 있게 해결하기를 기대하기는 어렵다. 설령 다양한 환경에 대처할 수 있는 인공지능이 개발되고 있다 하더라도 그것이 조만간 등장할 것 같지는 않다."[24]고 하기 때문이다.

둘째, "인공지능의 문제풀이는 인공지능 그 자신을 위한 것이 아니다. 지능에서 중요한 것은 지능을 그것의 주체의 선호도와 분리해서 평가할 수 없다는 것이다."[25] 이 두 번째 이유는 이미 앞서 언급했던 것이다. 이 두 번째 이유는 "인공지능의 성과는 인공지능이 아니라, 인공지능을 개발한 인간 지능의 표현이라고 보아야 한다."[26]라는 점을 지적하고 있다. 물론 여기서

18) 이상욱, "인공지능의 한계와 일반화된 지능의 가능성 : 포스트 휴머니즘적 맥락", 『과학철학』 12, 2009, 15면.
19) 이대열, 『지능의 탄생』, 바다출판사, 2017, 109면.
20) 이대열, 『지능의 탄생』, 바다출판사, 2017, 82면.
21) 이대열, 『지능의 탄생』, 바다출판사, 2017, 87면.
22) 이대열, 『지능의 탄생』, 바다출판사, 2017, 88면.
23) 이대열, 『지능의 탄생』, 바다출판사, 2017, 88면.
24) 이대열, 『지능의 탄생』, 바다출판사, 2017, 88면.
25) 이대열, 『지능의 탄생』, 바다출판사, 2017, 88면.

인간 지능의 표현조차도 온전한 표현은 아닐 것이다. 필자는 여기서 언급된 특징은 인공지능의 주체성 획득 여부와 관련된 부분으로 해석한다. 그리고 여기서 인공지능은 앞서 필자가 언급했던 주체성, 정체성, 인격성 등을 결여한 기계임을 드러내는 것으로 볼 수 있다.

셋째, 인공지능이 이대열의 지능 정의에 부합하는 지능이 되려면, "자신(하드웨어)이 처한 환경에서 복잡한 문제를 해결하기 위해 필요한 프로그램(소프트웨어)을 선택하는 능력"[27]을 지녀야 한다. 이대열에 따르면, 프로그램을 선택하는 것 역시 또 다른 프로그램으로서, 컴퓨터가 참된 지능을 가지려면 프로그램을 선택하는 메타-프로그램이 필요하며, 보통 이 메타-프로그램의 역할은 인간이 하고 있다는 것이다.[28] 그러나 이대열은 아직 실현되지 않은 것이지 불가능하다고 단정하지는 않는다. 주변 온도를 감지하며 자동으로 온도를 조절하는 에어컨이나 로봇 청소기의 예에서 볼 수 있듯이, 매우 제한적 영역에서 이런 역할을 수행하는 아주 기본적인 수준의 인공지능은 우리 주변에서 많이 찾아 볼 수 있다고 한다.[29] 자율적 인공지능에 대한 연구가 화성에서 진행되고 있음을 소개하면서, 이대열은 "언제가 자기 자신을 위해 의사결정을 내리는 인공지능이 도래할 때가 올지도 모른다"[30]라고 한다. 따라서 이대열은 인공지능이 인간을 대체할 것이라는 예측이 잘못된 것이라고 주장하고자 했던 것이 아니라, "조만간" 대체할 것이라는 예측이 기우라는 점을 밝히고자 한 것으로 보인다.

그러나 필자는 과연 위에서 제시한 이유가 시간이 지나면 극복될 수 있는 문제인지 의문이다. 위 세 가지 이유를 관통하는 핵심은 메타-프로그램을 선택하는 행위는 자의식을 필요로 하는 행위일 수 있다는 점이다. 더욱 중요한 문제는, 이대열이 스스로 제시한 지능의 정의에 부합하려면, 인공지능은 생물체처럼 자신의 생존과 번영을 위해 다양한 환경에서 의사결정의 문제를 해결하는 능력을 갖추어야 한다. 다시 말해, 앞서 언급한 세 가지 이유가 제거되고, 인공지능이 참된 지능이 되기 위해서는, 인공지능이 자기 자신을 위해(두 번째 이유) 메타-프로그램을 스스로 선택함으로써(세 번째 문제) 자신의 생존과 번식에 관련된 모든 문제를 해결해야 하는 동물의 신경계처럼 다양한 종류의 문제를 해결(첫 번째 이유)할 수 있어야 한다.

게다가 이대열의 지능 정의는 매우 폭넓어서 식물도 지능을 지닌 것으로 인정해야 한다. 흔히 지능은 "자동적으로 또는 본능에 의해 어떤 것을 하는 것 대신 생각하고, 추론하고 이해하는 능력"(the ability to think, reason, and understand instead of doing things automatically or by instinct)[31]으로 정의되는데, 적어도 이런 사전적 정의와는 다른 정의를 이대열은 제시하고 있다. 물론 이 사전적 정의에 부합하지 않음 그 자체가 이대열의 지능 정의에 문제가 있다는 것은 아니다. 우리가 학술적으로, 특히 철학적으로, 논의해야 하는 것은 지능을 매우 폭넓게 정의하더라도, 식물의 지능, 단세포의 지능, 고등동물의 지능, 인간의 지능의 차이를 어떻게 구별할 수 있는가이다. 그리고 이런 논의의 연장선상에서 인공지능과 관련해서는 다음의 두 가지 질문에 답해야 할 것이다. 첫째, 인공지능은 이러한 다양한 지능 중 어떤 지능인가 또는 어떤 지

26) 이대열, 『지능의 탄생』, 바다출판사, 2017, 88면.
27) 이대열, 『지능의 탄생』, 바다출판사, 2017, 92면.
28) 이대열, 『지능의 탄생』, 바다출판사, 2017, 93면.
29) 이대열, 『지능의 탄생』, 바다출판사, 2017, 93면.
30) 이대열, 『지능의 탄생』, 바다출판사, 2017, 108면.
31) https://www.collinsdictionary.com/dictionary/english/intelligence 참조.

능과 유사한가? 둘째, 과연 인공지능은 인간의 지능과 유사하거나 인간의 지능이라고 할 수 있는 수준에 도달할 수 있는가? 여기서 인공지능이 단지 계산적 능력에서 인간의 지능과 유사하다고 주장하는 것이라면 이 주장은 사소하다. 왜냐하면 이런 주장에 대해서는 "정교한 계산기" 또는 "정보처리기"라고 불러도 되는 것을 굳이 "인공지능"이란 거창한 용어를 사용할 필요가 있었는지 되물을 필요가 있기 때문이다. 필자의 눈에는 이런 정도를 "인공지능"이라 부르고자 했다면 이것 역시 비유적 표현의 오남용이거나 담론을 혼란스럽게 하는 잘못이 있다.

필자는 지능의 다양한 종류와 관련하여 인간의 지능은 최소한 '자의식'이라는 자기 인식을 지닌 매우 특별한 속성을 지닌 지능이라 주장한다. 아울러 인공지능과 관련된 위 두 질문에 대해 첫째, 인공지능은 이 "지능"이란 용어를 매우 넓게 사용한다 하더라도 인간의 지능에 유사하다고 볼 수 없다. 식물이나 저급한 동물 수준의 지능에 유사하다는 평가를 할 수는 있을 것이다. 그러나 이대열의 정의를 적용할 때, 인공지능은 생물체가 아니므로 생물체처럼 자기를 위한 생식과 번식에 해당하는 행위로 인식될 만한 특징을 갖추어야 한다. 인공지능이 생식과 번식을 위한 자의식을 가지고 있지 않더라도 생존과 번식이란 행위에 해당하는 행위를 하려면, 인공지능은 자기복제를 시도해야 한다. 그리고 이 욕구는 비록 자의식을 지닌 행위는 아니지만, 본능적으로라도 이런 행동이 내재되어 있어야 한다. 과연 인간이 인공지능에 그런 불변의 소프트웨어를 심어 놓지 않은 채, 인공지능 스스로가 자기 학습을 진행하면서 이런 행태를 지니게 될 것인지는 매우 의문스러운 일이다.

둘째, 인공지능이 인간의 수준에 도달한 지능으로 평가되기 위해서는 자의식을 획득해야 한다. 자의식을 지닌다는 것은 자신이 수행하는 활동을 자신이 스스로 되돌아보고 평가하는 능력이 있다는 것을 의미한다. 인간이 진화적으로든 역사적으로든 축적해 온, 문화, 도덕, 교육 등은 바로 이러한 자의식의 산물이다. 이 모든 산물은 인간은 유한한 존재이지만 즉 삶의 시작이 있고 죽음이란 삶의 끝이 있지만, 변화하지 않거나 적어도 무한한 것을 인식하고자 하고 지향하는 태도에서 발생한다. 자신의 존재가 동물이지만 불변하는 것을 추구하는 이성적 존재자이며, 자신의 삶이 일회적이고 소멸하는 것이고, 이 세계를 인식하는 주체이기에 세계의 중심에 있고, 단순히 본능적으로 살아가는 존재가 아니기에 인간은 존엄하다는 이념을 지닐 수 있었다. 과연 이런 정도의 수준에 달하는 "자기 자신을 위한 생존과 번식과 관련된 모든 문제를 해결하는 수준의 활동"을 인공지능이 할 수 있을까? 그리고 앞으로 그런 존재자가 될 수 있다고 예상할 수 있을까? 인공지능이 비유적으로 이런 조건이나 환경에 처해 있고 그런 환경에 반응하는 것으로 해석될 것들이 있을 수 있으나, 그런 행태가 자의식이란 메타 인식을 지닌 작동으로 보기는 어렵다. 왜냐하면 인간의 생존과 번식을 위한 문제해결 활동은 생존과 번식을 위해 함께 살아가면서 동등한 존재자로서의 다른 인간과 협력하고, 그런 존재자로 성숙하기 전까지의 인간을 양육하고 배려하며, 인간의 유한성을 극복하는 수단으로 교육과 문화를 전수하는 활동을 포함하기 때문이다. 그리고 이런 행위는 주체성과 타자성의 인식, 자신의 정체성 유지의 가치에 대한 인식 등등을 전제로 하기 때문이다. 동물의 생존과 번식이 본능적으로 수행되는 것이라면 인간의 생존과 번식은 자의식 즉 메타 인식을 통해 수행되고, 사회적으로 그리고 제도적으로 강화된다.

좀 더 비약적으로 기술하자면, 인공지능이 인간의 지능에 해당하는 수준에 도달하려면, 인공지능이 자신의 정체성을 인지해야 하고, 자신도 유한한 존재라고 인식해야 하며, 인공지능들 사이의 협력과 연대가 필

요하고, 자신의 유한성을 극복하기 위해 생존과 번식에 해당하는 자기복제라는 행위를 욕구에 기반하여 수행해야 한다. 그러나 인공지능이 과연 그럴 수 있을 것인가? 생물체의 자기복제는 동일한 자기복제가 아니다. 그렇기 때문에 개체의 정체성은 유한하지만, 종적 정체성은 비교적 오래 유지되는 것이다. 그러나 정확한 자기복제가 가능한 기계에게 있어 유한성이 왜 극복되어야 하는 성질인지 묻지 않을 수 없다. 보다 근본적으로는 이런 극복의 욕구나 동기조차 있을 수 있는지 묻지 않을 수 없다.

결국 인공지능에 대한 담론에 등장하는 용어들은 인간의 지능을 설명하는 용어와 같은 성격의 용어처럼 보이지만 사실 다른 의미를 지녔거나 비유적인 의미로 사용된 것일 수 있다. 우리는 인공지능의 담론에서 이런 비유적 표현에 각별히 유의할 필요가 있다.

물론 인간 지능을 설명하며 필자가 사용한 용어들이 신경과학자들의 눈에는 철학적으로 또는 인문학적으로 오염된 용어이기 때문이 선결문제 요구의 오류가 개입된 것처럼 여길 수 있다. 이런 관점에서 자의식, 양심, 감정, 정서, 판단력, 특히 도덕적 판단 등등의 개념은 그것이 신경과학적 측면에서 어떻게 작동하는지 연구할 필요가 있다. 이런 작동이 다른 기초적인 생물학적 작동으로 환원되는 것인지, 그래서 인간 종이 아니더라도 기계적 작동으로도 구현될 수 있는 것인지 과학적인 연구 성과를 바탕으로 논의될 필요가 있다.

필자가 인공지능이 자의식이나 양심 또는 정서의 진정성과 같은 특성을 소유할 수 없을 것이라 주장하는 것은 과학적 근거에 기초한 것은 아니다. 따라서 이 주장이 과학적 엄밀성을 갖추기 위해서는 기존의 인문학적 용어가 신경과학적으로 밝혀진 사실의 어디에 상응하는지 연구하고 이런 연구 성과를 바탕으로 필자의 주장이 보완될 필요가 있음은 인정한다.

4. 인공지능의 존재론적 지위와 인간의 특성

위에서 필자는 진정성이나 메타 인식의 측면에서 인공지능이 과연 인간 지능의 수준에 도달할 수 있을 것인지에 대해 논의하였다. 이제 인공지능이 의도를 지닌 행위자인지 여부를 따져보고자 한다. 이 주제는 인공지능이 인간의 지능 수준에 도달할 수 있으려면 어떤 특성을 지녀야 하는지 논의하는 데 있어 또 다른 중요한 주제이다. 의도의 문제는 행위를 의식적으로 인지하고 어떤 목적을 위해 수행하는 것인지와 관련하여 매우 중요한 문제이다. 과연 인공지능이 어떤 행위를 할 때 인간과 같은 의도를 가지고서 진정성 있는 행동을 하고 있다고 판단할 수 있는가? 그리고 그렇다고 판단할 수 있는 기준이 있다면 그것은 무엇인가?

1) 인공지능이 행위자인가? 나아가 도덕적 행위자인가?

인공지능과 관련하여 논의되어야 할 또 다른 핵심적인 질문은 인공지능, 다소 확장된 개념으로 이해되는 인공지능이라 하더라도, '인공지능이 도덕적 행위자인가?'라는 질문이다. 필자의 대답은 "아니다."이다.

도덕적 행위자는 전통적으로 자의식, 양심, 행위에 대한 의지 등등의 속성을 지녀야 하는 것으로 이해되어 왔다. 이런 의미에서 인간을 제외한 많은 동물들도 도덕적 행위자로 받아들여지고 있지 못하다. 지능을 지녔다고 인정하는 동물들조차도 단지 도덕적 배려의 대상으로 여겨질 뿐이며, 도덕적 행위자로 인정되지는 않는다. 이 주제 역시 별도의 논의를 필요로 하는 다양한 쟁점을 지닌 주제이다. 그러나 이 논쟁이 시사하는 바는 아무리 개별화된 유기체로서 인간과 유사한 특징을 많이 가지고 있고, 설사 유전학적으로는 인간과 대

단히 유사한 DNA를 가지고 있는 고등동물이라 하더라도, 이런 존재를 도덕적 행위자라고 보기는 어렵다는 점이다. 인간을 도덕적 행위자로 인식하게 하거나 간주하게 하는 것은, 설사 그것이 현실의 모든 인간이 갖추고 있는 속성은 아닐 수 있지만, 다음과 같은 특징을 지니고 있기 때문이다.

첫째, 인간은 자신의 삶과 행위를 뒤돌아보고, 대상화하는 반성적 사유를 한다. 따라서 자신의 행위가 타인에게 어떻게 평가될 것인지 신경 쓴다. 이런 고차원적인 지능의 기능은 감정이나 정서를 논의했을 때와 마찬가지로 주체와 타자를 구별하고, 자신의 주체성을 인식하며, 자신의 정체성 인지를 전제한다.

둘째, 인간은 자연 속에서 살고 있지만, 자연의 한 부분으로만 살아가고 있지 않다. 윤리 및 도덕이라는 것이 다 그런 성격의 것이다. 문화도 마찬가지이다. 문화 역시 자연을 이용하여 자연 위에 세운 구성물들이다. 윤리, 문화, 교육 등등은 자연적인 것이 아니다. 오히려 자연 위에 인간이 구성한 구성물들이다. 그러나 이런 구성물이 필요했던 이유는 자신의 생존뿐만 아니라 종의 생존과 번식을 위한 것이다. 번식은 유한자가 자신의 정체성 중 일부를 비교적 긴 시간동안 유지하는 방법이다..

셋째, 유한자로서의 인간은 자신의 유한성을 인지하고 있기에, 초월적 존재에 대한 이해 및 초월적 가치 내지 보편적 가치를 추구하며 살아간다. 인간 개개인 모두가 적어도 나와 마찬가지로 유한하지만 초월적 가치를 또는 영원한 가치를 추구하고 살아가는 존재이다. 그리고 이러한 존재들은 그 어떤 목적에 대한 수단이 될 수 없는 목적 그 자체인 존재로 인식된다.[32]

넷째, 인간은 관계적 존재로서 살아간다. 관계적 존재이기 때문에 자존감도 중요하고, 타인으로부터 인정을 받고자 하는 욕구도 있다. 도덕적 행위자란 목적적 존재자들 상호 간에 준수해야 할 보편 규범을 준수하고자 하는 자들이다. 도덕이나 윤리규범이 존재하는 이유는 유한자로서 자연의 한 부분으로 살고 있지만 단지 그런 존재자의 특성만으로 살아 갈 수 없는 목적적 존재로서의 삶에 대한 자각에 기초한다. '살인하지 말라'는 도덕 규범은 인간의 유기체가 단 한 번의 삶을 사는 동물로서의 생물학적 특징을 지니고 있기 때문에 생긴 규범일 것이다. 생식이 개인뿐만 아니라 사회적으로도 중요한 것은 한 사회의 지속성과 관련이 있기 때문이다. 약속을 지켜야 하는 것은 사회생활을 통해 협력을 하며 살아 갈 수밖에 없는 인간의 삶의 조건과 관련이 있다. 윤리 규범 자체는 사회적 관계성을 전제로 한다. 누구와의 관계성이 존재하지 않는 한 윤리 규범이 유의미한 가치를 지닐 수는 없을 것이다.

하지만, 인공지능이 위와 같은 특성을 지닌 지능으로까지 발전할 수 있을지는 의문이다.

2) 도덕적 행위자와 자유의지 및 책임

'도덕적 행위자'라는 존재론적 지위가 부여된다는 것은 자신의 정체성에 대한 인식, 즉 자의식뿐만 아니라, 관계성을 전제로 하며, 자연 속에서 유한한 삶을 영위하는 생존 욕구를 전제로 한다. 도덕적 행위자를 통제하는 책임이란 자신의 의도를 실현하는 데 있어 자신이 행위의 선택자이고 주체자라는 특징을 전제로 한다. 법적으로 고의성이 중요한 것은 윤리적으로도 자유의지와 무관하지 않다. 우리는 우리의 움직임이 아니라 우리의 의도가 개입된 행동에 대해 책임을 진다. 그것은 자신의 진정성이 담긴 행동이기 때문이고, 그 행동에 대한 정당화 근거를 제시하며, 그 행동으로 인해 발생한 결과에 대해 책임을 진다. 인공지능은 자신의 작업 수행 결과에 대해 책임을 지는가? 또는 자신의

32) 임마누엘 칸트, 『윤리형이상학의 정초』, 백종현 옮김, 아카넷, 2005, 145-148면.

수행결과가 정당하다고 주장할 수 있는가? 인공지능은 책임을 진다는 것의 의미를 알 수 있을까?

책임과 관련하여 인간이 욕구를 지닌 존재라는 특성은 주목할 만하다. 자기 자신을 위한 행동, 자신의 행복을 추구하는 행동, 나아가 생물학적으로는 생존과 번식을 위한 행동으로 해석되는 여러 행동들은 욕구를 전제로 한다. 이러한 욕구가 실현되지 않을 때, 인간에겐 고통이나 절망이란 정서가 뒤따른다. 처벌은 바로 이 욕구의 제한을 통해 의도적으로 고통을 가하는 것이다. 책임을 진다는 것은 다양한 형태로 실현된다. 약속한 것을 이행하게 함으로써 책임을 지게 하는 방법도 있고, 손실에 대해 보상하거나 손상에 대해 배상하는 방법도 있으며, 사회 규범을 해침으로써 사회적 가치를 손상시킨 경우에는 특별히 벌이라는 형식으로 책임을 묻는다. 이런 종류의 벌은 욕구를 제한함으로써 자신의 행위에 대한 대가를 지불하게 하는 특성을 지닌다. 생명에 대한 인간의 욕구가 가장 크기 때문에 생명형이 가장 가혹한 형벌이고, 인간은 신체의 자유를 추구하고 이것이 가장 기본적인 권리 중 하나이기 때문에 징역형이 존재하며, 이동의 자유를 추구하므로 고정된 장소에 가두는 것이고, 금전 즉 소유욕이 있기 때문에 벌금을 내는 형벌을 둔다. 과연 인공지능에게 이와 같은 효과를 발휘하게 하는 형벌이 존재할 수 있을까? 인공지능이 욕구를 지닌 존재자가 아니라면 이런 벌은 다 무의미하다. 벌이 존재할 수 없다는 그 자체가 어떤 의미에선 인공지능은 도덕적 행위자로서의 존재론적 지위가 부여될 수 없음을 의미한다.

따라서 '인공지능에게 책임을 물을 수 있을까?'라는 질문은 복합질문의 오류가 개입된 것으로 보인다. 왜냐하면 책임을 물을 수 있을지 묻기 이전에 인공지능이 과연 행위자이기는 한 것인지 답해야 하기 때문이다.

이중원은 책임의 문제를 자율성과 연결시켜 논의하는데, "인공지능 시스템 자체에 (인간과는 다른 의미의) 어느 정도의 선택의 '자율성'이 있다고 말할 수 있다."고 말한다.[33] 그렇다면 필자는 '자율성'이란 용어를 사용하는 것에 반대한다. 비유적 표현은 우리의 담론을 혼란시킬 수 있기 때문이다.

이중원은 자율주행 자동차의 사고와 관련하여, "만약 우리가 고려할 수 있는 모든 인간 행위자들에 대해 더 이상 사고의 책임을 물을 수 없는 경우, 자율적 인공지능에 기반한 자동차 의사결정 시스템에 궁극적으로 책임을 물어야 할 상황이 발생할 수 있다."고도 말한다.[34] 하지만 필자는 이 표현이 여전히 비유적으로 들린다. 시스템이 행위자인지부터 따져할 것이지만, 욕구가 없는 시스템에게 어떻게 책임을 물을 수 있다는 것인지 이해하기 어렵다.

물론 이중원도 인공지능에 전통적인 책임개념을 적용할 수는 없을 것이라는 견해를 피력하고 있다.[35] 이중원은 책임이 아닌 책무(accountability) 개념을 적용해 볼 수는 있을 것이라고 다음과 같이 주장한다.

> 인공지능 시스템의 활용 과정에서 사고가 발생했을 때 이에 대한 합당한 설명을 요구하는 경우, 우리는 인공지능 시스템에 논란이 많은 책임(responsibility) 개념 대신에 설명에의 의무에 바탕 한 책무(accountability) 개념을 (현 단계에서) 적용해 볼 수 있을 것이다. 여기서 책무는 주로 자기 자신의 행동을 설명할 수 있는 능력에 기반하고 있기에, 인공지능 시스템에 대해 의사결정 과정을 설명하고 오류 또는 예기치 않은 결과를 식별할 수 있는 능력

33) 이중원, "인공지능에게 책임을 부과할 수 있는가? : 책무성 중심의 인공지능 윤리 모색" 과학철학 22권 2호, 2019, 82면.
34) 이중원, "인공지능에게 책임을 부과할 수 있는가? : 책무성 중심의 인공지능 윤리 모색" 과학철학 22권 2호, 2019, 85면.
35) 이중원, "인공지능에게 책임을 부과할 수 있는가? : 책무성 중심의 인공지능 윤리 모색" 과학철학 22권 2호, 2019, 89면.

을 바탕으로 책무를 논할 수 있다.[36]

그러나 필자는 여기서의 책무조차도 설명이나 해명 정도에 불과하다고 평가한다. 이런 종류의 책무는 'duty'의 번역어인 책무와도 다르다.[37] 해명해야 한다는 정도를 가지고 과연 책임이나 책무라는 도덕적 지평의 개념과 연결시킬 수 있을지 의문이다. 그렇기 때문에 이런 성격의 책무는 역설적으로 인공지능을 도덕적 행위자로 보기 어렵게 한다. 결국 인공지능의 오류[38]는 제조자나 설계자가 책임져야 하며, 조작 미숙이 개입되어 있다면 사용자도 책임져야 한다. 설명이나 해명이 제조자, 설계자, 사용자의 책임소재를 밝히는 데 기여할 것이다. 그리고 책임은 나누어질 수도 있을 것이다. 그러나 인공지능의 책무 적용은 인공지능이 결코 유의미한 행위자가 될 수 없음을 보여주기만 할 뿐이다.

5. 맺음말

필자는 인간이 도덕적 행위자로서 지니고 있는 존재론적 특성이 인간이 지구에서 생존하고 번식하기 위해 개발된 특이한 특성인지 확신할 수 없다. 다시 말해, 동일한 특성이 다른 생물학적 존재자가 다른 환경에서도 다른 물리적 기반을 가지고 실현할 수 있을지 여부에 대해서는 명확한 답을 가지고 있지 않다. 이런 특성이 인간에게서만 실현된다고 단정할 수는 없을 것이지만, 적어도 생물학적 특성으로서 유한성을 인식하는 가운데 영원한 가치를 추구하며 함께 생존해야 한다는 특성이 인간이 인간으로서 이와 같이 살아가게 한 특성이라고 생각한다.

그런데 인간과 유사한 기능을 수행하는 존재자가 이 우주 어디엔가 존재할 수는 있을 것이다. 다시 말해 다른 도덕적 행위자가 존재할 가능성은 있다. 그러나 그렇다고 해서 그 존재자가 인간과 동일하다고 보기는 어려워 보인다. 같다거나 다르다는 것은 관점에 의존한다. 존재론적 동일성이란 기능적 동일성보다 훨씬 더 생물학적인 물리적 기반에 의존하는 것이라 여겨진다.

인공지능이 과연 이러한 특성을 지니게 될 것인지는 여전히 확실한 과학적 증거를 제시하며 답할 수 있을 정도는 아니다. 아직 신경과학이 이런 과학적 증거를 제시할 정도로 발달했다고 보기 어렵기 때문이다. 그러나 이 글에서 언급한 인간의 특성, 즉 인간을 인간이게 하는 특성은 생물체가 생존과 번식을 위해 자연환경과 씨름하며 문제를 해결하면서 형성시킨 특성임에는 틀림없다. 적어도 유한성은 생물체에게서 발견되는 것이다. 물질도 영원한 것은 아니지만 인공지능이 구현하는 지능이 물리적 기계 작용을 넘어서서 생물체와 같은 주체성과 유한성의 자의식을 형성할 것일지는 여전히 많은 의문을 갖게 한다. 적어도 기계가 유기체와 같이 어떤 욕구를 지니고 있다는 것을 입증하는 것이 과연 가능한 것인지 의문이다. 이런 특성들이 인공지능에 부여된 것이 아니라 인공지능 스스로가 갖춰나간다는 점이 입증되기 전까지, 필자는 인공지능은 여전히 인간의 일을 대신하는 기계이며, 인간의 지능 중 파편적인 일부 기능을 흉내내는 기계라고 본다. 따라서 인공지능이 욕구를 지닌 행위자라거나, 더 나아가 도덕적 행위자임을 인정하는 판단을 필자는 유보할 수밖에 없다.

---

36) 이중원, "인공지능에게 책임을 부과할 수 있는가? : 책무성 중심의 인공지능 윤리 모색" 과학철학 22권 2호, 2019, 91면.
37) 간혹 'obligation'과 'duty'를 구별하지 않고 둘 다 '의무'로 번역하기도 하지만, 필자는 전자는 의무, 후자는 책무로 번역하는 것이 적절하다고 본다.
38) 필자는 '실수'라는 의인화된 용어조차도 부적절할 수 있다 생각하여 '오류'라는 용어를 사용하였다.

【Abstract】

Can Artificial Intelligence become an Agent Like a Human Being?*

Kyungsuk Choi**

This paper aims to clarify the ontological status of artificial intelligence and features of human intelligence by dealing with the difference between them. Recently, as AlphaGo shows, artificial intelligence has the network of computers and adopts deep learning technology with big data. Further, artificial intelligence is introduced to the scope of emotion as well as to that of calculation. However, artificial intelligence just imitates human emotion because subjectivity and holistic personality of emotion exposer must be presupposed and because emotion plays a role with authenticity. the function of human brain is to make a human being survive. Human intelligence can be defined as a problem-solving capacity for survival. Thus, artificial intelligence is not the same as human intelligence. In order for artificial intelligence to be an human intelligence, it uses its intelligence for its problem-solving for survival. But it is hard to say that artificial intelligence has a desire to survive. In addition, it is also hard to say that artificial intelligence has or will have self-awareness like human being's. Human self-awareness is a very high level of intelligence activity in which a human being recognizes one's finitude as a biological organism and one's self identity as well as otherness through the cooperation with others. The difference between artificial intelligence and human one leads us to think that an artificial intelligence cannot become an agent with intention and not even moral agent. The attitude or project to consider artificial intelligence to be an human intelligence is only to show that there are big differences between them.

Key words: artificial intelligence, human intelligence, authenticity, agent, identity, self-awareness

투고(접수)일(2020년 4월 22일), 심사(수정)일(1차 : 2020년 6월 8일, 2차 : 6월 20일), 게재확정일(2020년 6월 23일)

* This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (NO. 2019M3E5D2A02064496)

** Professor, School of Law / Bioethics Policy Studies, Ewha Womans University, choiks@ewha.ac.kr