1. 인공지능과 인류의 미래

인공지능을 통해 인류가 발전을 이룰 수 있느냐, 오히려 퇴보하느냐 하는 논쟁이 과학계 안팎으로 활발하다. 과학기술자로 성장할 우리는 인공지능이 인류의 미래에 미칠 영향에 대해 어떻게 생각하고 있으며 어떠한 관점을 가지고 이 현상을 판단해야 하는가. 이에 대해 전공자로서, 보다 깊이 있는 읽기 자료를 통하여 고민해보도록 하자. 지금부터 미래로 이어지는 인류의 새로운 과학기술의 역사를 통찰할 수 있을 것이다.

다음에 제시한 예문은 2019년에 출판한 『포스트휴먼 오디세이』에 실린 글로, '인공지능이 사람을 대체할 수 있는가' 하는 질문에 대하여 인공지능의 한계를 논하고 있다. 글을 읽고 물음에 답해보자.

인공지능, 격렬한 논쟁의 핵이 되다

홍성욱*

초기 핵심적인 인공지능 연구자들은 인간의 두뇌를 닮은 컴퓨터를 만들어서 인공지능을 구현하려고 하는 대신에, 논리적 연산을 빨리함으로써 궁극적으로는 인간의 결정과 비슷한 결정을 내리는 기계를 만들려고 했다. 1956년 존 매카시가 조직한 인공지능에 대한 다트머스 학회에서 만난 매카시, 마빈 민스키, 앨런 뉴얼, 허버트 사이

* 서울대학교 생명과학부 교수. 과학사 및 과학철학 협동과정 전공주임. 저서로는 『크로스 사이언스』, 『홍성욱의 STS, 과학을 경청하다』, 『그림으로 보는 과학의 숨은 역사』, 『미래는 오지 않는다(공저)』, 『슈퍼휴머니티(공저)』, 『4차 산업혁명이라는 유령(공저)』 등이 있다.

먼은 불과 몇 주 동안 의견을 교환하고 헤어졌지만 이후 MIT(민스키), 스탠퍼드 대학교(매카시), 카네기 멜런 대학교(뉴얼, 사이먼)에 인공지능 연구 센터를 설립해 인공지능 연구에 박차를 가했다. 1950년대부터 1970년대까지 이들은 가까운 미래에 인간처럼 생각하고 판단하는 인공지능을 만들 수 있다고 낙관했다.

같은 시기에 인공지능을 둘러싸고 철학적 논쟁이 활발했다. 일군의 철학자들과 인공지능 연구자들은 인공지능에 대한 낙관론을 강력하게 비판했다. 인공지능은 인간과 같은 방식으로 생각하지 않으며, 따라서 결코 인간과 같은 수준의 지능을 가질 수 없다는 이유에서였다. 또 인공지능의 사회적 영향이 연구자들이 생각하는 것보다 부정적일 수 있다는 비판도 제기되었다. 이런 비판들에 대해서는 다시 반론이 이어졌고, 이들 중 일부는 인공지능 연구에 영향을 주었다.

지금 다시 르네상스를 맞고 있는 인공지능이 우리 사회를 어디로 인도할지 궁금한가? 1960~1970년대에 벌어졌던 논쟁에서 우리의 궁금증에 대한 답을 어렴풋하게나마 찾을 수 있다.

기호 인공지능 연구와 낙관론

매카시, 민스키, 뉴얼, 사이먼 같은 인공지능 선구자들은 한결같이 사이버네틱스에 기원을 둔 신경망 접근법을 거부하고 기호 논리에 근거한 기호 인공지능의 접근법을 택했다. 이들의 초기 연구는 크게 세 가지 방향으로 진행되었다. 하나는 인간을 이기는 체스 프로그램을 만드는 것이었다. 체스 프로그램은 컴퓨터 속도와 용량이 발전하면서 그 기술도 급속도로 발전했는데, 1967년에 한 체스 프로그램은 매사추세츠주의 챔피언 경기에 출전한 선수를 상대로 첫 승리를 거두었다. 두 번째 연구 방향은 논리적 결정을 내리는 인공지능에 맞춰졌다. 뉴얼과 사이먼의 '논리 이론가' 프로그램(1955)은 복잡한 논리 문제를 순식간에 풀었고, 이들은 이를 확장해서 일반적인 지적 결정을 내리는 '범용 문제 해결사(General Problem Solver)'라는 프로그램을 만들었다.

세 번째 방향은 컴퓨터가 인간의 자연언어를 이해하게 하는 것이었다. 이를 위해서는 인간이 사용하는 언어의 구조를 분석해서 이를 인공지능에 입력하는 것이 필요

했다. 사람이 적절한 말을 입력하면, 컴퓨터는 이미 입력된 언어 구조에 맞춰 이를 판독하고 적절한 방식으로 답을 했다. MIT의 컴퓨터 공학자 조지프 와이젠바움(Joseph Weizenbaum)은 1966년 언어를 이해하는 첫 번째 인공지능 프로그램 '일라이자(ELIZA)'를 개발했다. 일라이자는 채팅 프로그램이었는데, 특히 심리 상담을 하는 정신과 의사를 훌륭하게 흉내냈다. 예를 들어, 일라이자는 이렇게 사람과 대화를 나눌 수 있었다.

젊은 여성: 남자는 다 똑같아요.
일라이자: 어떤 의미에서요?
젊은 여성: 남자는 항상 우릴 들들 볶아요.
일라이자: 구체적인 예를 들어줄래요?
……
젊은 여성: 그는 내가 항상 우울하다고 해요.
일라이자: 우울하다는 얘길 들어서 안됐습니다.
젊은 여성: 사실이에요. 나는 불행해요.
일라이자: 여기 오면 불행하지 않을 거라고 생각해요?

이런 발전을 이룬 인공지능 분야의 선구자들 대부분은 낙관적이고 유토피아적인 비전을 공유했다. 이들은 가까운 미래에 인공지능이 인간의 주관이나 편견에 오염되지 않은 판단과 지식을 낳음으로써, 인간과 기계 사이의 갈등을 해소함은 물론 주관성과 객관성 사이에 존재하던 오랜 갈등과 긴장을 완전히 해소할 수 있을 것이라고 전망했다. '논리 이론가'를 만든 사이먼은 이 프로그램이 인공지능 컴퓨터 같은 '물질로 구성된 시스템이 어떻게 마음의 속성을 가질 수 있는지'를 보임으로써 서양 철학의 오래된 '마음/몸(mind/body)의 문제'를 완전히 해결했다고 주장했다. 논리 문제를 푸는 인공지능은 기계적 연산을 하는 컴퓨터의 연장이라고 볼 수 있지만, 사이먼은 자신의 인공지능이 과거의 컴퓨터와 전혀 다른 새로운 패러다임이라고 생각했다.

1958년 사이먼과 뉴얼은 향후 10년 안에 컴퓨터 프로그램이 세계 체스 챔피언을 이기고, 새로운 수학 명제를 만들어 증명할 것이라고 단언했다. 1967년 민스키는 10년 안에 인공지능과 관련한 모든 문제가 해결될 것이라고 전망했다. 같은 해 사이먼은

심리학 이론이 컴퓨터 프로그램과 비슷해질 것이라고 선언했다. 인간의 마음이 컴퓨터와 같아서 마음에 대한 연구인 심리학은 곧 컴퓨터 프로그램에 해당한다고 생각했기 때문이다. 1970년 민스키는 몇 년 뒤에 인간의 일반 지능 수준의 인공지능 기계를 갖게 될 것이라고 낙관했다.

연구자들의 낙관론은 단순히 인간처럼 생각하는 인공지능을 만드는 데 한정되지 않았다. 이들은 인공지능이 서양 철학의 오랜 주제였던 인식론과 존재론의 난제를 해결해줄 수 있다고 믿었다. 예를 들어, 인공지능이 가능한 선택지를 평가하고 이 중 하나를 고르는 과정에 대한 이해는 자유의지와 결정론 사이의 철학적 모순을 해결해줄 수 있다는 것이었다. 더 나아가서 매카시는 믿음, 지향(intentions), 욕구 같은 인간의 정신적인 특질을 인공지능에도 부여할 수 있다고 주장했다. 인공지능에 관한 연구가 인간 마음의 심연을 파헤치는 도구가 되는 것이다.

철학자들의 반론과 비판

철학자들은 인공지능에 대한 낙관론을 비판했다. 영국의 철학자 존 랜돌프 루카스(John Randolph Lucas)는 인간의 마음이라는 것이 부분으로 환원되지 않는다는 점에서 기계의 매커니즘과 본질적으로 다르며, 따라서 아무리 인공지능이 발전해도 인간의 마음을 흉내낼 수 없다고 주장했다. 마음은 나눌 수 없지만, 부품들로 조합된 기계는 분해할 수 있기 때문이다. 그는 인공지능이 인간을 따라올 수 없다는 점을 괴델의 정리에 근거해서 전개했다. 괴델의 정리에 따르면 수학적 명제들은 닫힌 형식 체계 안에서만 진릿값을 갖는데, 인간은 (괴델의 정리에서 보듯이) 형식 체계 밖에 위치할 수 있지만 하나의 체계에만 머무르는 컴퓨터는 그럴 수 없다는 것이다. 그래서 서로 다른 형식 체계를 상정하는 인간의 마음은 하나의 체계에만 머무르는 인공지능보다 뛰어날 수밖에 없었다.

미국의 철학자 휴버트 드레이퍼스(Hubert Dreyfus)는 하이데거나 모리스 메를로 퐁티(Maurice Merleau-Ponty)같은 유럽 철학의 전통에 근거해서, 인간의 사고가 마치 프로그램처럼 정해진 규칙에 따라서 상징을 조작하는 것과 흡사하다는 인공지능 연구자들의 기본 가정을 공격했다. 드레이퍼스에 따르면 인간의 사고는 인공지능

의 결정과 두 가지 점에서 달랐다. 그중 하나는 인간의 사고가 우리가 이미 알고 있는 복잡한 '맥락(context)' 속에서 일어난다는 것이다. 두 번째로 인간의 사고는 컴퓨터가 정보를 처리하는 선형적인 방식이 아니라 여러 단계를 건너뛰면서 거의 무의식적이고 직관적으로 이루어진다는 것이다.

이러한 직관적인 방식은 인간이 가진 '암묵지(tacit knowledge)'와 밀접하게 연결되어 있다. 사람들은 알고 있지만 글로 쓸 수는 없는 암묵지들을 가지고 있는데, 부호로 쓰인 알고리듬으로만 구성된 인공지능은 이런 암묵지를 가질 수 없으므로 인간과 같은 방식으로 세상을 이해할 수 없다는 것이 드레이퍼스의 생각이었다. 드레이퍼스는 인공지능이 이런 '맥락'과 '암묵지'를 포함하지 않는 한 인간과 비슷한 사고를 할 수 없다고 비판했다.

그는 또 인간의 사고가 인간의 몸과 밀접하게 관련되어 있다는 점을 지적하면서, 컴퓨터가 몸을 가지지 않는 한 인간처럼 사고할 수 없다고 주장했다. 드레이퍼스의 주장은 후설과 하이데거의 영향을 크게 받았지만, 마투라나와 바렐라의 급진적인 세계관에 닿아 있었다. 이것은 특히 인간의 인지가 외부 세계가 두뇌의 신경회로에 반영된 것이 아니라, 몸을 가진 인간이 세상을 향해 행한 것이라고 생각한 마투라나와 바렐라의 2차 사이버네틱스의 인지 이론과 상통했다.

또 다른 철학자 존 설(John Searle)은 유명한 '중국어 방(Chinese Room)' 논변을 폈는데, 이 논변은 이랬다. 중국어를 전혀 모르는 사람이 단어 카드가 잔뜩 있는 방 안에 있다. 이때 적절한 중국어 매뉴얼이 주어진다면, 그는 방 밖에서 입력하는 중국어 질문에 대해 적당한 중국어로 대답을 할 수 있을 것이다. 따라서 밖에서는 방 안에 있는 사람이 마치 중국어를 이해하는 듯 보이지만, 사실은 이 모든 과정이 중국어를 전혀 모르는 채로 매뉴얼에 따라서 진행된 것일 뿐이다.

설은 이것이 인공지능 컴퓨터가 질문에 답하는 방식이라고 보았다. 즉, 인공지능은 중국어를 전혀 '이해하지' 못하는데도 마치 중국어를 이해하는 것처럼 보일 수 있었다. 인공지능 연구자들은 인공지능이 마음을 가지고 '생각한다'고 주장했는데, 설은 이에 대해 인공지능이 '생각한다', '사고한다'고 말할 수 없다면서 이런 논변을 제시했다. 설은 인간처럼 생각하는 인공지능을 '강한 인공지능', 체스를 두는 인공지능을 '약한 인공지능'으로 구분해 인공지능 연구자들이 만들었다는 인공지능은 모두 약한 인

공지능에 불과하다고 주장했다.

인공지능 전문가 중 일부는 이런 철학자들의 비판에 응수하거나 이를 다시 역비판하기도 했지만, 다수는 이를 무시하거나 조롱했다. 컴퓨터 프로그램 한 줄도 써 본 경험이 없는 비전문가인 그들의 비판에 귀를 기울일 필요가 없다는 이유에서였다. 스탠퍼드 대학교의 인공지능 연구팀은 드레이퍼스에게 자신들이 개발한 체스 프로그램과의 대국을 제안했는데, 체스를 잘 둔다고 자랑하던 드레이퍼스가 패하자 야유를 퍼부었다. 인지과학자와 철학자 중에는 설의 '중국어 방' 논변에 대해 비판을 하거나 대안을 내놓는 사람이 많았지만, 인공지능 연구자들은 설이 인공지능을 이해하지 못하는 철학자에 불과하다고 평가하면서 '중국어 방' 논변을 무시했다. 철학과 같은 인문학과 컴퓨터 과학 사이의 거리는 1959년 영국의 작가이자 과학자인 찰스 퍼시 스노(Charles Percy Snow)가 얘기한 '두 문화(two cultures)' 사이의 간극을 그대로 드러낸 것이었다.

인공지능 학계 내부로부터의 비판

그러나 인공지능 프로그래머들이 무시할 수 없는 비판도 있었다. 그 비판은 놀랍게도 대화 인공지능인 '일라이자'를 개발한 와이젠바움에게서 나왔다. 와이젠바움은 자신의 간단한 프로그램과 대화를 나눈 MIT 직원들이 이 프로그램을 진짜 사람이라고 착각하며 자신들의 사생활을 털어놓은 사실에 충격을 받았다. 그들은 자신들과 대화를 나눈 '의사'가 프로그램이었으며, 이 기록을 와이젠바움이 쉽게 볼 수 있었다는 사실을 알고는 화를 냈다. 와이젠바움은 이런 경험을 통해 자신의 인공지능 프로그램이 인간의 삶에 위협이 될 수 있음을 깨달았으며, 1972년 『컴퓨터 권력과 인간의 이성(Computer Power and Human Reason)』을 출판해 인공지능의 발전 속도를 늦춰야 한다고 주장했다.

와이젠바움은 인공지능 연구 초기에 제너럴 일렉트릭사에서 '뱅크오브아메리카'의 전산화 작업에 관여한 적이 있었다. 이 과정에서 그는 자신이 수행한 전산화가 수백만 명의 삶에 큰 영향을 줄 수 있음을 깨달았다. 자신은 효율적인 은행 업무를 위해 만든 것뿐인데, 이로 인해 사람들이 해고되는 사태가 발생했기 때문이다. 나중에 일라이자

에 매혹된 사람들을 보면서, 그는 사람들이 컴퓨터가 제공하는 환영을 그대로 받아들이는 것이 충분히 가능하다고 생각했다. 일라이자 같은 인공지능은 실제로 아무런 감정이 없는데도 환자들은 실제 인간 의사가 마치 애정과 연민을 가지고 자신들과 상담하고 있는 것이라고 생각했기 때문이다.

인간만이 동정심 같은 감정적 요소를 포함한 지혜로운 선택을 내릴 수 있었다. 따라서 와이젠바움은 인간이 하는 중요한 결정을 인공지능이 대신하게 해서는 안 된다고 주장했다. 인공지능의 결정에는 감정이 결여되어 있기 때문이었다. 인공지능이 법관의 판결을 흉내낼 수는 있지만, 법관의 역할을 하게 해서는 안 된다는 것이 그의 입장이었다. 인간의 마음은 오랜 역사를 통해 만들어졌을 뿐만 아니라, 몸과 밀접한 연관을 맺으면서 형성된 것이다. 따라서 이런 인간의 마음은 인공지능에 의해 정확하게 모사될 수 없었다. 인공지능은 인간이 그 한계와 장점을 정확히 파악할 때 인간에게 유용한 도구가 될 수 있다는 것이 그가 내린 결론이었다.

인공지능 연구자들은 와이젠바움의 이런 비판을 당혹스러워했다. 그가 누구보다도 컴퓨터와 인공지능을 잘 아는 전문가였기 때문이다. 일군의 연구자들은 인공지능이 인간의 편견을 공유하지 않기 때문에 결정을 내리는 데 유용한 도구가 될 수 있다고 반박했다. 다른 한편에서는 완벽하게 결과를 예측하는 프로그램을 만드는 것은 불가능하다는 것을 인정하면서도, 인공지능의 발전을 멈추기보다는 이를 발전시키면서 그 영향을 단계마다 평가하는 것이 유일한 현실적인 방법이라고 했다. 와이젠바움의 비판은 인공지능 연구자들에게 자신들이 개발한 알고리듬의 한계를 다시 생각하게 했다.

현재와 미래의 인공지능

전체적으로 평가할 때, 1960년대부터 1970년대까지 인공지능 연구자들은 인공지능에 대한 비판을 심각하게 받아들이지 않았다. 이들은 새처럼 날지 않아도 날기만 하면 된다는, 즉 인간처럼 생각하지 않아도 생각만 하면 된다는 철학을 가지고 연구에만 전념했다. 1980년대 이후 인공지능에 대한 과도한 기대가 한풀 꺾이고 기호 인공지능의 문제점이 더 확연하게 드러나면서, 연구자 공동체 내에서도 이런 비판들이 재조명

되었다. 드레이퍼스나 설의 이름을 직접 언급하지 않더라도 인공지능 연구자들은 자신의 논의에 이러한 비판과 대안을 반영하기 시작했다. 이들은 인간의 지식과 행동의 맥락을 인공지능 프로그램에 삽입하려고 노력했다. 또 인공지능이 무엇을 '안다', '이해한다', '자각한다'라는 표현을 좀 더 신중하게 사용하기 시작했다.

2010년 이후 인공지능은 다시 르네상스를 맞았다. 자율주행자동차가 면허를 획득하고, 알파고가 이세돌 국수를 4 대 1로 이겼다. 사물인식 프로그램의 경우, 오류가 1.5%로 줄어 인간보다 정확해졌으며, 인공지능을 통한 번역도 훨씬 더 매끄러워졌다. 인공지능 비서, 인공지능을 이용한 (책, 영화 등의) 취향 판단, 인공지능을 이용한 결정이 여러 영역에서 사용되고 있다. 철학자 중에는 미래에 초지능이 도래해서 인류에게 해를 입힐 것을 걱정하는 사람들도 있다.

초지능의 문제는 먼 미래 얘기이며, 실제로 초지능이 도래하지 않을 가능성도 크다. 현재 우리가 당면한 문제는 인공지능의 발전으로 인한 실업의 가능성, 빅데이터의 편향에서 발생하는 인공지능의 편향과 편견 문제, 빅데이터와 인공지능에 의한 프라이버시 침해 등이다. 이 문제들은 과학기술자, 인문학자, 시민운동가, 정치인과 관료, 예술가, 그리고 시민사회 구성원들이 함께 참여해 해결해야 한다.

오늘날 인공지능이 낳은 이런 윤리적 문제들이 중요하고 심각한 문제로 평가되면서 와이젠바움의 비판이 주목받고 있다. 그는 우리가 인공지능의 장점과 한계를 정확하게 인식하는 것이 인공지능의 혜택을 취하는 지름길이라고 주장했으며, 사회가 인공지능을 받아들이지 못하는 상황에서는 인공지능 연구 속도를 늦추든가 아니면 잠시 멈추자고 주장했다. 이는 지금 우리에게도 시사하는 바가 크다. 시인 바이런이 얘기했듯이, "미래에 대한 최선의 예언자는 과거"이다.

인공지능에 대한 낙관론이 넘쳐나는 지금 와이젠바움의 주장은 우리 시대에 절실한 '느린 과학(slow science)'의 가치를 제공한다. 느린 과학이란, 과학기술의 발전이 낳는 여러 가지 사회적 결과에 대해 충분히 논의하고 과학의 발전 속도와 방향을 시민사회가 참여해서 정하는 것을 말한다. 과학기술의 발전을 멈추자는 얘기가 아니다. 과학기술의 발전이 우리 사회를 어떤 방향으로 몰아가고 있는지, 이 방향이 후세대를 위해 바람직한 것인지를 고민하면서 발전을 꾀하자는 얘기다. 이는 기술을 엔지니어의 전유물로 보고 기술을 통제하기만 하려는 입장을 극복하고, 기술과 인간이 공존하는

사회를 꿈꾸는 탈인간중심주의적 기술철학이다. 인공지능에 대한 포스트휴먼 감수성은, 인공지능의 한계와 장점을 잘 파악하고, 인간의 가능성과 한계를 고민하면서 인간과 인공지능의 상생적 관계를 조심스럽게 모색하는 것이다.

연습문제 1

1. 위 글에서는 인간의 사고와 인공지능의 결정에 차이가 있음을 밝히고 있다. 이를 토대로, 현재 존재하는 직업 중에서 인공지능이 대체하기 어려운 것이 있다면 무엇인지 제시하고 그 근거를 설명해보자.

2. 위 글에 나타난 인공지능의 장점과 한계를 생각해보고, 미래에 인간이 인공지능보다 경쟁력을 갖추기 위해 어떤 역량을 개발해야 하며 이를 위해 어떤 노력을 기울여야 하는지 한 편의 글로 적어보자.