【서평】 과학철학 21권 1호 (2018) pp. 69-79

인공지능의 진화는 불가능한가?

김재인, 『인공지능의 시대, 인간을 다시 묻다』 (동아시아, 2017)

전 진 권†

김재인 박사의 『인공지능의 시대, 인간을 다시 묻다』는 철학의 관점에서 인공지능의 문제를 다루는 저작이다. 이 책은 인공지능의 구조와 설계에 대한 상세한 설명을 제공하면서도 넓은 철학적 주제를 다루는 정보적으로 풍부한 작품이다. 목차를 보면 이 책이 얼마나 방대한 주제를 포함하는지 알 수 있는데, 2장은 지능적 에이전트라는 인공지능 기술의 핵심 프로젝트를 설명한다. 이어서 철학적 주제로 심신문제(3장), 인과와 시간(4장), 플라톤(5장), 데카르트(6장)가 서술된다. 마지막 7장은 이 책의 주제를 담은 핵심 장으로 인공지능 기술에 대한 인문학적인 성찰이 들어있다. 국내 인문학자가 방대한 문헌을 종합해서 이와 같은 작업을 발표했다는 점에서 이 책의 출판은 두 발 벌려 환영할 일이다. 그러나 다루는 주제의 범위가 넓은 만큼 내용 중에는 불분명한 부분도 있다. 그런 연유로 앞으로 인공지능에 대한 논의의 발전을 위해 이 글을 서평 보다는 비평에 가까운 글로 쓰고자 한다. 지면상 이 책에 포함된 넓은 주제를 모두 다루기란 불가능하므로, 이 글은 하나의 주제, '진화'에만 초점을 두려고 한다.

진화는 이 책의 전제 주장과 밀접한 관련이 있다. 이 책의 전체 결론은 초인공지능, 즉 인간과 동일한 지능을 가진 인공지능의 개발이 가능하지 않다는 것이다(서론과 7장). 이 결론을 입증하기 위해 이 책

† 고등과학원 초학제프로그램 <인공지능: 과학, 역사, 철학> 올해의 연구단, b613jk@gmail.com.

은 인공지능과 인간지능을 구분하고, 둘의 차이점을 보이는 방식으로 논증을 구성하고 있다. 여기서 인공지능은 알고리즘적인 것으로 설명이 된다(2장). 그런데 저자는 인간지능이 무엇인지 우리는 아직 모른다고 말한다.[1] 그렇다면 어떻게 두 지능이 다르다는 주장을 할 수 있을지 의문이 들 수 있는데, 여기서 저자는 진화에 의존한다. "지능은 생명 진화의 마지막에, 우주 진화의 마지막에 나타난 특별한 현상입니다. 따라서 지능을 얘기하려면 진화론적 접근을 해야만 합니다."[2] 즉, 인간지능의 본성을 우리는 아직 모르지만, 적어도 진화의 산물인 마음을 인공지능이 따라 할 수 없으므로 두 지능은 다르다는 것이 이 책의 논증 방식이다.

서평자가 여기서 다루려는 문제는 진화를 통해 인간지능을 규정하는 부분에 모호한 곳들이 있다는 것이다. 무엇보다 이 책에서 지칭하는 진화가 무엇인지 명확하지 않다는 문제가 있다. 진화(evolution)는 사실 여러 의미로 쓰이는 용어이지만, 이 책의 맥락에서는 생물학적 진화(biological evolution)로 한정하는 게 적절해 보인다(내용 중 종종 언급되는 우주의 진화는 생물학적 진화와는 다른 의미의 진화일 것이다). 일반적으로 생물학적 진화는 개체군이 가진 형질의 세대에 걸친 변화로 정의된다(Futuyma 2013). 하지만 이 책은 이와는 다른, 독특한 방식으로 진화를 규정한다. 크게 세 가지 규정이 등장하는데 하나씩 차례로 살펴보도록 하자. 첫 번째 방식은 베르그손(Henri Bergson)의 주장에 기반 한다.

> 진화란 문제의 발생과 문제의 포착, 그리고 문제의 해결 과정이라고 해도 과언이 아니지요...그리고 문제를 구성하는 능력은 생명의 고유한 능력이라고 해요. 베르그손이 진화를 바라본 관점이기도 합니다. 진화란 생명이 문제를 창조하고 해결해온 과정이라는 거죠[3]

---

1) 김재인 (2017), p. 14.
2) Ibid., p. 14.
3) Ibid., pp. 66-68.

그러나 베르그손의 이론은 엘랑 비탈(elan vital)과 같은 비과학적 개념에 의존하는 등, 과학 이론이라고 보긴 어렵다. 게다가 이 책의 내용과 상충되는 것 같다. 예를 들어, 본문에서 저자는 진화에 대한 의인화를 비판한다.[4] 그런데 베르그손의 규정은 생명에 대한 전형적인 의인화로 보인다. 문제를 창조한다는 말은 어떤 의미일까? 예컨데 특정 유전자의 진화를 설명하려면 위의 의인화된 설명 대신 분자생물학적인 설명이 필요할 것 같다. 또한 베르그손은 본능(instinct)과 지능(intelligence)을 구별하므로 그의 이론을 충실히 따르면 인공지능은 지능이 없고 본능만을 가진다고 해야 타당할 것이다. 그런데 저자는 인공지능 또한 지능을 가질 수 있다고 말한다. “나는 인간지능이나 마음이 인공지능의 기준이나 모델이 될 필요가 없다고 봅니다. 그럴 경우 인공지능이 다른 동물 종처럼 지능을 갖춘 새로운 종으로 출현하게 될지도 모르지요.”[5] 인공지능은 지능이긴 하지만 인간지능과 다르다는 것이 저자의 주장이다. 따라서 베르그송에 따른 진화 규정은 여러 문제로 이해하기가 어렵다. 더 설명이 필요한 부분이라고 생각한다.

이어서, 진화에 대한 두 번째 규정을 살펴보자. 이 규정은 진화의 결과물에 대해서 논한다.

> 그러한 급변의 와중에 어떤 성분이 살아남는데 도움이 될지는 미리 알 수 없습니다. 환경의 변화도 무작위적이니까요. 생명 전체, 종, 개체 등 어떤 관점에서 보건, 중요한건 각각 그 안에 얼마나 무작위성 또는 차이를 확보하고 있는가입니다. 다양성은 진화에서 살아남기 위해 가장 좋은 전략이에요.[6]

> 개체이건 개체군이건 아니면 특정한 형질을 지닌 유전자이건, 미리 아무리 노력해도 소용없어요. 전략을 짠다는 것 자체가 불가능합니다. 그저 변이와 다양성을 자기 안에 되도록 많이 확보하는 게 최선이에요.[7]

---

4) Ibid., p. 325.
5) Ibid., p. 121.
6) Ibid., p. 317.
7) Ibid., p. 324.

이처럼 이 책은 진화의 과정에서 항상 다양성을 확보하는 것이 유리하므로 진화의 결과물은 항상 가장 큰 다양성을 가질 것이라고 말한다. 그러나 이 규정에 따르면 이 책이 인간지능의 핵심이라고 보는 학습 능력이 진화할 수 없다. 진화의 맥락에서 학습은 이전 경험으로부터 특정한 정보를 보관하고, 다음 번에 같은 상황에 처하면 보관한 정보를 이용해서 이득을 얻는 것을 의미한다. 여기서 학습이 진화적으로 유리하려면 특정한 조건이 필요하다. 만약 환경이 완전히 무작위적이라면 학습은 소용이 없다. 왜냐하면 이전 환경과 지금의 환경이 완전히 다를 것이기 때문이다. 반대로 환경이 완전히 고정적인 경우에도 학습은 불리하다. 이 경우라면 경험으로부터 배울 필요 없이, 정보를 그냥 유전자 안에 저장해서 선천적으로 대응하는 게 더 유리하다. 학습을 가능케 하는 인간의 뇌가 얼마나 많은 에너지를 소모하는지 생각해보면 이해하기 쉽다. 학습 능력은 비용이 많이 드는 비싼 적응이다. 따라서 학습이 유리한 환경은 변동성이 양 극단의 중간 정도에 있는 경우여야 한다(Richerson and Boyd 2005).

진화를 보는 시간의 스케일을 극단적으로 넓힌다면 이 규정이 옳을 수 있다. 그러나 그렇게 하면 중간 과정에 발생하는 변동들을 설명할 수 없게 된다. 환경은 늘 변하지만 변하는 속도는 그때그때 다르다. 환경이 충분히 오랜 기간 안정적으로 유지된다면 그 변동성의 수준에 적합한 형질이 진화할 수 있다. 스터렐리(Kim Sterelny)는 인간의 사회적 학습이 바로 그런 상황에서 진화한 적응이라는 것을 잘 보인바 있다(Sterelny 2012). 그러므로 극단적인 다양성 혹은 무작위성은 학습의 실패를 의미한다. 따라서 두 번째 규정은 학습 능력을 인간 지능의 중요 특징으로 지목하는 이 책의 주제와 배치되므로 부적절하다.

세 번째 규정은 베이트슨(Gregory Bateson)의 이론에 기반 하는 걸로 보인다. 그는 진화를 스토케스틱(stochastic) 과정으로 정의했다.

> 어떤 사건들의 연쇄가 무작위 성분과 선별 과정을 결합해서 무작위적인 것 중 단지 몇몇 결과물만 견뎌 배기도록 허용된다면, 그 연쇄는 스토캐스틱 하다고 얘기된다...스토캐스틱 과정은 진화 과정을 다른 개념으로 표현한 거라고 이해해도 맞습니다.[8)]

이 정의는 자연 선택(natural selection)에 대한 일반적인 정의와 일치한다(Darwin 1859; Sterelny and Griffiths 2012).[9] 그러나 이 규정도 문제가 있는데 자연선택은 진화를 일으키는 여러 메커니즘 중 하나에 불과하기 때문이다.[10] 그렇지만 일단 이 문제는 놔두고, 이 정의에 따라 자연선택을 중심 메커니즘이라고 생각하고 문제를 다루도록 하자.

이 정의를 바탕으로 저자는 다음과 같이 주장한다. 서평자는 아래 논증이 이 책의 핵심 주장이라고 생각한다.

> 나는 현재 우리가 알고 있는 알고리즘의 본성상 인간의 생각을 프로그램으로 구현하는 건 불가능하다고 봅니다. 프로그램에 무작위성을 내장하는 것이 불가능하다는 점과 진화의 산물인 생각은 스토캐스틱 과정 없이는 발생할 수 없다는 점이 요점입니다.[11]

이 주장은 인간지능은 스토케스틱한 과정을 거쳐 진화한 결과물이고, 그러려면 무작위성을 필수적으로 포함해야 하므로 인공지능이 인간지능과 같을 수 없다는 말이다. 여기서 핵심 문제는 인공지능이 무작위성을 포함할 수 있냐는 것이 된다.

이 주장의 타당성을 논의하려면 먼저 알고리즘을 올바로 이해해야 한다. 이 책은 알고리즘을 다음과 같이 정의한다. “알고리즘은 어떤 문제를 해결하기 위해 작동이 일어나게 하는 단계들의 모임입니다.”[12] 여기서 서평자가 지목하고 싶은 점은 위의 정의처럼 알고리즘의 초점은 답을 내놓는 데 있을 뿐, 구체적인 과정을 제약하지 않는다는 것이다. 그래서 알고리즘은 무작위를 배제하지 않는다. 데넷은 이 문제에 대한 빼어난 설명을 내놓은 바 있다.

---

8) Ibid., p. 316.

9) 진화론 문헌에서 일반적으로 자연선택은 다음 세 조건의 결합으로 설명된다. 1) 변이의 발생, 2) 적합도의 차이, 3) 유전.

10) 다른 매커니즘의 예로 중립진화(neutral evolution) 등이 있다.

11) Ibid., p. 346.

12) Ibid., p. 70.

> 사람들은 우연이나 무작위(randomness)를 사용한 과정은 알고리즘이 아니라는 초보적인 실수를 하곤 한다. 그러나 장제법(long division) 조차도 무작위를 사용한다.[13]
>
> $$\begin{array}{r} 7 \\ \hline 47\ )\ 326574 \end{array}$$

장제법은 한국에서 초등학교를 다닌 사람이라면 누구나 아는 나눗셈 방법이다. 이 과정은 무작위로 아무 수나 위에 대입해보는 것으로 시작한다. 이어서 그 수와 나누는 수를 곱한 값이 나눠지는 수보다 크다면 더 작은 수를 넣고 작으면 더 큰 수를 넣는다. 이 과정을 반복하면 답이 구해지게 되어 있다. 그러므로 이 나눗셈 방법은 알고리즘의 일종이다. 그럼에도 무작위를 포함한다. 이외에도 정렬(sorting), 선별(winnowing) 등의 많은 알고리즘이 무작위를 사용한다. 따라서 알고리즘이 무작위를 포함하지 않는다는 것은 잘못된 주장이다.

> 그렇다 해도 무작위를 포함한 알고리즘을 컴퓨터가 실행하지 못하면 소용이 없을 것이다. 저자는 여기에 대해 확실한 의견을 표명한다. “진정한 무작위 성분이 발생함으로써...전혀 예측하지 않았던 동작이 일어날 수 있도록 만드는 것이 관건입니다. 그런데 컴퓨터는 난수를 스스로 생성하지 못합니다.”[14]

그러나 이 주장은 사실이 아니다. 난수 생성은 암호학(cryptography)의 핵심 연구 주제로 컴퓨터 과학에 의해 많은 연구가 이루어졌다. 과거에는 유사-난수(pseudo-random)나 시간과 연동하는 방식이 사용되었으나,[15] 암호의 중요성이 높아진 최근에는 상용 CPU도 자체적으로 하

---

13) Dennett (1995), p. 52.

14) 김재인 (2017), p. 350.

15) 예전에는 C 언어에서 사용되는 일반적인 방식이었다. 프로그램이 실행된 시간이나 키를 입력한 시간에 따라서 난수 테이블을 바꾸는 방법이다. 본문에서 저자는 알고리즘은 시간을 품을 수 없다는 언급을 하는데(p. 350을 참고할 것), 이 방식이야 말로 프로그램에 시간을 심는 방법의 사례가 아닐까?

드웨어 난수 발생기를 회로에 포함하고 있다.[16] 난수의 신뢰성이 중요한 경우에는 양자역학을 이용한 난수 발생기가 사용되기도 한다(Ma et al. 2016). 이런 장치가 없더라도 Random.org와 같은 서비스를 이용하면 인터넷을 통해 쉽게 하드웨어 난수를 이용할 수 있다. 이런 방법들은 호프스테터가 말한 대로 자연의 무작위성을 프로그램이 이용할 수 있게 해준다.

그런데 이 책은 위 문제와 관련해서 한 가지 논증을 더 제시한다. 컴퓨터의 경우 이런 무작위성을 감당하지 못한다는 것이다.

> 알고리즘은 '만일 ~라면 ~이다'의 복잡한 연쇄입니다. 순수 논리로 구성된 알고리즘은 모든 경우의 수와 모든 작동 경로를 인간이 미리 짜놓았습니다. 예외는 일어날 수 없고, 혹시라도 외부에서 예외가 개입하면 작동을 멈춥니다. 기계학습을 통해 만들어진 알고리즘이더라도 마찬가지 입니다.[17]

> 알고리즘은 사실상 그 안에 버그가 존재하면 작동하지 않습니다. 반면 생물은 버그나 고장에서 불구하고, 아니 어쩌면 그런 것들을 통해 작동합니다... 그에 반해 컴퓨터의 프로그램 안에 버그가 있다면 무슨 뜻이죠? 다들 잘 알거에요. 알고리즘은 간단한 지시들의 집합인데, 하나라도 고장 나면 작동이 멈춥니다.[18]

위의 인용문에서 저자는 무작위가 개입한 프로그램이 작동할 수 없다고 주장한다. 그런데 이 부분이 이 책에서 가장 이해하기 어려웠다. 여기서 작동이 "멈춘다"는 말은 무엇을 의미하는 것일까? 만약 이 말이 윈도우의 블루 스크린[19] 같은 것을 의미한다면 그것은 컴퓨터의 작

---

16) 예를 들어, 몇 년 전부터 intel의 CPU에는 열 엔트로피(thermal-entropy) 방식의 하드웨어 난수 발생기가 포함되고 있다. 참고: https://software.intel.com/en-us/articles/intel-digital-random-number-generator-drng-software-implementation-guide.

17) Ibid., p. 348

18) Ibid., p. 351.

19) 윈도우 OS가 시스템에 에러가 발생하면 파란 화면을 띄우고 경고 메시지를 보여주기 때문에 이런 이름이 붙었다.

동 방식을 오해한 것이다. DOS 시절 어셈블리나 C언어로 프로그래밍을 해본 사람이라면 쉽게 이해하리라 생각하는데, 악명 높은 포인터 문제 같은 오류가 있으면 프로그램 실행 뒤 십중팔구 검은 화면만 떠 있을 뿐 컴퓨터는 어떤 입력도 받지 않게 된다. 그러나 이 상태의 컴퓨터는 작동을 멈춘 것이 아니다. 단지 프로그램이 예측하지 못한 방향으로 진행돼서 작업을 끝마치지 못하기 때문에(즉, 멈추지 않기 때문에) 정상적인 입력을 받지 못하는 것뿐이다. 전원이 공급되는 한 프로그램은 작동을 멈추지 않을 것이다.

오류가 발생한 프로그램이 블루 스크린에 의해서 멈추는 이유는 현재의 고도화된 OS(operating system)가 프로그램과 메모리, 저장장치 등을 관리하기 때문이다. 예전에는 개별 프로그램이 직접 관리를 해야 했다. 만약 시스템을 모니터링 하는 OS가 없다면 오류가 발생한 프로그램은 멈추지 않을 것이다. 그런데 우리의 몸도 비슷한 관리를 받는다. 예를 들어, 세포사(cell death)는 오류가 발생하거나 수명이 다 된 세포를 제거하는 메커니즘이다. 이와 같은 메커니즘은 OS가 하는 것과 같은 역할을 담당한다. 이 메커니즘이 제대로 작동하지 않으면 발생하는 것이 바로 암세포이다. 그러므로 오류가 발생했을 때 컴퓨터에서 생기는 일과 생명체에서 생기는 일은 다르지 않다.

이 문제는 이 책의 다음 주장과도 관련된다.

> 작동이 일어나는 층위가 있고 동시에 한 단계 높은 층위에서 그 작동을 점검하는 겁니다... 그래야 반성이 성립하겠죠... 자기가 자기를 점검하고 평가하고 수정할 수 있을까요? 그런 일이 수학적으로 프로그래밍 될 수 있을까요? 그럴 수 없다는 것이 내 사색의 결론입니다. 하지만 인간은 그러고 있습니다.[20]

일반적인 PC에서도 OS는 물론 여러 겹의 아키텍처가 하드웨어, 소프트웨어의 여러 층위에서 작동을 점검하고 보완한다. 인터넷으로 오가는 정보 패킷(packet)도 마찬가지로 체크섬(checksum)을 비롯한 여러 보안 메커니즘이 올바로 정보가 전달되었는지를 교차 검증한다. 이

---

20) Ibid., p. 356.

런 여러 겹의 안전장치 덕분에 컴퓨터가 높은 정확도로 작동할 수 있는 것이다. 이와 같은 프로그램의 수행은 자기를 점검하는 프로그램의 사례이다. 인공지능의 경우도 같은 역할을 담당하는 장치가 있으며 책의 76페이지 그림에 나와 있는 "비평가"가 바로 그것이다. 비평가는 인공지능 에이전트의 출력을 검사해서 기준에 맞게 알고리즘이 작동하도록 조정한다.

따라서 이 책이 비판하는 것처럼 인공지능이 스토캐스틱 알고리즘을 수행하지 못할 이유는 없어 보인다. 알고리즘은 무작위를 포함할 수 있으며, 컴퓨터는 무작위에 의해 전체 프로그램이 파괴되지 않도록 관리할 수 있다. 그럼에도 궁극적인 의미에서 이런 조절 장치의 목적은 인간에 의해서 주어지지 않느냐고 반론을 할 수 있다. 저자도 같은 이야기를 한다.

> 비록 프로그램이 스스로 프로그램을 만들긴 하지만, 그렇게 만들어지는 프로그램은 인간이 지정한 수행 기준을 잘 따르도록 프로그램되니까요.[21]

> 인공지능에서 문제나 목표는 에이전트 바깥에서 (더 구체적으로는 인간에 의해) 주어지는 반면 인간지능은 문제나 목표를 스스로 정한다는 점에서 이 둘은 결정적으로 다릅니다. 인공지능과 인간지능의 원리상의 차이는 문제나 목표가 외적이냐 내적이냐에 있습니다.[22]

그러나 이 주장은 논쟁적이다. 첫째로 모든 인공지능이 인간이 지정한 기준이나 목표에 따라 작동하는 것은 아니다. 대표적인 예는 생성 모델(generative model)일 것이다. 생성 모델은 주어진 데이터를 학습한 결과로부터 샘플(sample)을 생성하는데 사용된다. 강화 학습(reinforced learning)과 비교하면, 이 경우에는 강아지 사진에 대한 기준을 인간이 프로그램에 지정하고 정답을 잘 맞출 때까지 반복 학습을 시킨다. 반면 생성 모델의 경우에는 강아지의 사진을 주고 학습 시킨

21) Ibid., p. 77.
22) Ibid., p. 66.

뒤, 인공지능으로 하여금 학습한 강아지의 그림을 그리게 한다.[23] 인공지능 예술에서 많이 사용되는 이런 종류의 인공지능은 인간이 수행에 대한 기준을 주지 않는다. 다만 데이터만 줄 뿐이다. 두번째 문제는 인공지능의 편이 아니라 인간의 편에 있다. 즉, 인간이 정말로 목표를 스스로 정하는가 혹은 인간이 자유의지를 가지는가의 문제이다. 그러나 이 질문에 대한 명쾌한 답이 제시된 적이 없으므로 논증을 지지하는 근거로 사용되기엔 무리가 있다.[24] 따라서 더 설명이 필요한 부분이라고 생각한다.

결론적으로 말해 진화에 의존해서 인간지능과 인공지능을 구분하려는 이 책의 시도는 성공적인 것 같지 않다. 세 종류의 진화에 대한 정의 모두 저자의 주장을 입증하는데 도움이 되지 않아 보이기 때문이다. 또한 진화에 근거해서 논의를 전개하는 데넷, 클락(Andy Clark) 등의 철학자들 대부분이 진화를 오히려 이 책의 주장에 대한 반대 논거로 사용한다는 측면에서도 그렇다.[25] 진화를 저자 입장의 논거로 사용하려면 적어도 이들 철학자들의 작업에 대한 논의가 필요하다고 생각한다. 그럼에도 서평자는 이 책의 시도가 무의미하다고 생각하지 않는다. 국내 인문학자가 방대한 과학기술 문헌을 소화하고, 자신의 주장을 펼친다는 것만으로도 중요한 의미가 있다. 다만 보다 정련된 논의가 필요해 보인다는 게 서평자의 생각이다. 서평자는 현재 한국 사회의 화두로 떠오른 이 분야에 활발한 토론과 학문적 진보가 이루어지길 바라는 마음에서 이 글을 쓴다.

---

23) 생성 모델의 이와 같은 응용에 대해서는 Ian Goodfellow의 tutorial 참고: https://arxiv.org/pdf/1701.00160.pdf

24) 그런데 저자는 자유의지에 대해 비관적인 것 같다(3.5절). 서평자는 이 시각과 인간 지능이 스스로 목표를 정할 수 있다는 주장이 어떻게 양립 가능한지 궁금하다.

25) 물론 진화를 자유의지나 인간 독특성의 근거로 드는 플랜팅가(A. Plantinga)와 같은 철학자도 있다.

참고문헌

Darwin, C. (1859), *On the Origin of Species by Means of Natural Selection, Or the Preservation of Favoured Races in the Strugggle for Life*, John Murray.

Dennett, D. (1995), *Darwin's dangerous idea*, Penguin Books.

Futuyma, D. J. (2013), *Evolution*, Sinauer.

Ma, X., Yuan, X., Cao, Z., Qi, B. & Zhang, Z. (2016), "Quantum random number generation", *Npj Quantum Information* 2:16021, doi:10.1038/npjqi.2016.21.

Richerson, P. & Boyd, R. (2005), *Not By Genes Alone: How Culture Transformed Human Evolution*, University of Chicago Press.

Sterelny, K. (2012), *The Evolved Apprentice*, MIT Press.

Sterelny, K. & Griffiths, P. (2012), *Sex and Death: An Introduction to Philosophy of Biology*, University of Chicago Press.

서평 투고일	2018. 03. 10.
게재 확정일	2018. 03. 23.