http://dx.doi.org/10.52723/JKL.48.007
『국문학연구』 제48호, 국문학회, 2023, 7~42면.

# 거대 언어모형에 대한 대학교육의 대응

나수호*



<국문초록>

작년 말에 출시된 ChatGPT로 인해 학계에 우려가 커지고 있는바 본고는 인공지능과 대학 교육의 미래에 대해서 탐색하고자 했다. 먼저 ChatGPT와 같은 거대 언어모형(large language model)이 어떻게 '훈련'하고 '학습'하는지를 설명하고 이러한 생성형 인공지능의 능력과 한계를 알아보았다. 즉, 그럴듯한 말을 생성하기는 하지만 실제로는 언어에 대한 이해가 없는 소위 '확률적 앵무새(stochastic parrot)'라고 할 수 있다.

이어서 생성형 인공지능의 위험성을 탐구했는데 허위 정보 생성 및 보급의 위험이 있음을 알 수 있었다. 악의로 사용하지 않더라도 훈련 데이터 집합이 지배적인 이념 등을 반영하는 인터넷에서 수집되기 때문에 사회적 불평등이나 주도권의 고정관념, 편견 등을 영속시키며 증폭시킬 수 있기도 하다. 반면에 훈련 데이터 집합에서 빠진 것을 생각하면 이미 소외된 목소리들이 계속해서 소외되기 마련이다. 사회적 취약층의 문제를 차치하더라도 거대 언어모형은 디지털화된 자료로만 훈련할 수 있어서 아날로그 자료가

* 서울대학교 국어국문학과 부교수

무시된다. 이는 일찍이 인공지능의 위험성에 대해 경고한 바이젠바움(Joseph Weizenbaum)이 말한 '역사의 파괴'라는 문제와 같은 맥락이다.

다음으로 인간과 기계의 관계에 관하여 논하면서 문제적인 관계와 올바른 관계가 무엇인지를 고려해보았다. 항공계 연구에서 일찍이 '자동화 편향(automation bias)'의 문제를 감지하였는데 사회 다른 분야에서도 볼 수 있는 현상이다. 즉, 기계에 대한 과신으로 인간이 의사결정의 책임을 기계에 전가하면서 기계에 고장이 나거나 문제가 발생할 경우 민첩하게 반응하여 문제를 해결할 수 없게 되는 것이다. 거대 언어모형은 항공기처럼 생명을 위협하는 기술은 아니지만 비슷한 편향이 보인다. 그러나 이것보다 더 중요한 문제는 사람을 온전한 인간이 아니라 기계로 본다는 점이다. 컴퓨터가 생기기 전에도 인간의 지능을 수량화하려는 지능지수시험 등 같은 맥락의 문제가 있어왔지만 인공지능 시대에 특별히 유의해야 할 것이다.

마지막으로 이러한 상황에서 대학 교육이 나아갈 방법에 대해서 고민했다. 특히 미국 교육자들이 ChatGPT의 문제를 다루면서 거대 언어모형 그 자체보다 교육의 '거래적(transactional) 성격'을 문제시하고 있음을 알 수 있었다. 즉, 교육이 진정한 배움의 과정보다는 그저 성적과 학위를 받는 '거래'가 된 지 오래되었다는 것이다. 이러한 거래적 환경 속에서 학생들이 ChatGPT를 과도하게 사용하는 것은 당연한 것인데 사고하는 방법을 제대로 배우지 못할 위험, 학문적 정보가 부족할 위험, '인공지능식 글쓰기'를 배울 위험 등과 같은 부작용이 있을 수 있다. 이 문제를 해결하기 위해서 미국 교육자들은 '채찍'(거대 언어모형을 사용할 수 없게끔 하는 전략)과 '당근'(학생이 알고리즘이 아니라 인간처럼 배우게 하는 전략)을 제안했는데 그 골자는 '책임감'이라고 생각된다. 학생이 스스로 책임감을 키울 수 있는 교육 환경을 조성하는 것이 인공지능 시대에 교육이 살길이다. 그렇게 한다면 거대 언어모형이 두려워할 적이 아니라 유용한 도구가 될 수 있을 것이다.

핵심어: 인공지능, 생성형 인공지능, 거대 언어모형, 챗GPT, 인공신경망, '확률적 앵무새', 기술이상주의, 자동화 편향, 인간과 기계의 관계, 대학교육, 교육학

## Ⅰ. 들어가며

작년 11월 말 GPT(Generative Pre-Trained Transformer)에 기반한 ChatGPT가 출시되자 곧바로 장안의 화제가 되었다. 이른바 '생성형 인공지능(generative artificial intelligence)'을 처음으로 구현한 것은 아니었으나[1] 대다수 일반인에게는 소위 '거대 언어모형(large language model)'과 접하는 첫 기회였다. 많은 이들이 인공지능의 비약적인 발전에 놀랐고 우려하는 목소리를 내기 시작했다. 특히 학계에 이러한 거대 언어모형이 대학 글쓰기 교육 등에 어떤 영향을 끼칠지 염려하는 사람들이 많다. 이제는 글쓰기 과제를 학생들에게 못 시키겠다거나 심지어 앞으로 AI가 논문을 다 쓰겠다는 등 동료들의 걱정을 필자도 많이 들었다. 과연 그 정도로 걱정할 상황인가? 더 중요한 것은 이 새로운 세상에서 교육을 어떻게 해야 할까?

상황의 심각성을 파악하기 위해 먼저 인공지능의 미래에 관한 발언 하나를 보고자 한다.

> 이제는 세상에 생각하고, 배우고, 창조하는 기계들이 있다. 더군다나 이런 것을 할 능력이 급격히 발전할 것이며 눈에 보이는 미래에 그들이 처리할 수 있는 문제의 범위는 인간의 뇌(mind)가 응용되는 범위와 동일하게 될 것이다.

이 발언을 좀 풀어서 볼 필요가 있다. '인간의 뇌가 응용되는 문제의 범위와 같은 범위로 문제를 처리할 수 있다'라는 표현은 소위 '인공일반지능(artificial general intelligence)'을 말하는 것이다. 다시 말해서 인간이 풀 수 있는 문제라면 인공지능도 풀 수 있는 데다 더 빠르고 효율적으로 할 수 있다는 뜻이라 하겠다. <스타워스>의 C-3PO, <스타트랙>의 Data, <인터스텔라>의 TARS등과 같은 공상과학에 나오는 인물이 연상되는데 결국 소설책이나 영화에서나 보던 인공지능이 튀어나와 '눈에 보이는 미래'에 경험할 수 있을 것이라는 기술이상주의적(techno-utopian) 발언이다.

그런데 위에서 든 인용문은 누구의 발언일까? OpenAI의 최고 경영자 샘

1) 이미지를 생성하는 인공지능인 Dall-E와 Midjourney는 각기 2021년 1월과 2022년 3월에 출시되었다.

올트먼(Sam Altman)이 ChatGPT를 출시하면서 했을까? 아니면 많은 이들이 과학기술의 선지자로 여기는 엘론 머스크(Elon Musk)가 했을까? 어림없다. 이 예측은 인공지능의 개척자 중 사이먼(Herbert A. Simon)과 뉴웰(Allen Newell)이 1958년에 한 것이다.[2] 인공지능에 대한 최근의 발언이 많을 텐데 왜 하필 옛날 말을 인용했을까 의문을 품을 수 있겠으나 그럴만한 이유가 있다. 인공지능의 가능성을 논할 때 미래로 눈을 향하곤 하지만 과거를 무시하면 안 되기 때문이다. 이처럼 인공지능의 여명기인 20세기 중반에 이미 인간을 초월하는 기계가 '눈에 보이는 미래'에 출현할 것이란 예언이 있었으나 예언의 당사자인 사이먼과 뉴웰이 자신의 눈으로 그러한 미래를 보지 못한 채 세상을 떠났음은 물론 65년이 지난 오늘날에도 인공일반지능이 어디에도 보이지 않고 있다. 물론 앞으로 개발되지 않을 것이라 단정할 수는 없지만, 인공지능의 발전사를 길게 보면 아직은 공황에 휩쓸리기에 이른 시점이다.

요즘 화제가 되는 생성형 인공지능은 인공일반지능과 거리가 상당히 멀다. 그러나 인공지능 연구자들의 홍보나 과잉 선전(hype)만 들으면 충분히 걱정될 만하다. 따라서 본 연구는 먼저 생성형 인공지능을 바로 알아보고, 우리가 의식해야 할 위험 요소를 지적하고, 인간과 기계의 관계에 대해서 논의한 후에 거대 언어모형이 교육계에 어떤 영향을 끼칠지, 그리고 우리가 어떻게 대응해야 할지 논하고자 한다. 요즘 한국에서도 이와 같은 문제에 대한 논의가 많이 이루어지고 있는데 본고에서는 ChatGPT의 본고장으로 첫 타격을 입은 미국에서의 논의를 중심으로 고찰하고자 한다.

## II. 생성형 인공지능의 사실과 허구

앞에서 언급했던 생성형 인공지능에는 Dall-E나 Midjourney 등 이미지를 만드는 인공지능을 비롯해 음악을 작곡하는 것 등 다양한 것이 있는데 국문학 교육은 물론 인문학 교육 분야와 가장 관련이 깊은 것은 아무래도

2) Herbert A. Simon and Allen Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research*, Vol. 6: No. 1, 1958, p.8.

거대 언어모형(LLM; Large Language Model)일 것이다. 기술적인 면이 복잡하기는 하지만 생성형 인공지능의 역량을 제대로 파악하려면 이런 모형이 어떻게 만들어지는지 일단 간략하게나마 알아보는 게 좋겠다.

주지하다시피 현재의 인공지능은 '인공신경망(artificial neural network)' 방식을 많이 이용한다. 인간 뇌의 수많은 신경 세포(neuron)가 접합부(synapse)를 통해 서로 소통하는 신경망을 이루는데 인공신경망은 바로 여기서 영감을 받은 구조다. 이것이 정확히 어떻게 '학습하는지' 여기서 자세히 설명하자면 이야기가 너무 길어지겠지만, 중요한 것은 분류되어 있지 않은 대량의 훈련 데이터를 입력해야 한다는 것이다. 거대 언어모형의 경우에는 그 데이터가 당연히 언어 즉, 텍스트이다. GPT 중에 'P'는 이런 사전훈련(pre-training)을 언급하고 있는데 이 작업이 끝나면 미세조정(fine-tuning)을 하여 모형이 구체적인 작업을 더 정확하고 효율적으로 수행할 수 있도록 한다. ChatGPT의 경우에는 다양한 질문을 골라서 인간이 바람직한 대답을 작성해주고 이와 같은 이상적인 대답을 모은 데이터 집합으로 모형을 다시 훈련한다. 그러고 나서 그 질문으로 모형을 여러 번 시험한 다음에 각 대답을 인간이 양적으로 평가하여 이러한 피드백 데이터로 모형을 조정하는 것이다.[3]

그런데 거대 언어모형은 정확히 무엇을 어떻게 학습하는가? 사전훈련 단계는 '무감독 학습'이라 그 과정이 다소 신비롭게 느껴질 수도 있다. 간단하게 말하자면 사전훈련 단계에서 기계는 '언어'가 무엇인지부터 배운다. 대량의 텍스트를 보고 스스로 '토큰(token)'을 정하는데 토큰이란 언어의 최소 단위이며 대체로 '단어'에 해당한다. 그리고 각 토큰을 백터(vector)로 변환하여 백터 공간에 배치한다.[4] 즉 단어의 연어(連語) 현상을 수량화하는 것인데 예를 들어 '나무'라는 토큰은 '뿌리', '가지', '~ㅅ잎' 등과 같은 토큰과 함께 결합되는 경우가 많다는 것을 배우는 것이다. 훈련 데이터의 분

3) Ryan Lowe and Jan Leike, "Aligning Language Models to Follow Instructions," *OpenAI*, January 27, 2022. https://openai.com/research/instruction-following

4) 이 과정에 대한 좀더 자세한 설명과 ChatGPT의 훈련 데이터 집합에 관해서 다음과 같은 자료를 참조할 수 있다. Jill Walker Rettberg, "ChatGPT is multilingual but monocultural, and it's learning your values," *jill/txt*, December 6, 2022. https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/

량이 적으면 잘 배우지 못하겠지만 데이터 집합이 충분히 크면 인간의 언어 사용법을 그럴듯하게 배울 수 있다. (그런 이유로 '거대' 언어모형이라고 한다.) 나중에 사용자가 질문을 하면 챗봇이 이러한 통계적 정보를 이용하여 대답하므로 흔히 '확률적 앵무새(stochastic parrot)'라고 부르기도 한다.

그렇다면 이제 기계가 스스로 생각하거나 자연언어를 이해할 수 있다고 봐도 되는가? 기계가 생각할 수 있느냐 하는 질문은 인공지능 연구 초반부터 논란이 많았다. 앞서 언급한 것처럼 사이먼과 뉴웰은 1958년에 과감하게 '생각할 수 있는 기계'가 있다고 선언했으나 컴퓨터 과학의 선구자인 튜링(Alan Turing)은 그 문제에 조금 더 조심스럽게 접근했다. 그는 컴퓨터와 지능에 대한 1950년의 논문에서 그 질문이 무의미하다면서 "생각이라고 표현되어야 하지만 사람이 하는 것과 매우 다른 것을 기계는 수행하지 않을까?"라면서[5] 20세기 말까지는 기계가 생각한다고 할 수 있을 정도로 "언어의 사용법과 교양있는 여론"이 충분히 형성될 것이라 믿었다.[6] 과연 그렇게 되었을까? 20세기의 끄트머리에 도나쓰(Judith Donath)는 다음과 같이 말했다. "컴퓨터가 생각할 수 있는지에 대하여 직접 질문하면 많은 이들이 '아니오'라고 대답하겠지만, 실제로 기계가 생각하는 존재인 것처럼 그것과 상호작용하며 기계에 자의(自意)가 있다고 생각하고 기계의 '의견'에 다른 사람의 의견인 것처럼 반응하기도 한다."[7]

문제는 '생각'이라는 개념이 너무 모호하므로 필자 역시 튜링과 마찬가지로 이 문제가 결국 무의미하다고 생각한다. 오히려 기계가 자연언어를 이해할 수 있느냐는 질문이 본 연구의 주제와 더 밀접하게 관련되며 더욱 명확한 답이 가능하다. 언어학자인 벤더(Emily Bender)와 콜러(Alexander Koller)는 그 질문에 대해서 "언어모형을 만드는 작업은 형식만을 훈련 데이터로 사용하기 때문에 원칙적으로 의미를 배우는 데까지 이어질 수 없다는 게 우리의 주장이다."라고 했다.[8] 다시 말하면 언어가 상징체계인데

5) Alan Turing, "Computing Machinery and Intelligence," *Mind: A Quarterly Review of Psychology and Philosophy*, Vol. 59: No. 236, p.435.
6) *Ibid.*, p.442.
7) Judith Donath, "Being Real: Questions of Tele-Identity," *The Robot in the Garden – Telerobotics and Telepistemology in the Age of the Internet*, ed. Ken Goldberg, Cambridge (MA): The MIT Press, 2000, p.301.
8) Emily Bender and Alexander Koller, "Climbing towards NLU: On Meaning, Form, and

기계는 상징만 알고 그 상징이 지시하는 대상을 모른다는 것이다. 그런데 사람들이 언어를 사용할 때는 언제나 '의사소통적 의도(communicative intent)'가 있으므로 사물, 감정, 개념 등과 같이 언어 외의 것을 지시한다. 다른 말로 표현하자면 거대 언어모형은 "상식 즉, 세상이 물질적으로나 사회적으로 어떻게 작동하는지에 대한 이해"가 결핍됐다고 할 수 있다.[9] 결국 사람들이 언어를 어떻게 사용하는지를 잘 알고 있어서 그럴듯한 대답을 제시하기도 하고, 언어가 세상을 반영할 때가 많아서 그 대답이 옳은 경우가 많기는 하더라도 "이러한 시스템은 세상과 그 세상이 어떻게 작동하는지에 대해 추리하지 않기 때문에, 말하는 바의 정확성이 어느 정도 우연의 문제가 된다."라는 것이 학자들의 일반적인 견해다.[10] 시애틀의 워싱턴 대학에 있는 컴퓨터 과학자 최예진은 "오늘 우리가 가진 것은 본질적으로 뇌가 없는 입이다."라고 한 바 있는데[11] 아마도 이를 가장 명료하게 표현한 말일 것이다.

요컨대 기계가 하는 것이 '생각'이라고 표현할 수 있을는지는 모르겠지만 확실하게 말할 수 있는 건 기계가 아직 자연언어를 이해하지 못한다는 것이다. 더군다나 현재의 훈련 방식으로는 그런 이해가 가능하지도 않다. 그렇다면 도나쓰가 지적한 대로 어째서 우리는 챗봇과 같은 기계를 인간처럼 대하는 것일까? 이 질문은 21세기에 들어서야 대두된 것이 아니다. 1960년대, 사이먼과 뉴웰이 인공일반지능을 예언한 지 얼마 지나지 않아 바이젠바움(Joseph Weizenbaum)이 첫 챗봇인 일라이자(ELIZA)를 개발했다. 사용자가 입력하는 텍스트에서 키워드를 뽑아 그 키워드에 대해 질문을 하는 아주 원시적인 프로그램이었는데 놀랍게도 이 사실을 아는 사람조차 마치 다른 인간과 대화하듯이 기계와 대화하는 것이었다. 이에 그는 1966년에 이처럼 '이해의 착각(the illusion of understanding)'이 얼마나 쉽게 만들어

---

Understanding in the Age of Data," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, July 5-10, 2020, p.5185.

9) Matthew Hutson, "The Language Machines," *Nature*, Vol. 591, 4 March 2021, p.25.

10) Gary Marcus, "AI Platforms like ChatGPT Are Easy to Use but Also Potentially Dangerous," *Scientific American*, December 19, 2022. https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/

11) Hutson, *op cit.*, p.23.

질 수 있는지, 그리고 그런 착각이 얼마나 위험한지를 경고하였다.[12] 그러나 이 경고가 먹히기는커녕 정신과 의사 중 몇 명이 일라이자를 보고 앞으로 인간 의사 대신에 기계가 정신과 치료를 해도 되겠다고 떠들썩거리자 바이젠바움은 깜짝 놀라 인공지능의 윤리에 관하여 논하기 시작했다.[13]

일라이자의 단순성에도 불구하고 사람들이 마치 컴퓨터를 다른 인간인 것처럼 대하며 친근하게 대화하는 모습을 보고 큰 충격을 받은 바이젠바움은 이러한 "강력한 망상(powerful delusional thinking)"을 한탄했다.[14] 그런데 이러한 현상을 그리 이상히 여기지 않는 이가 많다. 벤더는 ChatGPT를 논하면서 이러한 현상에 대해 다음과 같이 설명했다.

> 우리가 우리의 언어로 말하는 것처럼 보이는 존재를 마주치면, 우리도 모르게 다른 사람들과 의사소통하기 위해 그 언어를 사용한 것과 관련된 기술을 사용한다. 그런 기술은 핵심적으로 상호주관성과 공동 관심을 수반하기 때문에, 그 언어 배후의—실제로는 존재하지 않지만—의사(意思; mind)를 상상한다.[15]

쉽게 말하면 언어를 사용하면서도 의사소통적 의도가 없는 존재와 소통하기 위해 인간이 발달하지 않았다는 것이다. 극소수의 예외를 제외하면 동물도 언어를 이해하지 못하지만 우리가 언어를 사용해 동물과 의사소통을 시도하는 판에 언어를 생성하는 기계를 인간처럼 대하는 것은 그다지 놀랄 일이 아니라고 생각한다.

다만 기계가 우리와 같은 존재라고 착각하면 안 된다. 뻔한 소리일 것 같지만 대중문화에서 그리는 인공지능을 보면 우리가 그런 현상을 상상하는 것을 굉장히 좋아한다는 사실을 알 수가 있다. 최근의 예를 들자면 올 7월에 개봉한 <미션 임파서블: 데드 레코딩>에는 적국의 컴퓨터 시스템을 교

12) Joseph Weizenbaum, "ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine," *Computational Linguistics*, Vol. 9: No. 1, January 1966, pp.42~43.

13) 이와 관련한 대표적 저술은 Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, New York: W. H. Freeman & Co., 1976.

14) *Ibid.*, pp.6~7.

15) Emily Bender, "On NYT Magazine on AI: Resist the Urge to be Impressed," *Medium*, April 18, 2022. https://medium.com/@emilymenonbender/on-nyt-magazine-on-ai-resist-the-urge-to-be-impressed-3d92fd9a0edd

란하고 파괴하기 위해 개발된 인공지능이 어느 날 '지각력'이 생겨서 멋대로 굴기 시작하는 상황이 등장한다. 혹시 이러한 미래가 올까 봐 걱정하는 사람이 있을지 모르겠지만, 단언컨대 당장은 그런 일이 없을 것이다. 물론 관객 대부분은 이렇게 멋대로 구는 인공지능을 그냥 흥미 요소라 생각하고 영화를 재미있게 볼 것이다. 그러나 이렇게 현실과 거리가 있더라도 인공지능에 대한 미디어 표현이 중요하다. 우리가 그것을 보고 인공지능의 위험성이 공상과학 소설이나 액션 영화 같은 매체의 소재에 불과하다고 생각할 수도 있기 때문이다. 인공지능은 지각력을 얻지 않더라도, 악심을 품지 않더라도, 심지어 우리의 언어를 이해하지 못하면서 그저 확률적 앵무새 노릇만 하더라도 위험이 수반될 가능성이 있다.

## III. 생성형 인공지능의 위험성

물론 위험 중에 먼저 떠올릴 것은 아무래도 인공지능을 악용하는 나쁜 행위자일 것이다. 2023년 8월에 메타(Meta; 前 페이스북社)가 자사의 거대 언어모형인 라마2(Llama 2)를 오픈 소스로 공개하겠다고 발표했을 때 역시 우려의 목소리가 컸다.[16] 물론 메타가 사전 훈련된 모형을 '안전하게' 미세조정을 했으며 앞으로도 계속해서 미세조정을 할 계획이라고 했으나[17] 대중에게 공개가 된다면 누구나 마음대로 다시 미세조정을 해도 되기 때문에 혐오 발언이나 허위 정보를 생성하는 모형으로 만들 수 있다는 지적이다. 사실 오픈 소스가 아니어도 나쁜 행위자가 악심만 품으면 생성형 인공지능을 충분히 악용할 수 있다. ChatGPT는 공개되지 않았지만, 모형을 '탈옥'하기 위한 사용자들의 꾸준한 시도를 그런 맥락으로 볼 수 있다.[18] 또한 마커

16) 일례로 Kelsey Piper, "Why Meta's move to make its new AI open source is more dangerous than you think," *Vox - Future Perfect*, August 2, 2023. https://www.vox.com/future-perfect/23817060/meta-open-source-ai-mark-zuckerberg-facebook-llama2

17) 라나2의 훈련과 미세조정 과정, 특히 '유익함(helpfulness)'과 '안전성(safety)'에 최적화하는 과정에 대하여 다음과 같은 논문을 참조할 수 있다. Hugo Touvron, Thomas Scialom et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv*, Cornell University, 19 July 2023. https://arxiv.org/abs/2307.09288

18) '탈옥(jailbreaking)'이란 보안이나 안전을 위해 개발자가 프로그램에 일부러 가한 제약을

스(Gary Marcus)가 지적했듯이 “이러한 시스템은 말하는 내용의 사실을 확인하는 메카니즘을 전혀 포함하지 않기 때문에, 전례 없는 규모로 허위 정보를 생성하도록 쉽게 자동화될 수 있다.”[19]

한편 인공지능을 선의로 사용한다고 하더라도 예상치 못한 위험이 있을 수 있다. 사실 20세기 중반부터 바이젠바움과 같은 비판자들이 이러한 위험성에 대해 경고했던 것을 비롯하여 최근에 생성형 인공지능이 화제가 되기 이전에도 우려의 목소리가 계속 있었다. 시민의 개인 정보 등을 비밀리에 모으는 금융계나 기술업계의 알고리즘,[20] 다양한 사회 분야에 악영향을 미쳐 불평등을 초래하는 인공지능,[21] 심지어 인종차별주의적인 사고를 강화하는 검색 엔진의 알고리즘[22] 등에 대한 비판을 통해 인공지능의 위험성을 강조한 것 등이 그것이다. 물론 작년부터 거대 언어모형이 줄줄이 출시되면서 그 위험성에 관해 경고하는 이가 더욱 많아졌다.

디지털 문화 학자 렛버그(Jill Walker Rettberg)는 작년 말까지만 해도 ChatGPT가 아직 미국의 가치나 형식에 맞추어져 있었지만 이는 일시적인 현상일 뿐이라고 주장했다. 그 이유에 관해 “우리는 ChatGPT를 사용해보면서 우리의 가치에 더 부합하도록 훈련하고 있다. 그럼으로써 우리가 좋아하는 대답과 싫어하는 대답을 모아 인간이 표시한 방대한 데이터 집합을 OpenAI에게 제공하고 있는 것이다.”라고 설명했다.[23] 어느 정도 일리가 있는 말이다. 미세조정 과정에 사용자의 피드백을 활용하여 모형을 개선하는

---

피하는 작업을 말한다. ChatGPT의 경우에는 프롬프트 엔지니어링(prompt engineering)을 통해 탈옥하는 시도가 많이 이루어지고 있다. 물론 이렇게 하는 사람 중에 ‘나쁜 행위자’보다 그냥 재미로 하는 사람이 많겠지만 충분히 악용될 가능성이 있는 기술이다. Alex Albert, *Jailbreak Chat*, https://www.jailbreakchat.com/

19) Marcus, *op. cit.* 물론 마커스가 이 주장을 한 이후로 ChatGPT와 다른 거대 언어모형이 허위 정보를 생성하지 않도록 안전장치를 도입하기 시작했지만 출력하는 텍스트의 옳고 그름을 스스로 판단할 능력은 여전히 전무하다. 기계가 텍스트를 이해하지 못하는 한 이러한 판단도 불가능으로 남을 것이다.

20) Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge (MA): Harvard University Press, 2015.

21) Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Broadway Books, 2016.

22) Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: New York University Press, 2018.

23) Rettberg, *op. cit.*

것은 흔한 일이다. 그러나 우리가 GPT와 같은 거대 언어모형을 사용하면서 그 모형이 자연스럽게, 아니 스스로 향상된다고 생각하고 넘어가기 전에 물어야 할 질문이 있다. 바로 "'우리'가 정확히 누구인가?" 하는 질문이다. 2023년 7월 기준 통계에 의하면 세계 인구 중에 인터넷을 사용하는 인구는 64.5%라고 한다.[24] 다시 말하면 이 세상에 사는 사람 중에 삼분의 일 이상이 인터넷을 사용하고 있지 않다는 것이다. 그리고 인터넷 사용자 중에 몇 퍼센트의 사람이 ChatGPT를 사용하고 있을까? 게다가 ChatGPT 사용자 중에 과연 몇 퍼센트가 챗봇의 대답을 평가하는 피드백을 제공하고 있을까? 그런 통계는 잘 모르겠지만 ChatGPT에게 가치를 가르쳐주는 '우리'는 극히 소수일 것이다.

이와 같은 이유로 거대 언어모형이 배우는 인터넷의 언어는 '모두의 언어'가 아니라 '특권을 가진 소수의 언어'라고 할 수 있다. 더 정확히 말하자면 인터넷의 '대표적' 언어는 주도권의 언어인 것이다. 이런 텍스트를 훈련 데이터로 삼는 것의 위험에 대해 벤더와 게브루(Timnit Gebru)는 다음과 같이 요약했다. "대량의 웹 텍스트를 인류 '전체'의 '대표적'인 것으로 받아들임으로써 우리는 지배적인 관점을 영속시키고 권력 불균형을 증대시키며 불평등을 더욱 구체화할 위험을 무릅쓰고 있다."[25] 인종, 젠더 등과 같은 정체성 특징에 대한 고정관념이나 편견을 복제하듯이 그대로 나타내는 것이 대표적인 예이다. 물론 인공지능 개발자들이 이와 같은 해로운 내용이 생성되지 않도록 꾸준히 노력하고 있다. 이론적으로 보자면 폭력, 성적 학대를 묘사하는 텍스트나 혐오 발언 등과 같은 해로운 내용을 표시하고 다시 미세조정을 하면 되는 것이니 일견 어렵지 않아 보인다. 문제는 누군가 그런 해로운 내용을 검토하고 표시해야 한다는 것이며 역시 주도권이 주변화된 타자를 착취하는 일이 벌어질 수 있다. 올해 초에 보도된 바에 따르면

24) Simon Kemp, "Digital 2023 July Global Statshot Report," *DataReportal*, 20 July 2023. https://datareportal.com/reports/digital-2023-july-global-statshot. 흥미로운 점은 2분기의 통계에 따르면 4월에는 인터넷 사용자가 64.6%에 달했다. 물론 인터넷을 사용하는 인구수가 늘기는 했으나 세계 인구가 더 빨리 증가하는 바람에 인터넷 사용 인구의 비율이 오히려 낮아진 것이다.

25) Emily Bender, Timnit Gebru et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3–10, 2021, p.614. https://doi.org/10.1145/3442188.3445922

OpenAI가 그런 표시 작업을 수행하기 위해 케냐의 노동력을 이용했는데 보수를 시간당 $2 이하로 지불했다고 한다. 게다가 타임스지에 의하면 “그 텍스트의 상당한 부분은 인터넷의 가장 어두운 구석에서 끌어 놓은 것으로 보인다. 일부는 아동 성적 학대, 수간(獸姦), 살인, 자살, 고문, 자해, 근친상간과 같은 상황을 끔찍할 정도로 자세히 묘사한 것이었다.”[26]고 한다. 당연히 근로자 중에 정신적 피해를 본 사람이 있었으며 그들 중에 상담 치료를 제대로 받지 못했다고 호소하는 사람들도 있었다. 인공지능 기술이 이렇게 발전할 때마다 우리 눈에 보이지 않는 누군가가 그 대가를 치르고 있음을 잊어서는 안 된다.

물론 인공지능을 훈련할 때 데이터 집합에 무엇이 들어가는지도 중요하지만, 무엇이 빠지는지도 똑같이 중요하다. 인터넷에 올라와 있는 텍스트를 모두 훈련 데이터로 삼더라도 어쩔 수 없는 사각지대가 남아 있을 터이고 ChatGPT와 같이 엄청난 분량의 텍스트를 끌어모은 거대 언어모형에서도 앞서 언급한 것처럼 소외된 목소리들이 너무 많다. 이와 같은 지리적・사회적 사각지대 이외에도 거대 언어모형의 시간적 제한을 무시할 수 없다. 일단 사전훈련이 끝난 다음의 정보에 대해서는 모형이 전혀 모르기 때문에 최근 시사에 대해서는 완전 무지하다고 할 수 있다. 그러나 훈련 데이터의 시한성보다 훨씬 심각한 문제가 있다. 이 문제에 대해서는 거대 언어모형을 꿈도 꾸기 전에 바이젠바움에 의해 이미 경고된 바 있는데 바로 컴퓨터가 역사를 삭제한다는 사실이다. 그는 일찍이 1976년에 “컴퓨터는 역사를 파괴하기 위한 도구가 되기 시작했다. 사회가 ‘하나의 표준 형식’으로 되어 있으며 ‘기계에 쉽게 말할 수 있는’ 데이터만을 정당화하면 역사 즉, 기억 그 자체가 소멸하기 때문이다”라고 예견했다.[27] 다시 말해 디지털 양식으로 수량화된 데이터가 아니면 기계가 배울 방법이 아예 없다는 것이다. 예를 들면 바이젠바움의 저서에 대한 정보가 위키피디아와 같은 온라인 텍스트 보관소에 있으므로 ChatGPT가 책에서 다뤄진 중요한 주제를 요약할 수 있으나 자세한 내용에 관하여는 아무것도 모른다.

---

26) Billy Perrigo, “OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic,” *Time*, January 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/

27) Weizenbaum (1976), *op. cit.*, p.238.

물론 염두에 두어야 할 것은 컴퓨터가 이런 문제의 근원이 아니라는 것이다. 다시 말하면 새로운 매체나 기술이 생길 때마다 비판자들이 그 매체나 기술의 위험성에 대해서 경고해왔다는 것을 기억해야 한다. 잘 알려진 예를 들면 플라톤도 <파이드로스>에서 인간이 글쓰기를 배우면 지식을 되찾을 때 자신의 기억력이 아니라 외부의 표기에 의존하게 되므로 건망증이 생길 것이라고 주장한 바 있다.[28] 물론 플라톤은 틀리지 않았다. 구술 사회에서 기록 사회로 넘어가는 과정에 인간의 기억력이 나빠진 것이 사실이다. 대신에 인간 사회의 지식을 더 확실하게 보관하고 전할 수 있게 되었다. 컴퓨터와 인공지능도 비슷한 맥락으로 보면 잃을 것도 있고 얻을 것도 있을 것이다. 인공지능의 힘과 가능성에 휩쓸려 너무 귀한 것을 잃고 있지 않은가 심사숙고할 필요가 있다. 역사는 원래 승자 즉, 주도권을 쥔 자가 쓴다고 하는데 특히 인공지능 시대에 들어 역사가 잊히는 문제가 가속화되고 있다고 생각한다.

## Ⅳ. 인간과 기계의 관계

앞서 살펴본 인공지능의 위험성에도 불구하고 인간 사회는 가면 갈수록 기계에 의존할 수밖에 없는 사회가 되어간다. 더군다나 기계가 발달할수록 정확히 어떻게 작동하는지 그 원리를 알기 어려워지기도 한다. 그러나 이것은 새로운 현상이 아니다. 1950년에 튜링이 “학습하는 기계의 중요한 특징 중 하나는 교사가 학생의 행동을 어느 정도 예측할지라도 그 내부에서 어떤 일이 벌어지는지에 대해서는 대체로 무지할 것이라는 점이다”라고 지적한 바 있다.[29] 튜링은 학습하는 기계의 미래에 대해 이론상으로 추측한 것이지만 70년대에 이미 그가 예견했던 일이 벌어지고 있었다. 역시 광야에서 외치는 목소리인 바이젠바움은 “원래 사람들이 분석하고 결정하는 것에 ‘도움’을 주기 위한 것이었으나 사용자의 이해를 뛰어넘는 동시에 그들에게

28) Plato, *Plato's Phaedrus*, tr. R. Hackforth, Cambridge: Cambridge University Press, 1952, p.157.
29) Turing, *op. cit.*, p.458.

필수적인 것이 된 지 오래된 컴퓨터 시스템에 대한 우리 사회의 늘어나는 의존도는 아주 심각한 국면에 이르렀다"면서 경각심을 일깨우고자 했다.[30]

일반적으로는 잘 모르는 사람을 그다지 신뢰하지 않는 게 상식일 텐데 기계에 대해서는 조금 다른 듯하다. 기계의 작동 원리 등에 대한 이해의 결핍이 신뢰하는 데에 있어서 부정적 요인이 될 것 같지만 기계에 '홀리면' 그 결핍이 오히려 신뢰도를 높이며 기계를 맹신하게 만들 수 있다.[31] 아마 요즘 기계에 홀려 맹신하지 않는 사람이 별로 없을 것이다. 영국 공상과학 작가 클라크(Arthur C. Clarke)의 널리 알려진 법칙 중 "충분히 발전된 기술은 마법과 구분이 가지 않는다"라는 데에서 '충분히'라는 말을 '우리의 이해를 뛰어넘을 정도로'라는 뜻으로 보면 된다. 요즘 인간이 기계를 맹신하는 모습을 보면 '마법'보다 '종교'에 가깝다는 생각이 들 때도 있다.

기계가 이제는 인간 사회의 모든 영역에 침투했지만 그중 현대사회에 큰 영향을 끼친 영역 중에 항공 기술이 있다. 20세기 후반부터 군용기에 소위 '유리 조종석(glass cockpit)'을 도입하여 조종사의 업무량을 줄여주기 시작했고 얼마 후 여객기에도 그런 기술을 도입하여 요사이에는 비행기를 타는 것이 자동차를 타는 것보다 안전하다는 것이 상식이 되었다. 그러나 모든 것이 순탄하기만 한 것은 아니었다. 비행기 조종이 점차 자동화되면서 오히려 조종사의 실수가 늘어나는 일도 있었다. 그 원인을 연구자들은 '자동화 편향(automation bias)'이라고 부른다.[32] 항공에 대한 부담감을 줄이면서 인적 오류율도 내리기 위해 도입된 기계가 어찌하여 그 반대의 효과를 낼 수 있는 것일까? 모지어(Kathleen L. Mosier)와 그의 동료 연구자들에 의하면 문제의 핵심은 인간과 기계의 관계에 있다.

> 자동화와 자동화된 의사결정 보조구의 이용 가능성은 최소한의 인지적 노력의 길을 걸으려는 인간의 일반적 경향을 작동시킨다. 일반적으로 사람들은 최소한의 인지적 일을 하려고 하고 … 정당화하기 쉬운 전략을 선호할 것이며 … 자

30) Weizenbaum (1976), *op. cit.*, p.236.

31) Bonnie M. Muir, "Trust between humans and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, Vol. 27: Issues 5-6, 1987, p.533.

32) Kathleen L. Mosier et al., "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *The International Journal of Aviation Psychology*, Vol. 8: No, 1, 1998, pp.47~63 참조.

주 노력과 정보 부하를 줄이기 위해 휴리스틱(또는 인지적 지름길)을 활용하려 할 것이다.[33]

쉽게 말해서 '인간이 게으르다'라는 말인 것 같은데 그렇게 섣부르게 단언하기는 어렵다. 인간의 역사를 통틀어 현대사회처럼 인지적 부하가 큰 적이 있었을까. 특히 스마트폰의 출현 이후로 우리의 뇌가 처리해야 할 정보량이 급증하고 있다. 우리의 집중과 관심을 요구하는 요소가 우리가 실제로 처리하도록 진화된 것보다 훨씬 많아진 판이니 그런 요소에 우선순위를 매기는 것은 당연한 일이다. 위킨스(Christopher D. Wickens)와 딕슨(Stephen R. Dixon)이 지적했듯이 "선천적으로 인간은 자동화된 작업을 '2차적'으로 취급하고 '1차적' 비자동화 작업에 필수 자원을 할당함으로써 고성능을 유지해 왔다."[34] 다시 말해서 우리 인간은 단지 효율성을 높이기 위하여 기계에 맡길 수 있는 작업에 관심을 주지 않는 것일 뿐이다.

인지적 부하를 줄이는 건 좋지만 사람들이 책임까지 회피하여 기계에 전가하려고 할 때 문제가 된다. 비행기 조종사가 항공에 대한 책임을 기계에 전가하여 관심을 끄면 수백 명이 죽을 수도 있다. 바이젠바움이 기계에 대한 의존도가 증가 일변도인 현상을 걱정한 이유가 바로 여기에 있었다. 당시 미국에서 베트남 전쟁의 여파를 헤아리기 위해서 노력하고 있었는데 폭격할 목표물을 결정하는 작업을 장교들 중 누구도 제대로 이해하지 못했던 컴퓨터들이 했다는 사실이 알려졌다. 다시 말하면 무고한 시민 중에 살 가치가 있는 목숨과 살 가치가 없는 목숨을 분간한 주체가 인간이 아니라 기계였다는 것이다. 게다가 그 책임을 질 사람이 아무도 없었다. 심지어 어떤 장군은 인간이 컴퓨터의 노예가 된 상황을 한탄하기까지 했다.[35] 문제는 기계에 의존하는 것 그 자체가 아니라 기계를 이용한 결과에 대한 책임을 회피하려는 것이다. 목수가 못을 비뚤게 박고 나서 망치를 탓하지 않는 것처럼, 우리도 기계를 이용하여 좋지 않은 결과를 냈을 때 기계에 책임을 전

33) *Ibid.*, p.49.
34) Christopher D. Wickens and Stephen R. Dixon, "The benefits of imperfect diagnostic automation: a synthesis of the literature," *Theoretical Issues in Ergonomics Science*, Vol. 8: No. 3, 2007, p.209.
35) Weizenbaum (1976), *op. cit.*, pp.238~240.

가해서는 안 된다. 그리고 이 사실은 자동화 편향에 대한 해결책을 보여주기도 한다. 모지어 외에 따르면 “자동화된 시스템을 사용할 때 성과와 전략에 대한 책임감의 내재화는 자동화 편향에 의미 있는 영향을” 끼치며 “자동화된 시스템과의 상호작용에 대한 책임감은 그러한 상호작용 시 경계, 사전 예방적 전략, 그리고 모든 정보의 사용을 장려한다.”고 한다.[36]

다행히 국문학 교육은 잘못돼도 수백 명의 승객이 죽거나 어느 무고한 마을 시민들이 소이탄의 화염에 휩싸일 일은 없겠지만 그렇다고 인간과 기계의 관계에 대해서 무지해도 되는 것은 아니다. 기계를 과도하게 믿는 것이든 자동화 편향에 사로잡힌 것이든 그 뿌리에는 우리가 모두 고민해야 할 문제가 있기 때문이다. 앞서도 언급했지만, 첫 챗봇인 일라이자를 경험하고 나서 정신 치료를 기계에 맡겨도 될 미래가 곧 올 거라 주장하는 것에 바이젠바움이 적잖이 충격을 받았다. 그리고 그 충격이 어느 정도 사그라졌을 때 깨달은 바는 사람들이 그토록 기계가 인간적일 수 있다고 믿는 이유가 애초부터 인간이 일종의 기계에 불과한 존재라고 믿었기 때문이라는 것이었다. 그는 인간을 정보처리장치로 보는 것이 기본적으로 그릇된 생각은 아니지만, 그렇게만 볼 수 없다고 주장했다.[37] 그의 결론은 단호했다.

> 유기체는 주로 직면한 문제에 의해 정의된다. 인간은 어떤 기계도 직면할 수 없는 문제에 직면한다. 인간은 기계가 아니다. 인간이 당연히 정보를 처리하지만, 컴퓨터가 정보를 처리하는 방식으로는 하지 않는다는 것이 나의 주장이다. 컴퓨터와 인간은 같은 속의 다른 종이 아니다.[38]

인간을 기계처럼, 최소한 ‘온전한 인간’ 이하의 대접을 하면 결국 그 사람을 비인간화하는 것이라고 선언했다.[39] ChatGPT를 ‘안전하게’ 만들기 위해서 정신 건강에 해로운 데이터를 기계처럼 분류하고 표시하는 데에 이용된 케냐 근로자 경우를 떠올리지 않을 수 없다.

물론 이러한 현상은 어제 오늘의 일이 아니다. 인간의 역사를 보면 인간

36) Mosier et al., *op. cit.*, p.60.
37) Weizenbaum (1976), *op. cit.*, p.140.
38) *Ibid.*, p.203.
39) *Ibid.*, p.266.

을 비인간화하는 경우가 수없이 많겠지만, 여기서 교육과 관련된 예를 잠깐 논하고 넘어가고자 하는데 바로 인간의 지능을 수량화하려는 노력이다. 이런 노력은 19세기부터 시작되었다고 보는데[40] 우생학과 관련되어 인종차별주의적인 근거에 세워진 이론이었다.[41] 20세기에 들어서 이와 같은 연구의 유산으로 지능지수(IQ; intelligence quotient)가 개발되어 오늘날까지 사용되고 있다. 비슷한 시기에 학습 능력 적성 시험(SAT)도 개발되었는데 지능을 수량화할 수 있다는 믿음과 비슷한 논리로 학습 능력도 표준화된 시험으로 측정할 수 있다고 믿었다. 반면에 지능 검사와 학습 능력 적성 시험의 타당성에 대한 회의를 표한 학자도 많았는데 그중 심리학자인 맥클렐랜드(David C. McClelland)는 표준화 시험의 점수가 학교 성적이나 업무 능력 등과 아무런 통계적 상관성이 없다고 주장하는 한편, "현재 권력을 쥐고 있는 집단이 도입한 기준을 능력의 궁극적인 척도로 받아들이는 것에 대해 훨씬 신중해야 한다."라는 경고까지 덧붙이면서 '기준'이란 저절로 생긴 것이 아니라 누군가에 의해 정해진 것임을 상기시켰다.[42]

다행히도 요즘 미국에서 입학 과정에 표준화 시험 점수를 요구하지 않는 대학이 점점 늘어나고 있다. 많은 경우가 코로나19의 영향으로 시행한 정책이기는 했으나 팬데믹이 끝난 다음에도 계속 늘어나는 추세이며 2022년 말 보도된 바에 따르면 2023년 가을 학기 입학 과정에 4년제 대학 중 80% 이상이 SAT나 ACT와 같은 표준화 시험 점수 요건을 폐지할 계획이라고 했다.[43] 그런데도 인간을 기계로 여기는 사고방식은 뿌리가 깊고 가지도 널리 뻗어나가 우리 사회의 모든 영역에 큰 영향을 끼치고 있으며 물론 교

40) 초기 연구에 대해 심리측정학자인 골턴의 저서 참조. Francis Galton, *Hereditary Genius: An Inquiry Into Its Laws And Consequences*, London: Macmillan and Co., 1869.

41) 초기 연구자들은 인간의 지능은 사회경제적 지위와 같은 환경적인 변수에 의한 것이 아니라 바꿀 수 없는 선천적인 것이라 주장하면서 한 인종이 다른 인종보다 우월하다는 것을 논증하고자 했다. Melissa Nobles et al., "Science must overcome its racist legacy: Nature's guest editors speak," *Nature*, Vol. 606, 09 June 2022.

42) David C. McClelland, "Testing for Competence Rather Than for 'Intelligence,'" *American Psychologist*, Vol. 28: No. 1, January 1973, p.6.

43) Michael T. Nietzel, "More Than 80% of Four-Year Colleges Won't Require Standardized Tests for Fall 2023 Admissions," *Forbes*, Nov. 15, 2022. https://www.forbes.com/sites/michaeltnietzel/2022/11/15/more-than-80-of-four-year-colleges-wont-require-standardized-tests-for-fall-2023-admissions/ 참조.

육도 예외가 아니다.

## Ⅴ. 생성형 인공지능에 대한 대학 교육의 올바른 대응은 무엇일까?

여태까지 생성형 인공지능에 관하여 바로 알아보고 그 위험성과 나아가 인간과 기계의 관계에 대해 숙고했다. 그런데 위와 같은 내용이 대학 교육과 무슨 상관일까? 마지막으로 여러 생각을 정리하고 교육계에의 적용 가능성을 살펴보고자 한다.

앞에서 많이 인용한 바이젠바움은 컴퓨터 과학을 연구하고 인공지능을 비판하는 사람이었던 것만이 아니라 매사추세츠 공과대학(MIT)에서 학생을 가르치는 교수이기도 했다. 어떻게 보면 그가 내세운 모든 주장은 학생을 어떻게 가르쳐야 하는가 하는 결론에 도달한 것이라고 해도 과언이 아닐 것이다. 학생 교육에 있어서 대학의 역할에 대하여 그는 "대학의 기능은 단순히 입학 희망자들에게 선택할 수 있는 '기술'의 목록을 제공하는 것일 순 없다"고 하면서 "응당 대학은 학생과 교수진 등 그 구성원 모두를 진실 -달리 표현할 말이 있을까?-을 추구하는 즉, 자기 자신을 추구하는 인간으로 우선 바라봐야 한다."라고 선언했다.[44] 그 다음으로 교사의 의무에 대해 논했는데 교사가 스스로를 단지 '훈련'만 하는 사람으로 생각하고 남이 정한 목표를 달성하는 방법만 가르친다면 학생들에게 온전히 자율적인 개인이 아닌 "그저 남의 명령을 따르는 자가 되어, 결국 그 기능에서 언젠가 그들을 대체할지도 모를 기계보다 나을 게 없는 존재가 되도록 권하는 것이다."라고 했다.[45] 한 마디로 기계와 다름이 없는 학생을 육성할까 봐 걱정한 것이다.

ChatGPT가 출시된 후 얼마 지나지 않아 워너(John Warner)는 기계 같은 인간에 대한 문제를 상기시켰다.

44) Weizenbaum (1976), *op. cit.*, p.278.
45) *Ibid.*, p.279.

> 말 그대로 내용에 대해 아무것도 이해하지 못하는 알고리즘도 할 수 있다면 우리가 학생의 숙련도를 평가하기 위한 과제에 대해 무엇을 말해주고 있을까? … 학생들이 글쓰기를 배우는 골치 아프고 불쾌한 과정을 탐구하도록 하는 대신에 알고리즘처럼 행동하며 표면적 검사를 통과하는 시뮬레이션을 만들도록 인센티브를 주어왔던 것이다.[46]

'인센티브'란 어떤 경제적 관계를 암시하고 있는데 비슷한 맥락으로 많은 교육자가 생성형 인공지능을 논하면서 교육의 '거래적(transactional) 성격'을 언급하고 있다.[47] 골자는 생성형 인공지능이 갑자기 새로운 문제를 초래한 게 아니라 교육에 이미 존재하고 있는 근본적인 문제를 드러냈다는 것이다. 그 문제는 교육이 진정으로 자신과 세상에 관해 '학습하는 과정'이 아니라 요구된 작업을 수행하면 학위를 준다는 '거래적 관계'일 뿐이라는 점이다. 이 교육자들은 물론 미국의 현황을 이야기하는 것이지만 학원 생태계나 스트레스가 많은 입시제도 등으로 특징되는 한국의 교육 현황도 분명히 나을 게 없을 것이다.

교육의 거래적 성격을 고려하면 학생들이 생성형 인공지능을 사용하지 않는 게 오히려 놀랄 일이다. 그러면 생성형 인공지능 사용에는 어떤 부작용이 있을까? 가장 명백한 문제는 학생들이 사고하는 방법을 제대로 배우지 못할 위험이다. 교육자들이 지적한 것처럼 글쓰기란 단지 어떤 출력물을 생성하기 위한 작업이 아니라 생각하는 방식이기도 하다. 미국 소설가 플래

---

46) John Warner, "Freaking Out About ChatGPT – Part I," *Inside Higher Ed*, December 05, 2022. https://www.insidehighered.com/blogs/just-visiting/freaking-out-about-chatgpt%E2%80%94part-i

47) 다음과 같은 글에서 이런 언급을 볼 수 있다.
Beth McMurtie, "AI and the Future of Undergraduate Writing," *The Chronicle of Higher Education*, December 13, 2022. https://www.chronicle.com/article/ai-and-the-future-of-undergraduate-writing
Jordan S. Carroll, "Don't Blame Students for Using ChatGPT to Cheat," *The Nation*, January 20, 2023. https://www.thenation.com/article/society/chatgpt-plagiarism-ai-university/
Lauren M. E. Goodlad and Samuel Baker, "Now the Humanities Can Disrupt 'AI'," *Public Books*, February, 20, 2023. https://www.publicbooks.org/now-the-humanities-can-disrupt-ai/
Anna Mills and Lauren M. E. Goodlad, "Adapting College Writing for the Age of Large Language Models Such as ChatGPT: Some Next Steps for Educators," *Critical AI*, updated April 17, 2023. https://criticalai.org/2023/01/17/critical-ai-adapting-college-writing-for-the-age-of-large-language-models-such-as-chatgpt-some-next-steps-for-educators/

너리 오코너(Flannery O'Connor)는 "나는 내가 말하는 것을 읽기 전까지는 내가 무슨 생각을 하고 있는지 모르기 때문에 글을 쓴다"라는 명언을 남겼는데 이와 같은 맥락이다.[48] 필자도 본고를 작성하고 여러 차례 수정하기 전에는 생각이 제대로 정리되지 않았기에 이에 충분히 공감할 수 있다. 글을 쓰는 작업을 기계에 맡기면 학생들이 생각을 정리하고 구체화하는 법을 배우지 못하게 될 것이다. 두 번째 부작용은 역사의 파괴와 관련된 것이다. 앞에서 언급했듯이 역사의 파괴는 컴퓨터의 출현으로 인해 생긴 문제가 아니지만, 현대 사회생활을 하는 일반인이라면 인터넷에 올라와 있는 정보만 알아도 된다고 우길 수 있는 것이 우리의 현주소이다. 그렇다 하더라도 학자는 그럴 수 없다. 거인의 어깨 위에 올라서지 못하면 제대로 된 학문을 하지 못하는 법이다. 고전문학 연구야말로 더욱 그런 것 같다. 앞서 언급한 바와 같이 거대 언어모형은 디지털화가 된 자료가 아니면 알지 못하고 훈련 과정 특성상 훈련 데이터 집합에 대해서도 사실 알지 못한다. 2장에서 논한 바대로 사전훈련은 훈련 데이터 집합에서 토큰을 만들고 그런 토큰에 대한 백터를 계산하는 것인데 훈련이 끝나면 다시 훈련 데이터 집합에 접속할 수 없다. 따라서 출력하는 '주장'에 대한 출처를 밝힐 수 없고 심지어 없는 출처까지 지어내기도 한다.[49] 기계에 있어 언어는 의미가 있는 것이 아니라 단지 엮어낼 수 있는 토큰의 집합일 뿐이기 때문이다.[50] 마지막 부작용은 조금 더 장기적인 문제인데 생성형 인공지능 글쓰기가 일종의 부정적 피드백 루프를 이룬다는 것이다. 다시 말하면 훈련 데이터는 인터넷에 공개된 텍스트를 집합한 것인데 인공지능이 쓴 글이 많아지고 그것이 인터

48) 다음과 같은 글에서 재인용. 언어학 명예교수인 배론은 생각하는 방식으로서의 글쓰기에 대해 논하고 있다. Naomi S. Baron, "How ChatGPT robs students of motivation to write and think for themselves," *The Conversation*, January 19, 2023. https://theconversation.com/how-chatgpt-robs-students-of-motivation-to-write-and-think-for-themselves-197875

49) 출처를 밝히지 못하는 문제와 생성형 인공지능의 사용에 수반되는 다른 위험에 대해서는 현대어문학협회(MLA)와 대학 작문과 커뮤니케이션 회의(CCCC, Conference on College Composition and Communication)의 합동위원회가 언급한 바가 있다. Antonio Byrd et al., "MLA-CCCC Joint Task Force on Writing and AI Working Paper: Overview of the Issues, Statement of Principles, and Recommendations," MLA-CCCC Joint Task Force on Writing and AI, July 2023. https://aiandwriting.hcommons.org/working-paper-1/

50) 물론 인공지능으로 출처를 찾는 유용한 도구를 만들 수 있으나 거대 언어모형 그 자체에 이런 기능이 있을 수 없다.

넷에 공개되면 미래의 거대 언어모형은 인간이 쓴 텍스트보다 기계가 쓴 텍스트를 더 많이 접하고 그것을 배우게 될 것이다. 이처럼 올바른 글쓰기가 무엇인지에 관한 우리의 생각이 '오염'될 가능성이 없지 않다.

그렇다면 생성형 인공지능의 부정적인 영향을 최소화하려면 어떻게 대처해야 할까? 바라는 결과를 얻기 위해 다른 사람을 설득하고 장려하려면 당근도 있고 채찍도 있어야 하는 법이다. 먼저 채찍을 보자면 학생들이 생성형 인공지능을 사용하는 것을 단념하게끔 하는 전략이 없지 않다. 교육자들이 제안한 전략 중에는 플립러닝[51] 시행하기, 주로 멀티미디어 과제를 내기, 피드백과 수정 과정을 강조하기, 학생이 진정으로 관심 있는 주제로 글을 쓰게 하기 등 외에도[52] 거대 언어모형이 잘할 수 없는 주제로 글을 쓰게 하기, 확인할 수 있는 출처와 인용문을 요건으로 하기, 인공지능이 처리할 수 있는 것보다 긴 글에 대한 글을 쓰게 하기 등과 같은 것이 있다.[53] 그러나 이 전략을 제안한 교육자도 인정하듯이 완벽한 해결책은 아니다. 우선 기술이 발달할수록 지금이야 어려운 작업이 더 쉬워질 수 있는데 일례로 어떤 이미지에 대한 글을 쓰라는 과제를 주면 학생이 Midjourney라는 이미지생성 인공지능의 'describe(설명)' 기능을 이용한 후에 출력된 텍스트를 ChatGPT에 입력하는 것은 지금도 이미 가능하다. 물론 이 전략 중에 바람직한 대응책도 있지만 인공지능과 이른바 '군비 경쟁'을 벌이는 건 그리 현명한 길이 아닌 것 같다. 인공지능의 사용에 대해 처벌하는 조치에 관해서도 신중해야 할 것 같다. 최근에 MLA-CCCC 합동위원회가 지적했듯이 인공지능이 쓴 글을 탐지하는 도구가 허위양성 결과를 낼 수도 있고 너무 감시하는 분위기가 형성된다면 나름의 부작용도 생길 수 있다.[54] 루돌프(Jürgen Rudolph)와 그의 동료들도 "감시하는 접근법"보다 "학생 중심적 교수법에서 학생과의 신뢰 관계를 형성하는 접근법을 선호한다"고 한 바 있다.[55] 일상생활에서 인공지능을 쓰지 않는 사람이 없는데 학교에서만 억

51) '플립러닝(flipped learning)'이란 기존 교수방법을 뒤집어놓으려는 시도로 수업 전에 자료를 익히고 수업에 참여할 준비를 한 후 수업 시간에 과제를 수행하고 피드백을 받는 방식이다. 여기서는 '채찍' 전략으로 소개되고 있으나 학생들이 인공지능을 사용할 수 없게끔 하는 기능 이외에도 긍정적인 측면도 많으므로 교육에 대한 유망한 접근법이라고 생각된다.

52) McMurtie, *op. cit.*

53) Mills and Goodlad, *op. cit.*

54) MLA-CCCC Joint Task Force, *op. cit.*, p.7.

지로 사용할 수 없게 한다면 학생들이 느끼는 소외감이 심해질 수 있기 때문이다.

결국 위와 같은 '채찍' 전략은 요점을 놓치고 있다. 가장 근본적인 해결책은 교육의 거래적 성격을 바로잡는 것이며 이를 위해 우리가 어떤 교육을 원하는지에 대해 고민할 필요가 있다. 그 목적을 달성하기 위해 '당근' 전략을 모색하는 게 바람직할 것이다. 워너는 학생들이 알고리즘이 아니라 인간처럼 배울 수 있도록 몇 가지 제안을 했는데 기본적으로 글쓰기 과제를 재고하자는 것이다.[56] 하나의 결과물로 학습 과정을 평가하기보다 과정 그 자체를 평가한다든가 글쓰기 과제에서 요구하는 기준을 높여 개념을 종합할 수 있는 능력, 본인 고유의 입장과 체험을 이용하는 능력, 초인지적으로 반성하는 능력 등을 장려하는 것이 그 예다. 밀스와 굿래드는 위에서 살핀 '채찍' 전략보다 '당근' 전략을 추천하고 있는데 특히 내재적 동기를 키우는 것과 글쓰기 과정이 학습에 이바지하는 바를 강조하는 것이 효과적일 듯하다.[57] 캐롤은 글쓰기 수업을 가르친 체험을 공유하면서 글쓰기를 위한 연구 과정을 강조하고 있다. 특히 타당한 정보의 출처를 찾고 평가하며 이해하는 방법을 배움으로써 "학생들이 학문적 탐구를 세계에 대한 우리의 공유된 이해를 개선하기 위한 집단적 작업으로 간주하게 된다"고 했다.[58]

필자는 위와 같은 생각에 동의하며 그 핵심이 바로 '책임감'이라고 생각한다. 비행기 조종사처럼 자동화 보조구를 사용하든 안 하든 그 결과에 대한 책임이 결국 본인에게 있다는 믿음을 내재화하면 학문의 바른길에서 벗어나지 않을 것이다. 이러한 '내재화된 책임감'은 '내재화된 동기'의 이면이라고도 할 수 있다. 즉, 학습에 대해서는 학생 본인이 스스로 동기를 부여하고 학습 과정과 결과에 대해 책임을 져야 한다는 것이다. 그렇다면 인공지능 시대에 이런 책임감을 어떻게 키워야 할까? 일단 인공지능의 능력에 관하여 바로 아는 것이 중요하다. 위에서 언급했듯이 기계에 '홀리면' 그 기계를 맹신하게 되는 경우가 많다. 달리 표현하자면 기계에 대한 지식이 적

55) Jürgen Rudolph et al, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" *Journal of Applied Learning & Teaching*, Vol. 6, No. 1, 2023, p.354.
56) Warner, *op. cit.*
57) Mills and Goodlad, *op. cit.*
58) Carroll, *op. cit.*

을수록 기계의 능력을 제대로 판단할 수 없다는 것이다. 그러나 기계의 능력의 한계를 제대로 파악하면 그 기계를 어디까지 신뢰할 수 있는지도 알 수 있다. 따라서 학생들이 인공신경망이 어떻게 작동하는지는 상세히 모를지라도 ChatGPT와 같은 거대 언어모형의 기본 작동법과 그 한계를 알게 되면 맡겨도 되는 작업과 맡기면 안 되는 작업을 더 효과적으로 구분할 수 있을 것이다.

문제는 이러한 접근을 통해 책임감을 내재화할 수 있는 기반이 마련될지언정 그렇게 하도록 동기부여가 되지는 못한다는 점인데 책임감을 학생들에게 심어줄 좋은 방법은 없을까? 불행히도 책임감은 본질상 남에게 심어줄 수 있는 것이 아니라 다만 그것을 스스로 키워나갈 환경을 조성할 수 있을 뿐이다. 자동화 편견을 연구한 모지어가 내재화된 책임감에 대해서 논하면서도 그것을 어떻게 키울지에 대해 언급하지 않은 것을 보면 이를 특정 조종사 개인의 성격적 특징으로 간주한 것 같다. 역시 학생 중에도 책임감이 더 강한 학생이 있고 더 약한 학생이 있겠지만 교사가 학생이 스스로 책임감을 키우는 환경을 조성할 수는 있다고 생각한다. 우선 무엇보다 학생이 모험을 할 수 있어야 하는데 그러기 위해서는 실패가 허용되는 환경이 조성되어야 한다. 현 교육제도는 결과에 집착하고 있기 때문에 실패를 두려워할 수밖에 없겠지만 결과가 아닌 과정에 집중한다면 '실패'란 그저 그 과정의 일부가 되는 것이다. 공교롭게도 위에서 언급한 워너의 '당근 전략'과 같은 맥락인데 인공지능에 너무 의지하지 않도록 하는 것 이외에도 효과가 있을 듯하다. 시행착오를 거치면서 결국 원하는 목표에 도달하는 과정을 학습의 목표로 삼는다면 학생이 나름대로 책임감을 스스로 키울 수 있지 않을까 싶다. 교사는 이런 환경을 조성하고 학생들에게 이런 기회를 주면서 도와주는 조력자 역할을 맡으면 된다. 한국 사회의 위계질서적인 성격과 학부의 사제관계를 고려하면 실행하기가 쉽지 않을 듯하지만 꼭 필요한 것이라 생각한다. 물론 이렇게 하더라도 동기를 찾지 못하거나 지름길만 찾는 학생이 있겠지만 어쩔 수 없는 현실이다. 그래도 필자의 체험으로 말하자면 학생 대부분이 긍정적으로 반응하리라 확신한다.

여기서 필자의 경험에서 구체적인 예를 들겠다. 이는 완벽한 방법이라고 할 수 없고 필자도 어떻게 개선할 수 있을까 매학기 고민하고 있지만, 추상

적 논의만 이어나가기보다는 구체적인 실례를 제시하는 게 좋을 듯싶어 짧게 언급하고자 한다. 수업 시간에 학생에게 발표시키는 것은 기존에도 해오던 방식인데 필자는 한 걸음 더 나아가 학생들에게 책임을 질 기회를 부여하고자 학생들을 팀으로 나누어 한 팀이 한 수업 전체를 맡게 한다. 발표에 그치는 것이 아니라 스스로 텍스트를 선정하여 수업을 가르치는 것이다. 사실 필자도 학부 시절에 교환학생으로 해외에 나갔을 때 교수가 수업을 한 번 가르치라고 했다. 오늘날까지 그 경험을 생생하게 기억하고 있고 아마도 한 학기를 통틀어 배운 것보다 그날 하루에 배운 것이 더 많았을 것이다. 이제 교수가 되어 학생에게 그같이 하게 해보니 처음에는 중간에 끼어들어 '가르쳐주고' 싶은 마음에 안절부절못해 꽤 힘들었다. 그러나 몇 학기 동안 이렇게 해보면서 배운 것은 학생을 믿고 맡기는 만큼 대부분이 책임감을 느끼고 기대에 부응한다는 것이다. 물론 원래 책임감이 강한 학생이 더 잘하겠거니 할 수도 있겠지만 최소한 그 책임감을 강화하는 기회가 된다고 생각한다.

마지막으로 짧게 고민하고 싶은 것은 생성형 인공지능을 교육의 도구로 사용할 가능성이다. ChatGPT와 같은 인공지능은 좋은 것도 나쁜 것도 아닌 그저 도구일 뿐 관건은 우리가 그 도구를 어떻게 사용하느냐이다. 교육자 중에 당근과 채찍을 말하는 사람이 많으나 MLA-CCCC 합동위원회는 특히 문학 교육과 글쓰기 교육에서 거대 언어모형을 활용하는 방법을 모색하고 있다.[59] 문학 교육에는 토론의 출발점을 마련하기 위해 거대 언어모형에게 문학 텍스트에 대해 다양한 질문을 하기, 어떤 저자에 대해서 강의할 때 거대 언어모형으로 그 저자의 문체를 모방하거나 자주 다루는 주제를 정리하기, 비슷한 주제를 다루는 아주 다른 텍스트를 찾기 위해서 거대 언어모형에 문의하기 등과 같은 방법을 제안한다. 글쓰기 교육에는 발상을 찾거나 수정할 때 거대 언어모형을 사용하거나 다중 양식(multimodal) 글쓰기 과제를 위해 생성형 인공지능으로 이미지, 소리, 동영상 등을 생성하는 등의 방법도 있다. 필자는 솔직히 어떤 제안에 대해 회의적이기는 하지만 기본적인 의도에는 동의한다. 고백하건대 본고를 작성했을 때도 인공지능을 사용했다. 물론 거대 언어모형에게 글을 대신 써 달라고 맡기지는 않

59) MLA-CCCC Joint Task Force, *op. cit.*, pp.9~10.

았지만 필자가 ChatGPT에 다양한 질문을 던져보았다. '知彼知己 百戰百勝'의 지혜를 염두에 두고 ChatGPT가 질문에 어떻게 반응할지를 알아보고 싶었는데 그러한 작업을 통해서 ChatGPT의 강점과 약점을 파악하는 데에 도움이 됐다. 또한 ChatGPT가 좋은 '브레인스토밍' 도구로 사용될 수 있다. 사용자의 질문에 생성하는 출력물은 완제품은 물론 초고도 안 되겠지만 씨앗은 될 수 있다. 그 씨앗에 인간이 자신의 체험, 지능, 그리고 지혜로 영양과 물을 주면 의미 있는 열매를 맺을 수 있을 것이다. 거대 언어모형은 언어의 의미를 이해하지 못하므로 진정으로 창의적인 것을 만들 수는 없지만, 인간에게 창의력을 유발할 수 있다고 생각한다.

다만 다양한 학자의 논의를 살펴보면서 문제의 소지가 있는 제안도 발견했다. 소위 '지능적 개인교습 시스템(intelligent tutoring systems)'이나 자동화된 글쓰기 평가가 그런 것이다.[60] 전자는 생성형 인공지능을 이용하여 수업 시간 이외에도 학습 능력을 향상하는 시스템이며 후자는 학생의 글을 평가할 수 있도록 생성형 인공지능을 훈련하는 것이다. 물론 교육자도 첨단 기술을 사용하는 도구를 이용하여 일을 보다 효과적으로 하는 것이 나쁘다고 할 수 없지만 1960년대에 챗봇 일라이자를 보고 정신 치료를 자동화하자는 정신과 의사들의 주장이 떠오른다. 여기서 인공지능과 인간의 노동력을 논한 브린욜프쏜(Erik Brynjolfsson)의 경고를 유의할 필요가 있다.

> 인공지능의 분배 효과는 주로 인간의 노동력을 증강시키는 데 사용되는지 아니면 자동화하는 데 사용되는지에 따라 달라진다. 인공지능이 인간의 능력을 증강시켜 사람들이 이전에는 할 수 없었던 일을 할 수 있게 해줄 때 인간과 기계는 상호보완적인 존재가 된다.[61]

---

60) 지능적 개인교습 시스템과 자동화된 글쓰기 평가에 대해서 다음과 같을 자료 참조할 수 있다. Nabeel Gillani et al, "Unpacking the 'Black Box' of AI in Education," *Educational Technology & Society*, Vol. 26, No. 1, January 2023, pp.99~111. 그리고 지능적 개인교습 시스템의 논의와 구체적인 예에 대해서 다음과 같은 자료를 참조할 수 있다.. Gwo-Jen Hwang and Nian-Shing Chen, "Editorial Position Paper: Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions," *Educational Technology & Society*, Vol. 26, No. 2, April 2023, pp. i~xviii.

61) Erik Brynjolfsson, "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence," *Daedalus*, Vol. 151, No. 2, Spring 2022, p.273.

교육자도 노동자인데 교육자의 노동력을 자동화하면 인간이 기계로 대체되는 길을 걷는 것이지만 교육자의 능력을 향상하기 위한 목적으로 인공지능이 사용되면 인간 중심의 대학교육을 보호할 수 있을 것이다.

결국 우리는 인공지능이 존재하는 세상에 살고 있고 인공지능이 갑자기 사라지는 미래는 상상하기 어려우므로 인공지능을 유익하게 활용하는 방법을 모색해야만 한다. 그리고 착각하면 안 된다. 인공지능이 발달하더라도 기계는 기계이고, 인간은 인간이다. 교사는 학생에게 정보만 전달하는 기계나 실적물만 생성하는 기계가 아니며 학생 또한 정보를 흡수만 하는 기계나 요구된 과제물만 생성하는 기계가 아닌 것이다. 수량화하는 눈으로 보지 않고 서로를 '온전한 인간'으로 인식하고 대우하는 것이야말로 인공지능 시대에 교육의 바람직한 미래를 향한 첫걸음이 되지 않을까 싶다.

## Ⅵ. 나오며

위에서 교육의 현주소와 미래를 살펴보았으나 결론을 내리는 대신에 인공지능과 사회의 미래에 대한 몇 가지의 단상을 나누고자 한다. 앞서 논한 바처럼 바이젠바움은 인간과 기계를 구분하고자 했는데 물론 둘의 차이점이 많지만 근본적인 차이는 기계가 '결정(decide)'할 수 있으나 인간만이 '선택(choose)'할 수 있다는 것이다.[62] 다시 말해 기계에는 지혜나 도덕적 상상력이 있을 수 없으며 "기계가 아무리 지능적인 것으로 만들어진다 해도 오직 인간에 의해서만 시도되어야 하는 사고 행위들이 있다"고 주장했다.[63] 30년 전에 과학자의 오만으로 되살아난 공룡들이 인간을 공격하는 모습을 그린 영화 <쥬라기 공원>이 개봉됐는데 대사 중에 카오스 이론을 연구하는 이언 맬컴 박사의 "당신의 과학자들은 할 수 있는지 없는지에 대해 너무 몰두한 나머지 해야 하는가에 대해 곰곰이 생각하지 않았어요."라는 말이 생각난다. 뻔한 말인 것 같지만 할 수 있는 것과 해야 하는 것은 별다른 문제이다. 베트남 전쟁 당시 무고한 민간인이 죽고 사는 것에 대해 기

62) Weizenbaum (1976), *op. cit.*, pp.258~260.
63) *Ibid.*, p.13.

계가 결정할 능력이 있었다 하더라도 과연 기계에 맡기는 게 옳았을까? 물론 그 사건은 생각하기도 싫은 옛일이라며 넘어갈 수도 있겠지만 오늘날 법정에서 피고인의 재범 위험을 판단하는 일, 구직하는 사람의 업무 적합성을 판단하는 일, 대출 신청자의 신용 리스크를 판단하는 일 등을 기계에 맡겨야만 될 것인가?[64] 이미 하고 있기 때문에 그런 질문하기에 다소 늦은 감이 있기는 하지만 그래도 인간이 기계에 책임을 전가하는 사회에 살고 싶지 않으면 그런 질문을 던져야만 한다. 이런 일은 교육과 아무런 상관이 없다고 생각할 수도 있으나 인공지능이 사회의 모든 영역에 영향을 미치기 때문에 무관하다고만은 할 수 없다. 법정이나 은행에만 부작용이 있고 대학교에서는 멀쩡할 리가 없다. 그렇다면 우리는 어떤 미래를 원하는가? 기계와 달리 인간은 그것을 '선택'할 수 있다.

본고를 시작하면서 인공지능의 초기 선전자인 사이먼과 뉴웰의 예언을 인용했는데 마무리하면서 다시 한번 자세하게 들여다보고자 한다. 기계의 능력이 "급속히 증가"하며 인간과 "동일하게 될 것"이라고 했는데 자세히 보면 이렇게 만드는 행위자가 없고 그저 저절로 그렇게 될 것처럼 말하고 있다. 거대 언어모형에 대해서도 사람들이 이야기하는 것을 들어보면 비슷한 것 같다. 가면 갈수록 인공지능이 아무런 외부의 영향 없이 그냥, 자연스럽게 좋아질 거라는 것이다. 그러나 자연스럽게 되는 것도 없고, 저절로 일어나는 현상도 없으며, 배후에 자신만의 의도를 추구하는 행위자가 없는 일은 절대로 없다. '기술적 발달의 필연성'은 기술이상주의의 대사제들이 평신도에게 맹신을 불어넣기 위한 거짓일 뿐이다. 물론 바이젠바움도 이와 비슷한 경고를 이미 한 바 있다. 그는 '기술적 발달의 필연성'을 "양심의 강력한 진정제"라고 지칭하면서 "그것의 역할은 그것을 진정으로 믿는 모든 자의 어깨에서 책임을 덜어내는 것"이라고 했다.[65] 21세기에도 비슷한 목소리가 들린다. 민주주의에 대한 알고리즘의 위협을 연구한 오닐은 "만일 우리가 수학적 모델을 날씨나 조수처럼 중립적이고 필연적인 힘으로 취급한다면 우리의 책임을 포기하는 것이다"라고 경고했고[66] 벤더는 "많은 자원이 거

64) 오닐은 이와 같은 일을 자세히 다루고 있다. O'Neil, *op. cit.*
65) Weizenbaum (1976), *op. cit.*, p.241.
66) *Ibid.*, p.218.

대 언어모형(그리고 이미지 데이터 집합으로 훈련되는 거대 모형)에 쏟아지는 상황에서 이에 대해 미리 정해진 것은 아무것도 없다는 사실을 간과해선 안 된다."고 상기시킨 바 있다.[67] 노동력의 증강과 대체를 논한 브린욜프쏜 역시 "미래는 미리 정해져 있지 않다. 우리는 증강을 통해 인간의 기회를 확장하거나 자동화를 통해 인간을 대체하는 정도를 통제한다."고 했다.[68] 그리고 '교육을 위한 인공지능 권리 장전'을 고민한 콘래드(Kathryn Conrad)는 다음과 같이 인공지능 개발자들을 적나라하게 비판했다.

> 이러한 모델은 교육적 목표・관행・원칙을 고려하지 않고 설계된 것일 뿐만 아니라, 자주 고등 교육을 폄하하고, 교육이 대체로 자동화 가능한 작업이라고 상상하며, 인간 학습을 수익성 있는 기술의 습득으로 생각하고, 학생과 교사 모두를 무료 훈련 데이터의 원천으로 간주하는 기술주의적 환경에서 비롯된다.[69]

콘래드의 '교육을 위한 인공지능 권리 장전'을 보면 우리가 걱정해야 할 것은 학생들이 생성형 인공지능을 부적절하게 이용하는 것보다 오히려 학생들이 인공지능 때문에 받게 될 피해가 아닌가라는 생각이 든다. 예를 들면 교실에서 생성형 인공지능을 효과적으로 사용한다고 하더라도 학생의 창작물이 인공지능의 훈련 데이터로 사용될 가능성을 무시해서는 안 된다는 것이다. 물론 인간이 앞으로 인공지능에 의한 혜택을 많이 볼 수도 있겠지만 자칫하면 우리의 이익에 부합하지 않은 인공지능이 출현할 수도 있다. 우리가 인공지능 개발자는 아닐지라도 인공지능의 영향을 똑같이 받기 때문에 우리가 원하는 인공지능을 요구할 권리가 있다. 거대 언어모형에 의해 변형되어가는 교육의 지형을 교사와 학생이 함께 탐사하면서 이 사실을 염두에 두었으면 한다. 그리고 우리는 기계가 아니라 선택하고 책임질 수 있는 인간이라는 사실을 잊어서는 안 된다.

후기로 한마디를 덧붙이자면 최근 이러한 문제의 시급성을 단적으로 보

67) Bender (2022), *op. cit.*
68) Brynjolfsson, *op cit.*, p.281.
69) Kathryn Conrad, "Sneak Preview: A Blueprint for an AI Bill of Rights for Education," *CriticalAI*, July 17, 2023. https://criticalai.org/2023/07/17/a-blueprint-for-an-ai-bill-of-rights-for-education-kathryn-conrad/

여주는 사건이 인공지능 산업계에서 벌어졌다. 바로 11월 17일에 ChatGPT를 개발한 OpenAI사의 이사회가 최고 경영자인 올트먼을 사직시킨 일이다. 곧바로 올트먼이 마이크로소프트사에서 일하게 될 거란 소문이 퍼지고 OpenAI의 직원 90% 이상이 회사를 그만두겠다고 으름장을 놓았다. 결국, 22일에 올트먼이 복귀되었고 그를 몰아냈던 사람들이 이사직을 박탈당했다. 이렇게 하여 회사가 원상태로 돌아간 것 같으나 실상은 그렇지 않다. OpenAI는 원래 비영리 단체로 설립되면서 "인공일반지능이 온 인류에게 유익하게 되도록 보장하는 것"을 사명으로 한다고 헌장에서 밝힌 바 있다.[70] 이사회가 올트먼을 몰아낸 이유는 정확히 밝혀지지 않았지만, 이 사명에 대해 의견 차이가 있지 않았나 하는 추측이 무성하다. 만약 올트먼이 3월에 '인공지능 윤리와 사회' 팀 전원을 해고한 마이크로소프트사[71]로 전직할 준비가 되어 있었다면 최소한 인공지능을 인류에게 유익하게 개발하는 것을 최우선으로 생각하지 않았다고 할 수 있을 듯하다. 이번 일은 인공지능 개발을 주도하는 자들이 우리의 최선의 이익을 고려하기보다 첨단 인공지능을 최대한 빨리 개발하려 한다는 점을 여실히 보여준다. 따라서 위에서 말했듯이 우리가 원하는 미래를 위해 목소리를 내는 것이 어느 때보다도 중요하다.

70) OpenAI, "OpenAI Charter," *OpenAI*, April 9, 2018. https://openai.com/charter

71) Zoe Schiffer and Casey Newton, "Microsoft lays off team that taught employees how to make AI tools responsibly," The *Verge*, Mar 14, 2023. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs

## 참고문헌

1. 단행본

Donath, Judith, “Being Real: Questions of Tele-Identity,” *The Robot in the Garden – Telerobotics and Telepistemology in the Age of the Internet*, ed. Ken Goldberg, Cambridge (MA): The MIT Press, 2000.

Galton, Francis, *Hereditary Genius: An Inquiry Into Its Laws And Consequences*, London: Macmillan and Co., 1869.

Noble, Safiya Umoja, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: New York University Press, 2018.

O’Neil, Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Broadway Books, 2016.

Pasquale, Frank, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge (MA): Harvard University Press, 2015.

Plato, *Plato’s Phaedrus*, tr. R. Hackforth, Cambridge: Cambridge University Press, 1952.

Weizenbaum, Joseph, *Computer Power and Human Reason: From Judgment to Calculation*, New York: W. H. Freeman & Co., 1976.

2. 논문

Bender, Emily and Alexander Koller, “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, July 5-10, 2020, pp.5185~5198.

Brynjolfsson, Erik, “The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence,” *Daedalus*, Vol. 151, No. 2, Spring 2022, pp.272~287.

Gillani, Nabeel et al, “Unpacking the ‘Black Box’ of AI in Education,” E*ducational Technology & Society*, Vol. 26, No. 1, January 2023, pp.99~111.

Hutson, Matthew, “The Language Machines,” *Nature*, Vol. 591, 4 March 2021, pp.22~25.

Hwang, Gwo-Jen and Nian-Shing Chen, “Editorial Position Paper: Exploring the Potential of Generative Artificial Intelligence in Education: Applications,

Challenges, and Future Research Directions," *Educational Technology & Society*, Vol. 26, No. 2, April 2023, pp.i~xviii.

McClelland, David C., "Testing for Competence Rather Than for 'Intelligence,'" *American Psychologist*, Vol. 28: No. 1, January 1973, pp.1~14.

Mosier, Kathleen L. et al., "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *The International Journal of Aviation Psychology*, Vol. 8: No, 1, 1998, pp.47~63.

Muir, Bonnie M., "Trust between humans and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, Vol. 27: Issues 5-6, 1987, pp.527~539.

Nobles, Melissa et al., "Science must overcome its racist legacy: Nature's guest editors speak," *Nature*, Vol. 606, 09 June 2022, pp.225~227.

Rudolph, Jürgen et al, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?", *Journal of Applied Learning & Teaching*, Vol. 6, No. 1, 2023, p.354.

Simon, Herbert A. and Allen Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research*, Vol. 6: No. 1, 1958, pp.1~10.

Turing, Alan, "Computing Machinery and Intelligence," *Mind: A Quarterly Review of Psychology and Philosophy*, Vol. 59: No. 236, pp.433~460.

Weizenbaum, Joseph, "ELIZA－A Computer Program For the Study of Natural Language Communication Between Man And Machine," *Computational Linguistics*, Vol. 9: No. 1, January 1966, pp.36~45.

Wickens, Christopher D. and Stephen R. Dixon, "The benefits of imperfect diagnostic automation: a synthesis of the literature," *Theoretical Issues in Ergonomics Science*, Vol. 8: No. 3, 2007, pp.201~212.

3. 기타

Albert, Alex, *Jailbreak Chat*, https://www.jailbreakchat.com/

Baron, Naomi S., "How ChatGPT robs students of motivation to write and think for themselves," *The Conversation*, January 19, 2023. https://theconversation.com/how-chatgpt-robs-students-of-motivation-to-write-and-think-for-themselves-197875

Bender Emily, Timnit Gebru et al., "On the Dangers of Stochastic Parrots: Can

Language Models Be Too Big?" Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3-10, 2021, p.614 https://doi.org/10.1145/3442188.3445922

Bender, Emily, "On NYT Magazine on AI: Resist the Urge to be Impressed," *Medium*, April 18, 2022. https://medium.com/@emilymenonbender/on-nyt-magazine-on-ai-resist-the-urge-to-be-impressed-3d92fd9a0edd

Byrd, Antonio et al., "MLA-CCCC Joint Task Force on Writing and AI Working Paper: Overview of the Issues, Statement of Principles, and Recommendations," MLA-CCCC Joint Task Force on Writing and AI, July 2023. https://aiandwriting.hcommons.org/working-paper-1/

Carroll, Jordan S., "Don't Blame Students for Using ChatGPT to Cheat," *The Nation*, January 20, 2023. https://www.thenation.com/article/society/chatgpt-plagiarism-ai-university/

Conrad, Kathryn, "Sneak Preview: A Blueprint for an AI Bill of Rights for Education," *CriticalAI*, July 17, 2023. https://criticalai.org/2023/07/17/a-blueprint-for-an-ai-bill-of-rights-for-education-kathryn-conrad/

Goodlad, Lauren M. E. and Samuel Baker, "Now the Humanities Can Disrupt 'AI'," *Public Books*, February, 20, 2023. https://www.publicbooks.org/now-the-humanities-can-disrupt-ai/

Kemp, Simon, "Digital 2023 July Global Statshot Report," *DataReportal*, 20 July 2023. https://datareportal.com/reports/digital-2023-july-global-statshot

Lowe, Ryan and Jan Leike, "Aligning Language Models to Follow Instructions," *OpenAI*, January 27, 2022. https://openai.com/research/instruction-following

Marcus, Gary, "AI Platforms like ChatGPT Are Easy to Use but Also Potentially Dangerous," *Scientific American*, December 19, 2022. https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/

McMurtie, Beth, "AI and the Future of Undergraduate Writing," *The Chronicle*

*of Higher Education*, December 13, 2022.
https://www.chronicle.com/article/ai-and-the-future-of-undergraduate-writing

Mills, Anna and Lauren M. E. Goodlad, "Adapting College Writing for the Age of Large Language Models Such as ChatGPT: Some Next Steps for Educators," *Critical AI*, updated April 17, 2023.
https://criticalai.org/2023/01/17/critical-ai-adapting-college-writing-for-the-age-of-large-language-models-such-as-chatgpt-some-next-steps-for-educators/

Nietzel, Michael T., "More Than 80% of Four-Year Colleges Won't Require Standardized Tests for Fall 2023 Admissions," *Forbes*, Nov. 15, 2022.
https://www.forbes.com/sites/michaeltnietzel/2022/11/15/more-than-80-of-four-year-colleges-wont-require-standardized-tests-for-fall-2023-admissions/

OpenAI, "OpenAI Charter," *OpenAI*, April 9, 2018. https://openai.com/charter

Perrigo, Billy, "OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," *Time*, January 18, 2023.
https://time.com/6247678/openai-chatgpt-kenya-workers/

Piper, Kelsey, "Why Meta's move to make its new AI open source is more dangerous than you think," *Vox - Future Perfect*, August 2, 2023.
https://www.vox.com/future-perfect/23817060/meta-open-source-ai-mark-zuckerberg-facebook-llama2

Rettberg, Jill Walker, "ChatGPT is multilingual but monocultural, and it's learning your values," *jill/txt*, December 6, 2022.
https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/

Schiffer, Zoe and Casey Newton, "Microsoft lays off team that taught employees how to make AI tools responsibly," *The Verge*, Mar 14, 2023.
https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs

Touvron, Hugo, Thomas Scialom et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv*, Cornell University, 19 July 2023.
https://arxiv.org/abs/2307.09288

Warner, John, "Freaking Out About ChatGPT - Part I," *Inside Higher Ed*,

December 05, 2022.
https://www.insidehighered.com/blogs/just-visiting/freaking-out-about-chatgpt%E2%80%94part-i

# How Should College Education Respond to Large Language Models?

Charles La Shure

The release of ChatGPT to the public at the end of last year had many in the field of education worried. In response, this paper explored the future of college education and artificial intelligence (AI). First, a proper understanding of how large language models (LLMs) "train" and "learn," along with their abilities and limitations, was established. Simply put, while LLMs produce plausible linguistic output, they are "stochastic parrots" that have no actual understanding of language.

Next, we examined the dangers of generative AI and discovered that they might help in the creation and dissemination of misinformation. Even if these AI are not used with malicious intent, the fact that their training data sets are drawn from the internet—which reflects majority thinking—means that they can perpetuate and amplify social inequality and hegemonic stereotypes and biases. On the other hand, if we consider what is missing from the training data, it is only natural that marginalized voices should be even more marginalized. In addition, leaving the issue of the socially vulnerable aside, LLMs can only be trained on digital data, meaning analog data is ignored. This is in line with the idea of "the destruction of history" put forth by Joseph Weizenbaum, an early critic who warned of the dangers of artificial intelligence.

We then discussed the relationship between humans and machines and considered which relationships were problematic and which were desirable. Researchers in the aviation industry recognized the problem of automation bias from an early date, but this phenomenon can be seen in other areas of society as well. Put simply, if a human places too much trust in a machine, they abdicate their decision-making responsibility to that machine and thus fail to respond quickly to solve any problems that may arise should that machine malfunction. LLMs do not endanger lives in the same way that airplanes do, but a similar bias can be seen with them as well. A more important issue, though, is the fact that people are no longer seen as whole

human beings but as computers. This tendency was evident long before the advent of computers, for example in the attempts to quantify human intelligence through IQ tests, but it is a problem we must be particularly wary of in the age of AI.

Lastly, we considered means for college education to find its way in the present situation. Educators in the US in particular, while dealing with ChatGPT, have pinpointed not the LLMs themselves but the "transactional nature" of education as the problem. That is, they argue that education has long since become less a process of learning and more a transaction in which students receive grades and degrees. Given this transactional environment, it is no wonder that student would rely too much on ChatGPT. This over-reliance, however, comes with side effects: not learning how to think properly, a lack of sufficient academic information, and learning an AI-based writing style. In response, US educators have proposed both "stick" (strategies that make it difficult for students to use LLMs) and "carrot" (strategies that encourage students to learn like human beings, not algorithms) solutions, but the heart of the matter seems to be a sense of responsibility. Creating an educational environment in which students can develop a sense of responsibility for themselves is the path forward for education in the age of AI. If we do this, LLMs can become a useful tool rather than an enemy to fear.

Keywords: Artificial Intelligence, Generative Artificial Intelligence, Large Language Model,. ChatGPT, Artificial Neural Network, 'Stochastic Parrot', Techno-Utopianism, Automation Bias, Human-Machine Relationship, College Education, Pedagogy

접수일자: 2023. 9. 30.
심사기간: 2023. 10. 1.~2023. 11. 10.
게재결정: 2023. 11. 10.