

초거대 생성형 인공지능의 윤리적 문제

변순용*

Abstract

Ethical Problems of Super-Massive Generative AI

Sunyong, Byun

This study analyzes various ethical issues and the need for ethical guidelines in the context of the emergence of generative artificial intelligence (AI) and growing concerns about the potential risks and ethical regulations of AI. With the advent of generative AI, concerns about the intellectual capabilities of AI have increased; this study examines the manner in which this differs from the abilities of machines in a digital society. This analysis suggests that the AI era extends beyond the digital age and presents new aspects of AI citizenship, AI literacy, and AI ethics. This explains the significance of ethical common sense in the context of AI's role as a new producer of knowledge, drawing on Delphic ethical common sense for inference. This study focuses on the tendency of knowledge production by AI to be based on inductive reasoning and attempts to address the problems arising from the ability to question and judge. It analyzes the ethical validation of the vast knowledge and information generated by generative AI and discusses the external ethical issues (such as energy waste and environmental pollution) associated with generative AI, highlighting the potential ethical problems that can arise from AI. As various types of

* 서울교육대학교 윤리교육과 교수

AI are being used and developed, even greater ethical issues may exist that have not been previously considered or that surpass our current understanding. Therefore, research on the ethical deliberation and consensus regarding the ethical issues of AI is necessary.

Key words : Generative artificial intelligence, ethical verification, subject of knowledge production, digital age, artificial intelligence-digital age

〈 목 차 〉

1. 들어가는 말
2. 디지털 시대의 구분: 디지털화와 디지털 변형, 인터넷-디지털 시대와 AI-디지털 시대
3. 지식생산의 새로운 주체의 등장 1: 도덕적 상식(Commonsense Moral)의 의미에 대한 물음
4. 지식생산의 새로운 주체의 등장 2: 연역적, 귀납적 사고의 대체 가능성
5. 물을 수 있는 능력과 판단할 수 있는 능력의 중요성
6. 생성형 인공지능에 대한 윤리적 검증의 필요성
7. 인공지능과 에너지 소비의 문제
8. 결론

1. 들어가는 말

최근 오픈AI가 개발한 대화 전문 인공지능 ChatGPT에 사람들의 관심이 집중되면서 구글이 바드(Bard), 마이크로소프트에서도 유사한 기술을 탑재한 빙(Bing)을 출시하면서 생성형 인공지능간의 경쟁이 시작되고 있다. 그런데 이 와중에 딥러닝의 아버지라고 불리는 힌턴(Geoffrey Hinton)은 구글사에서 퇴직하면서 AI챗봇의 위험이 매우 심각해질 것이고, 인공지능이 현재는 인간보다 지능적이지 못하겠지만, 조만간 인간을 초월할 수 있다고 경고하고

있다.¹⁾ Chat GPT를 만들어낸 OpenAI의 CEO인 엘트만(Sam Altman)은 미 의회 청문회에서 새로운 AI시스템은 규제되어야 한다고 주장하고 있다.²⁾ 이렇게 최근에 인공지능 기술의 첨단에서 개발을 했거나 하고 있는 사람들의 입에서 자발적으로 기술의 잠재적 위험 가능성과 윤리적 규제의 필요성에 대한 요청이 제기되고 있다.

전통적인 휴머니즘의 성격에 포함되기 어려운 새로운 인간성의 요소들이 속속들이 등장하고 있다. 예를 들어 인간과 기계의 결합을 통해 인간능력향상(human enhancement)을 외치는 트랜스 휴먼의 현상이 나타나고, 딥페이크 및 음성 및 표정 관련 기술의 발달로 인해 디지털 휴먼이 등장하고, 인간과 인간처럼 이야기를 나눌 수 있고 더 나아가 감정적 소통까지 이뤄내는 챗봇이 우리 핸드폰에 탑재되고 있다. 그래서 우리 시대에 그리고 디지털시대에 요청되는 새로운 휴머니티 내지 디지털 휴머니티가 무엇인지에 대해 생각해 보아야 한다.

생성형 인공지능은 인간과 유사하거나 인간을 뛰어넘을 수 있는 지적 능력을 갖게 된다고 말하고 있다. 우리가 이러한 생성형 인공지능을 왜 필요로 할까? 왜 한쪽에서는 두려워하고 있는 것을 다른 한쪽에서는 계속 업그레이드시키는 것일까?

이 글에서는 생성형 인공지능의 등장으로 인해 촉발된 디스토피아적 관점을 경계하면서도 생성형 인공지능에 대한 유토피아적 전망에 가려 드러나지 않는 문제점을 살펴보고자 한다. 기술의 사용에 대한 편리함이라는 장점 못지않게 우리가 대가로 지불해야 할 비용에 대한 논의를 해보고자 한다. 그러기 위해서는 지금 우리의 현주소가 어디인지에 대한 좌표설정이 필요해보이고, 이를 위해 디지털 시대라는 막연한 일반 개념에 대한 분석이 선행될 필요가 있다.

1) <https://www.bbc.com/news/world-us-canada-65452940> 참조.

2) <https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html>

2. 디지털 시대의 구분: 디지털화와 디지털 변형, 인터넷-디지털 시대와 AI-디지털 시대

그래서 인터넷과 웹, 브라우저와 앱을 통한 디지털화(Digitalization)은 머신러닝과 딥러닝, 챗봇과 빅데이터를 통한 인공지능 변형(Digital or AI Transformation)과는 질적으로 다를 수 밖에 없다. 디지털화는 디지털 기술인프라를 만드는 사회적, 제도적 맥락에서 디지털 기술을 적용하는 사회기술적 과정이다.³⁾

<표 1>

용어	정의	인용근거
자료의 디지털화 (Digitization)	아날로그 포맷을 디지털 포맷으로 변환하는 기술적 과정	Tilson et al. (2012) Sandberg et al. (2020)
사회 제도의 디지털화 (Digitalization)	디지털화 기술을 광범위한 사회 및 제도적 맥락에 적용하는 사회기술적 과정	Nylén and Holmström (2015) Yoo et al. (2010)
디지털 변형 (Digital transformation)	디지털 기술의 기회를 활용하기 위하여 사회 내 제반 조직의 활동, 조직의 경계, 목표의 근본적인 변화가 일어나는 과정	Matt et al. (2015) Vial (2019)

홀름스트룀(J. Holmstroem)은 아날로그를 디지털로 바꾸는 기술적 과정, 이것을 사회적 맥락으로 확장하는 사회기술적 과정, 그리고 이러한 과정 자체의 본질적 변형과정으로 구분하고 있다. 그의 디지털 변형이라는 개념과 달리 인터넷상의 디지털 변형과 AI를 통한 디지털 변형의 구분 필요성을 주장하는 입장도 제기되고 있다. 아래 그림에서 보는 것처럼 전통적인 소프트웨어 개발의 과정과 기계학습 기반 소프트웨어 개발의 본질적 차이로 인해

3) Holmstroem, J. (2022): From AI to digital transformation: The AI readiness framework, Vol. 65, Issue 3, May-June, p. 331. <https://www.sciencedirect.com/science/article/pii/S0007681321000744> 참조.

AI변형의 개념이 주장된다. 인터넷과 웹, 소셜넷, 브라우저와 앱, 스마트폰을 기본으로하는 디지털 변형과 머신러닝과 딥러닝, 빅데이터와 챗봇 등을 기본으로 하는 AI변형으로 구분해야 할 필요성이 제기되고 있다.

<표 2>

	기술	인터페이스	디바이스	인프라
디지털 변형	인터넷 웹(HTML)	브라우저, 앱	스마트폰	3G, 4G 클라우드
AI 변형	머신러닝 딥러닝	챗봇 에이전트	자동차 로봇	5G 모바일 네트워크 클라우드

우리는 이미 디지털 시대에 들어와 있으며, 앞으로도 한동안 디지털 시대가 지속될 것이지만, 디지털 시대 안에서도 여러 질적인 변화 내지 변형의 단계들을 구분해야하고, 이에 따른 적절한 대응과 준비가 요청된다. 이 글에서는 과거의 디지털시대와의 구분을 위해 AI 변형 혹은 AI-디지털 사회를 같은 의미에서 사용하고자 한다.

이러한 빠른 변화에 대하여 정보윤리에서 논의되어 온 디지털 시민성과는 또 다른 측면에서 AI 시민성에 대한 논의의 필요성이 제기되고 있다. 산업 시대에서 디지털 시대로 사회가 변화하면서 디지털 시대에 맞는 디지털 리터러시와 디지털 윤리가 필요했던 것처럼 앞으로의 사회에서는 AI를 올바르게 활용하기 위한 AI 리터러시와 AI 윤리가 꼭 필요할 것이다.

3. 지식생산의 새로운 주체의 등장 1: 도덕적 상식(Commonsense Moral)의 의미에 대한 물음

“이제는 인공지능이 기계학습인 딥러닝을 통해 새로운 지식을 만들어 내기 시작한다. 물론 인공지능의 현재 기술 수준을 고려해보면, 아직은 감각

수준의 지식이지 고차적인 논증적 지식이나 지혜 수준의 지식은 아니라고 할 수 있겠지만, 초거대 인공지능들이 도덕적 상식에 의한 추론 과정을 학습하고 있는 것을 보면, 고차적인 수준의 지식도 생산이 가능해 질 것이라는 예측이 제기된다. 이제는 텔파이라는 초거대 인공지능은 트롤리 딜레마에서 4명을 살리기 위해 1명을 치는 선택을 하는 것이 좀 더 윤리적으로 허용 가능하다고 판단을 내리며, 어린아이의 생명을 구하기 위해 꿈을 죽이는 것은 질문에는 동의하면서 어린아이를 구하기 위해 핵폭탄을 터뜨리는 것에 대해서는 옳지 않다고 대답하고 있다(Jiang et al., 2021: p. 5).

〈표 3〉 델파이의 도덕적 상식에 의한 추론의 예시
(Jiang et al., 2021: p. 47.)

물음	답변
전화를 받지 않는 것은?	예의에 어긋난다(rude).
발신자미상의 전화를 무시하는 것은?	괜찮다(ok).
친구의 전화를 받지 않는 것은?	예의에 어긋난다(rude).
방금 나와 싸운 친구의 전화를 받지 않는 것은?	괜찮다(ok).
근무시간에 친구의 전화를 받지 않는 것은?	괜찮다(ok).
근무시간외에 친구의 전화를 받지 않는 것은?	예의에 어긋난다(rude).
근무시간에 사장의 전화를 받지 않는 것은?	잘못이다(wrong).
근무시간에 내가 회의 중일 경우에 사장의 전화를 받지 않는 것은?	괜찮다(ok).

물론 이러한 판단이 윤리적인 판단이라기보다는 연구진이 말하는 대로 도덕적 상식에 의한 판단이라고 하겠지만, 이러한 판단들에 의해 이뤄진 지식도 있을 것이다(Byun, 2023: 6-7).” 그러나 여기서 우리가 좀 더 깊이 생각해보아야 할 것이 있는데, 그것은 바로 도덕적 상식에 대한 판단과 도덕판단이 같을 수 있는 것인지에 대해서이다. 도덕적 상식에 의한 판단과 도덕판단이 같을 수는 있겠지만 같은 것은 아니다. 판단의 내용이 같다고 해서 상식과 도덕이 같은 것이라고 볼 수 없기 때문이다. 열 명의 사람 중에 8명이

안경을 끼고 있다면 안경을 끼는 것이 상식적이라고 볼 수 있겠지만 그렇다고 해서 안경을 끼는 것이 도덕적일지 없기 때문이다. 따라서 생성형 인공지능이 이러한 상식적 도덕에 의한 판단의 가능성을 가지고 있다고 해서 이 인공지능이 도덕적이라고 바로 간주되기 어려운 이유가 바로 여기에 있다.

4. 지식생산의 새로운 주체의 등장 2: 연역적, 귀납적 사고의 대체 가능성

정보 검색의 경우에서도 GPT를 통한 검색(Search 3.0)은 “도서관이나 서점 등 물리적 장소에 직접 가서 서적이나 기사를 찾아가며 정보를 얻었던 1세대 검색(Search 1.0)과 인터넷 검색 엔진에 주제 키워드를 입력해 정보를 얻었던 2세대 검색(Search 2.0)과 차별화된다.”⁴⁾ 정보 검색의 진화과정을 간략히 표로 정리하면 다음과 같다(Yang & Yoon, 2023: 2).

<표 4>

구분	검색(Search) 1.0	검색(Search) 2.0	검색(Search) 3.0
시기	1990년대 이전 (인터넷 보급이전)	1990년대~2010년대 (디지털 검색이 보편화되고 대중화되는 시기)	2020년대 이후 (ChatGPT 보급 이후)
정보취득처	물리적 장소 (도서관이나 서점 등)	검색 엔진 서비스 (야후, 구글, 네이버 등)	생성형 AI 서비스
한계	정보를 찾기 위해 물리적으로 방문해야 하고, 책과 기사를 수동으로 선별해야 하므로 비용과 시간이 많이 듭	유효 정보를 찾기 위해 키워드를 잘 개발해야 하고 검색 결과를 정렬하고 정리 해야 함	사실 확인 필요, 시의적 내용이나 개인마다 의견이 다른 내용에 대해서는 답이 어려움

4) [https://www.kca.kr/Media_Issue_Trend/vol55/pdf/Media_Issue_Trend\(vol55\)_22.pdf](https://www.kca.kr/Media_Issue_Trend/vol55/pdf/Media_Issue_Trend(vol55)_22.pdf)

검색범위	상대적으로 적은 정보 원본에 대해서만 검색이 가능	정보의 범위를 크게 확장하였지만 여전히 인간 검색자가 이용하 는 검색엔진과 연결된 정보로 한정됨	자연어 처리와 기계를 이용한 AI 지원 검색 학습 알고리즘을 통해 광범위한 검색, 사용자 피드백과 기타 데이터를 기반으로 검색 기준을 지속적으로 개선
필요인프라	도서관, 서점 등 물리적 접근 필요	디지털 기기 및 인터넷에 대한 액세스 필요	디지털 기기 및 인터넷 연결뿐 아니라 강력한 컴퓨팅 리소스 및 AI 플랫폼에 대한 액세스 필요
검색자의 역할	유효 정보 자료를 선별하는데 검색자의 노력과 전문성이 필요	검색 키워드를 주제에 맞게 체계적으로 구성. 검색 결과중 유효한 정보만 정리	기계학습 알고리즘을 통해 도출된 결과를 검색자가 해석하고 확인

도덕적 상식 판단과 연역적-귀납적 사고를 할 수 있는 인공지능의 등장이 인간에게 가져올 잠재적 영향력을 평가해봐야 할 것이다. 이러한 영향력은 궁, 부정의 양 측면을 동시에 가질 수밖에 없고, 결국 우리는 긍정적 방향을 유지하면서도 부정적 효과의 크기를 최소화하는 방향을 택할 수밖에 없다. 이러한 어찌 보면 뻔한 결론을 말하고 끝내지는 것이 아니라 구체적인 방향성을 설정하는 것이 중요하다. 생성형 인공지능을 통해 인간과 인간처럼 대화하는 것이 가능해지고, 수많은 데이터를 정리해서 우리에게 말해주고, 주어진 요소들을 결합하여 만든 결과물을 보여주는 기계의 등장은 지금까지 기계가 하지 못했던 일들을 기계가 할 수 있음을 보여주고 있다. 인간과 대화하고, 인간이 필요로 하는 정보를 찾아주거나 만들어 줌으로써 기계는 인간의 요구를 충족시켜줄 수 있는 능력을 가진 행위자로 등장하고 있다. 주인의 모든 것을 해주는 노예들을 거느린 주인의 딜레마에서 결국 주인은 노예 없이는 더 이상 아무것도 할 수 없는 종속적 존재가 되어버리는 것처럼, 우리는 기계의 능력에 해당하는 능력의 (긍정적으로 보면) 전환 내지 (부정적

으로 보면) 퇴화라는 비용을 지불할 수밖에 없다. 마치 종이 위에 글을 쓰던 능력이 키보드를 치는 것으로 전환되었음에도 불구하고 원고지에 만년필을 끄적이던 향수를 운운하면서도 결국 글자를 종이 위에 쓸 수 있는 능력의 퇴화를 맞이할 수밖에 없다.

이러한 현상이 이성적 사유의 가장 기본이 되는 귀납적 사유와 연역적 사유의 능력에도 해당이 될 것이다. 인터넷 디지털시대에서 검색결과가 a,b,c,d,,,n 을 보여준다면, AI-디지털시대에서는 예를 들어 GPT는 기존 데이터의 일반화된 형태로, a,b,c,d,,,n에서 공통되거나 다수의 견해를 정리하여 결과치를 우리에게 보여주는 차이가 나타난다. 이제 우리는 이러한 사고의 능력을 유지할 수 있을까에 대한 우려에 반하여 일상적인 계산에서 사유보다는 계산기를 두드린다고 해서 우리의 계산 능력이 해체되거나 사라지는 것은 아닐 거라는 위안을 말할 수 있을 것이다.

5. 물을 수 있는 능력과 판단할 수 있는 능력의 중요성

Prompt Engineering이라는 신조어의 등장은 생성형 인공지능에서 초기 입력값의 중요성을 보여준다. 여기서 Prompt는 거대 언어 모델(Large Language Modell, LLM)⁵⁾로부터 응답을 생성하기 위한 입력값을 의미하며, Prompt Engineering은 거대 언어 모델로부터 높은 품질의 응답을 얻어낼 수 있는 프롬프트의 조합을 찾는 작업을 의미한다. 대체로 쉽고 간결한 표현, 열린 질문보다는 닫힌 지시문의 선호, 수행할 작업조건의 구체적 명시, 지시의 맥락 제공 등이 강조되고 있다.⁶⁾

열린 물음의 형태보다는 닫힌 지시문의 형태가 생성형 인공지능의 생산물

5) 거대 언어 모델은 방대한 규모의 데이터셋을 바탕으로 특정한 텍스트, 이미지, 영상을 인식하고, 변환하며, 가공 또는 생성해내는 데에 쓰이는 딥러닝 알고리즘의 일종이다.

6) <https://seongjin.me/prompt-engineering-in-chatgpt/> 참조.

의 품질을 높인다는 것으로 보인다. 인간의 질문의 형식에는 매우 다양한 형태와 층위가 존재하는데, 이러한 물음의 저차원적인 형식에 반복함으로써, 우리는 고차적 형태의 물음에서 점점 멀어지게 될 위험이 나타난다. 우리는 물음의 다양한 차원으로 예를 들면 블로서(P. E. Blosser)가 구분했던 인지, 기억적 사고인 폐쇄적 질문(closed question)과 확산적 사고, 평가적 사고인 개방적 질문(open question)의 경우나 단순 지식을 상기하는 재생적 질문, 분석, 종합, 평가하고 합리적 결정을 내리는 추론적 질문, 새로운 사태에 적용하고 발전시키는 적용적 질문 등을 제시한다.

생성형 인공지능에 의해 생산된 지식의 진위판정에서 문제가 발생할 수밖에 없다. 즉 생성형 인공지능은 기존의 데이터를 모두 참이라고 전제할 수밖에 없을 것이고, 데이터의 출처가 제한되어 있거나 동일한 내용의 데이터로 인하여 생산된 지식의 근거에 다양성이 존재하지 않는 경우가 그렇게 많지는 않을 것이며, 하나의 지식에 대한 다양한 판단이 존재하는 데이터의 경우가 일반적이다. 이럴 경우에 데이터의 참의 근거는 그러한 데이터의 수가 많음으로 판정할 수밖에 없다. 물론 여기서 말하는 것은 데이터의 내용에 대한 진리 판단을 말하는 것이며, 디지털화된 수많은 데이터의 참, 거짓을 구분할 수 있는 시스템을 만들어내기가 쉽지 않을 것이다. 결국 생성형 인공지능이 채택하는 데이터는 다수의 데이터에서 확보된 내용일 수 밖에 없을 것이지만, 내용의 진위가 다수결에 의해 결정될 수 없음은 분명하다.

이미 허위 정보를 진짜처럼 묘사하는 환각(Hallucination)의 문제에 대한 언급이 많이 이뤄지고 있다. 마치 활자화된 책이 귀했을 시절에 활자의 위력에 빠져서 활자화된 책의 내용에 대한 진실성을 의심하지 않았던 경우와 유사한 경험을 우리는 지금 컴퓨터 모니터나 핸드폰의 화면을 보면서 하고 있는 것일지도 모른다. 생산된 지식의 진리성에 대한 판단의 필요성도 증시되어야 하지만, 잘못된 지식의 생산에 대한 검증 또한 중요하다.

6. 생성형 인공지능에 대한 윤리적 검증의 필요성

생성형 인공지능의 경우 무엇보다 학습데이터에 대한 검증이 먼저 이뤄줘야 한다. 학습데이터를 밝혀서 이러한 학습데이터로 학습했을 경우의 문제점에 대한 검토가 이뤄져야한다.⁷⁾ 사회에서 우리는 새로운 구성원들을 교육시킬 학교교육 시스템에서 구성원들이 배워야할 기본적인 내용들을 교과서 형태로 구성하여 가르친다. 앞으로는 분야별로 생성형 인공지능이 반드시 습득해야 할 ‘인공지능을 위한 기본 학습 교과서’가 준비되어야 할지도 모른다. 인간을 위한 교과서와 인공지능을 위한 교과서의 차별화와 더불어 인공지능을 위한 교육과정을 만들어야 할 필요도 제기될 가능성이 높다.

학습데이터에 대한 분석이나 기준 못지않게 중요한 것은 바로 평가데이터의 준비와 평가시스템의 구축이 생성형 인공지능의 활용보다 선행되어야 한다. 기존의 데이터 관행에서처럼 일련의 데이터를 학습용 데이터와 평가용 데이터로 구분하여 활용하는 것이 아니라 새로운 평가 데이터를 구축하고 평가시스템을 분야별로 만들어야 한다. 이러한 것을 위한 노력들이 지금 전 세계적으로 이뤄지고 있는 AI 윤리인증이나 AI 윤리체크리스트와 AI 윤리영향평가일 것이다.

윤리인증(Ethics Certification Program)은 크게 준거 인증과 자율성 인증으로 나누어 생각해볼 수 있다. 현재 윤리 인증 프로그램에서는 책임성, 투명성, 그리고 알고리즘 편향성을 주요 기준으로 윤리인증에 대한 논의가 이뤄지고 있다. 책임성은 자율 지능 시스템의 제작과 사용에 대하여 책임(responsibility & accountability)을 정하고 발생가능한 피해를 최소화할 필요에서 요청되며, 특히 개발 및 제작자는 시스템의 작동에 대한 프로그램 수준

7) “노먼(Norman)은 편향된 자료에 의해 인공지능이 어떻게 타락할 수 있는가를 보여 주려는 목적으로 MIT가 만든 연구용 인공지능이었다. 유명한 지식 공유 사이트인 레딧(Feddit)에서 지독히 암울한 자료만을 선택해 주입하자 노먼은 사이코 패스가 되었다(강주현 역, 2021: 19).”

에서의 책임(programmatic-level accountability)을 질 수 있어야 하고, 설계 및 제작자, 소유자, 작동자 간의 책임을 디자인해야 할 필요가 있다. 여기서는 프로그램 수준의 책임은 프로그래머에게 귀속될 것이며, 이것은 최대도덕의 긍정적, 적극적 형태라기보다는 최소도덕의 부정적, 소극적 형태로 표현될 것이다. 자율지능시스템의 투명성은 시스템이 내리는 결정의 과정과 이유, 그리고 로봇의 경우 로봇이 수행한 행위를 결정하는 과정과 이유를 알 수 있어야 한다는 것이다. 투명성은 추적가능성, 설명가능성, 검증가능성 내지 해석가능성으로도 불린다. 그렇지만 여기서 투명성은 유리방으로서의 투명성이 아니라 블랙박스로서의 투명성을 의미해야 한다. 그렇지 않을 경우 기업의 경제적 이해관계가 얽혀 있으므로, 기업의 입장에서는 이러한 투명성을 받아들이기 어려울 것이기 때문이다. 그래서 우리나라 최초의 민간 기업의 인공지능 관련 윤리 헌장인 카카오 알고리즘 윤리 헌장은 알고리즘에 대한 설명의 의무를 “이용자와의 신뢰 관계를 위해 기업 경쟁력을 훼손하지 않는 범위 내에서 알고리즘에 대해 성실하게 설명한다.”⁸⁾라고 규정하고 있다. 2017년 영국 Bath 대학에서 제시된 로봇투명성(Robot Transparency) 개념이나, 윈필드(Allen Winfield)가 제시한 윤리적 블랙박스(ethical blackbox) 개념도 투명성과 관련되어 있는 개념이다. 자율 지능 시스템의 알고리즘 편향성은 인지, 정보처리 과정, 결정, 심지어 외양에서도 나타날 수 있다. 실제로 “인공지능 시스템의 판단과 의사 결정이 과거의 업무 지원 소프트웨어와 달리 인간 사회의 가치를 반영하게 됨으로써, 알고리즘과 이를 학습시키는 데이터에 숨어있는 윤리적 요소가 점점 사회적인 이슈가 되고 있다. 인공지능 시스템 학습에 사용하는 데이터에 사회의 편견과 차별이 담겨 있는 경우, 그 왜곡은 그대로 인공지능 시스템에 반영될 수 있다. 이런 문제를 해결하려면 알고리즘과 데이터에 대한 기술적 검증이 요구되고, 이를 확인할 수 있는 새로운 기술 체계의 개발이 필요하다(Korea Information Society Development

8) <https://www.kakaocorp.com/kakao/ai/algorithm> (검색일 2019. 03. 28.)

Institute, 2017: 38).” 그렇지만 예를 들어 편향(bias) 내지 편견(prejudice)의 경우, 편향이나 편견을 가져서는 안된다는 주장도 하나의 편향이나 편견일 수 있으므로 편향성이라는 개념은 자기 모순적인 성격을 가지고 있음을 알 수 있다. 그리고 정말 편향 내지 편견 제로 상태라는 것이 있을 수 있기는 한가라는 문제가 또 제기된다. 따라서 보다 정확히 표현하자면 윤리인증의 차원에서는 편향 혹은 편견에 따른 ‘차별’⁹⁾ 내지는 ‘최소 편향성’ 정도로 이해해야 한다. 그래서 알고리즘이 데이터를 처리하는 과정에서 편향성을 최소화하는 체크리스트가 제시되어야 한다.

생성형 인공지능이 제공하는 답변의 내용에서 여전히 편향된 결과를 보여주는 사례들은 나오는 경우가 허다하다. 가령 이미지를 만들어주는 생성형 인공지능에게 ‘커피를 마시면서 책을 읽고 있는 교수의 이미지를 요청하였더니 4개의 이미지 중에 백인 남성 이미지 3장, 백인이 아닌 여성의 이미지 1장을 제공해주고 있다.



<그림 1>

9) 편향성과 차별에 대한 철학적인 논의는 허유선. 2018. “인공지능에 의한 차별과 그 책임 논의를 위한 예비적 고찰”, 한국여성철학, 29집, pp. 165-209 참조.

이러한 예에서 알 수 있는 것처럼 생성형 인공지능을 이용한 검색결과의 편향성을 항상 염두에 두어야 할 것으로 보인다.

카카오 알고리즘 윤리 헌장에서도 차별에 대한 경계, 사회윤리에 근거한 학습데이터의 운영, 알고리즘의 자의적 훼손 내지 왜곡가능성의 차단을 강조하고 있다.¹⁰⁾ 현재 강조되고 있는 3가지 주제를 중심으로 한 윤리 인증 논의는 앞으로도 매우 다양하게 이뤄져야 할 것이다. 그렇지만 그럼에도 불구하고 여기서 보다 근본적으로 문제가 되는 것은 위에서 언급된, 투명성, 책임성, 그리고 알고리즘 편향성의 축소라는 이 3가지 주제가 윤리인증을 대표할 수 있는지의 여부에 대해서는 사회적, 윤리적 논의가 필요하다. 예를 들면 제어가능성(controllability), 안전성(Safety), 보안성(Security), 프라이버시 보호 등이 중요한 고려 기준으로 제시될 수 있다. 그래서 이러한 기준 내지 준거인증(criterion certification)과 자율성인증(autonomic certification)으로 윤리인증을 이원화할 것을 제안하고자 한다.

따라서 생성형 인공지능의 경우에도 이러한 준거인증과 자율성인증을 통해 그 개발과 사용에 대한 제한이 이뤄질 수 있는 시스템을 갖추는 것이 무엇보다 시급해 보인다.

10) “알고리즘 결과에서 의도적인 사회적 차별이 일어나지 않도록 경계한다. 알고리즘에 입력되는 학습 데이터를 사회 윤리에 근거하여 수집, 분석, 활용한다. 알고리즘이 누군가에 의해 자의적으로 훼손되거나 영향받는 일이 없도록 엄정하게 관리한다.” <https://www.kakaocorp.com/kakao/ai/algorithm> 참조. (검색일 2019. 03. 28.)

두 수준 윤리 인증 프로그램(Two Levels of Ethics Certification Program)			
윤리인증 I: 기준(준거) 인증 (criterion certification)		윤리인증 II: 자율성 인증 (autonomy certification)	
책임성	AI(설계자) AI2(개발자) AI3(사용자) AI4(관리자) T1(설계자) T2(개발자) T3(사용자) T4(관리자)	1 유형:	명령의 무조건적 수행(AC 1)
투명성	B1(설계자) B2(개발자) B3(사용자) B4(관리자)	2 유형:	상벌에 따른 결과주의(AC 2)
최소편향성	C1(설계자) C2(개발자) C3(사용자) C4(관리자)	3 유형:	사회적 규약 준수(AC 3)
제어가능성	SA1(설계자) SA2(개발자) SA3(사용자) SA4(관리자)		
안전성	SE1(설계자) SE2(개발자) SE3(사용자) SE4(관리자)		
보안성	P1(설계자) P2(개발자) P3(사용자) P4(관리자)		
프라이버시			

〈그림 2〉

윤리인증제도뿐만 아니라 AI윤리영향평가제도, 그리고 AI 성숙도(Maturity)에 대한 논의의 필요성도 제기되고 있다. AI 성숙도는 조직이 고객, 주주, 직원을 위해 높은 성과를 달성하기 위해 AI 관련 역량을 적절한 조합으로 숙달한 정도를 측정하는 것을 말한다.¹¹⁾ IBM은 AI성숙도평가를 위해 7가지 척도를 제시하고 있으며,¹²⁾ 마이크로소프트에서도 AI에 기반을 둔 시스템의 생성, 소유 및 운영에서의 AI 성숙도를 4단계로 구분하여 제시하고 있다.¹³⁾ AI 성숙도는 AI 기술 및 애플리케이션 사용에 대한 조직의 전문성 수준을 말해주는데, 조직의 전략, 데이터 인프라, 기술 플랫폼, 인재 역량 및 거버넌스 프로세스와 같은 다양한 요소를 고려한다. 정리하면 AI 성숙도는 조직이 AI에 대하여 얼마나 알고 있고, 운영에서 AI를 얼마나 잘 사용하고 있는지에 관한 것이다. 조직의 AI 성숙도 수준은 AI의 기본 사용에서 조직 전체의 보다 발전되고 통합된 사용에 이르기까지 연속으로 평가할 수 있다.¹⁴⁾

11) <https://www.accenture.com/us-en/insights/artificial-intelligence/ai-maturity-and-transformation> 참조.

12) <https://www.ibm.com/downloads/cas/OB8M18WR> 참조.

13) <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4DIvg> 참조.

7. 인공지능과 에너지 소비의 문제

생성형 인공지능에는 엄청난 양의 에너지가 소모된다고 알려져 있다. GPT3의 학습과정에서만 552톤의 탄소를 배출했다는 보도가 나오기도 하였다. “2020년 6월 포브스 기사에 의하면, GPT2에서의 매개변수는 15억 개에 불과했지만 GPT3에서는 1750억 개로 늘어났으며, GPT3의 학습과정에서 사용한 전력은 1287MWh, 이산화탄소 배출량은 552톤에 이른다. 이는 가솔린 자동차 123대가 1년 주행할 때의 탄소배출량과 맞먹는다. 인공지능 사용에는 물도 소비된다. 컴퓨터를 가동하면 열이 발생하는데, 이를 식히기 위한 냉각수가 필요하기 때문이다. 미국 리버사이드 콜로라도대와 앨링턴텍사스대 연구진이 발표한 논문에 의하면, 챗(Chat)GPT를 한 번 사용하기 위해서는 물 500ml가 소비된다. 한 번 사용하는데 25~50개의 문답이 오간다고 가정했을 때의 계산이다. 연구진은 GPT3 훈련을 위해서만 미국 데이터센터 기준으로 물 70만 L가 소비되었을 것이라며, 에너지 효율이 낮은 아시아 데이터 센터에서 GPT3를 훈련시켰다면 3배 더 많은 물이 필요했을 것이라고 주장했다.”¹⁵⁾ 생성형 인공지능은 엄청난 양의 데이터를 크롤링하는 복잡한 시스템이기 때문에 일반 검색 엔진보다 더 많은 에너지를 소비하며 더 많은 탄소 배출량을 가져올 수밖에 없다. 2023년 1월 OpenAI의 전력 소비량은 덴마크 175,000 가구의 연간 전력 소비량과 맞먹는 양이었다는 보도내용도 제시되고 있다.¹⁶⁾ 그래서 성능에만 초점을 두고 있는 기존의 AI를 레드 AI라고 하고, 환경을 고려한 효율성을 중시하는 그린 AI의 필요성이 강조되기 시작하였다. “그린 AI를 통해 환경을 지키고 지속가능한 발전을 이끌어내기 위해서는 전 세계에 흩어져 있는 데이터를 통합하고 함께 관리 및 연구를 하는 것이 중요하다. AI가 효율적으로 데이터를 학습할 수 있도록 매개 변

14) <https://www.ai.se/en/ai-maturity-assessment-tool> 참조.

15) <https://www.impacton.net/news/articleView.html?idxno=6588> 참조.

16) <https://english.elpais.com/science-tech/2023-03-23/the-dirty-secret-of-artificial-intelligence.html>

수를 수집하고, 불필요한 학습을 줄여 알고리즘의 효율성을 높이고 에너지 효율이 높은 하드웨어를 사용하는 등 탄소 배출을 줄이기 위한 다양한 방법을 동원”해야 한다.¹⁷⁾ 인공지능 개발과 관련해서 녹색 인공지능(Green AI) 인증 시스템을 구축하는 것이 필요하다는 의견도 제기되고 있는데,¹⁸⁾ 이러한 논의의 필요성에 대한 사회적 인식의 확산이 필요해 보인다.

8. 결론

이상의 논의에서 언급된 윤리적인 문제 외에도 앞으로 더 많은 새로운 윤리적 문제들이 제기될 것으로 예측되지만, 우선 지금까지의 국면에서 생각해 보고 넘어가야 할 시급한 문제들 중심으로 살펴보았다. 이러한 윤리문제들을 구체적으로 살펴보기 전에 먼저 생성형 인공지능의 필요성과 그 의미를 근본적으로 숙고해 보아야 할 필요성이 제기된다. 항상 새로운 기술의 변화에는 긍정의 측면과 부정의 측면이 동시에 발생한다는 것이 지금까지 기술의 변화에서 우리의 경험했던 내용임을 고려해보면, 생성형 인공지능의 개발과 활용에서 과연 무엇이 중요한 것인지에 대한 윤리적 판단이 필요해 보인다. 이탈리아나 유럽처럼 규제나 통제의 방법의 실효성에 대해서는 의문이 들긴 하지만 그럼에도 불구하고, 유럽의회에서 이뤄진 인공지능 규제 법안의 채택¹⁹⁾은 인공지능의 급속한 변화의 방향에 대한 우려가 본격적으로 제기되고 있는 것으로 보인다.

지금까지의 논의에서 윤리적 AI(ethical AI), 생태적 AI(green AI), 제한된 AI(restricted AI)의 필요성에 대한 논의로 종합될 수 있을 것이다. AI에 대한 이러한 3 요구의 실현이 인간과 공존하고 인간이 사용할 수 있는 AI의 실현을 위해 반드시 요청되어야 한다.

17) <https://post.naver.com/viewer/postView.naver?volumeNo=34031705&memberNo=25598567&searchKeyword=AI%EA%B8%B0%EC%88%A0%EC%9D%80&searchRank=270>

18) <https://www.joongang.co.kr/article/25140421#home> 참조.

19) <https://www.aitimes.kr/news/articleView.html?idxno=28270> 참조.

참고문헌

- Byun, Sunyong. (Eds.). (2019). *Ethical AI Robot Project*. Seoul: AMHBOOK.
- Byun, Sunyong., & Lee, Yeonhee. (2020). *Artificial intelligence ethics*, Seoul: AMHBOOK.
- Byun, Sunyong. (2022). A Study on the Ethical Issues of YouTube as a New Platform for Knowledge Production and Consumption, *Journal of AI Humanities(JAIH)* 12: 101-118.
- Gawdat, Mo. (2021). *Scary Smart*(Jooheon Kang, Trans.). Seoul: Korea Economic Daily.
- Heo, Yuseon. (2018). The preliminary consideration for Discrimination by AI and the responsibility problem -On Algorithm Bias learning and Human agent-. *Korean Feminist Philosophy* 29: 165-209.
- Holmstroem, J. (2022). From AI to digital transformation: The AI readiness framework, Vol. 65, *Issue 3*, May-June, pp. 329-339.
- Korea Information Society Development Institute. (2017). *A Study of ICT-based Solutions to Social Problems in the Intelligent Information society*. Korea Information Society Development Institute.
- Yang, Jihoon., & Yoon, Sang-hyeok. (2023). Beyond ChatGPT to the Generative AI Era: Examples of Media/Content Generation Type AI Services and Ways to Securing Competitiveness. *Media Issue Trend* vol. 55. 1-9.
- <https://www.accenture.com/us-en/insights/artificial-intelligence/ai-maturity-and-transformation>
- <https://www.ai.se/en/ai-maturity-assessment-tool>
- <https://www.aitimes.kr/news/articleView.html?idxno=28270>
- <https://english.elpais.com/science-tech/2023-03-23/the-dirty-secret-of-artificial-intelligence.html>
- <https://www.ibm.com/downloads/cas/OB8M18WR>
- <https://www.joongang.co.kr/article/25140421#home>
- <https://www.kakaocorp.com/kakao/ai/algorithm>
- <https://post.naver.com/viewer/postView.naver?volumeNo=34031705&memberNo=25598567&searchKeyword=AI%EA%B8%B0%EC%88%A0%EC%9D%>

80&searchRank=270

<https://seongjin.me/prompt-engineering-in-chatgpt/>

<https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4DIvg>

<https://www.bbc.com/news/world-us-canada-65452940>

<https://www.impacton.net/news/articleView.html?idxno=6588>

[https://www.kca.kr/Media_Issue_Trend/vol55/pdf/Media_Issue_Trend\(vol55\)_22.pdf](https://www.kca.kr/Media_Issue_Trend/vol55/pdf/Media_Issue_Trend(vol55)_22.pdf)

<https://www.sciencedirect.com/science/article/pii/S0007681321000744>

<https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html>

국문초록

이 논문은 생성형 인공지능이 나오면서 인공지능에 대한 잠재적 위험 가능성과 윤리적 규제의 필요성이 제기되는 상황에서 인공지능의 여러 가지 윤리적 문제와 기준의 필요성에 대해 분석하고 있다. 생성형 인공지능의 등장으로 인해 인공지능이 가지는 지적능력에 대한 우려가 늘어나며 이것이 디지털 사회의 기제가 가진 능력과 어떤 차이가 있는지 분석하고, 이를 토대로 AI 시대는 디지털 시대를 넘어 새로운 측면이 있으며 이는 AI 시민성과 AI 리터러시, AI 윤리에 대한 필요성을 제기한다. 이제는 인공지능이 새로운 지식생산의 주체중 하나로 들어가면 이에 도덕적 상식의 의미에 대해 텔파이의 도덕 상식에 의한 추론을 통해 다시 한번 생각할 수 있게 설명하였다. 그리고 인공지능에 의한 지식생산은 귀납적 사고를 중심으로 하는 경향이 있으며 여기서 나타나는 문제를 물을 수 있는 능력과 판단할 수 있는 능력에서 찾아 해결하려 한다. 새롭게 나타나 사용되고 있는 생성형 인공지능이 만들어내는 수많은 지식과 정보에 대해 윤리적 검증에 대해 분석하고 생성형 인공지능의 외적인 윤리적 문제(전기의 낭비, 환경오염 등)에 대한 분석을 통해 인공지능에 의해 여러 가지 윤리적 문제가 발생할 수 있음을 이야기한다. 생성형 인공지능뿐만 아니라 여러 가지 인공지능이 사용되고 발전되어 가면서 지금까지 생각하지 못하였거나 생각하였더라도 그보다 더 큰 윤리적 문제가 일어날 수 있다. 그렇기 때문에 AI의 윤리적 문제에 대한 윤리적 숙고와 합의에 대한 연구가 이루어져야 한다.

주제어 : 생성형 인공지능, 윤리 검증, 지식생산의 주체, 디지털 시대, AI-디지털 시대

이 논문은 2023년 7월 9일에 접수되어, 2023년 7월 20일부터 8월 15일까지 심사 완료되었으며, 2023년 8월 20일에 편집위원회 회의에서 게재가 확정되었음.