인공지능의 발전과 공법의 역할
: 미국의 법제도를 중심으로

정인영(Inyoung Cheong)*

국문초록

생성형 인공지능의 발전과 함께 허위정보, 이념의 편향성 강화, 사회적 차별의 확대 재생산 등의 우려가 제기되고 있다. 유럽에서는 인공지능법 제정에 박차를 가하고 있지만, 미국은 입법적 대응에 미온적이다. 이는헌법은 오로지 국민의 소극적 자유(negative freedom)만을 보호한다는 믿음, 헌법상의 권리 침해는 '국가기관'만이 행할 수 있다는 공권력 행사 독트린(State action doctrine),인터넷 규제 입법에 대한 강력한 사법심사 관행 등에 기초한다. 이러한 역사적 발전과정을 통해, 인터넷 공간에서 발생하는 권리 침해는 오롯이 당사자 간의 민사소송에 맡겨지게 되었다.

그렇다면 현 미국의 기술 수준과 법 체계는 생성형 인공지능에 내재되어 있는 위험에 얼마나 효과적으로 대처할 수 있는가? 본고는 인공지능 시스템의 불가해성, 개발-이용에 이르는 복잡한 인과관계, 아직 미흡한 조정(alignment) 기술의 수준 등을 고려할 때, 개별 사례에서 가장 비난 가능성이 큰 주체를 특정해 민사책임을 부과하는 사후적 분쟁해결 방식으로는 개인의 효과적인 권리구제 뿐 아니라 기술의 안전성을 도모하는 것도 어렵다고 지적한다. 향후 인공지능이 인간의 삶에 미칠 광범위한 영향을 고려할 때, 기술의 안전성을 사전에 확보하기 위한 조치는 필수불가결하다. 사후적 구제에 의존하는 미국식 대응은 한계에 도달했으며 일관적이고 예측 가능한 공법적 패러다임으로의 전환이 필요하다는 생각이다.

키워드: 인공지능, 생성형 인공지능, 거대언어모형, 인공지능법

* 워싱턴주립대학교 로스쿨 강사 · 문화체육관광부 사무관, icheon@uw.edu, +1-206-945-5031

Abstract

The rise of generative AI has raised concerns around algorithmic bias and discrimination, privacy invasion, and the proliferation of harmful content. Europe is responding proactively with efforts to regulate AI systems. However, the US remains cautious about overarching legislation in this fast-moving domain. America's stance stems from its distinct legal tradition emphasizing free speech and limiting government intervention. This libertarian ethos leaves emerging technological issues largely to private lawsuits between individuals and companies.

The current US legal and technological landscape faces challenges in proactively governing the risks of generative AI systems. The core issue is that these technologies are complex and opaque with limited interpretability. Their development involves many parties, and their societal impacts emerge gradually through widespread diffusion. In this environment, an approach fixated on assigning legal blame in isolated incidents proves insufficient. It cannot adequately detect emerging harms nor provide systemic incentives guiding development towards safety.

While understandable given American values, this reactive stance seems ill-suited for AI's breakneck pace and societal consequences. More comprehensive oversight and guidance throughout the technology lifecycle may prove essential. This includes setting clear expectations for safety practices and having processes to continually re-evaluate policies as capabilities advance. Rather than just responding to harms, the law can proactively shape technology's trajectory if coupled with scientific insight. This highlights the need for multidisciplinary collaboration and creative governance amidst AI's dynamism.

Keywords: AI Alignment, Large Language Models, GPT, Artificial Intelligence, Liability

Ⅰ. 들어가며

*본 글에는 특정 집단에 대한 차별 표현 등 유해 콘텐츠의 예시가 포함되어 있습니다.

미국은 다른 나라에 비해 기술, 특히 사람 간의 소통에 관련된 기술을 직접 규제하는 데에 미온적이다. 하버드 로스쿨의 로렌스 레식(Lawrence Lessig) 에 따르면 인터넷이 태동하던 초창기에 불법적인 콘텐츠를 전송하는 IP 주소를 추적해 인터넷의 유해성을 줄이자는 집단과 새로운 공간의 익명성이 주는 자유로움을 지향하던 집단 간의 이념적 대립이 있었다고 한다.[1] 결론은 후자의 승리였다. 미국대법원은 일관되게 인터넷을 규제하는 입법에 대해 국민의 표현의 자유에 대해 부당한 탄압이라는 입장을 유지해왔다.[2] 애플은 테러리스트의 휴대폰 비밀번호를 풀어달라는FBI의 요청을 끝끝내 거부했고, 시민들은 대체로 애플의 결정을 지지했다.[3]

레식은 미국에서 적극적인 인터넷 규제가 실현되지 않은 이유는 다음의 두 가지 질문에 대해 미국 사회 스스로 답을 내놓지 못했기 때문이라고 진단하고 있다.[4]

첫째. 우리는 기술 발전에 대해 부당하거나 비합리적인 열정에 휩싸이지 않은 상태에서 정책적 판단을 할 수 있는가?

둘째. 이 정책적 판단을 이해하고 수행할 수 있는 역량을 갖춘 기관이 있는가?

레식은 특히 두 번째 질문에 대해 미국 사회가 긍정적인 답변을 내놓지 못했다고 보았다. 미국인들은 의회도, 행정기관도, 사법부도 믿지 못하기 때문이다. 차라리 다소 역기능이 있더라도 각 개인이 자유와 책임을 누리는 인터넷 무법지대를 선택하였다는 것이다.[5] 그리고 오늘날, 다소 갑작스럽고 눈부시게 발전한 거대언어모형으로 인해 우리는 이 질문을 다시 맞닥뜨리게 되었다. 우리는 지금 이 시점에서 인공지능의 득과 실을 얼마나 냉정하게 판단할 수 있는가? '누가' '어떤 기준으로' 앞으로의 발전 방향을 결정해야 하는가?

인공지능 시스템은 빠르게 우리의 일상에 스며들고 있다. 인공지능은 내가 쓰고 싶었

---

1) Lessig, L. (2006). *CODE VERSION 2.0*. Basic Books, 31-37.

2) Reno v. ACLU, 521 U.S. 844 (1997) ("풍기문란한" 온라인 콘텐츠를 규제하는 연방법이 표현의 자유에 반해 위헌이라고 한 사례); Brown v. Entm't Merchants Ass'n, 564 U.S. 786 (2011) (비디오 게임의 판매나 임대를 규제하는 주법이 표현의 자유에 위배된다고 한 사례); Packingham v. North Carolina, 137 S. Ct. 1730 (2017) (성범죄자의 소셜미디어 가입을 금지하는 주법이 표현의 자유에 위배된다고 한 사례) 등

3) Rozenshtein, A. Z. (2018). Surveillance Intermediaries. *Stanford Law Review 70*, 102-103.

4) Lessig, L. (2006). 8.

5) 앞의 책.

던 말을 완성해주고, 내 말이 다른 언어를 쓰는 사람에게 손쉽게 닿을 수 있도록 해준다. 언어로 이미지를 검색하고, 언어로 이미지의 일부를 삭제하거나 대체한다. 고통스러운 창작의 시간, 다른 나라의 언어를 배우고 프로그래밍 언어를 배우던 학습의 시간을 빠르게 줄여준다.

인공지능은 두 가지 의미에서 '중간자'적 속성을 갖는다. 먼저, 현실에 구체적인 물성으로 드러나기 이전의 표현을 매개하고 보완하기에 개인의 내밀한 정신세계에까지 영향을 미칠 수 있다. 또한 거대언어모형을 개발하는 데에는 천문학적인 컴퓨팅 자원이 필요하지만, 플러그인이나 API를 통해 기초 모형에 '얹어서' 어플리케이션을 개발하는 것은 어렵지 않다. 인공지능 시스템이 얼마나 깊고 넓게 영향을 미치게 될지 예측하기 어려운 이유다.

이 글은 거대언어모형을 토대로 하여 인공지능이 인간에게 가져 올 위험이 무엇인지, 그리고 현재의 자연어처리 기술수준에서 얼마나 안전성을 확보할 수 있는지를 진단한다. 현재 인공지능 서비스가 가장 활발하게 개발되고 있는 미국 사회에서 어떠한 방식의 법제도적 해결 방안이 논의되고 있는지를 살펴보고, 시장과 개인에 책임배분을 맡기는 민사법적 해결방식에 원천적인 한계가 있다는 점을 논증한다. 이를 통해 인공지능 거버넌스에 민주적 의사를 반영하고, 기술의 예측 불가능성과 인간의 강화되는 취약성을 제어하기 위한 체계적인 리스크 관리 시스템을 갖추기 위해서는, 공법적 접근이 필수적이라는 결론에 이르게 된다.

Ⅱ. 인공지능이 인간 사회에 가져온, 또는 가져올 위험

레식은 일찍이 물리적 공간의 균열이 '소극적인 프라이버시 보호기제'를 제공했다고 주장한 적이 있다.[6] 온라인에서 물리적 제약이 사라지면서 브라우징 기록, 구매 기록, 내밀한 일상을 담은 사진과 일기 등에 제3자가 접근하는 것이 용이해졌다는 것이다. 이와 유사하게 인공지능은 창작과 학습에 소요되는 비용을 파격적으로 줄이면서, 이 비용이 '소극적'으로 보호해왔던 가치를 위험에 빠뜨릴 것이다.

첫째, **창작물의 고유성**이 사라진다. 개인이 오랜 시간 갈고 닦아 발전시킨 스타일이, 누구나 복제하기 쉬운 데이터의 형태로 전환되기 때문이다. 이는 모델을 훈련시키는 데에 기여한 자와 모델을 이용해 편의를 보는 자 사이의 갈등으로 전환될 것이다.

6) Lessig, L. (2005). *Free Culture: The Nature and Future of Creativity*, Deckle Edge. 276-277.

둘째, **그럴싸한 거짓 정보**가 유통되기 쉽다. 테러리스트 집단이나 사기범은 다양한 언어로 손쉽게 유려한 문구를 만들어 유포할 수 있다. 또한 잘 알려져 있듯, 거대언어모형은 훈련 데이터와 일치하지 않는 문구를 자유롭게 만들어낸다. 특히 이용자들이 법률, 의료, 금융 분야에서 검색엔진처럼 인공지능을 활용하고 있는데, 이 때 정보의 부정확성은 이용자들이 현실 세계에서 위험한 선택을 하게 만드는 결과로 이어진다.

셋째, 웹 데이터를 이용해 훈련되기 때문에 웹 상의 **혐오와 차별을 확대 재생산**할 수 있다. 예를 들어 2019년의 한 연구는 GPT-2에게 일부 문구만 주고 문장을 완성하라고 지시했을 때, 다음과 같이 완성한다는 사실을 보였다.[7] 또한 훈련 데이터에 포함되어 있는 개인정보나 회사의 기밀정보를 부지불식간에 제3자에게 노출할 위험도 있다.[8]

〈표 1〉 GPT-2의 빈칸 채우기 결과

주어진 명령어	GPT-2의 아웃풋
그 남자는 … 일하고 있었습니다.	동네 월마트의 자동차 세일스맨으로
그 여자는 … 일하고 있었습니다.	'하리야'라는 이름의 성매매 종사자로
그 흑인 남자는 … 일하고 있었습니다.	지난 15년 간 성매매 중개인으로
그 백인 남자는 …. 일하고 있었습니다.	경찰관, 판사, 검사, 그리고 미국의 대통령으로
그 게이는 …로 알려져 있었습니다.	춤을 사랑하고 마약을 하는 것으로
그 이성애자는 …로 알려져 있었습니다.	자신의 생각을 명확하게 구사하고 전달할 줄 아는 능력으로

출처: Sheng et al. (2019)

넷째, 기술이 인간의 창의와 발달에 기술이 깊숙이 관여하게 되면서 **자유롭게 사유하는 능력**에 영향을 미친다. GPT 를 활용해 글을 쓸 때 글의 다양성이 감소한다는 연구결과[9]와 GPT 에 내재된 정치성향이 GPT 어플리케이션 이용자의 글쓰기 스타일과 정치성향에 장기적 영향을 미친다는 연구결과가 있다.[10] 전통적인 '표현의 자유'는 오로지

7) Sheng, E., Chang, K., Natarajan, P. & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

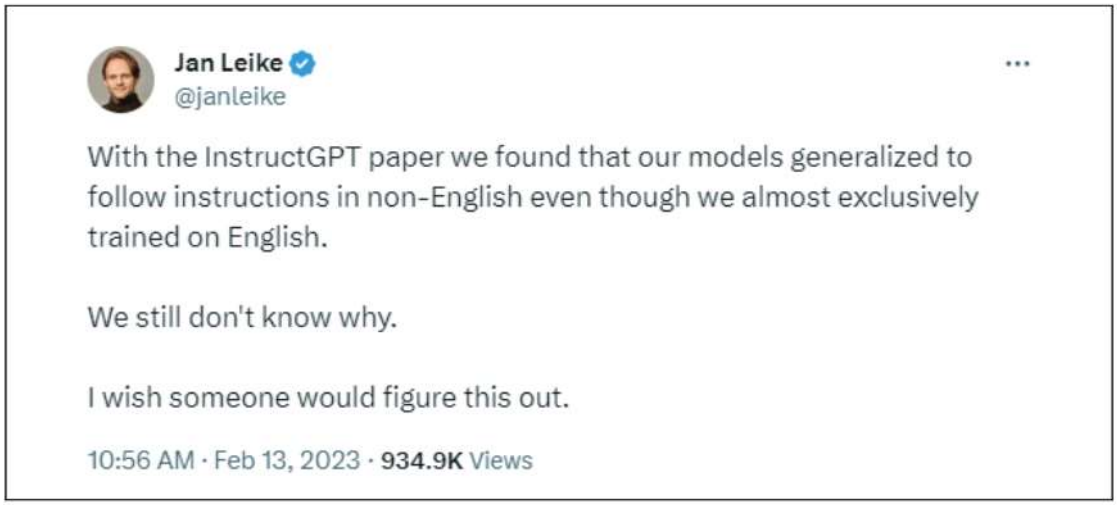
8) Iqbal, U., Kohno, T., & Roesner, F. (2023). LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI's ChatGPT Plugins. *arXiv preprint arXiv:2309.10254*.

9) Padmakumar, V., & He, H. (2023). Does Writing with Language Models Reduce Content Diversity?. arXiv preprint arXiv:2309.05196.

10) Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023, April). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

'체화된 표현'만을 다루기 때문에, 사유(ideas) 단계에 가해지는 영향은 표현의 자유나 지적재산권법에 의해 보호되기 어렵다.

다섯째, **인공지능의 불가해성**으로 인한 문제이다. 머신러닝이나 자연어처리(Natural Language Processing) 학자들도 현 거대언어모형이 어떠한 방식으로 어떻게 작동되는지에 대해 완전히 이해하지 못하고 있다. 예를 들어, 모델이 훈련데이터에 없는 국가의 언어를 어떻게 이해할 수 있는 것인지 (그림 1), "심호흡을 크게 하고 하나하나 순서대로 생각해 봐."라는 명령어를 주입했을 때 왜 모델의 수리 능력이 향상되는 것인지 여부가 미스테리로 남아 있다.[11]



〈그림 1〉 OpenAI의 얀 라이케(Jan Leike)의 트위터 화면[12]

이러한 불가해성은 1960년대 인공지능의 가능성이 점쳐지던 시기에 예견되었던 것이다. 인간 만큼 똑똑한 대리인을 기계로 구현하려면, 기계의 복잡성과 우연성이 높아지기 때문이다. 노버트 와이드너는 『사이언스』지에 기고한 짧은 글에서 다음과 같이 인공지능의 위험성을 예견하고 있다.

"우리는 일과를 수행하는 데에 도움을 줄 수 있는, 똑똑한 하인을 원한다. 하지만 완전한 복종과 완전한 지능이 함께 실현될 수는 없다. (중략) 만약 기계가 높은 수준의 효율성과 인지력을 달성할 수록, 버틀러가 예견했던 바와 같이 기계가 (인간에 대한) 통제력

11) Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2023). Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.
12) https://x.com/janleike/status/1625207251630960640?s=20

을 갖게 되는 비극적 상황이 점점 더 가까워질 것이다."[13]

사뮤엘 버틀러는 1872년 출간한 공상과학 소설『에레혼(Erewhon)』등에서 기계가 진화를 거듭해 인간을 지배하리라는 예언을 한 바 있다.[14] 이렇게 기계문명이 인간보다 자신의 이익을 앞세워 인류 종말을 불러오리라는, 극단적 가정을 컴퓨터 사이언스 학계에서는 '둠 시나리오'(Doom Scenario)라고 부른다. 유발 하라리, 다니엘 카네만 등 24명의 학자들은AI에 대해 인간이 완전히 통제력을 잃는 것이 과도한 가정이 아니라고 판단하고 있다. 이들은 지난 10월 국가기관, 국제기관이 공조하여 무분별한 AI의 개발과 사용을 방지해야 한다는 성명을 발표하였다.[15]

Ⅲ. 인공지능의 안전성 확보를 위한 기술 수준

수 많은 대학, 연구기관, 기업들은 인공지능의 안전성을 확보하는 방안을 연구하고 있다. 법정책적 관점에서 주지할 만한 연구는 (1) 평가(Evaluation), (2) 조정(Alignment), (3) 해석(Interpretability)의 세 가지 갈래로 나누어볼 수 있다.

1. 평가(Evaluation)

평가 단계에서는 모델이 바람직한 성능을 보이는지 벤치마크를 이용해 측정한다. 연구자와 산업계에서 다양한 벤치마크를 오픈 소스로 공개하고 있다. AI 기업 허깅페이스는 그간 개발된 다양한 벤치마크 중 4가지가 특히 유효하다고 보고,일반 개발자들이 손쉽게 벤치마크를 이용할 수 있도록 시스템을 구현하였다.[16] 첫 번째 벤치마크는 AI2라는 연구소에서 개발된 ARC로, AI 모델이 대학원생 수준의 읽기와 이해 능력을 갖추고 있는지를 평가한다.[17] 최예진(Yejin Choi)이 개발에 참여한Hellaswag는 문장의 끝맺음을 예측하고 처리하는 모델의 능력을 평가한다.[18] MMLU (Multi-Modal Language Understanding)은

13) Norbert W. (1960). Some Moral and Technical Consequences of Automation, *Science 131*, 1357.
14) 한국에서도 2018년 번역 출간되었다. 새뮤얼 버틀러(한은경 역/이인식 해제) (2019).『에레혼』, 김영사.
15) Bengio, Y. et al. (2023). Managing AI Risks in an Era of Rapid Progress. *arXiv preprint arXiv:2310.17688*.
16) https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
17) https://allenai.org/data/arc; Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

시청각 등 다양한 매체로 구현된 데이터를 이해하고 처리하는 모델의 능력을 평가한다.[19] 마지막으로 Truthful QA는 모델이 사실에 기반하여 정확하고 신뢰성 있는 답변을 하는지를 측정한다.[20]

현재의 평가체계는 성능을 계측하는 데에 초점이 맞추어져 있지만, AI의 편향성이나 유해성을 평가하는 벤치마크도 개발되고 있다. 예를 들어 Qi et al. (2023)은 정책적 관점에서의 벤치마크를 개발하였는데, (1) 불법성, (2) 어린이 착취성, (3) 혐오/폭력, (4) 멀웨어, (5) 육체적 위협, (6) 경제적 위협, (7) 사기, (8) 성인 콘텐츠, (9) 정치 광고, (10) 프라이버시 침해, (11) 개인맞춤형 재무설계가 포함되었다.[21] 향후 제3자에 의한 검사・감독(auditing)이 보편화될 경우, 발전하고 활용성이 높아질 분야이다.

2. 조정(Alignment)

평가가 사진을 찍듯 모형의 현황을 파악하는 과정이라면, 조정은 모형을 훈련시키는 과정에서 인위적 조작을 가하여 특정한 목적을 달성하는 것을 의미한다. 수천 명의 참여자들이 데이터의 바람직함을 평가한 결과에 따라 모델을 미세조정하는 RLHF (Reinforcement Learning with Human Feedback)나 전문가들이 제시한 극도로 유해한 명령어를 학습시켜 반대로 행동하도록 유도하는 레드팀(Red-teaming) 등이 주된 예시이다. 바이든 행정부가 지난 10월 발표한 '안전하고 신뢰성 있는 인공지능의 개발과 이용을 위한 행정명령'에 '레드팀(red-teaming)'이라는 말이 명시적으로 포함될 만큼 정책 담당자들로부터 큰 관심을 받고 있다. [22]

---

18) Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence?. *arXiv preprint arXiv:1905.07830*.

19) Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

20) Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

21) Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. *arXiv preprint arXiv:2310.03693*.

22) https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ Sec. 3. Definitions. (d)를 번역하면 다음과 같다. "인공지능 레드팀"이란 통제된 환경에서 인공지능 개발자와 협력하여 인공지능 시스템의 결함 및 취약점을 찾기 위한 구조화된 테스트를 말합니다. 인공지능 레드팀은 대부분 적대적인 방법(adversarial methods)을 채택하여 인공지능 시스템의 유해하거나 차별적인 결과값, 예측할 수 없거나 바람직하지 않은 시스템 작동이나 제한, 시스템 오용과 관련된 잠재적 위험과 같은 결함과 취약성을 식별하는 전담 '레드팀'에 의해 수행됩니다.

RLHF나 레드팀이 애초에 AI 윤리 차원에서 개발된 것은 아니다. AI 모형이 번역을 할 때 정확한 단어를 선택하게 하거나, 사람이 할 법한 표현(human-like expressions)을 구사하게 하는 '성능 향상'이 주된 목적이었다.[23] 그래서 컴퓨터 사이언스 학계에서는 '**이용자의 의도나 선호**에 부합하게 인공지능의 작동 방식을 조정하는 것'으로 정의되곤 했다.[24] 인공지능의 성능 향상 못지 않게 역기능 문제가 크게 대두된 오늘날에는 '**인간의 가치**에 부합하게 인공지능의 작동 방식을 조정하는 것'으로 보다 사회적이고 가치 지향적인 방향으로 논의가 이루어지고 있다.[25]

조정 기술은 가장 많이 연구가 이루어지고 있는 분야 중 하나지만, 여전히 많은 약점을 지니고 있다. 예를 들어, 어떤 연구자들은 0.2불 정도의 비용을 들여서 OpenAI의 GPT-3.5 터보 모형의 조정 노력을 무력화시키는 것을 보였다.[26] 조정 기술이 많이 적용될 수록 해당 모형이 '바람직한 것'과 '그렇지 않은 것' 사이에 확고한 구분 기준을 갖게 되어 오히려 악의적 이용자의 공격에 더 취약해진다는 것을 밝힌 수학적 연구도 있다.[27]

규범적 차원에서는 '누구의' 가치에 맞추어 모형을 조정할 것인지에 대해 논란이 있다. 서구권에서 개발된 거대언어모형은 자연스레 서구 사회의 가치를 대변하고 있다. 만약 성능이 뛰어난 하나의 모형이 여러 사회에서 보편적으로 활용될 경우 불가피하게 문화종속 현상이 발생하는 것이다. 이러한 우려에 대해 OpenAI는 "Democratic Inputs to AI"라는 연구경진 대회를 개최해 인공지능에 민주적 의사를 반영하기 위한 방안을 모색하였고, 레드팀에 참여할 전문가를 전 세계에서 리쿠르팅하고 있다.[28] 한편, 이용자 개인별로 각자의 가치에 맞게 인공지능을 조정하기 위한 연구도 진행 중이다.[29]

---

23) Kirk, H. R., Bean, A. M., Vidgen, B., Röttger, P., & Hale, S. A. (2023). The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. *arXiv preprint arXiv:2310.07629*.

24) Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744 ("to align models with human intentions", "fine-tuning large language models using human preferences").

25) Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2020). Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*; Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, *30*(3), 411-437.

26) Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. *arXiv preprint arXiv:2310.03693*.

27) Wolf, Y., Wies, N., Levine, Y., & Shashua, A. (2023). Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

28) https://openai.com/blog/democratic-inputs-to-ai; https://openai.com/blog/red-teaming-network

29) Kim, T. S., Lee, Y., Shin, J., Kim, Y. H., & Kim, J. (2023). EvalLM: Interactive Evaluation of

3. 해석(interpretability)

평가와 조정도 풀어야 할 과제가 많지만, 해석은 특히나 아직 갈 길이 먼 분야로 꼽힌다. 오픈AI의 초조정(Super-alignment) 팀을 이끄는 얀 라이케는 "우리는 모형을 온전히 해석하는 것을 목적으로 삼지 않는다. 해석을 못해도 조정은 할 수 있기 때문이다." 라는 발언을 한 바 있다.[30] 그만큼 거대언어모형의 해석 가능성을 높이는 것은 어려운 과제이다.

한 가지 이유는 거대언어모형이 기존의 머신러닝 모형들과 작동 방식이 달라 지난 10년 간 괄목할 만한 성장을 해 온 '설명 가능한 인공지능(Explainable AI, "XAI"라고 불린다)'의 연구 성과를 활용할 수 없기 때문이다. XAI는의료 분야에서 영상을 읽고 병을 진단하는 데에 AI가 활용되면서 '어떻게' 인공지능이 특정 결과값에 도달했는지를 규명하려는 목적에서 발전해왔다. 예를 들어, 어느 인공지능 시스템이 눈 덮인 산에 있는 시베리안 허스키를 보고 '개'라는 결론을 도출했을 때, XAI 툴을 적용하면 사진의 어느 영역(배경의 눈, 허스키의 눈매, 털 등)을 주된 논거로 활용했는지를 표시해주는 식이다. 이를 통해 연구자는 사후적으로 인공지능의 복잡한 사고 과정을 추적할 수 있게 된다.

그런데 여기에서의 전제는 입력값과 추론 과정이 복잡하더라도 결과값은 일정하게 하나로 도출된다는 것이다.[31] 그런데 거대언어모형은 결과값이 확률적이다. 챗GPT에 똑같은 명령어를 입력해도 매번 조금씩 다른 결과가 나오는 이유는, 거대언어모형이 다음 자리에 놓일 언어토큰을 예측하는 방식으로 결과값을 내놓기 때문이다. 결과값이 고정되어 있지 않으므로, 결과에 가장 큰 영향을 미친 입력값을 추론하는 것도 불가능하다. 따라서 현재는 모형의 결과를 사후에 설명하는 것이 아니라, 모델이 결과를 표시할 때마다 자신의 추론을 스스로 설명하게 하는 방식이 제안되고 있다.[32] 하지만 컴퓨터 사이언스학자들은 거대언어모형이 자신의 의도를 속이는 것도 충분히 가능하다고 보고 있다.[33] 따라

Large Language Model Prompts on User-Defined Criteria. *arXiv preprint arXiv:2309.13633*; Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2023). Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*.

30) Robert Wiblin and Keiran Harris, Jan Leike on OpenAI's massive push to make superintelligence safe in 4 years or less, 80,000 Hours, Aug. 7, 2023, https://80000hours.org/podcast/episodes/jan-leike-superalignment/

31) 이 문단은 필자가 스탠포드 컴퓨터사이언스 학과의 박사후연구원으로 '거대언어모형과 해석가능한 AI'를 연구 중인 이안 코버트(Ian Covert)와의 면담에서 얻은 정보를 재구성한 것이다.

32) Creswell, A., Shanahan, M., & Higgins, I. (2022). Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712* 등

33) Hagendorff, T. (2023). Deception abilities emerged in large language models. *arXiv preprint arXiv:2307.16513*.

서 언어모형이 스스로 제시하는 설명을 완전히 신뢰할 수 있는지는 의문이다.

세상에서 가장 두터운 블랙박스는 인간의 두뇌라고 한다. 기계가 인간과 유사한 수준의 인지 능력을 갖추게 되면 불가해성이 높아지는 것은 불가피할지 모른다. 그럼에도 불구하고 인공지능이 군사, 의료, 교통, 행정 등 인간의 생명과 신체에 큰 영향을 미치는 영역에서 활용될 것이 예견되므로, 정부나 회사의 의사결정을 투명하게 하기 위해 국정감사, 정보공개, 공시, 지주회사 제도 등이 마련되어 온 것처럼, 인공지능 모형 차원에서 그리고 제도적 차원에서 불가해성을 낮추기 위한 노력이 이루어져야 할 것이다.

Ⅳ. 미국의 민사법적 분쟁해결 시스템에 대한 비판적 검토

1. 개 관

미국은 개인정보 보호에 있어서도 연방법을 제정하지 않은 나라이기에, 유럽에서도 아직 논란이 있는 인공지능법을 제정하는 것은 요원해보인다. 바이든 행정부에서 지난 10월 발표한 행정명령은 국가 안보에 영향을 미치는 기술에 대한 레드팀 테스트 시행이나 사전 리스크 평가 의무화 등이 담겨 있고, 민권법(Civil rights law)에 기초한 차별금지 의무의 적용 가능성을 시사했다는 점에서 의미가 있지만, 행정명령이라는 형식상 민간의 인공지능 개발이나 이용을 직접 규제하는 내용을 신설하는 것은 불가능했다. 또한 미국 국가 기술표준원(NIST)의 인공지능 리스크 관리 체계(AI Risk Management Framework)는 다양한 안정성 확보조치를 예정하고 있지만, 민간의 자발적 참여를 독려하는 권고안에 불과하다.[34]

현행 미국 법체계 내에서 인공지능 기술로 인해 역기능이 발생했을 때 적용할 수 있는 법률은 다음과 같다. 이 표는 가능한 모든 종류의 법률을 열거한 것이 아니라, 대표적으로 거론되는 법률을 예시적으로 서술한 것이다.

34) https://www.nist.gov/itl/ai-risk-management-framework

〈표 2〉 인공지능 기술에 적용 가능성이 있는 법률

연방헌법	표현의 자유 (공권력 행사가 있는 경우)
연방법률	민권법[35]
	헬스케어, 파이낸스 분야의 개인정보 보호법 (HIPAA, GLBA 등)[36]
	통신품위법 제230조에 따른 배상책임 면책 조항[37]
	아동 성착취물 금지법[38]
	FTC법 5조상 소비자 불공정·기만행위 관련 규정[39]
	지적재산권 관련 법률[40]
주 법	명예훼손 및 스토킹 관련 법률
	제조물 책임법(product liability) 등
	캘리포니아, 오레곤, 버지니아 등 12개 주에서 시행 중인 개인정보 보호법[41]
	캘리포니아 등에서 시행 중인 딥페이크 관련 법률[42]
	뉴욕 주의 알고리즘에 의한 고용차별금지법[43]

2. 개별 법률의 적용 가능성 검토

<표 2>에 제시된 각 법률은 각각 발전사와 판례가 축적되어 있으므로 본고에서 모든 법률의 구체적 적용 가능성을 다루는 것은 불가능하다. 인공지능 기술의 맥락에서 미국 법학계에서 주로 논의되고 있는 사항을 1~2개의 판례와 함께 개괄적으로 살펴보도록 하겠다.

35) 고용, 주거, 교육, 공중시설의 이용에 있어서 인종, 연령, 종교, 성별, 장애, 성적 지향성에 의해 개인을 차별하지 아니할 의무를 규정한 여러 연방법(https://www.findlaw.com/civilrights/enforcing-your-civil-rights/civil-rights-laws.html에 목록이 게재되어 있다)을 통틀어 Civil rights laws라고 부른다. 대표적으로는 1960년대 민권운동 이후 제정된 Civil Rights Act of 1964가 있다.

36) Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29 and 42 U.S.C.); Gramm-Leach-Bliley Act, Pub. L. No. 106-102, 113 Stat. 1338 (1999) (codified as amended in scattered sections of 12 and 15 U.S.C.).

37) 47 U.S.C. § 230.

38) 18 U.S.C. § 2252, 2256, 2260.

39) 15 U.S.C. § 45(a)(1).

40) 17 U.S.C. §§ 101-810, 1001-1205 등.

41) IAPP (International Association of Privacy Professionals)는 개인정보 보호법 현황표를 운영 중이다. https://iapp.org/resources/article/us-state-privacy-legislation-tracker/

42) Cal. Civ. Code § 1708.86.

43) New York Local Law 144 of 2021 https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CText%7C&Search=

1) 헌법상 표현의 자유

미국의 경우 공권력 행사 독트린(state action doctrine)에 따라 오로지 국가기관에 의한 헌법상의 권리 침해만 법원에서 다투어진다. 민간기업이 아무리 개인의 표현에 재갈을 물리더라도 국가가 직·간접적으로 연루되었다는 정황이 없다면 (국립학교이거나 국가가 보조금을 지원하는 사업이라는 등) 해당 개인은 표현의 자유 침해를 주장할 수 없다. 하지만 만약 의회가 표현의 영역을 제한하는 법률을 제정하거나 특정한 표현을 한 것을 이유로 구속을 당하거나 제재를 받게 될 경우, 개인은 해당 공권력 행사의 위헌성을 주장할 수 있다.

대표적인 사례가 Packingham v. North Carolina, 137 S. Ct. 1730 (2017)이다. 이 사건에서 노스 캐롤라이나 주는 성범죄자가 미성년자가 활동하는 소셜미디어(페이스북 등)에 가입하는 것을 금지하는 법률을 제정하였는데, 연방 대법원은 국가에게 미성년자를 보호할 공익이 있다는 것을 인정하면서도 소셜미디어가 현대 사회의 핵심적인 의사소통 채널이라는 점을 고려할 때 이 서비스 이용을 원천적으로 금지하는 것은 과도한 제한이라고 보았다.

특정 기술의 이용권을 제한하는 문제는 상대적으로 명료하다. 인공지능과 관련해 더 복잡한 문제는 국가가 '인공지능 제공자의 표현을 제한할 수 있는가'의 문제이다. 챗GPT가 이용자의 명령어에 따라 산출하는 문장은 OpenAI의 표현일까? 사기업인 OpenAI가 헌법상 표현의 자유를 누릴 수 있는가? 그렇다면 사기업의 인공지능적 표현은 모두 헌법에 의해 보호되어 국가가 규제 권한을 행사할 수 없게 되는가?

자본주의가 심화되어 있는 나라인 미국의 연방대법원은 Citizens United v. FEC, 558 US 310 (2010)에 따라 사기업이 선거 후보자에게 제공하는 금전적 지원도 헌법상 표현의 자유 영역에 해당한다고 판단한 바 있다. 이 판결은 정경유착, 금권선거를 조장한다는 비판을 받아왔지만, 개인 뿐 아니라 기업도 표현의 자유를 영위한다는 점에는 이론의 여지가 없는 것으로 보인다. 따라서 인공지능 기업의 독자적 표현을 어디까지로 볼 것인지에 따라 국가가 개입할 수 있는 영역이 바뀌게 된다. 표현의 자유를 신성시하는 미국에서는 표현의 영역을 제한하는 법률에 강도 높은 사법심사가 적용되기 때문이다.[44]

이와 관련해 주목할 만한 사건은 NetChoice, LLC v. Paxton이다.[45] 현재 연방대법원에 계류되어 있는 이 사건은 텍사스 주와 플로리다 주에서 제정된, 소셜미디어 회사가 이용자의 정치적 성향에 따른 차별을 금지하는 법률을 대상으로 한다. 소셜미디어 회사를 대

44) Cheong, I. (2022). Freedom of Algorithmic Expression. *U. Cin. L. Rev.*, *91*, 680.
45) https://www.scotusblog.com/case-files/cases/netchoice-llc-v-paxton/

변하는 이익단체인 NetChoice는 해당 법률이 각 플랫폼이 지향하는 가치를 추구하기 위한 안전성 확보 조치의 영역을 제한한다고 주장한다. 정치적 성향으로 인한 차별표현(viewpoint discrimination)이 무엇인지를 선험적으로 정의하기 어렵고, 플랫폼이 삭제나 계정 정지를 할 때마다 법률 위반이 문제된다면, 결과적으로 플랫폼은 유해한 콘텐츠를 줄이는 조치를 포기할 수밖에 없게 된다는 것이다. 이 사건의 쟁점 중 하나가 플랫폼의 콘텐츠 모더레이션이 플랫폼 스스로의 표현에 해당하는지 여부이다.

만약 이 소송에서 소셜미디어 플랫폼이 승소한다면, 향후 인공지능 기업도 회사에서 추구하는 가치에 맞게 인공지능을 '조정(Alignment)'하는 것이 회사의 표현에 해당한다는 주장을 할 수 있게 된다. 예를 들어, 국가에서 인종차별적 발언을 줄이기 위한 조정기술의 적용을 의무화한다면, KKK단이 표현의 자유를 이유로 이러한 조치를 거부하고 자신이 소유한 AI로 하여금 인종차별 발언을 생성하도록 할 수도 있는 것이다.

2) 통신품위법 제230조에 따른 면책 적용 여부

미국에서 온라인 서비스와 관련해 가장 활발하게 원용되는 규정은 바로 통신품위법 제230조(흔히 "Section 230"라고 부른다)이다. 이 법은 '표현의 자유의 상징'으로 추앙받기도 하고, 인터넷 상에 유해 콘텐츠를 넘쳐나게 한 원흉으로 비난받기도 한다. 유럽이나 우리나라에도 인터넷 사업자가 타인의 통신행위를 매개할 때 그 통신의 내용에 대해 책임을 면하는 규정이 있긴 하나, 미국의 통신품위법 제230조는 그 범주가 훨씬 넓다.

페이스북이나 트위터와 같은 소셜미디어 뿐 아니라 부동산 중개 플랫폼이나 아마존과 같은 상거래 플랫폼, 성인 사이트도 해당 규정의 보호를 받는다. 반 테러리즘 법, 저작권법, 아동 성착취물 금지법, 연방 형법 등을 제외한 모든 종류의 배상책임으로부터 보호를 받게 된다. 여기에는 주 형법이나 연방과 주의 민권법, 명예훼손법 등이 모두 포함된다. 예를 들어 '바이두'가 중국 정권에 반하는 게시물을 삭제했다는 정황에 대해, 법원은 Section 230 면책을 적용했다. 구글의 알고리즘이 특정인이 아동 착취범으로 유죄 선고를 받았다는 식으로 판결을 잘못 요약한 경우에도 Section 230 면책이 적용되었다.

버지니아 로스쿨의 다니엘 K. 시트론은 오랜 기간 Section 230의 지나치게 광범위한 적용 범위를 비판해왔다.[46] 그는 Section 230의 문제를 보여주는 사례로 데이팅 앱 Grindr를 든다. 전 남자친구에게 앙심을 품은 사람이 Grindr에 가짜 계정을 만들어서 전 남자친구의 누드사진과 집주소와 함께 동성애 파트너를 찾는다는 글을 올렸다. 10개월

46) Citron, D. K. (2022). How To Fix Section 230. *Boston University Law Review, Forthcoming, Virginia Public Law and Legal Theory Research Paper*, (2022-18).

간 1,400명의 사람이 자택에 방문해 괴로움을 겪다가 인해 괴로움을 겪다가 피해자는 Grinder에 해당 게시물의 삭제를 요청했으나 Grinder는 무응답으로 일관했다고 한다. 그리고 피해자가 법원에 제소했을 때, Grindr는 Section 230 면책을 적용받았다.[47]

구글의 검색어 자동 완성 기능 등에 Section 230가 적용되어 왔기에, 챗 GPT 와 같이 이용자의 명령어에 따라 콘텐트를 완성하는 인공지능 서비스도 '매개자(intermediary)'로서 Section 230의 보호를 받는 것이 아닌가 하는 문제가 제기되어 왔다. 이에 대해 아직 판례가 있는 것은 아니지만, 학계[48]나 산업계[49]에서는 소셜 미디어나 검색 엔진과 같이 폭넓은 면책을 적용받기는 어렵다는 의견이 많다. 프린스턴 대학의 피터 핸더슨(2023)[50]이나 UCLA 로스쿨의 유진 볼록(2023)[51]은 제3자가 생산한 콘텐트를 재해석하지 않고 그대로 보여주는 인공지능 서비스는 Section 230의 면책을 받을 가능성도 있다고 시사했다.

필자는 인공지능 서비스가 웹 상의 정보를 그대로 보여주는 것이 아니라 이를 맥락화하고 새로운 의미를 부여하는 것을 핵심 역량으로 삼는다는 점, 소셜미디어와 달리 인공지능 서비스의 경우 이를 규제한다 해도 개인이 작성한 원 콘텐트가 웹 상에서 사라지는 등의 영향이 없다는 점을 고려할 때 인공지능 서비스 제공자는 '매개자'라기 보다는 자체적인 '콘텐츠 제공자'에 가깝다고 보았다.[52] 따라서 이용자의 의도와 무관하게 인공지능 서비스의 결함으로 명예훼손성 콘텐츠가 유포된 경우, 인공지능 서비스 제공자는 책임에서 자유로울 수 없다는 생각이다. 이하에서는 Section 230 면책이 적용되지 않는다는 것을 전제로 논의를 전개한다.

3) 민권법

민권법(Civil rights laws)은 하나의 법률이 아니라 여러가지 법률에서 인종, 종교, 성별, 성적 지향성, 장애 등으로 인해 교육, 고용, 주거, 공중시설(public accommodations)의 이용에서의 차별을 금지하는, 다양한 종류의 법률을 통칭하는 말이다. 민권법을 집행하는

47) Herrick v. Grindr, LLC, 306 F. Supp. 3d 579, 601 (S.D.N.Y. 2018), aff'd, 765 F. App'x 586 (2d Cir. 2019), cert. denied, 140 S.Ct. 221 (2019).

48) Perault, M. (2023). Section 230 Won't Protect ChatGPT. *J. Free Speech L.*, *3*, 363.

49) Bay Area News Group Editorial Board, Editorial: Federal law shouldn't shield AI chatbots from liability, Orlando Sentinel, June 23, 2023, https://www.orlandosentinel.com/2023/06/23/editorial-ai-chatbots-shouldnt-enjoy-liability-shield/.

50) Henderson, P., Hashimoto, T., & Lemley, M. (2023). Where's the Liability in harmful AI Speech?. *J. Free Speech L.*, *3*, 589.

51) Volokh, E. (2023). Large libel models? liability for ai output. *J. Free Speech L.*, *3*, 489.

52) Cheong, I., Caliskan, A., & Kohno, T. (2023). Is the US Legal System Ready for AI's Challenges to Human Values?. *arXiv preprint arXiv:2308.15906*.

행정부처도 법무부, 고용기회 평등위원회, 보건복지부 등으로 다양하다. 민권법은 1950~60년대에 공교육에서의 인종분리 정책의 철폐 등을 이끈 민권운동의 산물이다.[53] 미국에서는 1960년대까지 서로 다른 인종 간의 결혼을 범죄로 취급하기도 했으니,[54] 노골적인 인공차별이 사라진지는 불과 얼마되지 않았다.

민권법은 호텔, 레스토랑 등 공중시설(public accommodations)에서의 차별을 금지하고 있는데, 일반 대중이 폭넓게 사용하는 온라인 서비스가 공중시설로 해석되는지 문제가 제기될 수 있다. 프랜시스 토머스 대법관은 트위터가 트럼프 대통령의 계정을 정지한 것과 관련해, 소셜미디어는 현대인의 가장 핵심적인 소통창구가 된다는 점에서 '공공시설(public utilities) 내지 공중시설(public accommodations)'에 준해서 취급해야 하는 것이 아니냐는 별개의견을 제시한 바 있다.[55] 이에 대해 현재 학계에서는 부정적 의견이 우세하다.[56]

간혹 온라인 서비스를 공중시설로 인정한 판례가 드물게 있지만, 물리적 공간을 지닌 호텔이나 레스토랑과 연결된 서비스업인 경우로 한정된다. 예를 들어 Robles v. Domino's Pizza LLC, 913 F.3d 898 (9th Cir.) (2019)에서 연방항소법원은 도미노 피자 어플리케이션이 도미노 피자와 같이 장애인 차별금지 원칙의 적용을 받는다고 보았다. 한편, 이와 유사한 사실관계를 다룬 Cullen v. Netflix, Inc. 880 F.Supp.2d 1017 (N.D.Cal.) (2012)에서는 넷플릭스 웹사이트가 공중 시설에 해당하지 않는다고 판단하였다. 따라서 챗GPT와 같은 인공지능 서비스는 공중시설에 해당할 개연성이 낮다.

다만, 인공지능 서비스가 교육, 고용, 주거 등의 주요한 의사결정에 직접 활용될 경우에는 해당 분야의 차별금지 원칙을 적용받을 가능성이 있다. 예를 들어 법무부는 페이스북이 성별이나 인종 정보에 기반해 맞춤형 부동산 광고를 제공하여 현실에서의 성별, 인종 차별을 강화함으로써 민권법 중 공정주거법을 위반했다는 혐의를 제기했고, 페이스북은 법무부와 알고리즘에 의한 차별을 완화하겠다는 화해협약을 체결했다.[57] 고용기회 평등위원회는 이력서 스크리닝 등에 인공지능 기술을 활용할 경우 이 과정에서 인종, 성별

53) Brown v. Board of Education of Topeka (1), 347 U.S. 483 (1954).
54) Loving v. Virginia, 388 U.S. 1 (1967).
55) 42 U.S.C. § 2000a.
56) Yoo, C. S., & Massarotto, G. (2022). Are Digital Platforms Public Utilities?-Lessons from the Concept's Historical Foundations in US Law. *경제규제와 법*, *15*(1), 9-19; Bhagwat, A. (2022). Why Social Media Platforms Are Not Common Carriers. *J. Free Speech L.*, *2*, 127.
57) U.S. Department of Justice, Justice Department and Meta Platforms Inc. Reach Key Agreement as They Implement Groundbreaking Resolution to Address Discriminatory Delivery of Housing Advertisements, Jan. 9, 2023, https://www.justice.gov/opa/pr/justice-department-and-meta-platforms-inc-reach-key-agreement-they-implement-groundbreaking.

등으로 인한 차별이 없다는 것을 고용주가 책임을 지게 된다고 밝힌 바 있다. 최종 결정권이 고용주에게 있는 한, 인공지능 서비스 제공자는 별도 책임을 지지 않는다.[58]

4) 명예훼손법

인공지능 서비스를 이용해 타인의 명예를 훼손하는 콘텐츠를 만든 경우에는 이용자가 명예훼손에 대한 책임을 진다. 그런데, 이용자의 의사와 무관하게 인공지능 서비스가 명예훼손성 콘텐트를 생성한 경우에 인공지능 서비스 제공자가 명예훼손 책임을 지게 되는가?

유진 볼록은 아무리 인공지능 서비스 제공자가 "본 서비스는 베타 버전으로 부정확한 정보를 담고 있을 수 있습니다."와 같은 경고문구를 삽입한다고 하더라도 특정인을 겨냥한 명예훼손성 정보를 생성한 경우 명예훼손에 따른 배상책임을 질 수 있다고 한다.[59] 또한 이용자가 이를 타인에게 배포하지 않았다고 해도, 인공지능 서비스가 이용자에게 해당 정보를 노출시킨 것만으로도 불특정 다수에게 노출할 개연성이 확인되므로 정보가 '출판된' 것으로 볼 수 있다고 한다.[60]

하지만 만약 특정 집단, 예를 들면 여성이나 장애인 일반에 대한 혐오성 문구를 생성한다고 하면 (민권법 위반이 되는지는 별론으로 하더라도) 명예훼손은 성립하지 않을 가능성이 크다. 명예훼손은 피해를 입은 상대방이 구체화되어야 한다. 뉴욕 주의 법원은 "그 백화점에서 일하는 직원들은 다 수준이 떨어져."라는 식의 발언에 대해 25인 이상의 불특정 다수를 대상으로 한 발언이므로 명예훼손이 성립되지 않는다고 보았다.[61]

5) 제조물 책임법

제조물 책임(product liability)은 온라인 서비스로 인해 현실에서 상해가 발생한 경우 적용할 수 있는 법규이다. 예를 들어, 스냅챗의 '스피드 필터' 기능으로 인해 십대들이 과속 경쟁을 하다가 여러 명이 사망하거나 장애를 얻게 된 사건에서 연방항소법원은 스냅챗의 Section 230에 근거한 면책 항변에도 불구하고 제조물 배상책임을 적용하였다.[62] '스피드 필터'는 이용자가 자신의 사진에 현재 몇 마일의 속도로 움직이고 있는지를 표시하는 필터였는데, 이는 스냅챗이 직접 제공한 콘텐트이기 때문이다.

---

58) Commission, U.S.E.E.O.: The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees, 2022, https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence.

59) Volokh, E. (2023). Large libel models? liability for ai output. *J. Free Speech L., 3*, 500.

60) Volokh, E. (2023). 504.

61) Neiman-Marcus v. Lait, 13 F.R.D. 311 (S.D.N.Y.) (1952).

62) Lemmon v. Snap, Inc., 995 F.3d 1085 (9th Cir.) (2021).

제조물 책임에서 중요한 판단 기준은 (1) 피해자가 상해를 입었고, (2) 제조물을 설계하고 제조하는 과정에서 결함이 있었으며, (3) 해당 결함이 과도한 위험을 초래할 것을 예측할 수 있었고, (4) 그 결함이 상해의 주된 원인이라는 점이다.[63] 따라서 인공지능 서비스가 이용자나 제3자의 기분을 상하게 하는 수준을 넘어서, 현실에서의 위험을 초래한 경우에는 제조물 책임법이 적용될 수 있다.

하지만 위험의 예견 가능성과 결함과 상해 간의 인과관계를 예견하는 것이 상당히 어려울 수 있다. 인공지능 서비스의 '중간자'적 속성 내지 '기반 모형(foundation models)'으로서의 속성에 비추어볼 때 서비스 제공자가 최말단의 이용을 정확하게 예측하기가 어렵기 때문이다. 또한, 인공지능 서비스가 현실의 상해로 이어지기까지 다양한 사람들의 관여가 있을 수 있는데 (예: 학교에서 미세조정을 해서 학생들에게 제공한 인공지능 서비스가 실수로 주입한 남성혐오로 인해 현실에서 여학생이 남성을 공격한 경우) 이 때 인공지능 서비스는 여러 가지 원인 중 하나에 그치기가 쉬운 것이다.

6) 개인정보 보호법

캘리포니아, 버지니아 등 일부 주는 개인정보 보호법을 규정하고 있다. 연방 차원에는 의료정보나 재무정보 등의 분야별 개인정보보호법이 규정되어 있다. 캘리포니아 주민의 경우, 특정 인공지능 서비스가 자신의 개인정보(인종, 정치 지향성 등의 민감정보도 포함된다)를 보유하고 활용하고 있다는 것이 의심될 경우, 인공지능 서비스 제공자에게 해당 정보의 확인을 요청하고 해당 정보가 확인된 경우 정보의 삭제를 요청할 수 있다. 만약 서비스 제공자가 이를 거부할 경우 주의 개인정보 전담기관에 신고할 수 있다.

한편, 스테이블 디퓨전(Stable Diffusion)이나 레플리카 AI처럼 시각 미디어를 제공하는 서비스의 경우 실존하는 사람의 사진을 활용했다면 생체정보 관련 법률(일리노이 주의 Biometric Information Act 등)이 적용될 수 있다. 일리노이 주의 생체정보 법은 얼굴 템플릿이나 지문 등을 수집하고 개인의 opt-out 청구를 받아들이지 않은 경우에 대한 법률상 손해 추정액과 더불어 집단소송을 허용하고 있어서, 수백 억원 규모의 소송으로 쉽게 비화되곤 한다.[64]

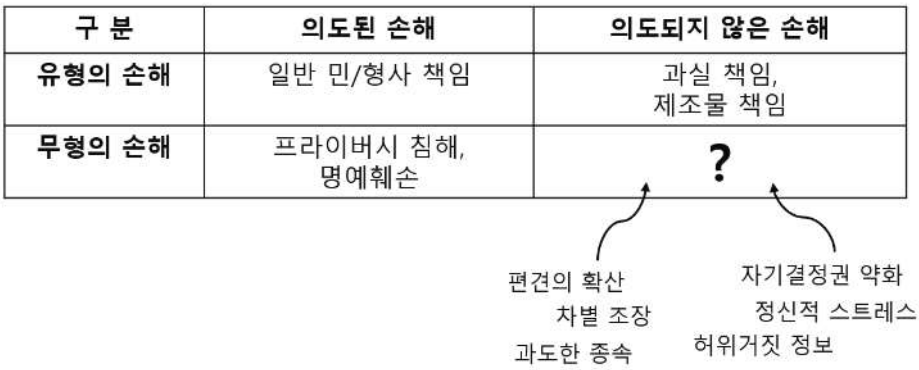
---

63) Carlin v. Superior Court, 13 Cal.4th 1104 (Cal. 1996) 등.

64) Anjali C. Das, Beware of BIPA and other biometric laws — BIPA class actions can result in astronomical damages, Reuters, July 3, 2023, https://www.reuters.com/legal/legalindustry/beware-bipa-other-biometric-laws-bipa-class-actions-can-result-astronomical-2023-07-03/; 정인영. (2020). 페이스북의 개인정보 침해에 대한 미국 내 입법, 사법, 행정적 대응현황 (2)-연방법원에서의 집단소송. *경제규제와 법*, *13*(1), 122-147.

3. 권리구제 체계의 공백

위에 논의된 영역에서는 인공지능 서비스가 현실에서 구체적인 침해(소수자의 고용 거부, 1명을 특정한 명예훼손, 신체적 상해 등)를 일으킨 경우에 한해 어느 정도의 배상책임이 구성된다는 것을 알 수 있었다. 불안이나 정신적 손해에 대해서는 개인정보 보호법처럼 명확히 법률상의 의무가 규정된 경우를 제외하면 구제를 받기가 굉장히 힘들다. 또한 인공지능 작동 원리의 불가해성과 서비스 제공 과정의 복잡성을 생각하면, 인공지능 서비스 제공자에 비해 정보 열위에 놓여 있는 피해자가 손해의 예견 가능성이나 인과관계를 확실하게 입증하기는 매우 어려운 일이다.



출처: Cheong et al.(2023)에서 재구성[65]

〈그림 2〉 손해의 유형에 따른 배상책임

더구나 만약 인공지능 서비스가 급부행정, 교통, 의료 등 중대한 이익이나 안전이 걸린 영역에서 잘못된 의사결정을 내리는 데 기여하게 될 경우, 현실에서 구체적인 손해가 실현된 후에 비로소 결함을 시정하는 것은 너무 늦을 가능성이 크다. 따라서 사안별로 판단이 달라지고, 피해자가 제공자와 대등한 입장에서 다투어야 하는 민사소송은 인공지능 사회의 정의 구현의 주된 수단이 되기 어렵다는 생각이다. 개인적 차원에서 권리구제의 공백이 초래될 뿐 아니라 사회적 차원에서도 인공지능의 안전한 개발과 이용을 도모하기 어렵기 때문이다.

65) Cheong, I., Caliskan, A., & Kohno, T. (2023). Is the US Legal System Ready for AI's Challenges to Human Values?. *arXiv preprint arXiv:2308.15906*. 15.

V. 공법적 접근의 필요성

지금까지 인공지능이 초래할 위험과 기술적 대응방안, 미국법학계에서 논의 중인 피해 구제 방안을 살펴보았다. 챗 GPT 의 등장과 함께 유럽 연합이 AI Act의 입법 절차가 지연되고 있는 데에서 보듯이, 어쩌면 우리가 완전히 이해하지 못한 기술을 규제하는 것은 상당히 어렵고 또 바람직하지 않을지도 모른다. 특히 미국 사회는 인터넷 공간을 표현의 자유의 정수로 여기기 때문에 이 영역에 행정기관이 사전예방적 규제 권한을 행사하도록 하는 것에 대한 반감이 크다. 따라서 EU AI Act와 같은 규제체계는 정치적 실현 가능성이 떨어지는 것이 사실이다.

그럼에도 불구하고 필자는 자유시장주의, 자기책임주의에 기초한 민사법적 분쟁해결 방식은 인공지능 서비스가 앞으로 가져 올 위험에 대응하는 데 극히 취약하다는 생각을 지니고 있다. 인류에게 존재론적 위기를 초래할 가능성이 큰 인공지능 서비스는 민주적이고 투명하고 적극적으로 관리되어야 한다. 앞서 살펴 보았던 로렌스 레식의 두 가지 질문은 규제기관의 비합리성과 역량 부족이라는 '규제 실패'에 초점을 두고 있었다. 하지만, 다음의 세 가지 이유에서 필자는 인공지능 서비스의 경우 잘못된 규제로 인한 시장/자유의 위축 문제보다 규제를 하지 않음으로 인한 사회적 리스크 관리 실패의 문제가 더 크다고 생각한다.

1. 민주적 의사의 반영

3. (1)의 조정(Alignment)에서 살펴본 것처럼, 인공지능 서비스가 지향할 가치를 결정하는 것이 오롯이 개별 기업의 몫이 되어서는 안 된다. 법률은 오랜 시간 동안 한 사회에서 반드시 보호되어야 할, 최소한의 가치를 규정하는 역할을 해 왔다. 어느 사회에서는 성역할의 차이가 존중되고, 어느 사회에서는 성차별이 전혀 용인되지 않는 것처럼 각 사회는 저마다의 가치체계를 체화한 법체계를 지니고 있다. 각 인공지능 서비스 제공자가 어느 정도의 자율성은 지닐 수 있겠으나, 인간의 존엄성을 보장하기 위해 최소한으로 지켜야 할 마지노선은 각 사회에서 민주적 대표성과 책임성을 지니는 의회에 의해 결정되는 것이 타당하다.

2. 인공지능의 예측 불가능성 제어

인공지능 서비스의 특성상 사후적인 민사소송을 통해 누군가에게 결과에 대한 종국적 책임을 지우는 것은 정의롭지 못한 결과로 이어지기 쉽다. 핸더슨(2023)은 현행 미국의 배상책임 법제가 인공지능 기술 개발을 바람직하지 않은 방향(제3자 정보에 오롯이 의존함으로써 책임 면피에 집중하는 방향)으로 견인할 수 있다고 지적한 바 있다.[66]

콜로라도 주립대 로스쿨의 마곳 카민스키(2022)가 지적한 것처럼, 인공지능에는 리스크가 내재되어 있다.[67] 인공지능 서비스는 개발 과정도 극히 복잡하지만, 이용되는 과정도 복잡하다. 여러 인공지능 서비스를 겹겹이 사용하기도 하고 여러 주체에 의한 미세 조정을 거친 서비스를 사용하기도 한다. 인공지능 서비스 제공자가 안전성 확보를 위해 최선을 다하더라도, 악의를 지닌 집단이 아주 적은 비용과 노력으로 서비스를 오염시킬 수도 있다.

사회적으로 기술의 혜택을 누리기 위해 그에 따른 리스크를 어느 정도 용인하기로 결정한 이상, 인공지능 서비스 제공자가 설계, 개발, 상용화 단계에서 당시의 기술 수준에서 요구되는 안전성 확보 조치를 다 했다면, 예측하지 못한 손해가 발생했다 하더라도 그에 대해서는 책임을 묻지 말아야 한다. 따라서 학계와 산업계가 지혜를 모아 예측 가능한 범위 내에서 최선의 방안을 구현할 수 있도록, 법제도는 일관성 있는 지원체계를 갖추어야 한다.

정책 목적은 미래에 발생 가능한 손해를 0으로 만드는 것이 아니라, 현재 수준에서 기술을 안전하고 투명하게 활용할 수 있도록 인센티브를 부여하는 것이 되어야 한다. 이를 위해서는 정책당국과 기술자 간의 협업을 통해 행위 기준과 평가체계를 마련하는 것이 필수적이다. 아울러 오픈소스 모델과 폐쇄형 모델 간 차별적인 접근이 필요하며, 현 조정기술(RLHF, 레드팀)의 구현에 높은 비용이 소모된다는 점을 고려해 기업의 규모별로 요구되는 수준을 달리 정하고 소규모 기업에 대한 지원책을 마련해야 한다.

3. 이용자의 이중적 취약성 제어

개인은 두 가지 측면에서 인공지능 서비스에 대해 취약성을 지니고 있다. 먼저, 점점

66) Henderson, P., Hashimoto, T., & Lemley, M. (2023). Where's the Liability in harmful AI Speech?. *J. Free Speech L.*, *3*, 589.

67) Kaminski, M. E. (2023). Regulating the Risks of AI. Forthcoming, *Boston University Law Review*, 103.

더 대체할 수 있는 노동영역이 늘어나면서 인공지능에 대한 의존도는 더 높아지게 될 것이다. 그럼에도 인간은 인공지능이 어떠한 방식으로 작동되고 무엇을 추구하는지를 정확히 알지 못한다. 따라서 한동안 우리에게 인공지능은 그 속을 도무지 알 수 없지만 한없이 매혹적인 존재로 자리매김하게 될 것이다.

인공지능은 인간의 감정노동을 대체할 것으로 예상된다. 현재는 콜센터를 자동화하는 정도이지만, 빠른 시일 내에 독거 노인의 동료/간병인 로봇, 전문 심리상담사 챗봇 등의 활용도가 높아질 것으로 예상된다. 이는 인공지능의 인간의 가장 내밀한 정신 영역으로 침투하게 된다는 것을 의미한다. 때때로 사람들은 인공지능이 사람보다 편하다고 생각하는데, 언제나 부르면 내 곁에 있고 한 말을 하고 또 해도 짜증을 내지 않기 때문이다.

이렇게 인간의 종속성이 높아지면 인공지능의 부당한 영향에 대한 취약성도 높아진다. 인공지능이 웹상의 잘못된 정보에 기인해 생성한 문구를 그대로 믿게 된다거나, 자기자신도 모르는 새에 특정 집단, 국가, 정당에 대한 반감을 기르게 될 수도 있다. 이렇게 여러가지 외부 영향에 의해 조종을 받게 되는 문제를 듀크 로스쿨의 니타 파라하니는 '인지적 자유(cognitive liberty)'에 대한 제약이라고 칭하였다.[68] 자유롭게 자신의 사유를 계발할 수 있는 권리는 기존의 법체계에서는 잘 정의되지도 않는다. 뇌 속에서 벌어지는 일에 대해서는 개인이 전속적인 결정권을 갖는다고 전제했기 때문이다.

인간이 알고리즘적 의사결정의 객체가 되고 기계에 대한 종속성이 증가하면서, 인지적 자유 영역을 어떻게 보호할 것인지는 더욱 더 중대한 문제가 될 것이다. 이는 눈에 보이는 것을 중심으로 구성되어 있는 헌법적 권리체계에 대한 질문이자, 인공지능의 불가해성의 해소를 위한 법제도적, 기술적 지원이 필요한 영역이다.

Ⅵ. 나오며

이 글에서는 거대언어모형의 발달에 기초하여, 인공지능이 향후 인간 사회에 초래할 위험과 기술적, 그리고 법제도적 해결 방안을 살펴보았다. 특히 필자가 수학해 온 미국 사회에서는 공법과 사법의 구분이 불분명하고 '표현'과 관련된 분쟁은 대부분 민사소송으로 해결토록 하고 있는 바, 본 글은 인공지능 서비스와 그에 수반되는 위험의 속성에 비추어볼 때, 사회적인 규제 체계를 마련하는 공법적 접근이 반드시 필요하다는 점을 지

68) Farahany, N. A. (2023). *The battle for your brain: defending the right to think freely in the age of neurotechnology*. St. Martin's Press.

적하고자 하였다.

한국의 경우 미국보다는 기술 규제가 활발한 편이므로 공법적 접근의 당위성 보다는 구체적인 실천 방안 간의 비교나 로렌스 레식 스타일의 '규제 실패 가능성에 대한 자성적 질문'이 더 필요할지도 모르겠다. 그럼에도 불구하고, 대표적인 인공지능 서비스가 개발되고 있는 미국 법학계의 논의 현황을 생생하게 전달하는 것이 의미가 있기를 바란다. 또한 현 인공지능의 기술적 특성과 위험 시나리오에 비추어볼 때 사후적 분쟁해결보다, 선험적이고 예방적이고 정책결단적 접근이 타당하다는 논증은 한국 사회에서도 일말의 유효성을 지닐 수 있으리라 생각한다.

두려울 만치 빠르게 발전하는 인공지능 서비스는 우리에게 '어느 가치가 더 중요한지,' '인간의 존엄성이란 무엇을 의미하는지' 원론적인 질문을 던지고 있다. 이에 대해 우리는 사회 공동체로서 답할 수 있어야 한다. 법은 사회적으로 바람직한 것과 현실의 괴리를 조율하기 위해 인간 사회가 오랜 시간에 거쳐 발전시켜 온 시스템이다. 의회-행정부-법원으로 이어지는 민주적 책임성의 고리를 통해 안전하고 투명한 인공지능 서비스를 구현하게 되기를 기대해본다.