2023년 4월 26일, 3권 22호 (통권 112호)

아시아 브리프

Current Issues and Policy Implications

SNUAC
Seoul National University Asia Center
서울대학교아시아연구소

특집: 초거대 AI의 미래(2)

생성모델과 AI 거버넌스

〈그림 1〉 AI와 법
출처: https://oecd.ai/

Summary Of Article

마경태 법무법인 태평양

최근 챗GPT를 비롯한 다양한 인공지능(AI) 생성모델(Generative Model)이 우리의 일상생활에 근본적인 변화를 가져올 것으로 기대되고 있다. 하지만 동시에 저작권 침해 가능성, 거짓 정보를 그럴듯한 문장으로 출력하는 '환각' 현상 등 다양한 법적, 윤리적 이슈들도 제기된다. AI의 잠재적 위험을 관리하기 위한 수단으로써 그동안 기업 내부적으로 AI 거버넌스 체계를 구축하는 방안이 활발하게 논의되어 왔다. 현재 국내외에서는 AI 거버넌스 체계 구축을 법제화하기 위한 노력도 진행되고 있다. 다만 지금까지 논의되어 온 AI 거버넌스 체계는 생성모델의 작동 방식과 잠재적 위험에 대하여 충분한 연구와 검토가 이루어지지 못했다. 이에 새로운 AI 거버넌스를 마련하여 생성모델로 발생할 수 있는 새로운 위험 요소에 대응하는 방안이 필요하다.

생성모델의 급부상과 법적·윤리적 이슈

2023년은 모두가 놀랄 만한 새로운 인공지능(AI) 서비스 출시 소식이 쏟아지며 AI 기술과 산업에서 기념비적인 한 해가 되고 있다. 현재 혁신을 이끌고 있는 AI 모델은 챗GPT(ChatGPT)와 스테이블 디퓨전(Stable Diffusion)과 같은 생성모델(Generative Model)이다. 생성모델이란 문장, 이미지, 음성을 학습한 데이터와 유사하게 생성해내는 모델이다. 생성모델은 방대한 양의 데이터를 학습하는데, 이제는 인터넷에 공개된 거의 모든 정보를 학습할 정도로 엄청나게 큰 규모의 데이터셋을 통해 훈련이 이루어진다.

그런데 이처럼 대규모의 학습데이터를 구축하고 이를 생성모델이 학습하는 과정에서 대량의 저작물 복제·전송과 개인정보의 수집·이용이 이루어지게 된다. 그리고 그 과정에서 모든 저작권자와 정보주체의 허락을 받을 수 없기 때문에 저작권 침해 및 개인정보 무단 수집·이용이 문제되고 있다. 실제로 2022년 11월에 미국에서 자동으로 코드를 완성하는 GPT 기반 생성모델 서비스인 깃허브(GitHub)의 코파일럿(Copilot)에 대하여 저작권 침해 소송이 제기되었고, 9월에는 스테이블 디퓨전이 학습데이터로 활용하고 있는 LAION 데이터셋에 대하여 환자들의 의료 시술 전후 비교 사진이 포함된 개인정보 무단 수집·이용이 문제된 바 있다. 최근 이탈리아 정보보호국(Data Protection Authority)은 챗GPT의 학습 과정에서 개인정보가 법적 근거 없이 대규모로 수집되었다는 이유로 이탈리아 내에서 챗GPT의 이용을 일시 차단하였고, 다른 유럽연합(EU) 국가들도 챗GPT에 대하여 EU 개인정보보호법(GDPR) 위반 여부를 조사할 계획이다. 이외에 학습 데이터뿐만 아니라 생성모델의 결과물에 대해서도 저작권 인정 여부, 개인정보 유출 가능성 등 법적으로 불명확한 여러 이슈들이 발생하고 있다.

법적 문제 외에도 생성모델이 거짓 정보를 그럴듯한 문장으로 출력하는 '환각'(Hallucination) 현상도 문제이다. 현재 챗GPT에 활용되고 있는 언어모델은 실제 사람이 쓴 것과 같이 그럴듯한 문장을 출력하는 것을 목표로 작동한다. 이러한 언어모델은 애초에 정확한 정보 제공에 특화된 AI 모델이 아니다. 그런데 현재 많은 이용자들이 챗GPT를 대화를 위한 챗봇을 넘어서서 검색엔진을 대체하는 데이터베이스와 같이 사용하다 보니 부정확한 생성정보를 신뢰하게 되는 새로운 문제가 발생하고 있다.

AI 시스템 위험 관리 방안: AI 거버넌스 체계

그렇다면 이처럼 AI 시스템의 활용이 개인과 사회에 발생시킬 수 있는 잠재적 위험에 대응하기 위한 방안은 무엇인가. 최근 AI 시스템의 위험을 관리하기 위해서 기업 내부적으로 AI 거버넌스 체계를 구축하는 방안이 활발하게 논의되고 있다. 현재 논의 중인 AI 거버넌스 체계들은 그 이름과 세부적인 내용은 조금씩 다르지만 공통적으로 AI 시스템의 잠재적 위험을 사전에 평가하고, 위험을 체계적으로 관리하기 위한 조직과 내부 절차를 마련하는 것을 내용으로 한다.

AI 기술이 사회에 본격적으로 도입됨에 따라 2018년경부터 각국 정부, 다국적 기업, 국제기구 등은 AI 시스템이 준수해야 하는 윤리원칙들을 발표해왔다. AI 윤리원칙은 주로 공정성, 투명성 등 추상적인 원칙으로의 성격을 갖는데, 이 원칙들은 AI 거버넌스의 구축을 통해 실무 현장에서 이행 가능한 형태의 지침으로 구체화된다. 예를 들어 공정성 원칙과 관련하여서는, 기업이 AI 시스템을 운영하기 전에 개인에게 부당하게 차별적인 결정을 내릴 위험이 있는지를 검토하고, AI 시스템의 편향성을 확인할 수 있는 통계 지표를 설정하여, 편향성을 관리하고 완화하는 절차를 마련하는 것이다.

아직까지 AI 윤리원칙과 그에 관한 AI 거버넌스 체계는 대부분 법률이 아닌 연성 규범(Soft Law)의 형태로 존재하고 있다. 이러한 연성 규범으로는 대표적으로 정부기관들이 발표하고 있는 AI 윤리 '가이드라인'을 들 수 있다. 특히 금융분야에서는 금융 기관들이 AI의 생애 주기별로 준수해야 할 사항들을 상세히 정리한 가이드라인들이 발표되고 있어 주목을 받고 있는데, 해외에서는 싱가포르 통화청(Monetary Authority of Singapore)의 'Veritas 프레임워크'가 대표적이고, 국내에서는 작년 8월 금융위원회에서 발표한 '금융분야 AI 개발·활용 안내서'가 있다.

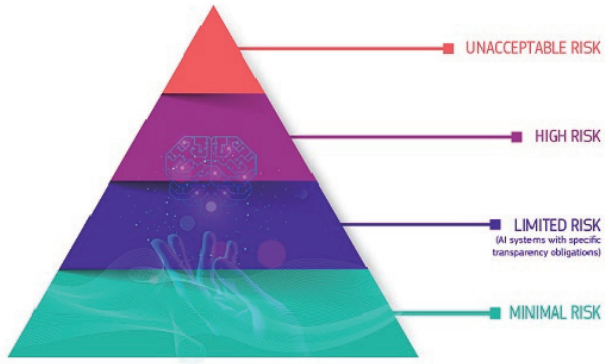
AI 거버넌스 체계를 법률의 형태로 만들어야 한다는 논의도 이루어지고 있다. 유럽연합(EU)은 2021년 4월 EU 집행위원회(European Commission)에서 AI 법안(Regulation laying down harmonised rules on artificial intelligence)의 초안을 발표하였고, 이후 2022년 12월 EU 각료 이사회(Council of EU)에서 개정안을 발표하였다. 현재는 EU 의회에서 논의 중인데, EU AI 법안은 2023년 말이나 2024년 초에 입법이 이루어질 것으로 예상되고 있다. 미국에서는 2022년 3월에 상·하원에서 각각 인공지능 책임 법안(Algorithmic Accountability Act of 2022)이 발의되었고, 12월에는 EU와 AI의 신뢰성 확보와 위험 관리에 관하여 정보를 공유하고 협력을 증진하는 내용의 공동 로드맵이 발표되었다(TTC Joint Roadmap for Trustworthy AI and Risk Management). 국내에서도 EU AI 법안과 유사한 형태의 인공지능의 신뢰성 확보를 위한 여러 법안이 발의되어 현재 국회에서 논의 중이다.

국내외 AI 거버넌스 체계의 주요 내용: EU AI 법안과 금융위 AI 안내서

이하에서는 국내외 대표적인 AI 거버넌스 체계인 EU의 'AI 법안'과 금융위원회의 '금융 분야 AI 개발·활용 안내서'의 주요 내용을 살펴본다. 그리고 기존 AI 거버넌스 체계가 생성모델에도 적용될 수 있는지 검토해보겠다. 두 거버넌스 체계는 모두 AI 시스템의 위험 관리를 위한 구체적인 방안을 담고 있어서 향후 국내 AI 거버넌스 체계에 상당한 영향을 미칠 것으로 예상된다.

(1) AI 위험 수준 평가

EU AI 법안은 위험 기반 접근 방식(risk-based approach)을 적용하여 AI가 개인의 기본권을 침해하거나 중대한 위험을 발생시킬 수 있는지에 따라 (i) 용인할 수 없는 위험(unacceptable risk), (ii) 고위험(high risk), (iii) 제한된 위험(limited risk), (iv) 최소한의 위험(minimal risk)으로 구분된다. 각 단계별로 다른 수준의 규제를 적용하고 있는데, 용인할 수 없는 방식으로 사용되는 경우를 제외하면 AI 시스템의 위험 수준은 3단계로 구분된다. 또한, AI 시스템이 사용될 목적에 따라 AI 시스템의 위험 수준을 사전 정의하고 있다. 즉 개인신용평가, 채용·인사평가, 학교 입시 등의 목적으로 사용될 AI 시스템은 개인의 기본권을 침해하거나 중대한 위험을 발생시킬 수 있다고 보아 고위험으로 지정된다. 고위험 AI 시스템은 엄격한 기술적·관리적 요건을 충족해야 한다.



〈그림2〉 EU AI 법안의 AI 시스템 위험 수준 분류
출처: EU 집행위원회(European Commission)

금융위 AI 안내서도 EU AI 법안과 유사하게 위험 기반 접근 방식을 적용하고 있다. AI 서비스가 금융소비자에게 미치는 영향, 위험 요소 등을 평가하여 고·중·저위험 등으로 분류하고, 위험 수준별로 준수사항을 정하도록 안내하고 있다. AI 시스템의 위험 수준을 사전에 정하고 있는 EU AI 법안과는 달리, 금융위 AI 안내서는 금융기관으로 하여금 제공하려는 AI 서비스별로 개인의 권익과 안전, 자유에 대하여 미치는 위험을 평가하여 위험 수준을 정하도록 하고 있다. 다만, EU AI 법안과 금융위 AI 안내서는 AI 시스템이 사용될 목적 및 방법에 따라 사전에 위험 수준을 정하고 준수사항을 차등적으로 정하도록 하고 있다는 점에서 접근 방식이 유사하다고 볼 수 있다.

(2) 데이터 관리

EU AI 법안은 AI 시스템의 성능 및 공정성 확보를 위해 몇 가지 요건을 요구한다. 그 내용으로는 고위험 AI 시스템에 대하여 위험 관리 시스템 구축, 고품질의 데이터 마련, 시스템 기술 명세서 작성, 시스템 운영 내역 기록, 이용자에 대한 정보 제공, 사람에 의한 통제 확보, 높은 시스템 성능·안정성·안전성 확보 등이 있다. 그 중 데이터의 경우, 수집·전처리 과정 관리, 데이터의 대표성·무오류성·완전성 확보, 편향성 검토 등 고품질의 데이터 확보를 위한 요건을 충족하여야 한다. 금융위 AI 안내서도 AI 생애주기 중 개발 단계에서 EU AI 법안과 유사한 데이터 관리 절차를 요구하고 있다.

(3) AI 공급자와 이용자간 통제

EU AI 법안은 고위험 AI 시스템에 관하여 공급자(provider)가 1차적인 책임을 부담하도록 한다. 공급자는 AI 시스템의 사용 목적과 방법이 기재된 사용설명서(instructions for use)를 작성하여 활용자(user)에게 제공하고, 활용자는 사용설명서에 따라 고위험 AI 시스템을 사용하고 모니터링하는 구조이다. 즉 활용자가 고위험 AI 시스템을 이용하여 고객에게 서비스를 제공하는 경우에도 공급자가 AI 시스템의 잠재적 위험을 발견하고 관리할 의무가 있는 것이다. 반면 금융위 AI 안내서의 경우, 활용자가 AI 시스템의 위험을 관리할 1차적인 책임을 부담하는 구조다. 금융기관이 AI 시스템을 이용하여 고객에게 서비스를 제공할 때, 직접 위험관리정책을 수립하고 AI 시스템의 공급자(수탁기관)에게 이를 준수할 것을 요구할 책임이 있다.

생성모델에 대한 AI 거버넌스 적용 가능성

위와 같이 기존에 논의되어 온 주요 AI 거버넌스 체계는 AI 시스템이 내린 의사결정이 개인의 권리 및 안전에 미칠 수 있는 위험을 관리하는 것을 목적으로 한다. 따라서 주로 대출 심사나 채용 결정과 같이 개인에 대한 의사결정 과정에서 사용되는 AI 시스템이 거버넌스 적용의 주 대상이다. 그리고 이러한 의사결정 과정에 사용되는 AI 시스템은 주로 AI 모델 중에서 판별모델(discriminative model)을 사용한다. 여기서 판별모델이란 사과와 바나나를 구분하듯이 입력된 데이터를 일정한 기준에 따라 분류하는 모델을 의미한다.

그런데 앞서 본 생성모델은 사람의 의사결정 과정을 대체하거나 보조하기 위하여 사용되는 판별모델과 비교하여 크게 다음과 같은 차이점이 있다.

- 생성모델은 판별모델에 비해 최종 활용 단계에 이르기 전까지 이용 목적과 방법을 쉽게 확정하기 어렵다. 따라서 개발 단계에서 이용 목적과 방법을 어느 정도 확정할 수 있는 판별모델과 달리 생성모델은 다양한 목적과 방법으로 사용될 수 있다.

- 생성모델은 판별모델보다 이용자 수가 많다. 생성모델은 다양한 목적과 방법으로 사용될 수 있는 반면, 판별모델은 주로 사람 한 명 한 명에 대해 의사결정을 내리기 위한 용도로 사용되기 때문이다.

– 생성모델은 학습데이터로 주로 인터넷에 공개된 데이터를 활용하기 때문에 학습데이터 구축 과정에서 저작권 침해 등 다양한 위법행위가 발생할 수 있다.

– 생성모델은 판단의 정확성을 우선시하지 않는다. 판별모델은 성능의 1차적인 척도가 판단의 정확성이지만, 생성모델은 AI 시스템이 사람이 만든 문장이나 이미지 등 학습데이터와 얼마만큼 유사한 품질의 결과물을 출력하는지가 중요하다.

이에 생성모델은 다음과 같이 기존 AI 거버넌스 체계에 부합하기 어려운 성격을 갖고 있다.

우선 생성모델은 선제적으로 위험 수준을 상정하기 어렵다. EU AI 법안과 금융위 AI 안내서는 AI 시스템이 사용될 목적 및 방법에 따라 사전에 위험 수준을 정하는 방식을 적용하고 있는데, 생성모델은 최종 활용 단계에서 사용될 목적과 방법이 정해질 수 있기 때문이다. 위험 수준이 늦게 확정될수록 생성모델의 위험을 관리하기 위한 보호조치도 그만큼 늦게 정해지고 미흡해질 수밖에 없다.

그리고 기존 AI 거버넌스 체계는 사회 구성원들에게 광범위한 위해를 가하는 경우에 대해서 충분한 대응 절차를 마련하고 있다고 보기 어렵다. 기존 체계는 주로 개별 의사결정으로 인하여 개인에게 발생할 수 있는 위험을 관리하는 데에 초점을 맞추고 있는데, 생성모델이 거짓 또는 저속한 정보를 출력한다면 개인이 아닌 사회 단위에서 문제가 발생할 수 있다.

또한, 앞서 본 학습데이터의 저작권 침해나 환각 현상 등 생성모델에서 지적되고 있는 위험 요소에 대해서는 해결 기준과 대응 방안에 대한 연구가 필요한 상황이다. 여기서 저작권 침해와 같이 법적으로 불명확한 영역에 대해서는 법제도 정비를 통한 해결 방안과 이를 AI 시스템에 적용하기 위한 AI 거버넌스 방안이 함께 논의될 필요가 있다.

마지막으로 금융위 AI 안내서와 같이 활용자인 일반 사업자가 AI 서비스 공급자에게 거버넌스 준수를 요구하는 것은 서비스 제공 현실과 거리가 있다. 현재 생성모델 AI 서비스는 고성능 컴퓨팅이 필수적이어서 글로벌 사업자들 주도로 클라우드 기반 서비스 형태로 제공되고 있는데, 일반 사업자가 공급자인 글로벌 사업자에게 자사의 위험관리정책을 준수하도록 요구하는 것은 어렵기 때문이다.

생성모델에 관한 AI 거버넌스 연구의 필요성

이처럼 기존의 AI 거버넌스 체계는 최근 사회적으로 급부상한 생성모델과 부합하지 않는 부분들이 있지만, 그렇다고 효용이 없어지는 것은 아니다. 국내외 AI 거버넌스 체계들은 AI 윤리 개념이 등장한 이후 이를 실제 현장에 적용하기 위해 오랜 기간 전세계 전문가들이 연구한 결과물이다. 그리고 민간영역과 공공영역 모두 판별모델 기반의 AI 시스템이 빠른 속도로 사람의 의사결정을 대체해 가고 있기 때문에 AI 거버넌스 구축의 필요성은 갈수록 중요해지고 있다. 다만 판별모델과 생성모델의 작동 방식과 잠재적 위험의 차이를 고려하면서 정합성을 확보할 수 있는 새로운 AI 거버넌스를 마련해야 한다는 커다란 도전과제가 놓인 것이다.

생성모델은 스마트폰의 등장처럼 우리의 일상 생활을 근본적으로 변화시킬 것이라고 기대되고 있다. 하지만 우리 삶의 일부가 되어가는 만큼 그로 인하여 발생할 수 있는 위험들을 관리할 수 있는 AI 거버넌스 방안에 대해 논의가 진행되어야 한다. 최근 EU 의회는 허위 정보를 출력할 우려가 있는 생성모델을 고위험으로 분류하는 등 범용 AI 시스템에 관하여 엄격한 관리 요건을 부여하는 내용의 EU AI 법안 개정안을 준비하고 있다. 국내에서도 AI 산업 현장의 목소리를 반영하고 실증적인 데이터에 기반한 새로운 AI 거버넌스 체계가 수립될 수 있도록 활발한 연구와 지원이 필요한 시점이다.

* 이 글의 내용은 아시아연구소나 서울대의 견해와 다를 수 있습니다.

최신 관련 자료

- 금융위원회 (2022). 금융분야 AI 개발·활용 안내서.
- Council of the European Union (2022). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – General approach.*

Tag: 챗GPT,생성모델,언어모델, AI거버넌스, AI법

마경태(kyungtae.ma@bkl.co.kr)

현) 법무법인(유한) 태평양 파트너 변호사, 방송통신위원회 지능정보사회 이용자 보호 민관협의회 위원

주요 저서와 논문: "AI 투명성 거버넌스와 법제 정비 과제" 『DAIG 매거진』 (2), 2021.
『2020 지능정보사회 이용자 보호 환경조성 총괄보고서』 (공저), (정보통신정책연구원, 2020)

발행처: 서울대학교 아시아연구소, HK+메가아시아연구사업단

발행인: 박수진 **편집위원장:** 박수진 **편집위원:** 이명무, 김윤호
편집간사: 김정희 **편집조교:** 박효진, 전민규, 민보미, 최태수, 김용재 **디자인:** 박종홍

연락처: 02-880-2087, snuac.issuebrief@gmail.com

아시아 브리프의 목표

- 아시아의 현안 분석과 정책적 함의 제시
- 한국의 아시아 진출 전략 개발
- 메가 아시아 건설을 위한 공론장