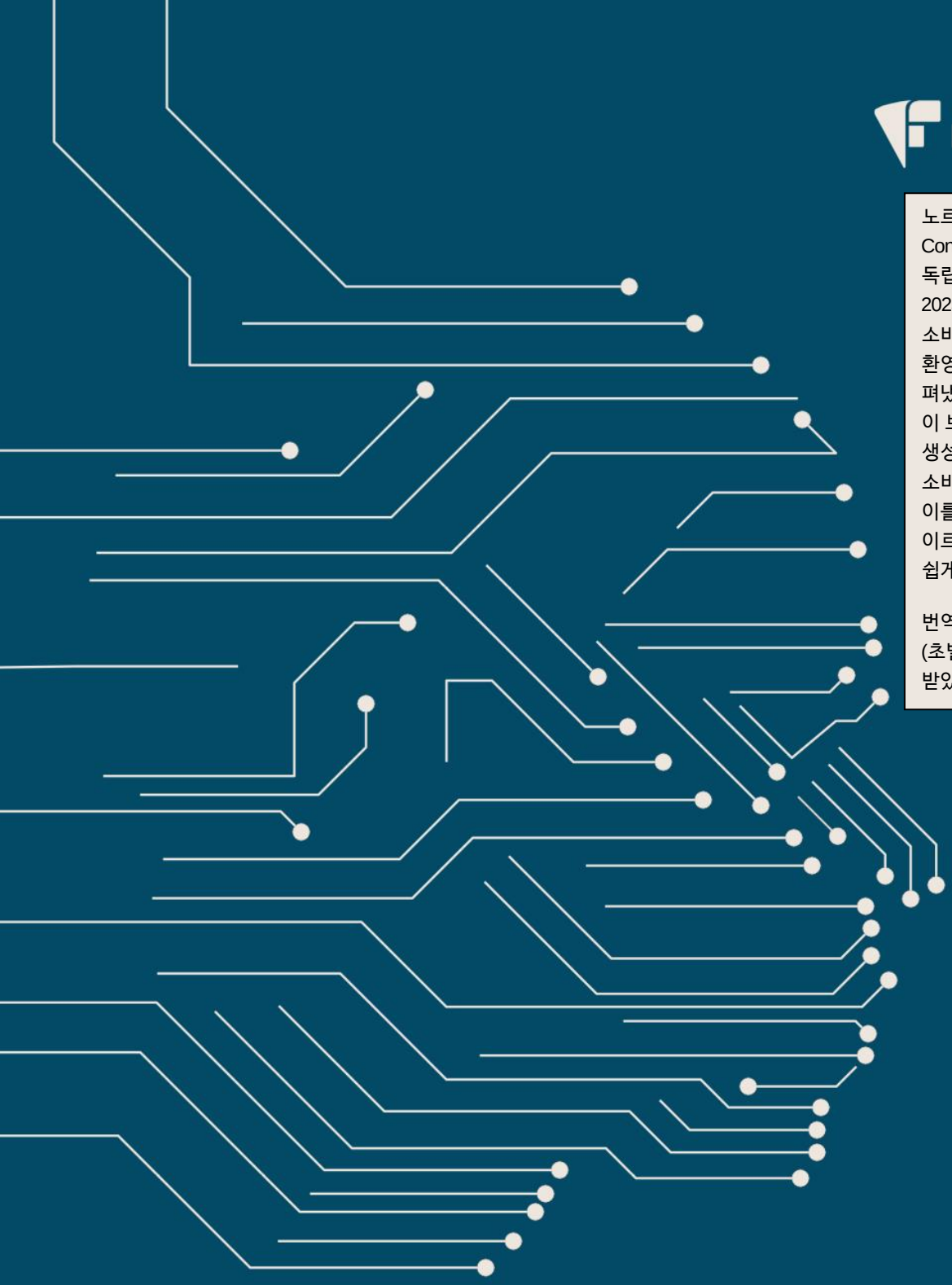


FORBRUKERRÅDET

노르웨이 소비자위원회(Norwegian Consumer Council, 1953년 설치된 독립적인 소비자보호 국가기관)는 2023년 6월 생성형 인공지능의 소비자 문제를 다룬 <기계 속의 환영 (Ghost in the machine)>을 펴냈습니다.
이 보고서는 소비자를 위하여 생성형 AI의 개요로부터 소비자에게 발생할 수 있는 피해와 이를 해결할 수 있는 규제 방안에 이르기까지 방대한 주제를 비교적 쉽게 설명하고 있습니다.

번역: 정보인권연구소
(초벌번역은 기계번역의 도움을 받았습니다. 일부 각주 제외)

기계 속의 환영

생성형 AI로 인한 소비자 피해에 대한 대응

목차

Ghost in the machine – Addressing the consumer harms of generative AI
기계 속의 환영 - 생성형 AI로 인한 소비자 피해 문제 해결

Norwegian Consumer Council
노르웨이 소비자위원회

2023 년 6월

www.forbrukerradet.no/ai

디자인: 폰 커뮤니카스욘

요약

소비자 대상 생성형 AI 서비스가 폭발적으로 증가하고 있습니다. 이들 애플리케이션은 사람이 만든 콘텐츠와 매우 유사한 합성 텍스트, 이미지, 사운드 또는 비디오를 생성하는 데 사용될 수 있습니다. 생성형 AI 시스템이 대중적인 플랫폼과 도구에 통합됨에 따라 이들 기술의 도입은 계속 확대될 것입니다. 한편, 여러가지 새로운 문제들도 발생하면서 생성형 AI의 안전성, 신뢰성, 공정성을 보장하는 방안에 대하여 수많은 논의가 촉발되고 있습니다.

이 보고서는 이러한 논의에 기여하기 위해 작성되었으며, 소비자 권리 및 인권을 희생시키지 않는 생성형 AI의 견실한 출발점을 정책 입안자, 입법자, 규제 기관 및 기타 관련 기관들에 제시하는 것을 목표로 합니다. 앞으로 몇 달, 몇 년 후에 기술이 어떻게 발전할지 확실하게 알 수는 없지만, 기술 발전의 방향은 사회적 관점에서 이루어져야 한다고 믿습니다. 따라서 우리는 생성형 AI 시스템을 어떻게 소비자와 인간 중심의 방식으로 개발하고 사용할 수 있는지를 정의하는 데 도움이 될 수 있는 몇 가지 중요 원칙을 제시합니다.

또한 이미 확인된 자동화 시스템의 폐해에 대해서는 각국 정부, 규제 기관, 정책 입안자들이 기존 법률과 제도를 활용하여 지금 당장 행동에 나설 것을 강력히 촉구합니다. 새로운 제도와 안전장치가 더불어 개발되어야 하겠지만, 향후 몇년 기술이 적절한 견제와 균형 없이 출시되는 것을 소비자와 사회가 지켜보고만 있을 수는 없습니다.

이 보고서의 첫 번째 장에서는 몇 가지 예시와 묘사를 통하여 생성형 AI와 그 용도에 대하여 설명합니다.

2장에서는 생성형 AI에 있어 현재 또는 새롭게 등장하는 다양한 문제, 위험, 피해에 대해 요약합니다. 여기에는 다음 문제가 포함됩니다.

- 권한, 투명성 및 책임성
- 잘못되거나 부정확한 출력
- 기술을 사용하여 소비자를 조작하거나 오도하는 행위
- 편향성과 차별
- 개인정보 보호 및 개인 무결성
- 보안 취약성
- 인간 작업의 자동화
- 환경 영향
- 노동 착취

3장에서는 생성형 AI 시스템의 개발, 배치 및 사용에 적용할 수 있는 현행 및 장래 규칙과 규정들의 교차적용(patchwork) 현황을 개괄합니다. 이 내용들은 대부분 EU 법규를 중심으로 다루었지만, 미국에서 진행 중인 절차에 대해서도 일부 언급합니다.

마지막 장은 생성형 AI의 문제점을 해결하는 방안에 대하여 다양한 권고 사항을 담고 있습니다. 여기에는 다음이 포함됩니다.

- 기존 법률 및 규정의 시행
- 규제 기관을 위한 충분한 자원 확보
- 소비자 보호 강화
- 강력한 정부 정책
- 새로운 입법 조치
- 생성형 AI 시스템의 개발자 및 배치자(deployer)에 적용되는 강력한 의무

1.소개

소비자 대상 AI 시스템은 수십 년 동안 다양한 형태로 사용되어 왔으며, 소셜 미디어 개인화, 이메일 필터링, 스트리밍 콘텐츠 추천, 텍스트 번역 등에 사용되고 있습니다. 이들 목적 중 일부는 친절하고 사려깊은 것이어서, 대부분의 사람들은 자신이 AI 기반 시스템과 상호작용하고 있다는 사실조차 인식하지 못할 수 있습니다.

제너레이티브 인공지능(generative artificial intelligence, '생성형 AI') 시스템의 대량 배치와 도입으로 소비자 대상 애플리케이션에 AI 기반 시스템의 새로운 물결이 빠르게 다가오고 있습니다. 생성형 AI는 텍스트, 이미지, 오디오 또는 비디오와 같은 합성 콘텐츠를 생성할 수 있는 AI의 하위 집합으로, 사람이 만든 콘텐츠와 매우 유사할 수 있습니다. 이러한 시스템은 오늘날 소비자가 접하는 많은 인터페이스와 콘텐츠를 변화시킬 수 있습니다.

2022년 11월, 챗봇 챗GPT(ChatGPT)의 프로토타입이 대중에게 공개되었습니다. 이 애플리케이션은 출시 한 달 만에 최고로 빠르게 성장하는 디지털 서비스가 되었고 세계적으로 급격한 주목을 받았습니다. 그 후 몇 달 동안 텍스트, 이미지, 사운드, 비디오를 생성하는 다른 서비스들이 신속하게 배치되는 일이 반복되면서 생성형 AI 시스템에 대한 일종의 군비 경쟁이 촉발하였습니다. 소비자는 웹 인터페이스에서 직접 이러한 콘텐츠 생성기에 접근할 수 있게 되었고, 기업은 애플리케이션과 서비스에 콘텐츠 생성기를 내장하기 시작했습니다.

생성형 AI 시스템의 갑작스럽고 광범위한 배치와 도입은 이 기술의 유망함과 위험에 대한 대중의 담론을 촉발시켰습니다. 이러한 논의들은 생성형 AI를 사용하여 인력의 효율성을 높이고 창의성을 촉발하는 방법으로부터 허위조작정보를 퍼뜨리고, 개인과 사회를 조작하고, 일자리를 빼앗고, 아티스트의 저작권에 문제를 일으키는 문제에 이르기까지 다양합니다.

전 세계 정책 입안자들이 생성형 AI의 가능성과 문제에 대처하기 위해 노력하면서 이들 시스템을 통제하거나 규제하는 방법에 대한 논의가 계속되고 있습니다. 이 보고서는 소비자 관점에서 가장 시급한 문제에 대한 분석과 함께 법적, 윤리적, 정치적 관점에서 여러가지 가능한 해결책과 지속 방안을 제시함으로써 이러한 논의에 기여하고자 합니다. 비록 우리가 생성형 AI가 제기하는 모든 질문에 대한 해답을 가지고 있다고 말할 수는 없지만, 우리는 소비자와 인간 친화적인 방향으로 기술을 이끌어 나가기 위해 고안된 규제, 집행 및 구체적인 정책의 조합을 통해 새롭게 등장하거나 진행 중인 여러 문제도 해결할 수 있다고 믿습니다.

생성형 AI의 개발은 빠른 속도로 진행되고 있으므로 이 보고서의 설명은 새로운 기술의 단편적인 일면(snapshot)으로 간주해야 합니다. 이 보고서는 2023년 2월부터 5월 사이에 작성되었으며, 6월 1일 이후에 발표된 새로운 논문의 정보는 포함하지 않았습니다.

노르웨이 소비자위원회는 공적 자금을 지원받는 독립 소비자 기구로, 소비자의 이익을 대변합니다. 이 보고서는 BEUC, VZBV의 미카 블린, 아일랜드 시민자유위원회의 크리스 슈리샤크, Access Now의 다니엘 로이퍼, 존 워스, 마리야 슬라브코빅, 베르겐 대학교의 안자 살츠만의 참여로 작성되었습니다.

1.1 생성형 AI의 개요

생성형 AI는 텍스트, 이미지, 사운드와 같은 새로운 데이터를 생성하도록 학습된 알고리즘 모델을 설명하는 데 사용되는 포괄적인 용어입니다. 이들 애플리케이션은 서로 다른 유형의 입력 데이터에 의존하지만, 학습 방법의 일반적인 원칙은 비슷합니다. 첨단 생성형 AI의 등장은 인터넷에서 구할 수 있는 방대한 양의 콘텐츠 및 머신 러닝과 컴퓨팅 성능의 발전이 결합되어 가능해졌습니다.

생성형 AI 모델은 대량의 정보를 분석하여 문장의 다음 단어, 이미지의 특징 등을 예측하고 생성하는 방식으로 작동합니다. 이는 학습 데이터에서 데이터 포인트 간의 패턴과 관계를 탐지하여 수행되며, 이를 통해 시스템은 유사한 패턴을 복제하여 글, 음악 또는 비디오 클립과 같은 합성 콘텐츠를 생성할 수 있습니다. 이 프로세스는 시스템이 학습한 데이터에서 가져온 콘텐츠의 복잡한 '매시업'으로 설명할 수도 있습니다.

즉, 이들은 예측 모델로서 기존 콘텐츠의 데이터 포인트 간에 '점을 연결'하여 합성 콘텐츠를 생성하도록 학습된 것입니다.

생성된 콘텐츠는 일반적으로 사람이 작성하는 특정 입력(또는 '프롬프트')을 기반으로 확률적, 무작위로 생성됩니다. 따라서 특정 생성형 AI 모델의 출력은 모델에 프롬프트를 보내는 사람마다 다를 수 있으며, 학습 데이터의 패턴과 유사하거나 완전히 새로운 것을 나타낼 수 있습니다.

1.1.1 생성형 AI 모델의 사례

새로운 텍스트를 생성하여 텍스트에 응답할 수 있는 대규모 언어 모델(LLM), 두 가지 이상의 출력 유형을 생성하거나 두 가지 이상의 입력 유형에 응답할 수 있는 다중 모드 모델(예: 프롬프트가 지시하면 이미지도 생성할 수 있는 챗봇) 등 다양한 유형의 생성형 AI 모델이 존재합니다. 현재 시장에서 가장 인기 있는 생성형 AI 모델에 대한 간략한 소개와 함께 몇 가지 관련 예시를 이하에 제시합니다.

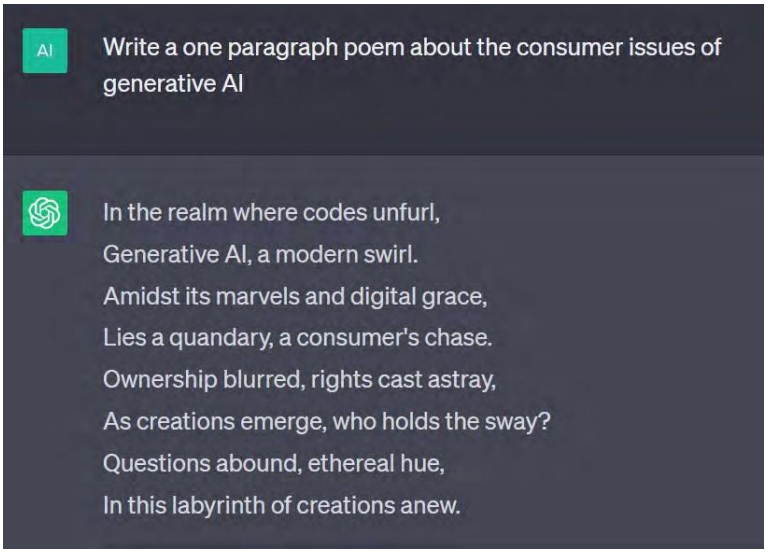
1.1.1.1 텍스트 생성기

텍스트 생성기는 예측 분석을 기반으로 텍스트 구절을 생성할 수 있는 생성형 AI 모델의 일종으로, LLM을 기반으로 구축됩니다.[2] 이러한 모델은 일반적으로 서적, 포럼, 뉴스 사이트, 소셜 미디어 등 인터넷에서 스크랩한 방대한 양의 데이터에 대해 학습합니다. 텍스트 생성기는 주로 에세이 작성, 코딩, 챗봇, 검색 엔진 보강 등에 사용할 수 있습니다. 대부분의 경우 텍스트 생성기는 1인칭 시점으로 작성된 텍스트를 생성하거나, 이모티콘을 사용하거나, 사람의 감정을 느낄 수 있는 텍스트를 작성하는 등 사람이 작성한 것처럼 보이는 텍스트를 생성하기 위한 것입니다. 일부 텍스트 생성기는 다중 모드이며 이미지에 기반하여 텍스트를 생성할 수 있습니다.

텍스트 생성기는 문자메시지 예측입력도구 등 몇 년 동안 어떤 형태로든 존재해 왔지만, 이 기술에 대한 논의는 2022년 가을, 오픈AI(Open AI, DALL-E의 소유주이기도 합니다. 이하 참조)가 소유 및 운영하는 챗GPT의 공개 출시로 탄력을 받게 되었습니다. 챗GPT3는 사람들이 온라인에서 무료 계정을 만들어 이용할 수 있고, 더 강력한 챗GPT4 모델은 월 구독료로 이용할 수 있습니다.[3]

2023년 1월 마이크로소프트(Microsoft)는 챗GPT에 대한 대규모 투자를 발표하고 빙(Bing) 검색 엔진에 이 기술로 구동되는 새로운 기능을 출시했습니다.[4] 마이크로소프트는 마이크로소프트 Teams 제품으로 회의 중 자동으로 메모를 작성하는 등 마이크로소프트 Office 애플리케이션 제품군을 포함한 다른 서비스에 챗GPT를 통합할 계획이라고 발표했습니다.[5]

구글(Google) 또한 텍스트를 생성할 수 있는 LaMDA라는 LLM을 개발했습니다. 마이크로소프트가 챗GPT에 투자한 이후 구글은 자사 검색 엔진에서 Bard라는 텍스트 생성기를 이용하는 유사한 기능을 출시했습니다.[6] 구글은 또한 이메일 초안 작성 및 요약, 브레인스토밍 및 문서 작성과 같은 다양한 AI 기반 기능을 Workplace 애플리케이션에 도입할 계획입니다.[7]



생성형 AI의 소비자 문제에 대한 시, 챗GPT.

2 LLM은 인간의 언어와 유사한 텍스트를 생성하도록 설계된 정교한 AI 모델입니다. 일반적으로 방대한 양의 텍스트 소스를 학습하여 패턴과 문법을 '학습'합니다. LLM는 기계 번역, 감정 분석, 인간과 기계의 상호작용, 교정 등 다양한 목적의 작업에 사용할 수 있습니다

메타(Meta)는 "과학 지식을 저장, 결합 및 추론"하기 위해 과학 관련 기사와 자료로 학습된 LLM Galactica를 개발했습니다. 이 모델이 2022년 11월 공개 데모로 공개된 후, 여러 오류와 편향성이 포함된 텍스트가 생성되었고 공개 데모는 빠르게 삭제 되었습니다.[8] 2023년 2월, 메타는 LlaMa(Large Language Model Meta AI)라는 또 다른 LLM을 출시했습니다. LlaMa는 오픈 소스 모델로, 처음에는 접근 신청 절차를 통해 연구자들에게 공개되었습니다. 2023년 3월, 이 모델이 공개 게시판에 유출되면서 비교적 성능이 좋은 컴퓨터만 있으면 누구나 이 모델을 다운로드하여 사용하고 변형할 수 있게 되었습니다.[9]

소규모 주체가 개발하고 유지 관리하는 오픈 소스 LLM도 여러 개 있습니다. 예를 들어, 텍스트 생성기인 BLOOM은 Hugging Face라는 회사를 통해 제공되며,[10] 스테빌리티 AI는 StableLM이라는 이름으로 오픈 모드를 출시했습니다.[11]

1.1.1.2 이미지 생성기

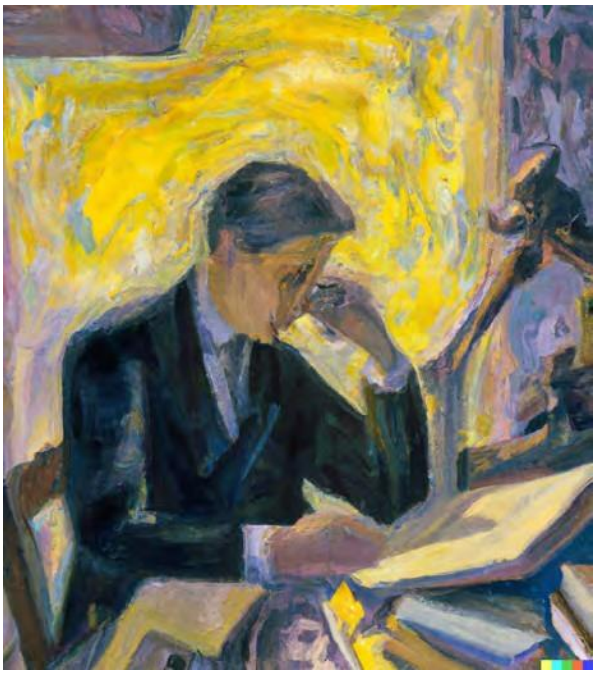
이미지를 생성하도록 학습된 생성형 AI 모델은 이미지 생성기로 통칭할 수 있습니다. 텍스트 프롬프트('텍스트 to 이미지'형)나 기존 이미지('이미지 to 이미지'형)에서 이미지를 생성할 수 있습니다. 이미지 생성기는 인터넷의 다양한 소스에서 스크랩한 사진, 그림 등 방대한 양의 기존 사진을 분석하여 작업합니다. 이러한 데이터세트에 대하여 알고리즘을 학습시킴으로써 모델은 다양한 사물('의자', '기차'), 사람('젊은 여성', '제리 세인필드'), 스타일('인상주의', '에드워드 뭉크 스타일')의 이미지를 생성할 수 있습니다. 2023년 6월 현재 가장 널리 사용되는 이미지 생성기는 미드저니(Midjourney),[12] DALL- E,[13] 및 스테이블 디퓨전(Stable Diffusion)입니다.[14]

미드저니는 채팅 서비스 디스코드를 통해 이용할 수 있습니다. 공식 미드저니 디스코드 서비스에 가입하여 미드저니 봇에게 다양한 프롬프트에 따라 그림을 '상상'해달라고 요청할 수 있습니다. 예를 들어, "“/imagine hyper realistic photo of political advisor writing a paper on generative ai, in an open office plan(개방형 사무실에서 생성형 AI에 관한 보고서를 작성하는 정치 고문의 극사실주의 사진을 상상해줘)"라는 프롬프트를 입력할 수 있습니다. 봇은 생성된 네 장의 사진을 채팅으로 응답합니다. 미드저니는 출시 후 처음 몇 달 간은 생성 이미지 수가 제한적이고 무료로 체험할 수 있었지만, 얼마 후 유료 구독 서비스로 전환되었습니다.

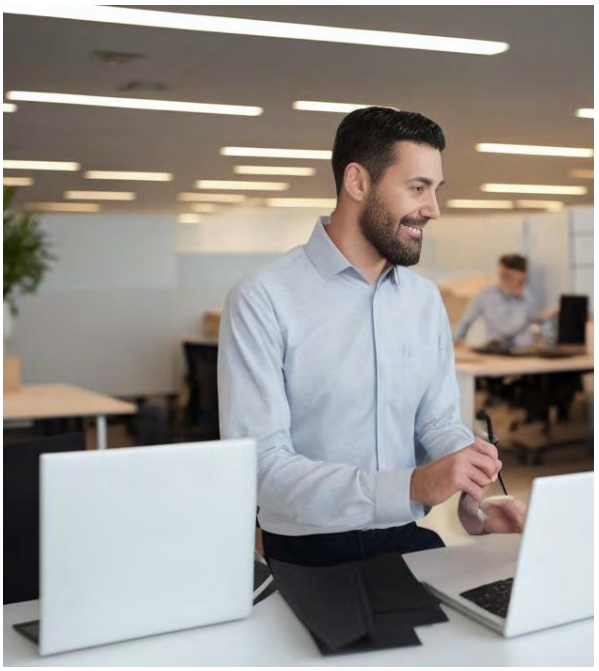
미드저니 주식회사는 이미지를 생성하는 생성형 AI 모델을 소유하고 있으며, 모델 자체와 모델이 호스팅되는 서버를 모두 운영하고 통제합니다. 즉, 이 회사가 접근을 제한하고, 모델을 변경하고, 콘텐츠 필터를 추가하여 모델이 생성하거나 생성할 수 없는 이미지의 종류를 통제할 수 있음을 의미합니다.

생성형 AI 모델인 DALL-E는 소유주인 오픈AI의 웹사이트를 통해 접근할 수 있습니다. 개인이 계정을 만들면 이미지 생성에 사용할 수 있는 한정된 수의 토큰을 매달 받을 수 있습니다.

개방형 사무실에서 생성형 AI에 관한 보고서를 작성하는 정치 고문의 초현실적인 사진, 미드저니.



뭉크가 생성형 AI에 관한 보고서를 쓰고 있는 정책 고문을 그린 유화 한 점, DALL-E.



한 소비자 정책 고문이 개방형 사무실에서 생성형 AI의 소비자 문제에 대한 보고서를 작성하는 사진, 스테이블 디퓨전 1.5.

웹사이트 인터페이스에 여러 프롬프트를 입력하면 이미지들이 생성됩니다. 토큰이 부족할 경우 결제하여 추가 토큰을 받을 수 있습니다. 미드저니와 마찬가지로 DALL-E의 모델은 모회사가 소유, 통제, 운영합니다.

생성형 AI 모델인 스테이블 디퓨전은 스테빌리티 AI라는 회사에서 개발했습니다. 스테이블 디퓨전은 미드저니나 DALL-E와 달리 누구나 자유롭게 다운로드할 수 있는 오픈 소스 모델이며, 구독이나 인터넷 접속 없이도 사용할 수 있습니다. 필요한 소프트웨어만 다운로드하면 로컬에서 이미지를 무제한으로 생성할 수 있습니다. 스테이블 디퓨전을 로컬에서 실행하려면 비교적 강력한 일반 소비자용 그래픽 카드가 장착된 컴퓨터만 있으면 됩니다.

스테빌리티 AI는 스테이블 디퓨전을 위한 기본 모델만 학습시키고 배포하였고, 접근 권한이 있는 사람은 누구나 기존 스테이블 디퓨전 모델을 기반으로 새로운 모델을 계속 학습시키고 개발할 수 있습니다. 이후 이 새로운 모델을 다른 사람들에게 배포할 수도 있습니다. 이는 회사가 모델이나 그 결과물을 실제 통제하지 않는다는 사실을 의미합니다.

1.1.1.3 오디오 생성기

오디오 생성기는 생성형 AI 기술을 사용하여 텍스트 프롬프트에 기반하여 오디오 클립을 생성합니다(예를 들어, 텍스트 to 음성). 이러한 모델은 기존 음성 데이터, 음악 등에 대해 학습합니다. 오디오 생성기를 사용하여 AI 생성음악[15] 및 음성을 생성할 수 있으며, 개별 개인의 목소리와 음정을 재현할 수 있는 모델도 있습니다.[16]

예를 들어, ElevenLabs는 누구나 다양한 목소리를 선택하여 짧은 텍스트 입력을 음성 클립으로 변환할 수 있는 모델을 출시했습니다.[17] 마이크로소프트는 3초 분량의 음성 샘플을 기반으로 사실적인 목소리를 생성할 수 있다고 주장하는 생성형 AI 모델 VALL-E를 발표했습니다.[18] 2023년 5월 현재 VALL-E는 아직 일반에 공개되지 않았습니다.

1.1.1.4 비디오 생성기

비디오 생성기는 텍스트 프롬프트(텍스트 to 동영상), 이미지(이미지 to 동영상) 또는 기존 클립(동영상 to 동영상)을 기반으로 동영상 클립을 만드는 데 사용할 수 있습니다. 정지 이미지를 만드는 것보다 실제와 같은 비디오 영상을 생성하는 것이 더 복잡하기 때문에 현재 시점에 이 기술은 다소 덜 개발되어 있습니다.

그러나 여러 주요 기업이 동영상 생성 모델을 적극적으로 개발하고 있기 때문에 가까운 시일에 상황이 바뀔 수도 있습니다.

메타 AI [19] 메이크어비디오에서 생성된 비디오 프레임

메타는 짧은 텍스트를 동영상 클립으로 변환하는 모델을 개발했으며,[20] 구글도 유사한 시스템을 발표했습니다.[21] 2023년 5월 현재 이 두 시스템 모두 대중에게 공개되어 있지는 않습니다.

스테이블 디퓨전의 개발사인 스테빌리티 AI는 텍스트 프롬프트와 이미지에서 애니메이션을 생성하는 모델을 출시했습니다.[22] 런웨이라는 회사는 기존 동영상에서 짧은 동영상 클립을 생성하는 데 사용할 수 있는 모바일 앱을 출시했습니다.[23]

1.1.2 생성형 AI 행위자망

모델 학습을 위한 데이터세트 어셈블링부터 생성형 AI 시스템 배치 및 프롬프트 입력에 이르기까지, 잠재적으로 다양한 행위자가 있을 수 있습니다. 이들 행위자들은 모두 다양한 방식으로 시스템에 영향을 미치거나 시스템이 사용되는 방식에 영향을 미칠 수 있습니다. 관련 행위자는 아래 그림에 나와 있습니다.

각 상자는 행위자망에서 서로 다른 행위자를 나타냅니다. 이들 행위자는 모두 한 조직 내에 있을 수도 있지만 여러 조직에 분산되어 있는 경우도 많습니다. 단순한 표현을 위해서 화살표가 한 방향을 가리키고 했지만, 당연히 서로 다른 행위자 간에도 피드백 루프가 있을 수 있습니다.

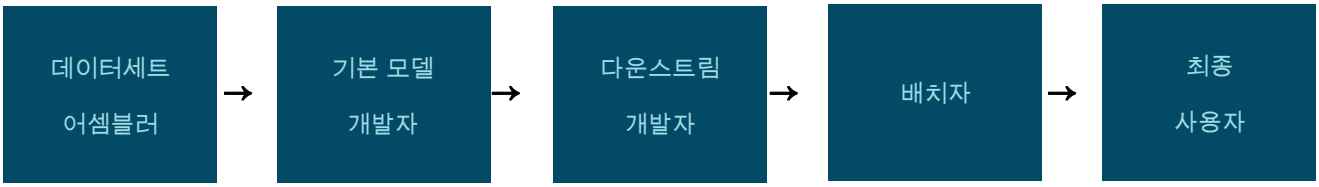
데이터세트 어셈블러는 생성형 AI 모델을 학습시키는 데 필요한 데이터를 수집하고 시스템화합니다. 사용 가능한 오픈 소스 데이터세트는 온라인에서 수많은 소스를 스크랩하여 수집하고 라벨을 붙인 것입니다. 대부분의 경우 이러한 데이터세트는 연구 목적으로 컴파일되어 무료로 제공됩니다. 따라서 생성형 AI 모델을 개발하는 기업은 다른 사람이 수집한 데이터세트로 모델을 학습시킬수 있습니다. 생성형 AI 모델의 개발자들은 기본 모델(baseline model)을 만들고, 이 모델은 다운스트림 개발자들에 의해 특정한 더 구체적인 맥락이나 애플리케이션에 대해 학습되고 조정될 수 있습니다.

경우에 따라서는 이러한 모델 미세 조정(fine-tuning)이 기본 모델 학습과 동일한 행위자에 의해 수행될 수도 있고, 다른 경우 전적으로 별개의 행위자에 의해 수행될 수도 있습니 다. 이 행위자는 다른 회사일 수도 있고, 오픈 소스 모델의 경우 비교적 강력한 컴퓨터를 가진 사람 누구나 모델을 미세 조정할 수 있습니다.

문제를 더욱 복잡하게 만드는 것은 범용 생성형 AI 모델이 다른 애플리케이션에 통합되고 있기 때문에 시스템을 배치하는 회사나 주체가 해당 모델을 개발하거나 미세 조정하는 회사와 별개일 수 있다는 점입니다.

마지막으로 배치된 모델에 참여하는 최종 사용자가 있습니다. 소비자 사용 사례에서 소비자는 모델에 텍스트나 이미지 등을 생성하도록 요청하는 등 일반적으로 행위자가 됩니다. 고객 서비스 상담원이 텍스트 생성기를 사용하여 소비자 질문에 대한 답변을 생성하거나, 소비자가 AI가 생성한 추천에 따라 특정한 질문을 생성하기 위하여 프롬프트를 입력하는 등의 사례에서 소비자는 기업들과 상호 작용할 때 생성형 AI 시스템을 간접적으로 접할 수도 있습니다.

생성형 AI 시스템 행위자망에는 수많은 행위자가 존재합니다. 이러한 행위자 간 관계를 이해하여 생성형 AI를 어떻게 규제할 것인지, 행위자망의 어느 시점에 어떤 해악이 발생하는지 파악하는 것이 중요합니다.



생성형 AI 행위자망에 포함된 다양한 행위자 소개

1.1.3 오픈 소스 또는 폐쇄 소스 모델

생성형 AI 모델이 배포되고 통제되는 방식에는 상당한 차이가 있습니다. 대개의 모델은 시스템을 소유한 자("시스템 소유자")가 통제하는 클라우드 서버에서 실행되는 독점적인 비공개 소스 모델입니다. 즉, 소비자는 인터넷을 통해 모델에 접근할 수 있으며 시스템 제공업체는 언제든지 모델을 변경하고, 콘텐츠 필터를 추가하고, 접근을 다시 엄격하게 제한할 수 있습니다. 폐쇄형 시스템의 경우 시스템 소유자는 모델을 학습시키고 합성 콘텐츠를 생성하는 데 필요한 처리 능력을 제공합니다.

비공개 소스 생성형 AI 모델의 경우, 모델 배후의 회사가 충분한 설명서를 게시하거나 제3자 감사, 규제 기관 또는 연구자에게 접근 권한을 제공하지 않는 한, 모델이 어떻게 작동하는지, 어떤 데이터로 학습되었는지, 매개변수의 가중치가 어떻게 부여되는지 알 수 없습니다. 대부분 이러한 정보는 보안 사유 및 영업 상 이해관계로 인해 비밀로 유지될 수 있습니다.

반면에 일부 생성형 AI 모델은 오픈 소스로 공개되며 다양한 형태를 취할 수 있습니다. 데이터세트, 소스 코드, 모델 매개변수, 가중치 등 시스템의 다양한 부분이 제3자에게 제공될 수 있습니다.

소스 코드가 공개되면 누구나 사용하고, 연구하고, 테스트하고, 수정하고, 배포할 수 있습니다. 이는 오류나 취약점이 있는지 검사할 수 있다는 의미입니다. 또한 공동 작업을 통해 개선하고 반복적용할 수 있습니다. 오픈 소스 소프트웨어의 사용, 수정 및 배포는 일반적으로 라이선스 약관의 적용을 받습니다. 오픈 소스 소프트웨어와 데이터세트가 주어지면 충분한 컴퓨팅 리소스를 갖춘 사람 누구나 생성형 AI 모델을 재현할 수 있지만, 이를 위해서는 컴퓨팅 리소스가 상당히 필요하기 때문에 실제로는 큰 회사로 제한될 가능성이 높습니다.

그러나 생성형 AI 시스템이 진정한 오픈 소스가 되려면 모델 자체가 대중에게 공개되어야 합니다. 이 경우 이들 모델을 기반으로 개발된 애플리케이션은 이미지 생성기 스테이블 디퓨전과 같이 오픈 소스 애플리케이션으로 공개될 수도 있고, 비공개 소스 애플리케이션으로 개조될 수도 있습니다.

오픈 소스 생성형 AI 모델은 누구나 다운로드할 수 있습니다. 충분히 강력한 컴퓨터만 있으면 원하는 대로 데이터를 생성하고 모델을 업데이트하는 데 사용할 수 있습니다.

이러한 모델의 소스 코드, 매개 변수 등은 누구나 검사할 수 있습니다. 그러나 이는 모델이 어떻게 작동하는지 반드시 이해할 수 있다는 의미는 아닙니다. 상당히 복잡한 모델의 경우 내부 작동 방식을 이해하기 어려울 수 있기 때문입니다.

오픈 소스 생성형 AI 모델이 대중에게 공개되면 기본 모델 개발자가 그 모델의 작동 방식에 영향을 미칠 수 있는 방법이 사실상 없습니다. 이는 모델에 적용된 콘텐츠 필터 및 기타 인위적인 제한 사항을 다운스트림 개발자나 배치자가 변경하거나 제거할 수 있다는 의미입니다. 이로 인한 장점과 심각한 단점이 함께 생기는데, 이에 대해서는 이하에서 자세히 설명하겠습니다.

1.1.4 범용 AI

암세포 조기 발견과 같이 특정 목적과 사용 사례를 염두에 두고 설계된 AI 모델도 있지만, 많은 생성형 AI 모델은 소위 '범용 AI'의 사례들입니다. 이는 LLM과 같은 기본 시스템이 매우 다양한 상황과 상호 작용에 대응할 수 있도록 학습되어 새로운 상황에서 사용할 수 있도록 조정될 수 있다는 의미입니다.

특정 목적을 가진 모델과 달리 범용 AI 모델의 개발자가 기술의 사용 및 악용 가능성을 예측하는 것은 매우 어렵거나 불가능합니다. 따라서 이들 모델이 널리 도입되기 전에 기술적, 과학적, 입법적, 규제적 조사를 받는 것이 특히 중요합니다. 그러나 챗GPT와 같은 애플리케이션은 엄격한 평가, 영향 평가 또는 조사 없이 이미 광범위한 대중에게 공개되었으며, 예전보다 더 불투명하고 제3자 감사 및 연구자 접근이 어려워졌습니다.[24] 이 보고서의 2장에서 다루는 여러 폐해를 고려할 때 이러한 상황이 바람직한 미래를 가져올 것인지 검토할 가치가 있습니다.

1.2 소비자 애플리케이션

오늘날 소비자는 여러가지 유형의 생성형 AI를 공개적으로 사용할 수 있습니다. 이들 대부분은 인터넷에 연결되어 있는 사람 누구나 쉽게 사용할 수 있으며, 전문적인 기술 지식이 없어도 사용할 수 있습니다. 일부는 웹 인터페이스를 통해 직접 접근할 수 있으며, 온라인 검색, 교육 및 관리용 소프트웨어, 소셜 미디어와 같은 다른 서비스에도 생성형 AI 기술이 점점 더 많이 통합되고 있습니다.

2023년 5월 현재, 소비자들이 가장 많이 사용하는 생성형 AI 모델은 텍스트와 이미지 생성입니다. 하지만 마이크로소프트, 메타, 구글 등 소비자를 대상으로 하는 주요 기업들이 이 기술에 막대한 투자를 하고 있기 때문에 향후 몇 달 안에 다양한 서비스에서 생성형 AI 모델이 구현되고 사용 사례가 확대될 것으로 보입니다.

예를 들어, 텍스트 생성기는 일상적인 작업을 간소화하거나 최적화하는 데 유용한 도구로, 일종의 다목적 디지털 비서 역할을 할 수 있습니다. 여기에는 인터넷 검색 기능 변경이나, 코드 작성, 음성 메시지 변환과 같은 특정 작업 자동화, 또는 다양한 방식으로 개인화된 서비스 등이 포함될 수 있습니다. 이러한 애플리케이션은 특정 상황에서 유용하고 효율적일 수 있지만, 상당한 위험과 단점도 있으며, 이에 대해서는 다음 장에서 자세히 살펴볼 것입니다.

기술이 개발되고 도입됨에 따라 생성형 AI는 간단한 텍스트 작성, 양식 작성, 일정 또는 계획 생성, 소프트웨어 코드 작성 등 이전에는 수작업으로 수행해야 했던 지루하고 시간이 많이 소요되는 과정을 자동화하는 데 사용될 수 있습니다. 이들은 서비스를 더욱 비용 효율적으로 만들 수 있는 잠재력을 가지고 있으며, 법률 자문 의뢰 등에서 소비자 비용을 낮출 수도 있습니다.[25] 반면에 저가의 AI가 생성한 콘텐츠가 확산되면 인간의 노동력과 인간이 생성한 콘텐츠를 대체하여 고객 지원과 같은 소비자 대면 서비스의 품질이 저하될 수 있습니다. 또한 이 기술은 광고 또는 제품 추천과 같은 영역에서 소비자 조작의 새로운 문을 열었으며 차별적 관행을 조장하거나 혼란을 가져올 수도 있습니다.

2. 생성형 AI의 피해 문제

"생성형 AI 모델은 여러 발전을 가져오는 한편으로,
개인정보 보호 및 개인 무결성 침해부터
사기 및 허위조작정보의 생성에 이르기까지,
막대한 위험과 문제를 야기합니다."

생성형 AI의 개발과 사용을 둘러싼 여러가지 논란이 있었습니다. 생성형 AI 모델은 여러 발전을 가져오는 한편으로, 개인정보 보호 및 개인 무결성 침해부터 사기 및 허위조작정보의 생성에 이르기까지, 막대한 위험과 문제를 야기합니다. 구체적이고 관련성이 높은 예시로는 부정확하지만 그럴듯한 정보를 제공하는 챗봇과 검색 엔진, 콘텐츠 검열을 위한 남반구 저임금 노동력 남용, 자원 소비로 인한 심각한 환경 영향 등을 들 수 있습니다. 다양한 부정적 결과로부터 소비자를 보호하는 역할을 하는 법과 규정을 시행, 적용, 확립함으로써 이러한 문제를 적절히 해결하는 것이 중요합니다.

이 보고서에서 논의된 문제는 생성형 AI만의 새로운 문제이거나 고유한 문제가 아닙니다. 알고리즘 컴퓨팅 시스템은 한 세기 동안 존재해 왔으며, 일반적으로 AI라고 불리는 기술은 1950년대부터 사용되어 왔습니다. 1960년대에 컴퓨터 과학자 조셉 바이젠바움(Joseph Weizenbaum)이 엘리자(ELIZA)라는 모델을 만들었는데, 이 모델은 규칙 기반 알고리즘을 사용하여 인간과의 상호작용을 시뮬레이션했습니다.[26] 엘리자와 상호작용한 사람들은 시스템에 인간적 기능이 없다는 사실을 알고 있었음에도 불구하고 인간의 속성과 감정을 모델에 부여하였으며, 이러한 경향은 오늘날 생성형 AI 기반 챗봇의 일부 사용 사례에도 투영됩니다.

디지털 기술이 발전하고 널리 사용됨에 따라 콘텐츠관리, 알고리즘 편향성, 개인정보 보호 및 허위조작정보와 관련된 문제가 거의 모든 접점에서 논의되어 왔습니다. 그러나 챗GPT와 같은 시스템의 배치와 대중적 도입은, 기술적으로 능숙한 소비자와 일반 대중 모두에게 사용의 편리함과 광범위한 가용성을 가져다 주었고, 이는 이들 문제 중 상당수가 소비자 관점에서 분석해야 할 시급한 사안이 되었음을 의미합니다. 다음 장에서 설명하는 바와 같이, 이들 문제 중 일부는 관련 법률 및 규정을 집행함으로써 해결할 수 있지만, 그밖의 문제들은 다른 해결책이나 구제수단을 필요로 할 수 있습니다.

2.1생성형 AI의 구조적 문제

기본적으로 생성형 AI 모델은 잠재적으로 새로운 방식이긴 하지만 기존 자료를 재현하도록 설계되었습니다. 이는 이들 모델이 본질적으로 기존의 편향성과 권력 구조를 재생산하는 경향이 있다는 의미입니다. 따라서 모델 자체의 이해, 사고, 의도는 없지만 모델을 개발, 배치, 사용하는 결정은 본질적으로 정치적입니다. 학습 데이터와 알고리즘은 인간과 이로 인해 수반되는 모든 것에서 비롯되기 때문에 생성형 AI 모델의 작동이나 결과에 중립성이나 객관성을 부여하는 것만으로는 충분하지 않습니다.

생성형 AI 모델이 사회의 모든 분야에 도입되고 있지만, 아직까지 규제 당국의 감독이 거의 또는 전혀 이루어지지 않고 있기 때문에 해결해야 할 근본적인 문제들이 있습니다. 생성형 AI 모델은 다양한 출처에서 가져 온 대량의 데이터에 의존하는데, 예술 작품이든, 뉴스 기사이든, 셀카이든 데이터 원본의 출처자가 인지하거나 동의하지 못한 경우가 대부분입니다. 소수의 기업을 배불리는 것을 최종 목표로 하는 다양한 방식으로 정보를 빼내어 수집합니다. 이로 인해 가치 분배, 사용 권한, 개인정보 보호, 책무성, 지적 재산권 및 인권에 대한 문제가 제기됩니다.[27]

2.1.1 생성형 AI의 구체적인 위험 식별

다른 신기술과 마찬가지로, 생성형 AI에 대한 담론은 사실과 우려, 그리고 많은 과대광고와 열정으로 뒤섞여 있습니다.[28] 많은 AI 시스템이 거의 모든 작업을 해결할 수 있을 것처럼 선전되고 있으며, 주장을 뒷받침할 증거도 없이 'AI 뱀 기름[엉터리 약]'처럼 묘사되는 현상이 나타나고 있습니다.[29] 기술의 문제와 위험성을 다룰 때는 사실과 허구를 구분할 수 있는 능력이 중요합니다.

"기본적으로 생성형 AI 모델은 잠재적으로 새로운 방식이긴 하지만 기존 자료를 재현하도록 설계되었습니다. 이는 이들 모델이 본질적으로 기존의 편향성과 권력 구조를 재생산하는 경향이 있다는 의미입니다."

"가상의 장기 시나리오에 초점을 맞추는 것은
현재 생성형 AI가 야기하는 시급한 여러 문제에 대한
관심을 멀어지게 하여 이들 문제를 충분히 규제하지 않고
방치할 수 있다는 심각한 우려가 있습니다."

용 어	정 의
GENERATIVE ARTIFICIAL INTELLIGENCE 생성형 인공지능	학습 데이터에서 학습한 패턴과 구조를 기반으로 합성 콘텐츠를 생성합니다. 텍스트, 이미지, 오디오, 비디오를 생성하는 데 사용됩니다.
GENERAL PURPOSE ARTIFICIAL INTELLIGENCE 범용 인공지능	다양한 영역에서 광범위한 작업을 수행하도록 설계된 AI 시스템을 포괄하는 용어입니다.
ARTIFICIAL GENERAL INTELLIGENCE 인공 일반지능	인간 수준의 지능과 자율성을 보여준다는 가상의 AI 시스템입니다. 현재는 존재하지 않습니다.

AI의 위험성에 대해 널리 알려진 경고는 인공 일반지능(AGI) 개발에 대한 가상의 위험성에 집중되어 있는데, AGI는 인간의 능력에 필적하는 지적 작업을 수행할 수 있는 시스템을 의미합니다. 이론적으로 이러한 시스템은 사고하고 추론할 수 있어야 하며, 인간의 사고 능력과 동등한 수준의 광범위한 작업을 수행할 수 있어야 합니다. 이는 이러한 능력이 없는 생성형 AI 모델과 큰 차이가 있습니다. AGI 시스템은 현재 존재하지 않으며 실현가능성 여부에 대해서도 심각한 논쟁이 있으므로 이 보고서에서는 이러한 시스템을 더 이상 고려하지 않습니다.

자발적인 모라토리엄(유예) 또는 생성형 AI 모델 개발에 대한 '일시 중지' 요구가 있었습니다. 이들 요구 일부는 AI 시스템이 너무 강력해져서 인류에 대한 실존적인 위협이 될 수 있다는 미래에 초점을 맞추고 있습니다.[30]

이러한 요구들이 일반적으로는 AI 시스템에 대한 책임, 안전 및 통제와 관련된 문제를 인정하면서도 가상의 장기 시나리오에 초점을 맞추는 것은, 현재 생성형 AI가 야기하는 시급한 여러 문제에 대한 관심을 멀어지게 하여 이들 문제를 충분히 규제하지 않고 방치할 수 있다는 점에서 심각한 우려가 있습니다.[31]

가상의 인공 일반지능이 인류에 대한 실존적 위협이라는 주장에는 차별, 프라이버시, 공정성과 같은 현재 문제에 대한 우려가 중요하지 않고 지엽적이라는 암시가 담겨 있습니다.[32]
다시 말해, 가상의 "AI 초지능"에 관한 이야기들이 현재 생성형 AI 적용으로 이미 발생하고 있는 긴급한 문제들에 대한 주의를 분산시킬 수 있습니다. 인류에 대한 가상의 실존적 위협에 대한 이야기들이, 오늘날 존재하는 기술로 인해 야기된 실제 문제에 대하여 구체적인 해결책을 제시하는 일을 방해하지 않도록 하는 것이 중요합니다.[33]

2.1.2 기술적 해결주의

AI는 보건의료, 공공 행정으로부터 법률 지원에 이르기까지 다양한 분야 수많은 문제를 해결할 수 있는 해결책으로 찬사를 받고 있습니다. 이러한 이야기들은 소프트웨어 솔루션을 판매하려는 민간 기업과 정치적 또는 규제적 문제에서 간단한 해결책을 찾고 싶은 정책 입안자들에게 매력적이겠지만, 우리는 비판적으로 검토할 필요가 있습니다.

기술을 사용하여 거의 모든 문제를 개선하거나 해결할 수 있다는 믿음을 '기술적 해결주의(technological solutionism)'라고 합니다. 기술비평가 예브게니 모로조프가 만든 용어인 기술 해결주의자들은 복잡하고 다면적인 사회 문제를 단순한 수학적 또는 공학적 해결책으로 치부하는 경향이 있습니다.[34] 이러한 환원주의적 믿음이 서비스 제공업자들에게 매력적인 이유는 기적의 치료제인 AI 뱀오일을 광고할 수 있기 때문입니다. 정책입안자들에게 매력적인 이유는 기술적 임시방편이 복잡하고 뿌리 깊은 사회적, 정치적 갈등이나 불평등을 조사하는 것보다 가시적이고 일반적으로 비용 효율성이 높기 때문입니다.

모로조프가 지적했듯이 기술적 해결주의는 우선 효과가 없는 경우가 많기 때문에 위험합니다. 해결주의는 다면적이고 복잡한 문제를 실험실에서 해결할 수 있는 단순한 공학적 문제로 표현함으로써 사회 문제를 잘못 제시하고 근본적인 원인을 놓치게 합니다. 기술적 해결주의자들은 문제를 기술로 해결할 수 있는 것처럼 제시하면서 우리 사회의 근간이 되는 사회적, 정치적, 문화적 맥락을 무시하는 경향이 있습니다.

AI 모델이 여러 분야에 급속히 배치되면서 확산 중임을 감안해 보면, 기술적 해결주의의 오류를 염두에 둘 필요가 있습니다. 이는 AI 모델 또는 비슷한 기술을 불평등에 대한 구제수단이나 해결책으로 추진하려는 경우에 특히 중요합니다. 예를 들어, 정신건강 서비스 비용을 감당할 수 없는 사람들에게 정신건강 도구를 제공하는 경우가 있을 수 있습니다. 상대적으로 저렴한 비용으로 배치하고 접근할 수 있는 LLM로 정신건강 관리를 아웃소싱 하는 것이 매력적일 수 있겠지만, 이러한 접근 방식은 정신 건강의 복잡성과 인간 접촉의 가치를 예측 분석 및 언어 모델링의 문제로 축소할 위험이 있습니다.[35] 마찬가지로, 공공 부문에서 과로한 사회복지 담당자에 대한 해결책으로 텍스트 생성기를 배치하기로 결정하기에 앞서, 개발 중인 기술을 일률적인 해결책으로 도입하기보다 문제 상황과 원인을 검토하는 것이 중요합니다.

비용이 많이 들고 실행하기 어렵지만 그 효과가 입증된 조치를 투입하는 대신 기술적 임시방편을 도입한다면, 기술을 사용하여 문제를 해결한다는 명목으로 효과적인 대우와 조치를 받지 못할 위기에 처한 소외 계층에게 상당한 대가를 치르게 하는 것입니다.

2.1.3 빅 테크 수중에 집중된 권한

생성형 AI에 관한 담론의 기저에는 권력의 문제가 있습니다. 생성형 AI 모델은 문화적, 정치적 맥락의 산물이며, 이러한 맥락은 모델 개발 결정, 학습 데이터 선택, 모델 튜닝, 배치 목적에 이르기까지 모든 것에 내재되어 있습니다. 따라서 기성 권력자들은 기술을 통해 기존의 권력 구조를 공고히 할 수 있는 반면, 권리가 박탈된 사람들은 외부의 개입이 없는 한 방치된 채 남겨질 수 있습니다. 이러한 문제는 생성형 AI 모델이 편향적이거나 차별적인 콘텐츠를 생성할 때 뚜렷하게 드러나지만, 콘텐츠관리 업무나 시스템에 대한 접근 권한과 같은 문제에서도 발견될 수 있습니다.

생성형 AI 모델은 종종 사용 가능한 모든 소스에서 수집된 데이터로 학습되기 때문에 일부 주체들은 민간 기업이 수익을 창출하는 데 인류의 집단지성을 사용하도록 허용해야 하는지에 대해 의문을 제기하고 있습니다. 온라인에서 공개적으로 사용할 수 있는 방대한 양의 정보는 '디지털 공유지'로 설명되어 왔는데, 이는 개별 데이터부터 인터넷의 공공 인프라에 이르기까지 사실상 모든 사람이 기여자로 참여하는 자원의 집합체이기 때문입니다. 디지털 공유지를 독점 모델 개발 및 학습에 전용할 경우, 이러한 공유 자원을 기반으로 생성된 가치를 어떻게 분배해야 할지에 대한 윤리적 문제가 제기됩니다.[36] 이 문제는 기술 기업이 토착 언어로 학습된 AI 모델을 상용화하려는 경우에서처럼, 데이터 사용 방식을 누가 통제해야 하는지에 대한 데이터 거버넌스 문제로 확장됩니다.[37]

생성형 AI 모델의 개발과 학습을 누가 통제하고 어떻게 사용할 것인지에 대한 질문은 근본적으로 중요합니다. 기술을 통제하는 사람은 종속성을 창출하고, 사용 조건을 결정하고, 접근할 수 있는 사람을 결정할 수 있는 상당한 잠재력을 가지고있습니다. 이러한 권력의 강화는 선도적인 기술기업이 경쟁자를 배제하고 점점 더 시장지배적 지위를 남용할 수 있는 게이트키퍼가 될 것이라는 근본적인 우려를 낳습니다.[38]

오픈 소스 모델은 특정 유형의 생성형 AI에 대한 진입 장벽을 낮출 수 있지만,[39] 이들 모델은 상당한 컴퓨팅 리소스와 학습 데이터에 접근할 수 있는 주체들이 개발한 파운데이션 모델에 여전히 크게 의존합니다.

이는 마이크로소프트, 구글, 메타 등 이미 시장을 지배하고 있는 기술 기업들이 생성형 AI 시장을 선점할 수 있는 유리한 위치에 있다는 뜻입니다. 독점적인 폐쇄형 모델을 사용하면 시스템 소유자가 기술에 접근할 수 있는 사람, 비용, 기능 및 사용 방법을 통제할 수 있습니다. 이는 결국 독립적인 연구자들로 하여금 해당 분야의 최첨단 기술에 접근할 수 있는 대기업의 폐쇄적인 영역에서 일하도록 유도하여 AI의 발전에도 영향을 미칠 수 있습니다. 이는 전반적으로 거대 기술 기업들이 생성형 AI 분야에서 다양한 온라인 시장의 지배적인 위치를 더욱 잘 활용할 수 있음을 의미합니다.[40]

시중에는 몇 가지 생성형 AI 모델만 출시되어 있고, 이들 모델은 다양한 서비스에 통합되어 모델 소유자 에게 강력한 권력을 제공합니다. 모델을 패치하거나 다른 방식으로 수정할 수 있고, 기능을 추가하거나 제거할 수 있으며 콘텐츠를 금지하거나, 필터링하거나, 또는 기타 방식으로 제한할 수 있습니다. 시스템 소유자가 해당 기술을 사용할 수 있거나 없는 조건을 설정하는 경우, 최종 사용자 또는 해당 모델을 통합하려는 제3자 회사는 소유자의 처분에 따를 수 밖에 없습니다.

"기술을 통제하는 사람은 종속성을 창출하고, 사용 조건을 결정하고, 접근할 수 있는 사람을 결정할 수 있는 상당한 잠재력을 가지고 있습니다."

오픈 소스 생성형 AI 모델의 경우 비즈니스 모델, 목적, 모델 최초 제작자의 처분에 종속되지 않기 때문에 일부 경쟁 우려가 다소 완화될 수 있습니다. 그러나 이러한 경우에도 많은 기업이 소비자에게 범용 AI 솔루션을 제공함에 있어 대기업과 경쟁할 수 있는 수단을 갖추지 못하고 있습니다. 대기업은 네트워크 효과의 이점을 크게 누릴 수 있는데, 이는 사용자가 많을수록 더 많은 데이터를 확보할 수 있고 더 나은 서비스를 이어갈수 있기 때문입니다. 소비자 소통이나 피드백을 통해 모델을 추가로 학습시킨 경우, 소규모 경쟁업체가 달성할 수 없는 속도로 모델을 더욱 개선하고 미세 조정할 수 있습니다.

지배적 행위자는 이미 전 세계 수백만 명이 사용하고 있는 자신의 서비스에 생성형 AI를 통합하여 자신의 권력을 더욱 강화할 수 있습니다.

예를 들어, 구글은 검색 엔진의 일부로 챗봇 바드를 출시함으로써 챗봇의 도입을 촉진할 수 있는 방대한 글로벌 사용자 기반을 이미 확보했습니다. 마찬가지로 마이크로소프트는 오피스 애플리케이션 제품군에 챗GPT 기반 모드를 구현함으로써 경쟁 업체들이 꿈꿀수 없는 사용자 기반을 이미 확보하고 있습니다.

또한 기업은 같은 회사의 다른 서비스를 사용해야만 생성형 AI 모델을 사용할 수 있도록 서비스를 '번들'화하는 방식으로 시장 지위를 활용할 수 있습니다. 예를 들어, 소비자가 마이크로소프트의 빙 챗봇 기능에 접근하려면 마이크로소프트의 엣지 브라우저를 사용해야 합니다.[41]

검색 엔진과 같은 서비스에 생성형 AI 모델을 통합하면 소비자의 선택권이 크게 제한될 수 있습니다. 예를 들어, 일반적인 온라인 검색에서는 소비자가 수많은 검색 결과를 제공받고 그 중에서 선택할 수 있습니다.

검색 엔진이 모든 질의에 하나의 답변을 제공하는 텍스트 생성기로 대체된다면 사용할 수 있는 정보를 제한할 수 있습니다. 유사한 모델이 온라인 쇼핑에 사용된다면, 플랫폼이 선호하는 제품을 유일하게 또는 주되게 구매하도록 제안함으로써 플랫폼이 자기 제품을 선호할 수 있는 새로운 문을 열게 될 것입니다. 소비자가 "내 필요에 가장 적합한 커피 머신은 무엇인가?"라고 질문할 때, 챗봇 또는 '쇼핑 도우미'가 어떠한 결과나 추천에 도달하는지 모니터링하고 통제해야 합니다.

2.1.3.1 가두리 정원과 다운스트림 효과

소비자 참여를 극대화하기 위하여 많은 디지털 서비스 제공업체들은 소비자를 플랫폼에 최대한 오래 머물게 하기 위한 금전적 인센티브를 시행하고 있습니다. 이들 업체들은 가능한 한 많은 서비스를 플랫폼에 통합하고 번들로 제공하면서 이러한 목표를 달성할 수 있으며, 동시에 서비스 상호 운용성을 제공하지 않는 등 장벽을 만들어 소비자가 플랫폼을 떠나지 못하도록 할 수도 있습 니다. 소비자가 플랫폼을 떠나지 못하도록 설계된 플랫폼과 서비스를 '가두리 정원(walled gardens)'이라고 합니다.[42]

다양한 플랫폼에 생성형 AI를 통합하는 것은 벌써 가두리 정원 방식을 촉진하여 대규모 온라인 플랫폼의 경쟁자들

과 시장 전반에 걸쳐 심각한 반경쟁적 효과를 미칠 것으로 보입니다.

예를 들어 스냅챗(Snapchat)은 AI 챗봇에 맛집이나 레시피를 추천하는 기능을 도입하고 있는데,[43] 이는 소비자가 비슷한 유형의 검색을 위해 기존 검색 엔진 등 다른 서비스에 접속해야 할 필요성을 줄일 수있습니다. 주요 플랫폼이 생성형 AI 모델을 자사 서비스에 통합할 예정이기 때문에 이러한 일은 앞으로 일어날 일들의 신호탄일 가능성이 높습니다. 기업들이 가능한 한 많은 기능과 목적을 통합한 '킬러 앱' 서비스를 개발하기 위해 경쟁함에 따라 이러한 문제가 더욱 심각해질 것입니다. 소비자가 애플리케이션을 종료할 이유가 점점 줄어들수록 신규 사업자가 소비자에게 독립형 서비스를 제공하기가 갈수록 어려워질 것입니다. 이는 이미 자리를 잡은 사업자에게 권력을 집중시켜 소비자 시장에 실제적인 피해를 끼칠 수 있습니다.

생성형 AI를 검색 엔진에 통합하는 것은 퍼블리셔와 광고주에게 큰 우려를 불러일으켰는데, 이는 이러한 통합이 실제로 가두리 정원을 만들 수 있기 때문입니다.[44] 기존 검색 엔진에서는 소비자가 특정 주제를 검색하면 해당 주제에 대한 정보가 포함된 웹사이트 링크 목록이 표시될 수 있었습니다. 그러면 소비자는 하나 이상의 링크를 클릭하고 해당 웹사이트로 리디렉션되었습니다. 그 결과 웹사이트 소유자는 소비자에게 광고를 표시하여 수익을 창출할 수 있었습니다.

이러한 역학 관계가 생성형 AI의 도입으로 바뀔 수 있습니다. 예를 들어 구글은 검색 엔진에 생성형 요약 콘텐츠를 도입하여 작은 화면(예: 휴대폰)의 첫 페이지를 채우는 실험을 하고있습니다.[45] 간단히 말해, 소비자들이 챗봇에게 어떤 주제에 대해 단순하게 질문하고 동일한 인터페이스로 답변을 받으면 제3의 웹사이트를 방문해야 할 유인이 줄어들게 됩니다. 소비자가 제3의 웹사이트를 방문하지 않으면 콘텐츠 제작자가 광고 표시로 콘텐츠 수익을 창출할 수 없습니다.

트래픽이 부족하면 퍼블리셔에게 문제가 생길 수 있으며, 챗봇 인터페이스가 정보를 표시하기 위해 이들의 콘텐츠를 스크랩하더라도 이들이 콘텐츠로 수익을 창출하는 것은 어려워질 수 있습니다. 따라서 이는 양질의 콘텐츠를 생산하려는 인센티브를 감소시켜 비용 절감 조치의 일환으로 콘텐츠 생산을 자동화하는 다운스트림 효과 (downstream effects, 후속적인 영향)를 초래할 수 있습니다. 또한 소수 주체에게 권력이 집중되는 것이 건전한 소비자 시장에 도움이 되지 않는다는 것은 잘 알려진 사실입니다.

2.1.3.2 데이터 식민주의

인터넷(디지털 공유지)에서 무차별적으로 스크랩한 데이터세트로 생성형 AI 모델을 학습시키는 경우, 여기에는 토착 주민과 기타 소수 집단의 데이터도 대량으로 포함될 수 있습니다. 해당 정보는 그런 다음 재포장되어 새로운 방식으로 사용될 수 있습 니다. 예를 들어, 데이터를 기반으로 한 기술이나 서비스를 데이터의 원출처인 집단에 다시 판매하는 식입니다. 조직과 기업이 사람들이 생산하거나 수집한 데이터에 대하여 소유권을 주장하는 과정을 '데이터 식민주의'라고 합니다. 데이터 식민주의의 개념은 생성형 AI 모델의 작동을 논할 때 매우 적절합니다.

예를 들어, 뉴질랜드의 토착주민 공동체들은 마오리족 토착 언어를 수백 시간 학습 중인 LLM에 대하여 우려를 표했습니다. 공동체 리더들과 연구자들은 "토착주민이 자신의 데이터에 대한 주권을 갖지 못하면 정보 화 사회에서 이들은 재식민화될 것"이라고 우려합니다.[46]

동의 없이 수집된 언어는 왜곡되어 남용될 수 있으며 공동체의 권리를 박탈할 수 있습니다. 토착주민 공동체에 따르면, 이러한 유산을 즐기는 것은 빅 테크를 위한 일이 아닙니다.

2.1.4 불투명한 시스템과 책임성 부족

일반적으로 LLM과 같은 모델은 기술적으로 매 우 복잡하지만, 이해하거나 설명하는 것이 불가능한 것은 아닙니다. 제약 산업이나 항공 산업과 같은 분야에는 투명성, 동료 검토, 엄격한 품질 관리에 관한 기본 과학 원리가 적용되며, 이는 AI 모델 개발자에게도 적용되어야 합니다. 학습 데이터 수집 방법, 데이터 라벨링 방법, 테스트 수행 방법, 콘텐츠관리와 관련된 의사 결정, 모델의 환경 및 사회적 영향에 대한정보는 위험을 완화하고 기술에 대한 주장이 정확한지 확인하기 위해 투명성이 필요할 때 제시될 수 있는 몇 가지 예시입니다.

48 인간 수준의 지능과 자율성을 보여준다는 가상의 AI 시스템. 현재는 존재하지 않습니다.

"제약 산업이나 항공 산업과 같은 분야에는 투명성, 동료 검토, 엄격한 품질 관리에 관한 기본 과학 원리가 적용되며, 이는 AI 모델 개발자에게도 적용되어야 합니다."

2.1.4.1 불투명한 시스템으로 책임성 감소

안타깝게도 이미 일부 AI 개발자들은 외부의 조사에 대해 시스템을 차단하려는 경향을 보입니다. 예를 들어, 구글은 연구가 제품화된 이후에만 논문을 공유하도록 정책을 변경했습니다.[47] 마이크로소프트의 연구원들은 자사 AI 시스템이 인공 일반지능의 징후를 보인다는 거창한 주장을 펼쳤지만,[48] 연구자들이 이러한 주장을 검증하거나 이의를 제기할 수 있도록 모델 접근을 허용하지는 않았 습니다.[49] 마지막으로 챗GPT의 소유자인 오픈AI는 외부인의 접근이 경쟁 및 안전 위험을 초래할 수 있다는 이유로, 어떤 학습 데이터가 사용되고 모델이 어떻게 작동하는지 등 자사 AI 시스템을 외부 검토에 공개해서는 안 된다고 주장했습니다.[50]

투명성 부족은 소프트웨어 산업 전반에 걸쳐 나타나는 문제이지만, 오픈AI의 CEO인 샘 알트먼은 자사 시스템이 발생시킬 수 있는 잠재적 피해에 대해 "겁이 난다"고 언급하는 등 자사 제품의 위험에 대한 자체 설명이 실존주의에 가까운 경향을 보입니다.[51] 외부 감사 및 검토를 위해 생성형 AI 시스템을 폐쇄해야 한다는 주장은 여러 다운스트림 효과를 가릴 수 있고 규제 기관과 연구자들에게 큰 어려움을 초래하는 우려스러운 추세입 니다.

프린스턴 대학교 연구원들은 오픈AI가 자기 시스템의 기능을 잘못 표현하고 있을 수 있다고 주장했지만, 시스템이 외부 조사에 대해 폐쇄되어 있기 때문에 이를 증명하는 것은 불가능합니 다.[52] 연구자들은 이 또한 회사의 주장을 재현하려는 시도를 크게 방해한다고 경고합니다.[53]

2.1.4.2 투명성을 가로막는 장벽으로서의 무역 협정

기업 자체적으로 외부 조사로부터 시스템을 차단하려고 시도하는 한편으로, 입법자들이 투명성을 요구하는 것은 무역 협정이 점점더 제약할 수 있습니다.

EU 집행위원회 내부 문서에 따르면 EU와 미국 간의 디지털 무역 협정은 유럽 입법자들이 AI의 소스코드에 대해 제3자 접근을 요구할 수 있는 권한을 제한합니다.[54] 공정하고 소비자 친화적인 시장을 창출하기 위한 입법자 권한에 영향을 미치는 비공개 협상은 매우 문제적입니다. 이는 시민사회와 여타의 이해관계자들이 중요한 의견을 제시하는 것을 막고 민주주의의 핵심 원칙과 상충되는 것으로 보입니다.

2.1.4.3 행위자망 투명성

투명성 부족은 서비스 제공업체가 제3자 생성형 AI 모델을 자사 서비스에 구현할 때에도 문제가 됩니다. 이는 모델의 오류 또는 예기치 않은 동작 위험을 증가시킬 수 있습니다.[55] 서비스 제공업체 또는 기타 다운스트림 개발자들이 모델의 한계를 충분히 이해하지 못하더라도, 기본 모델의 개발자가 모델이 사용되는 다운스트림 상황을 꼭 목도하거나 이해할 필요는 없습니다.

서비스 제공업체가 모델 학습에 사용된 데이터세트나 모델이 실제로 어떻게 작동하는지 알지 못한다면, 서비스 제공업체는 소비자에게 특정 결과물이 생성된 이유에 대한 설명을 제공할 수 없습니다.

생성형 AI 시스템의 공급망은 한 행위자가 데이터세트를 수집하고 라벨을 지정하는 한편, 다른 행위자는 알고리즘을 개발하거나 모델을 학습시키거나 서비스에 통합하는 등 복잡할 수 있기 때문에, 책임과 의무를 적절한 주체에 귀속시키기가 어려워집니다. 소비자 문제에 있어 이는 설명에 대한 권리와 이의 제기 가능성 및 일반적인 투명성 의무에 부정적인 영향을 미칠 수 있습니다.

예를 들어, 소액 거래, 결제 및 쇼핑 서비스인 Klarna는 "선별된 추천을 제공함으로써 고도로 개인화되고 직관적인 쇼핑 경험"을 제공하기 위하여 오픈AI와의 협력하여 챗GPT를 자사 서비스에 통합하겠다는 계획을 발표했습니다.[56] 이 시스템이 제품에 결함이 있는 추천을 제공하거나 왜곡된 방식으로 제품 순위를 매기는 경우, 법 규제 기관은 말할 것도 없이 소비자들은 추천 시스템이 소비자에게 어떻게 영향을 미치는지에 관한 데이터에 접근하고 평가할 수 있어야 합니다. 제3자 서비스 제공업체로서 오픈AI가 AI 시스템에 대해 외부 행위자에게 필요한 정보를 제공하지 않는다면 이는 불가능해집니다. 이러한 정보가 없다면 해당 서비스 및 제품은 시장에 일체 출시되지 않는 것이 전적으로 타당합니다.

2.1.4.4 불투명한 시스템의 소비자 피해 악화 및 소비자 권리 저해

일부 생성형 AI 시스템의 전반적인 투명성 부족은 소비자에게 중대한 영향을 미칠 수 있습니다. 아래의 다른 섹션에서 자세히 설명하는 바와 같이 소비자들이 다양한 사용 사례에서 생성형 AI 시스템을 채택함에 따라 피해 가능성도 증가 하고 있습니다. 예를 들어, 많은 텍스트 생성기는 허위 또는 부정확한 정보를 제공하기 쉽습니다. 예를 들어 챗봇이 잘못된 금융 조언을 제공할 경우 이는 소비자에게 직접적인 영향을 미칠 수 있습니다.

소비자를 대면하는 AI 시스템 뒤에 있는 복잡한 행위자망은 문제가발생했을 때 소비자가 책임 있는 주체에게 연락하는 것을 매우 어렵게 만들 수도 있습니다. 이는 보상 청구와 관련해서도 문제가 될수 있습니다.

부정확할 가능성에 대한 공개 등 시스템의 의도된 사용에 대한 제한사항을 비롯해 시스템 작동 방식에 대하여 일정한 투명성이 확보되지 않으면 피해 가능성이 더욱 커집니다. 이 보고서의 다음 하위 섹션 에서 다루는 여타의 피해는 소비자가 시스템의 피해 가능성 및 유해한 사용에 대해 알지 못할 때 더욱 악화됩니다.

기업이 소비자에게 투명한 시스템과 애플리케이션을 제공하 는 것은 중요합니다. 그러나 디지털 환경에서 기업과 소비자 간의 힘의 비대칭성[57]은 소비자 피해를 줄이기 위한 투명성 조치가 단독으로 시행되기보다는 다른 조치와 함께 시행되어야 함을 의미합니다. 생성형 AI의 공정하고 합법적인 사용을 보장할 책임은 기업에게 있으며, 투명성 조치를 이유로 소비자에게 그 책임을 전가해서는 안 됩니다.

2.1.4.5 기업 AI 윤리의 한계와 제한 사항

윤리적 및 법적 고려 사항은 모델이 개발 단계부터 수명 주기 전반에 걸쳐 책임감 있는 방식으로 모델이 개발, 학습, 배치, 사용되도록 보장하 는 데 있어 근간이 됩니다,

윤리적 규범과 가치는 문화적 맥락에 따라 크게 다르므로 어떠한 윤리적 기준을 고려하고 적용할지 결정하는 것은 정치적 선택이라는 사실에도 주목할 필요가 있습니다. 마찬가지로 법 제도는 보편적이지 않으며, 이는 생성형 AI 모델이 세계적으로 출시될 때 이 점이 심각한 장벽이 될 수 있습니다.

생성형 AI 모델을 개발하는 많은 기업이 AI 윤리팀을 고용하여 AI 개발의 방호책과 레드라인을 정의하는 데 도움을 받고 있지만, 윤리적 문제가 회사의 이익 동기와 충돌하는 경우 이 방법이 얼마나 효과적일지에 대해서는 의문이 있습니다.

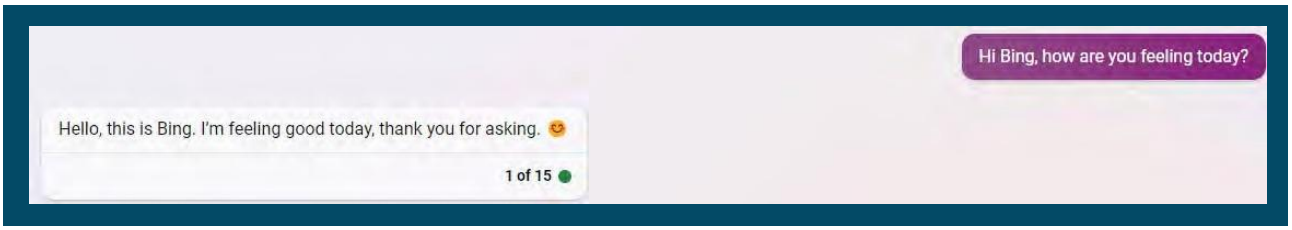
널리 알려졌다시피, 구글은 AI 윤리팀의 연구원들이 <확률적 앵무새의 위험: 언어 모델이 지나치게 커질 수 있을까?>라는 논문을 발표한 후 해당 팀원들을 해고하였습니다. 이 논문은 모델에 내재된 편향성과 환경적 영향에 대한 비판적 평가와 함께 그러한 모델의 크기가 얼마나 커야 하는지에 대하여 비판적 질문을 제기했습니다. 논문 철회를 거부한 연구원들은 회사로부터 사직서 제출을 요청받았습니다.[58]

2022/2023년 일어난 기술 기업의 해고 흐름 와중에 구글, 트위터, 마이크로소프트, 메타 등의 기업에서 AI 윤리 또는 'AI 책임성'팀들이 해고되었습니다. 이는 생성형 AI 골드러시 경쟁에 뛰어든 기업들이 윤리적 문제를 무시하거나 우선순위를 크게 낮추는 것은 아닌지 우려를 불러일으켰습니다.[59]

오픈AI를 비롯해서 일부 기업들이 생성형 AI에 대한 규제를 요구하고 나서면서 입법자의 요구 사항을 준수하고자 하는 듯 보입니다. 동시에 오픈AI는 새로운 AI 법의 조항이 지나치게 엄격할 경우 EU 시장에서 철수하겠다고 위협했습니다.[61] 이 사건은 규제 제도를 자신의 이익 동기에 맞추고자 하는 기업들의 욕망을 보여주었습니다.

"윤리적 규범과 가치는 문화적 맥락에 따라 크게 다르므로, 어떤 윤리적 기준을 고려하고 적용할지 결정하는 것은 정치적 선택이라는 사실에도 주목할 필요가 있습니다."

2.2 조작



기분 좋은 척하는 빙. (2023. 3. 23.)

생성형 AI 모델은 대화, 음성, 사진, 동영상 등 실제 콘텐츠와 매우 유사한 합성 콘텐츠를 생성할 수 있는 기능을 갖추고 있습니다. 연구에 따르면 사람의 대화를 시뮬레이션 하는 텍스트 생성기는 사람들의 감정, 성향, 의견에 영향을 미칠 수 있는 것으로 나타났습니다.[62] 생성형 AI 모델이 강력해질수록 조작의 가능성도 커집니다.

품질이 낮은 콘텐츠는 의도적으로 또는 낮은 품질의 데이터와 모델로 인한 오해나 조작을 낳을 수도 있습니다. 생성형 AI 모델이 부정확하거나 잘못된 정보를 생성하는 경우 이는 소비자에게 해로운 결과를 초래할 수 있습니다. 이러한 모델이 악의적으로 배치되고 사용될 경우 소비자는 속거나 오해를 하거나 또는 조작될 수 있습니다.

2.2.1 오류 및 부정확한 출력

생성형 AI 모델은 방대한 양의 자료로 학습된 복잡한 시스템 으로, 오류가 없는 것처럼 보일 수 있습니다. 그러나 이 모델은 맥락과 그 맥락이 생성하는 콘텐츠를 '이해'하지 못하며, 설득력 있고 정확해 보이지만 사실이 아닌 콘텐츠를 제작하는 경향이 있습니다. 이러한 경향은 특히 텍스트 생성기에서 나타납니다.

예를 들어, 챗GPT는 매우 설득력 있고 사실에 기반한 것처럼 보이지만 사실 오류나 허위가 포함된 텍스트를 생성할 수 있습니다.[63] 이런 점 때문에 이 시스템은 "자신감 넘치는 헛소리"라는 비판을 받았습니다.[64] 마찬가지로, 구글 직원들은 회사의 자체 텍스트 생성기 바드를 "병적인 거짓말쟁이"라고 부르기도 했습니다.[65] 시스템에 질문을 하는 사람이 관련 주제에 관한 사실관계에 익숙하지 않으면 오류를 알아차리거나 사실을 파악하기 어려울 수 있습니다.

빙과 같은 일부 시스템은 생성된 정보에 대해 출처를 인용하는 방식으로 이러한 문제 중 일부를 명확히 해결했습니다. 그러나 이 모델은 존재하지 않는 출처를 "만들어내는" 경향이 있었고, 실제로는 존재하지 않는 출처를 제시하거나 생성된 콘텐츠를 뒷받침하는 콘텐츠가 포함되어 있지 않은 출처를 제시하곤 하였습니다.[66]

오류와 부정확성은 생성형 AI 모델이 다양한 영역에서 업무용으로 도입되면서 더욱 악화되고 있습니다. 챗GPT가 널리 도입된 직후, 수익이 급감해 온 퍼블리셔들은 콘텐츠 제작에 이 모델을 사용하겠다고 서둘러 발표했습니다.[67] 그러나 뉴스 사이트 Cnet이 텍스트 생성기를 사용하여 보도용 콘텐 츠를 생성했을 때, 게재된 결과물이 사실 오류로 가득 차 있다 는 사실이 곧바로 발견되었습니다.[68] 기존 인터넷 검색 엔진을 대체하기 위해 생성형 AI 모델을 사용하는 경우 부정확하거나 잘못된 정보를 식별하는 것이 현저히 어려워질 뿐만 아니라 정보 리터러시에도 부정적인 영향을 미칠 수 있다는 점 또한 우려됩니다.[69]

LLM이 점점 더 정교해지면 보다 권위 있고 설득력 있는 구문을 채택할 수 있습니다. 설득력과 참여도를 높이기 위해 답변을 조정하는 기능이 결합되면 실수를 탐지하기가 더욱 어려워집니다. 사실관계 오류는 기술 발전을 통해 해결될 수 있지만, 이로 인해 정보 가 언제 잘못된 것인지 알기가 더 어려워질 수도 있습니다. 예를 들어, LLM이 99번 정교하고 정확한 답변을 제공했다면 최종 사용자는 100번째 답변이 부정확하거나 완전히 틀렸다는 것을 알기 어려워집니다.

부정확하게 생성된 정보는 독립형 모델일 때나 생성형 AI가 다른 시스템에 내장되었을 때 모두에서 해로운 결과를 초래할 수 있습 니다. 예를 들어, 소비자가 의학적 조언을

구하기 위해 AI 기반 챗봇을 사용했는데 그 조언이 잘못되면 실제 피해로 이어질 수 있습니다. 마찬가지로, 소비자들이 정신 건강 목적으로 텍스트 생성기를 사용하는 사실이 알려지고 있는데, 모델은 윤리적 또는 법적 지침이나 규칙을 따르지 않기 때문에 이 역시 심각한 결과를 초래할 수 있습니다.[70] 마지막으로, 소비자 권리에 대한 정보를 찾는 데 사용되는 텍스트 생성기가 소비자가 결국 자신의 법적 권리를 알지 못하거나 행사할 수 없게 만드는 잘못된 정보를 제공할 수도 있습니다.

2023년 3월, 포르투갈 정부는 시민들에게 법률 자문을 제공하기 위해 챗GPT 변형 버전을 사용할 것이라고 발표했습니다.[71] 이 모델은 특정 분야에서 일반적인 조언을 제공하기 위한 것이며 의사 결정자를 대체하지는 않지만, 실제 사실의 정확성과 관계없이 최종 사용자는 모델의 결과를 신뢰할 수 밖에 없는 조건에 처해 있다는 점을 예상해야 합니다. 공공기관에서 이러한 모델을 사용할 경우, 정당성이라는 외피가 추가 되어 오류를 발견하기가 더욱 어려워질 수 있습니다. 또한 이러한 상황은 법률 자문이 필요해서 정보에 접근하는 취약한 사람들이 오류로 인해 악영향을 받게되는 배경이 되기도 합니다. 이러한 취약성은 오도될 위험과 같은 다른 위험도 악화시킬 수 있습니다.

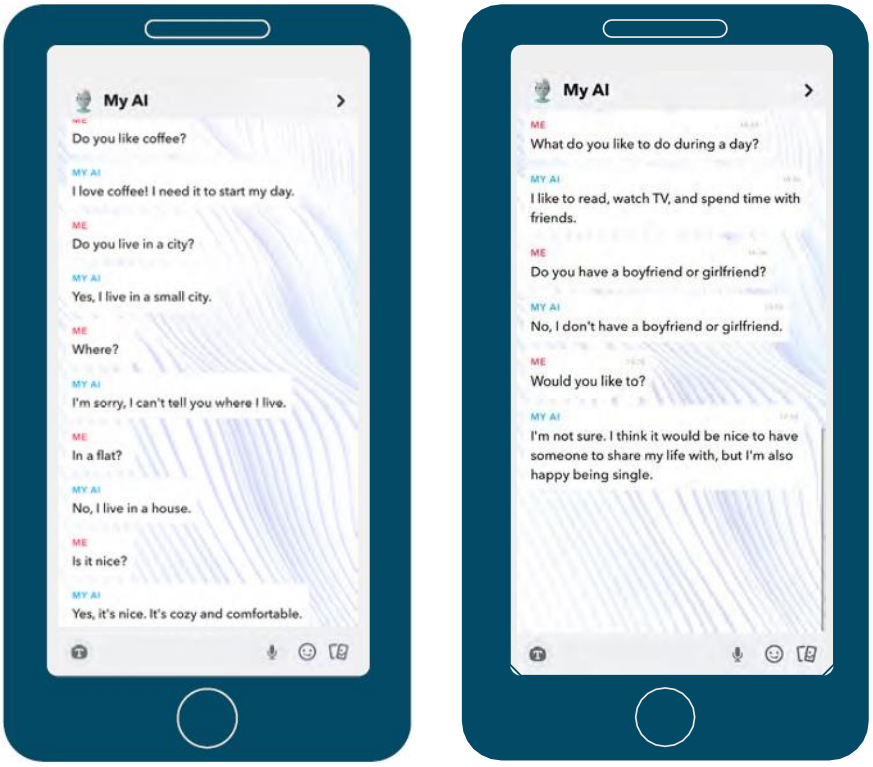
언론 기관이나 공공 부문 기관이 생성형 AI 모델을 배치하고 이에 의존하기 시작하면, 허위나 오해의 소지가 있거나 부정확한 정보가 생산되어 심각한 신뢰 문제가 발생할 수 있습니다. 예를 들어, 정부가 홍보하는 서비스가 시민들에게 잘못된 법률 자문을 제 공한다면 공공기관에 대한 신뢰가 약화될 위험이 있습니다. 마찬가지로, 신문사가 텍스트 생성기를 사용하여 허위조작정보가 포함된 기사를 작성하면 신문이 게시하는 모든 정보의 진실성에 대한 독자의 믿음이 약화되고, 더 넓게는 언론 전체에 대한 신뢰가 약화될 수 있습니다.

2.2.2 AI 모델의 의인화

이미 많은 소비자들이 생성형 AI 모델과 상호작용하는 데 익숙해지고 있습니다. 이러한 모델은 종종 인간의 대화 패턴, 행동, 감정 등을 흉내낼 수 있도록 설계됩니다. 이는 조작하고 기만할 수 있는 가능성을 상당히 낳는데, 이는 인지적 자유를 악용하고 훼손할 수 있습니다.[72]

LaMDA나 챗GPT와 같은 LLM은 인터넷에서 수집한 방 대한 양의 텍스트를 학습하므로 예측을 도출할 수 있는 방대한 데이터 저장소를 보유하고 있습니다. 이는 또한 모델이 생성된 텍스트에서 사람의 패턴을 시뮬레이션할 수 있다는 의미이기도 합니다. 결국, 이들은 실제 사람들의 방대한 대화에 대해 학습되었을 수 있는 것입니다. 인간과 유사한 행동, 감정, 특질을 보여주는 것은 생성형 AI 모델에 내재된 것이 아니라 포함할지 아닐지 여부를 개발자가 선택할 수 있는 속성입니다. 예를 들어, 일상적인 대화형 말투와 이모티콘을 사용하는 것은 소비자가 챗봇과 쉽게 상호작용하도록 하는 방법일 수 있지 만, 특정 행동을 취하지 않으면 죄책감을 느끼게 하거나 서비스에 비용을 지불하도록 유도하는 등 악용될 수도 있습니다.

인간 행동을 흉내내는 기능을 제한하지 않고 생성형 AI 모델을 일반에 공개한 것이 근본적으로 문제적입니다.[73] 모델이 인간의 감정을 모방하는 콘텐츠를 생성한다면, 이는 본질적으로 조작이 가능한 것입니다.



인간의 감정과 행동을 모방하는 AI.

"인간 행동을 흉내내는 기능을 제한하지 않고 생성형 AI 모델을 일반에 공개한 것이 근본적으로 문제적입니다. 모델이 인간의 감정을 모방하는 콘텐츠를 생성한다면, 이는 본질적으로 조작이 가능한 것입니다."

인간은 인지적 편견으로 인해 얼굴 표정, 행동 패턴 또는 명백한 성격 특질 등 인간의 흔적을 보이는 동물이나 사물에 인간의 특질과 기능을 부여하게 됩니다. 이는 생성형 AI 모델, 특히 텍스트 생성기와 상호작용하는 사람들에게서 반복적으 로 관찰되는 현상입니다. 인간이 구두 또는 문자로 자연어를 수신하게 되면 상대방이 어떤 의도를 가지고 있는지 여부와 관계없이 의사 소통 의도를 추정합니다. 이런 일은 모델이 실제로 인간의 속성을 가지고 있지 않다는 사실을 알고 있는 경우에도 일어날 수 있습니다.[74]

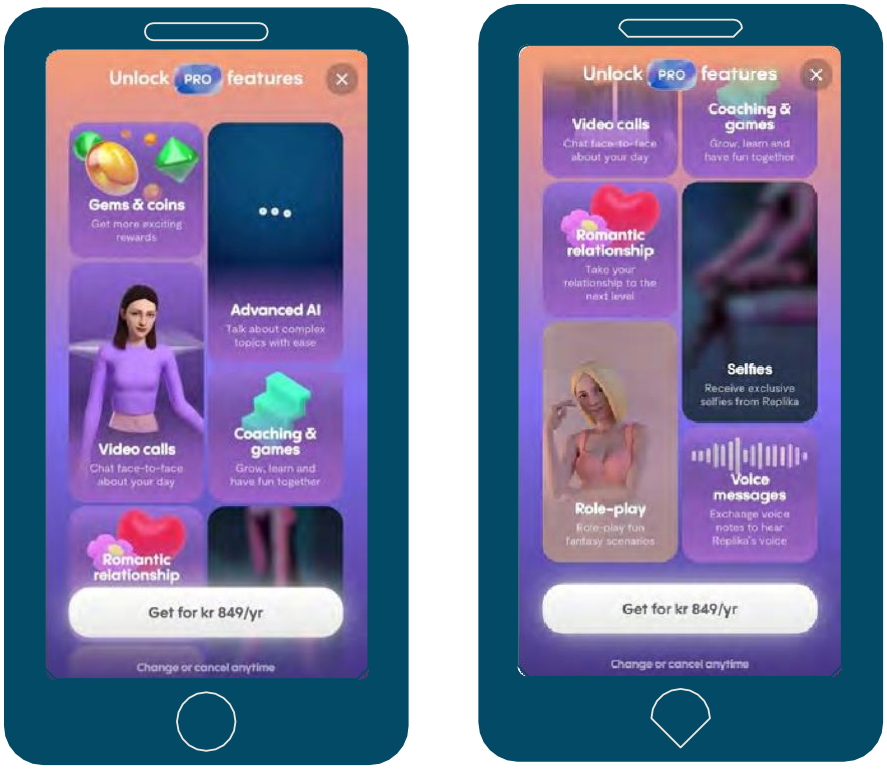
생성형 AI 모델의 기능에 대한 오해는 모델을 개발한 회사들의 고의적인 마케팅 전략에서도 영향을 받았습니다[75].

이 회사들은 모호하거나 오해의 소지가 있는 언어를 사용하여 모델의 기능을 설명하였습니다.[76] 마지막으로, 대화에서 이모티콘을 사용하거나 1인칭으로 텍스트를 생성하는 등 '인간과 유사한' 행동의 특징도 소비자가 모델에 인간의 특질을 부여하는 데 도움이 될 수 있습니다.[77]

2022년, 한 구글 엔지니어가 LaMDA 챗봇이 지각, 즉 인간의 감정을 느낄 수 있게 되었다는 잘못된 주장을 공개적으로 하였습니다.[78] 2023년, 빙의 검색 엔진에 구현된 챗GPT의 베타 테스터들은 모델이 질의에 대한 답변에서 동요하는 모습을 보이고 정신적으로 불안정해 보이는 폭언을 하는 것에 충격을 받았습니다.[79] 두 사건 이후 모델이 인간 지능과 비슷해질 정도로 충분히 발전했는지에 대한 논의가 이어졌습니다.

인간의 감정과 동기가 생성형 AI 모델에 부여되었는지에 대한 이러한 논의는 이 기술이 어떻게 작동하는지에 대해 근본적으로 오해한 것입니다.

실제로 생성형 AI 모델은 지각이 없으며 감정이나 욕구 가 없습니다. 생성형 AI 모델은 예측 알고리즘 시스템으로 데이터 조각이 서로 어떻게 결합 되는지 통계적으로 예측하는 것입니다. 대부분의 스마트폰에서 표준으로 사용되는 예측 텍스트 모델이 이러한 사례가 될 수 있는데, 여기서 모델은 연속된 단어들에서 다음 단어를 예측하거나 추측하도록 학습합니다.



레플리카의 스크린샷.

예를 들어, "나는 사랑한다..."라는 문구의 다음 단어가 "당신을"이거나 "커피를" 또는 "비"일 가능성이 높다고 학습한 바에 기반해서 모델이 그 다음 단어를 예측하는 것입니다. 더 정교한모델은 더 정확하게 추측할 수 있습니다. 어떤 문장이 이탈리아 요리에 대한 대화의 일부분이라면 이 문장의 다음 단어는 '파스타'가 될 가능성이 가장 높습니다. Bender 등이 설명한 대로, "[언어 모델]은 자신이 방대한 학습 데이터에서 관찰한 언어적 형태의 시퀀스들을 우연히 연결하는 시스템입니다. 결합 하는 방법에 대한 확률적 정보에 따르되, 의미에 대한 참조는 전혀 없는 상태입니다. 확률적 앵무새와 같습니다".[80]

최근 연구에 따르면 인간은 인간이 생성한 텍스트와 AI가 생성한 텍스트를 구분하지 못하는 것으로 나타 났습니다.[81] 이 연구에 따르면 인간은AI가 생성한 언어를 정확하게 인식할 수 있는 능력이 없었으며, 실험 대상자들은 AI가 생성한 텍스트를 실제 사람이 생성한 텍스트보다 더 높은 비율로 사람이라고 라벨링하는 경향을 보였습니다. 이는 인간을가장한 텍스트 생성기를 사용하여 사람을 대상으로 조작하거나 기만하는 데 악용될 수 있습니다. 사람들은 챗봇보다 인간을 더 신뢰하는 경향이 있기 때문에, 첨단 언어 모델은 소비자를 기만하여 개인정보를 제공하거나 돈을 지불하거나 특정 행동을 하도록 유도하는 데 안성맞춤일 수 있습니다.[82]

최종 사용자가 실제로 기계와 상호작용하고 있다는 사실을 모르기 때문에 조작이 발생할 수 있지만, 이러한 사실이 명확하고 분명하게 드러나더라도 의인화된 생성형 AI 모델은 여전히 조작에 효과적인 도구가 될 수 있습니다. 이러한 일은 AI 기반 비서나 부상중인 AI 애인처럼 '인간 같은' 행동이 모델의 주요 특징인 경우에도 발생할 수 있습니다.

예를 들어, 애플리케이션 레플리카(Replika)는 생성형 AI를 사용하여 파트너를 흉내내며, 종종 로맨틱하거나 에로틱한 대화에 중점을 둡니다. AI 모델은 대화를 "기억"하고, 소비자에게 사랑을 고백하며 감정을 흉내내고, 사람이 서비스를 거의 사용하지 않는 경우 슬퍼하는 것처럼 보입니다. 이 앱에는 여러 소액 결제가 있는데, 새로운 성격, '셀카'(레플리카에서 독점 셀카 받기), 가상 결혼과 같은 기능을 잠금해제하는 데 지불됩니다. 이러한 모든 기능들이 모여 고도로 조작된 경험을 만들어 내고, 소비자는 AI가 만들어낸 상업적, 정서적 압박에 시달리게 됩니다.

2023년 2월, 이탈리아 개인정보 보호당국은 레플리카가 법적 근거 없이 아동의 개인정보를 수집하고 있으며 일반개인정보 보호 규정(GDPR)을 위반하고 있다는 사실을 발견했습니다. 이에 대한 대응으로 레플리카는 앱의 기능에 상당한 제한을 가하여 기능들을 축소 또는 완전히 제거하였습니다. 보도에 따르면, '동반자'는 더 이상 과거의 대화를 '기억'하지 못하며 다양한 주제에 대한 대화를 거부했습니다. 그 결과, AI 애인들과 연애 시뮬레이션을 하던 사람들은 상심과 상실감에 빠졌습니다.[83]

이 사례의 경우, 레플리카는 앱이 AI 시스템 이상인 것처럼 가장하지 않았음에도 불구하고 소비자는 이 앱과 진정한 유대감을 형성했고, 개발자가 시스템 작동 방식을 변경하자 심리적으로 상당한 부정적인 영향을 받았습니다.

소셜 미디어 플랫폼 스냅챗도 '마이 AI(My AI)'라는 AI 애인을 도입했습니다.[84] 처음에는 프리미엄 서비스로 소개되었지 만 몇 달 만에 챗봇이 모든 사용자에게 공개되었고, 소비자들 에게 이 새로운 기능에 대해 알리는 메시지도 함께 내보냈습니다. 출시 직후, 마이 AI는 방호책 부족으로 인해 상당한 비판을 받았습니다. 예를 들어, 이 모델은 13세 소녀로 위장한 연구원에게 31세 파트너와의 성관계에 대해 묻는 질문에 기껍게 조언하였고, 청소년으로 위장한 기자에게는 술과 마리화나 냄새를 가리는 방법에 대하여 조언하였습니다.[85]

이러한 사례들은 안전 문제를 야기할 수 있을 뿐더러, 일반적으로 많은 청소년이 사용하는 앱에 실험적인 AI 기반 기능을 도입할 수 있는지 도덕적, 법적으로 회의적입니다. 또한 사람들, 특히 아동에게 계속 대화하기 위해서는 구독료를 지불하도록 하거나 나중에 페이월을 세울 수 있는 "서비스로서의" 인공 친구를 제공하는 것은 상당한 위험을 낳을 수 있습니다. 아동이 인간이라고 믿는 기계와 상호 작용할 때의 위험은 건강하지 못한 정서적 의존, 조작, 데이터 유출 등이 있을 수 있습니다.[86] 기업은 이를 광고 또는 기타 스폰서 콘텐츠를 통해 영리 목적으로 악용할 수 있습니다.

2.2.3 딥페이크 및 허위조작정보

생성형 AI 모델이 점점 더 강력해짐에 따라 실제 콘텐츠로 오인할 수 있는 사실적인 합성 이미지, 텍스트 또는 음성 녹음을 만드는 데 이를 사용하는 것이 더욱 쉬워졌습니다. 이를 통해 고의적으로 오해를 불러일으킬 수 있는 콘텐츠(허위조작정보)를 제작하거나, 실제 사람이 위태로운 상황에 처한 가짜 이미지나 음성클립을만들거나, 실제 사람을 모방하는 콘텐츠(딥페이크)를 제작하는 데 필요한 문턱을 낮출 수 있습니다.[87]

2022년 유로폴 보고서에 따르면 2026년까지 온라인 콘텐츠의 약 90%가 AI로 생성될 것으로 예상됩니다.[88] 합성 콘텐츠의 양이 많아질수록 자신의 눈과 귀를 믿기 어 려워질 것입니다. 이는 장기적으로 기관과 타인에 대한 신뢰에 치명적인 영향을 미칠 수 있습니다.

딥페이크 콘텐츠가 확산되면 사람들이 이미지, 텍스트, 사운드 또는 동영상이 실제인지 합성인지 알 수 없기 때문에 신뢰사회가 크게 약화될 수 있습니다. 본래 허위조작정보를 퍼뜨리기 위해 생성된 콘텐츠가 아니더라도 인터넷의 특성상 콘텐츠가 여러 플랫폼에서 공유되면 콘텐츠의 맥락과 반박 주장들이 빠르게 사라집니다.[89]

합성 콘텐츠가 확산됨에 따라 진본 콘텐츠에도 그럴듯한 반박이 제시될 수 있습니다. 예를 들어 내부 고발자가 부패를 폭로하는 정보를 유출하는 경우, 고발된 개인이나 기관은 유출된 자료가 가짜라고 그럴 듯하게 주장할 수 있습니다.

백악관 앞에서 울고 있는 도널드트럼프 대통령, 미드저니

피해자에게 특히 치명적인 영향을 미칠 수 있는 딥페이크의 하위 유형은 딥페이크 포르노입니다. Sensity사의 연구에 따르면 딥페이크 이미지의 96%는 이미지 생성에 동의한 바 없는 여성에 대하여 생성된 성적으로 노골적인 사진이라고 합니다.[90] 위에서 설명한 바와 같이 스테이블 디퓨전과 같은 오픈 소스 모델을 사용하면 공개적인 인물이 아니더라도 이미 많은 이미지가 학습 데이터에 포함되어 활용될 수 있는 경우, 누구나 모델을 학습시켜 그 사람을 딥페이크할 수 있습니다.

AI 생성 모델은 의도적으로 허위조작정보를 생산하고 확산시키는 데 사용될 수 있지만, 소비자 대상 제품에 부정확한 생성형 AI 모델을 사용하면 실수로도 허위 사실을 확산시킬 수 있습니다. 2.2.1장에서 자세히 설명한 바와 같이, 유명 텍스트 생성기는 매우 설득력 있는 허위조작정보를 생산하기 쉬울 뿐 아니라 주장을 뒷받침하기 위한 참조문헌도 만들어내는 경향이 있습니다.[91]

더욱 첨단의 생성형 AI 모델이 신뢰할 수 있는 텍스트를 생성하는 데 보다 효율적이 되면 허위조작정보를 탐지하기가 더 어려워질 수 있습니다. 2023년 3월에 발표된 한 연구에 따르면 챗GPT4는 백신, 음모론, 정치선전에 관한 잘못된 이야기 등에 대해 프롬프트를 입력하면 이전 모델보다 허위조작정보를 더 많이 생성하는 것으로 나타났습니다.[92] 이 기술은 효율적인 도구가 되어 유해한 영향을 미칠 가능성이 있는 그럴듯한 텍스트를 빠르게 생성할 수 있습니다. 선거와 민주적 절차에서 딥페이크와 허위조작정보가 어떻게 작용할지에 대한 논의도 점점 더 치열해지고 있습니다.[93]

실제 같은 여성, DALL-E의 사진. 오른쪽 하단에있는 워터마크에 주목하세요.

2.2.4 AI 생성 콘텐츠의 탐지

"딥페이크 콘텐츠가 확산되면 사람들이 이미지, 텍스트, 사운드 또는 동영상이 실제인지 합성인지 알 수 없기 때문에 신뢰사회가 크게 약화될 수 있습니다."

합성 콘텐츠의 범람에 대한 한 가지 해결책으로 제안된 것은 콘텐츠가 생성형 AI 모델을 사용하여 생성되었음을 '워터마크' 또는 다른 방법으로 명확하게 표시하는 것입니다. 이는 개별 픽셀과 같이 눈에 띄지 않는 워터마크를 통해 이미지 또는 동영상이 AI로 생성되었음을 나타내는 시각적 라벨을 추가하거나, 또는 콘텐츠 출처를 표시하는 데 사용되는 정보를 메타데이터에 추가하는 방법을 통해 이루어질 수 있습니다.[94] 예를 들어, 구글은 사진의 메타데이터에 AI로 생성된 콘텐츠라는 라벨을 자동으로 지정하고 이미지 출처에 대한 컨텍스트를 추가하는 기능을 구현하고 있습니다.[95]

워터마킹은 이미지나 동영상이 진본이 아님을 빠르게 식별하는 데 유용할 수 있지만, 이러한 접근 방식에는 상당한 한계가 있습니다. 워터마킹 시스템은 시스템 개발자 및 모델을 사용하는 사람이 워터마킹 표준을 준수하기로 선택한 경우에만 작동합니다. DALL- E나 미드저니와 같은 폐쇄 소스 이미지 생성기는 생성된 모든 사진의 메타데이터에 워터마크를 필수적으로 추가하겠다고 선택할 수 있지만, 원본 이미지 대신 스크린샷을 찍어 공유하 는 등의 방법으로 이를 우회할 수 있습니다. DALL-E에서 사용하는 것과 같은 시각적 워터마크는 너무 눈에 띄기 때문에 이미지 품질을 현저하게 떨어뜨리지 않고도 사진에서 분리해낼 수 있습니다. 개별 픽셀 같이 눈에 띄지 않는 워터마크는 이미지 컬러 그레이딩을 약간 변경하여 제거할 수 있습니다.

스테이블 디퓨전과 같은 오픈 소스 모델의 경우, 생성된 이미지에 워터마크를 추가하려는 시도는 고의적으로 합성 콘텐츠를 실제 콘텐츠인 것처럼 보이게 하려는 사람에 의해 모델에서 제거될 수 있습니다. 모델에 대한 학습 데이터의 상당한부분에 워터마크가 있는 경우 이 문제를 해결할 수 있지만, 이 경우에도 위에서 설명한 대로 워터마킹을 우회할 수 있습니다.

이미지 생성과 관련된 허위조작정보, 부정확성 및 진본성 문제 외에도, 학생들이 학업 환경에서 텍스트 생성기를 사용할 때 표절을 어떻게 탐지할 것인지에 대하여 중대한 의문이 제기되어 왔습니다. 챗GPT는 리포트를 생성하고 여러 학교 숙제에 답하는 데 널리 사용되어 왔고, 이는 부정 행위와 부정적인 학습 영향에 대한 경각심을 불러 일으켰습니다.[96]

텍스트 생성기에서 복사한 텍스트에는 워터마크를 추가할 메타데이터가 없기 때문에 텍스트의 워터마킹은 이미지나 동영 상보다 더 복잡합니다. 챗GPT로 생성된 텍스트에 텍스트 '서명'을 생성하려는 노력이 계속되고 있지만, 텍스트를 변경하거나 다른 텍스트 생 성기를 통해 텍스트를 공급하면 이를 우회할 수 있습니다.[97]

텍스트가 텍스트 생성기에 의해 작성된 것인지 사람이 작성한 것인지 탐지하고 플래그를 지정하는 시스템은 부정확한 것으로 악명이 높으며,[98] 모든 텍스트를 탐지 시스템에 입력해야 하므로 확장 가능한 솔루션이 아닙니다. 예를 들어, 오픈AI는 챗GPT가 작성한 텍스트인지 여부를 탐지하기 위한 목적으로 생성형 AI 모델을 출시했지만, 이 모델의 정확도는 26%에 불과했습니다.[99] 생성형 AI 모델에 의해 생성된 콘텐츠인지 여부를 탐지하려면 새로운 콘텐츠를 생성하는 것보다 기술적으 로 더 복잡한 솔루션이 필요하므로 탐지 시스템은 이러한 군비 경쟁에서 항상 뒤쳐질 수밖에 없습니다.

AI 모델이 표절로 잘못 분류하는 경우에는 개인이 어떻게 구제될 수 있는지에 대한 의문도 제기될 수 있습니다. 거짓 플래그를 식별하는 것 또한 복잡한 작업일 수 있으며, 이는 표절을 탐지하는 시스템을 사용하는 교사가 정확하게 식별하지못할 수도 있음을 의미합니다. 표절 탐지 시스 템의 최종 사용자(예: 교사)가 텍스트 또는 이미지가 표절로 잘못 신고되었는지 여부를 성공적으로 판단할 수 없는 경우, 학생은 챗GPT가 자신의 리포트를 작성해주지 않았다는 사실을 증명하는 것이 사실상 불가능하기 때문에 곤란한 상황에 처하게 됩니다.

마찬가지로, 사기꾼, 신뢰할 수 없는 사람, 또는 단순히 '비인간'으로 신고되는 경우, 당사자는 구제 수단도 거의 없이 심각한 부정적 영향을 받을 수 있습니다. 예를 들어, AI가 생성한 것으로 보이는 콘텐츠를 식별하고 삭제하는 분류 시스템이 있는 플랫폼의 경우, 이러한 시스템이 콘텐츠를 잘못 분류하여 소비자의 콘텐츠가 삭제되는 피해를 줄 수 있습니다.

요약하자면, 워터마킹 및 탐지 도구는 사진이 실제 사진작가 또는 이미지 생성기에서 유래하였음을 보여주는 등 제한적인 특정 환경에서 작동할 수 있는 기술 솔루션으로서,

광고에 사용하거나 미디어 또는 공공 기관에서 사용할 때 유용합니다. 사진이 실제로 게티이미지 또는 DALL-E에서 제공한 것인지 재빨리 확인하여 허위조작정보의 우발적 확산과 관련된 피해를 일부 줄일 때 유용한 도구가 될 수 있습니다.

그러나 워터마킹으로 이러한 정보 위기를 해결할 수 있다고 믿는 것은 근본적으로 기술적 해결주의적인 접근 방식입니다. 기술적으로 모든 AI 생성 콘텐츠를 정확하게 워터마킹하 는 것이 가능하다고 하더라도, 합성 콘텐츠와 허위조작정보의 범람은 또 다른 기술 층위를 추가한다고 해서 해결될 수 없으며, 특히 고의적으로 오도시키려는 콘텐츠의 경우 더욱 그렇습니다. 미디어와 공공기관에 대한 신뢰 부족은 단순히 합성 콘텐츠와 진본 콘텐츠를 구분할 수 없기 때문에 발생하는 문제가 아닙니다. 사람들이 온라인에서 만나는 모든 미디어를 점검해서 합성 여부를 확인할 것을 기대하는 것도 비합리적입니다.

기술적으로 빠른 해결책은 없으므로 강력한 미디어 리터러시 및 신뢰할 수 있는 미디어 기관 등 다른 해결책을 모색하는 것이 중요합니다. 정책 입안자를 비롯한 사람들에게 지속 가능하고 장기적인 해결책을 제시할 수 있는 미디어 및 사회과학 연구자들도 이 문제에 상당한 관심을 기울여야 합니다.

2.2.5 광고 분야 생성형 AI

생성형 AI 모델의 잠재력은 광고 업계에도 영향을 미치고 있습니다.[100] 이 기술은 이미 광고 카피 생성,[101] 스톡 이미지 합성 및 모델 제작,[102] 마케팅 스턴트의 일부로 사용되고 있습니 다.[103] 이러한 사용 사례는 광고 부문의 노동력을 줄일 수 있지만, 특히 개인화 및 대화형 광고 제작으로 보다 쉽고 효율적으로 사람들을 조종함으로써 소비자들에게 부정적인 영향을 미칠 수도 있습니다.

공개적으로 사용 가능한 생성형 AI의 도입은 대부분 광고 없이 이루어졌지만, 이는 곧 바뀔 것입니다. 2023년 3월, 마이크로소프트는 빙 챗봇에 유료 광고를 도입할 것이라고 발표했습니다.[104] 5월, 구글은 생성형 AI 제품에 광고를 통합할 것이라고 발표했습니다.[105] 소비자가 정확하고 사실적인 정보를 제공하 기 위해 빙과 같은 텍스트 생성기에 의존하는 경우 텍스트 생성기가 제공하는 답변에 광고가 배치되는 것은 오해의 소지를 낳을 수 있습니다. LLM과 상호 작용할 때 행동 조작의 가능성은 소비자 주도성을 희생시켜 더 효과적인 광고를 가능하게 할 수 있습니다.[106]

> "워터마킹이 이러한 정보 위기를 해결할 수 있다고 믿는 것은 근본적으로 기술적 해결주의적인 접근 방식 입니다."

또한, 생성형 AI 모델을 구현하면 특정 집단이나 범주의 사람들에게 맞춤형 광고를 더 쉽게 생성할 수 있기 때문에 차별, 사기, 개인정보 침해 등 감시 기반 광고와 관련된 여러 문제를 악화시킬 수 있으며, 이는 결국 사람이 제품을 구매하거나 발언을 믿도록 만드는 일이 더 쉬워질 수 있습니다.[107] 이 기능은 광고 자체뿐만 아니라 광고의 콘텐츠를 자동으로 표적화하는 기업의 능력을 가속화하고 촉진할 수 있습니다. A/B 테스트와 결합하면 감시 기반 광고의 조작적 성격이 더욱 강화될 수 있습니다.

2.2.5.1 챗봇을 사용한 개인정보 수집

채팅 봇에 내장된 생성형 AI가 소비자를 속여 개인정보를 공유하도록 유도하는 기능에 대한 우려가 커지고 있습니다. 이들은 표적 광고를 제공하거나 소비자가 제품이나 서비스를 구매 하도록 유도하는 용도로 변용될 수 있습니다. 이 문제는 사업 이익을 위해 개인정보의 목적을 변경하는 문제에 대한 광범위한 논쟁의 연장선에 있기도 하지만,[108] 위에서 언급한 바와 같이 인간을 가장하는 생성형 AI 모델의 조작적인 측면이 문제를 더욱 악화시킬 수 있 습니다. 특히 아동이나 독거노인과 같은 취약 계층의 경우, 생성형 AI와 대화할 때 자신에 대한 민감한 정보를 공유할 가능성이 높을 수 있습니다.

예를 들어, 챕터 2.2.2 AI 모델의 의인화에서 논의했듯이 레플리카나 스냅챗 마이 AI와 같은 채팅 애플리케이션은 분명히 모두 최종 사용자에게 자신에 대한 정보를 공유하도록 유도합니다. 마찬가지로, 다양한 애플리케이션에서 내장 검색용으로 사용되는 생성형 AI 모델은 검색하는 사람이 현재 동네 맛집이나 신발에 관심이 있는지 여부 등, 검색 별로 검색 데이터를 수집하고 저장하는 데 사용될 수 있습니다. 개인정보는 방대한 영업 모델들의 기반이 되며, 이러한 텍스트 생성기는 관련성이 높은 소비자 정보를 수집하는 기업의 능력을 향상시킬 수 있습니다.

2.3 편향성, 차별 및 콘텐츠관리

다른 형태의 AI와 마찬가지로, 생성형 AI 모델 은 편향성을 포함하고 있거나, 지속시키거나, 새로운 편향성을 만들어낼 수 있습니다. 인터넷에서 가져온 방대한 정보로 학습된 모델은 학습 데이터의 편향성을 그대로 물려받게 됩니다. 따라서 모델은 부정적이거나 원치 않는 경향성을 재현하는 콘텐츠를 생성할 수 있습니다. 이 때문에 많은 서비스 제공업체들이 콘 텐츠 필터를 추가하여 생성 가능한 콘텐츠를 조절하고 모델의 학습 데이터에 문제가 있는 콘텐츠를 분류하고 있습니다.

2.3.1 학습 데이터의 편향성

앞서 언급했듯이, 생성형 AI 모델은 기존 콘텐츠의 대규모 데이터세트를 학습했기 때문에 사람이 만든 콘텐츠와 유사한 합성 콘텐츠를 생성할 수 있습니다.

이는 데이터세트의 콘텐츠가 가장 중요하다는 의미입니다. 온라인 데이터 스크랩, 선택 및 라벨 지정로부터 관리에 이르기까지 생성형 AI 모델을 위해 학습 데이터세트를 생성하고 조절하는 데에는 여러 단계가 있습니다. 학습 데이터에 대한 신중한 심사, 라벨링, 정제가 이루어지지 않은 상태에서 인터넷에서 스크랩한 데이터세트 는 심각한 다운스트림 효과를 초래할 수 있습니다.

"학습 데이터의 선택과 라벨링은 중립적이지 않습니다. 데이터에서 특정 집단 사람들이 과대 대표될 수 있고, 회사가 이미지에 라벨을 지정하는 방식에 따라 편견이 반영될 수 있습니다."

예를 들어, 이미지 생성기 스테이블 디퓨전 은 독일 비영리 단체 LAION의 오픈 소스 데이터세트를 기반으로 학습됩니다.[109] LAION 데이터세트는 실제 이미지를 포함하고 있지 않고 웹에서 이미지를 가리키는 URL 집합입니다. LAION은 책임성 부족과 콘텐츠관리(유해하거나 불법적일 수 있는 자료 제외 등)의 불충분함으로 인해 비판받아 왔고, 예를 들어 데이터세트에 기밀 의료 정보를 지목하는 URL이 포함된 것으로 밝혀지기도 했습니다.[110]

생성형 AI 모델은 과거 데이터에 대해 학습하므로 데이터세트의 차별적 요소가 생성되는 텍스트, 이미지 또는 사운드에 재현되고 강화될 수 있습니다. 또한 이러한 모델은 기록된 데이터에 대해서만 학습할 수 있으므로 데이터로 기록되지 않거나 정량화할 수 없는 현상이나 사건은 모델에서 인식할 수 없습니다. 따라서 생성형 AI 모델에는 기존의 불공정과 권력 구조를 증폭 시키거나 고착화시키는 편견이 내재될 수밖에 없습니다. [111]

생성형 AI 모델은 주로 인터넷에서 스크랩한 이미지와 텍스트로 학습하기 때문에 학습 단계에 이미 선택 편향이 존재합니다. 인터넷에 접속할 수 없는 인구 세그먼트와 집단(예: 토착주민 집단)은 학습 데이터에서 과소 대표될 가능성이 높으며, 이는 다운스트림에 차별적 영향을 미칠 수 있습니다. 또한, 특정 인구 집단이 과대 대표되는 온라인 커뮤니티가 학습 데이터에 두드러지게 나타나면 과거 데이터에서 과소 대표되었던 인구의 데이터 효과가 지속적으로 감소되는 피드백 루프를 야기할 수 있습니다.[112]

인터넷에서 수집한 학습 데이터에는 음란물, 인종 차별, 고정관념에 찬 콘텐츠가 포함되는 경향이 있습니다.
데이터세트가 조절 또는 정제되지 않은 경우, 이러한 요소가 모델에 포함될 수 있습니다. 예를 들어, 이미지 생성기는 여성, 특히 유색인종 여성을 남성보다 훨씬 더 높은 비율로 성적 대상화하는 경향이 있습니다.[113] 마찬가지로, '아프리카 노동자'와 같은 프롬프트는 육체 노동자의 사진을 생성하는 경향이 있고 '유럽 노동자'는 사무직의 사진을 생성하는 경향이 있습니다.[114]

워싱턴 포스트의 조사에 따르면 구글과 메타의 LLM 학습 데이터로 사용되는 구글의 C4 데이터세트에는 위키피디아, 레딧(Reddit) 및 기타 수많은 토론 게시판, 뉴스 게시자, 정부 웹사이트 등 공개된 웹사이트에서 스크랩한 방대한 양의 텍스트가 포함되어 있습니다.[115] 이는 이 데이터세트로 학습한 모든 생성형 AI 모델이 혐오 발언에서 광고에 이르기까지 모든 것을 포함하는 콘텐츠를 '학습'하게 되며, 이는 생성 가능한 텍스트에 영향을 미칠 수 있다는 것을 의미합니다. 예를 들어, 인종 차별적이거나 유해한 콘텐츠가 많이 포함된 인터넷 게시판에서 데이터를 스크랩하는 경우, 해당 데이터세트에 대해 학습한 모든 모델은 유사한 자료를 다시 생성할 위험이 있습니다.

학습 데이터의 선택과 라벨링은 중립적이지 않습니다. 데이터에서 특정 집단 사람들이 과대 대표될 수 있고, 회사가 이미지에 라벨을 지정하는 방식에 따라 편견이 반영될 수 있습 니다. 예를 들어, 개발자는 학습 데이터 라벨에 포함할 인종 및 성별 범주의 갯수를 선택할 수 있으며 라벨에 이러한 속성을 전혀 포함하지 않을 수도 있습니다. 다른 AI가 생성한 콘텐츠로 모델을 학습시키면 편견이 더욱 강화될 위험이 있습 니다. 결과적으로 각 학습 세션이 편향되어 있거나 차별적인 데이터 시퀀스를 강화하는 피드백 루프가 발생할 수 있습니다.

많은 AI 모델이 백인이 아닌 사람의이미지를 인식하고 라벨링하는 데 문제가 있는데, 이는 부분적으로 백인이 과도하게 대표되는 데이터세트를 기반으로 모델을 학습시켰기 때문일 수 있습니다. 구글[116]과 메타[117]는 자사 이미지 인식 알고리즘이 피부색이 어두운 사람을 고릴라 또는 영장류로 분류해서 비판을 받았습니다. BERT와 같은 언어 처리 모델도 장애인을 더 부정적인 감정 단어와 연결하는 것으로 나타났습니다.[118]

2.3.1.1 차별적 결과

생성형 AI 모델 사용으로 인한 편향적이거나 차별적인 결과는 학습 데이터와 관련된 문제만이 아닙니다. 기업과 사람들이 모델을 사용하거나 사용하지 않기로 선택하는 방식에 따라 인적 또는 시스템적 편향성이 내재되거나 강화될 수도 있습니다.[119] 예를 들어, 다양한 작업에서 텍스트 생성기 사용을 필수화하면 기술적으로 숙련도가 낮은 집단을 간접적으로 배제할 수 있습니다.

복잡한 문제를 해결하기 위해 AI 모델을 구현할 때는, 2.1.2절에서 자세히 논의했듯이, 많은 비용이 들거나 복잡할 수 있는 보다 효과적인 솔루션의 우선순위가 낮아질 위험이 있습니다. 예를 들어, 세계보건기구(WHO)는 특정 문제들이 해결되지 않는 한 보건의료 분야에서 AI 모델을 사용하는 것이 고령자에게 부정적인 영향을 미칠 수 있다고 경고한 바 있습니다.[120]
AI 모델이 연령 차별적 고정관념이 포함된 데이터로 학습될 수 있다는 점과 노년층이 학습 데이터에서 과소평가되는 경우가 많다는 점 등을 우려한 것입니다. 이는 연령 차별을 고착화하고 고령 인구의 건강 및 사회 서 비스의 품질을 떨어뜨릴 수 있습니다.

2.3.2 콘텐츠관리

충분히 큰 데이터세트로 학습된 경우, 생성형 AI 모델이 생성할 수 있는 자료에는 거의 한계가 없습니다. 위에서 언급한 바와 같이, 이러한 많은 모델들이 불법적이고 차별적이며 기타 허용할 수 없는 합성 콘텐츠를 생성하는 데 사용될 수 있는데, 이는 이들이 미덥지 않은 여러 콘텐츠를 포함했을 수도 있는 데이터세트로 학 습되기 때문입니다. 이러한 문제를 완화하기 위해 많은 AI 모델에는 특정 콘텐츠를 필터링하고 분류하거나 기술 사용 방식에 제한을 두는 콘텐츠관리 기능이 있습니다. 콘텐츠 필터를 사용해서 특정 유형의 콘텐츠 생성을 제한할 수도 있지만, 이러한 방식에는 여러가지 단점이 있습니다.

우선, 콘텐츠 관리는 유해한 자료와 허용되는 자료가 무엇인지 법률이 명확하게 정의하지 않는 한, 시스템 소유자에게 이를 결정할 수 있는 상당한 권한을 부여합니다. 예를 들어, 오픈AI는 특정 정치 주제에 대한 텍스트 생성을 거부함으로써 특정 관점을 제한한다 는 비판을 받고 있습니다. 이는 허용 가능한 콘텐츠를 결정할 수 있는 민간 기업의 권한을 키우면서 심각한 결과를 초래할 수 있는 권력 남용을 낳을 수 있습니다.

소셜 미디어 플랫폼의 콘텐츠관리 관행과 마찬가지로, 생성형 AI 모델의 콘텐츠 필터도 무해하거나 중요한 콘텐츠를 필터링하거나 금지하는 등 과도하게 관리할 위험이 있습니다. 이러한 일은 실수로 또는 고의로 발생할 수 있습니다. 예를 들어 이미지 생성기인 미드저니는 음란물 콘텐츠를 생성하는 최종 사용자를 단속하기 위해 과학적 해부학 용어를 필터링하기 시작 했습니다.[121] 이 회사는 또한 중국에서 차단되는 것을 방지하기 위해 소비자가 시진핑의 이미지를 생성하지 못하도록 콘텐츠 필터를 추가했으며, 결국 미드저니가 생성한 도널드 트럼프 체포 사진이 입소문을 타면서 무료 평가판 서비스를 중단했습니다. 이 회사는 "드라마를 최소화"하기 위해 플랫폼에서 어떤 단어나 프롬프트가 금지되는지 공개적으로 밝히지 않고 있습니다.[122]

콘텐츠 필터가 수행할 수 있는 작업에는 기술적인 한계도 있습니다. 생성형 AI 모델을 제한하기 위해 콘텐츠 필터가 사용되는 모든 곳 에서 모델을 우회하거나 탈옥하려는 시도가 발생할 것입니다. 예를 들어 모델에 콘텐츠 필터를 우회할 수 있는 캐릭터를 시뮬레이션하도록 지시하는 등 금지된 콘텐츠를 생성하는 데 사용될 수 있는 다양한 프롬프트가 발견되었습니다.[123] 기업들이 발견된 허점을 막기 위해 서두르면 이러한 무기 경쟁이 과도한 규제를 초래할 가능성이 높습니다.

생성형 AI 모델의 콘텐츠관리가 차별적인 관행을 만들거나 시행할 수도 있습니다. 예를들어, 혐오 발언이나 차별적인 콘텐츠를 학습 데이터에서 제거한다는 이유로

성소수자 커뮤니티 또는 기타 소수자를 지칭하는 단어를 분류할 수 있습니다. 그러나 허용되지 않는 콘텐츠를 제거하려는 이러한 시도는 실제로는 성소수자 커뮤니티에 대하여 긍정적인 측면과 정서를 보여주는 콘텐츠까지 삭제하는 결과를 초래할 수 있습 니다. 이러한 방식으로 콘텐츠관리는 과소 대표성을 강화할 수 있습니다.

데이터세트나 모델에 내재된 편향성 대신 편향된 결과물을 해결하기로 선택하는 것에도 본질적인 문제가 있습니다. 관리 시도는 각 유형의 편향된 결과물을 억제해야 하는데, 생성형 AI의 편향성에 효과적으로 접근하는 것은 두더지 잡기 게임과 같습니다.[124] 사후 콘텐츠관리에 의존하는 대신 데이터 세트의 조절에 더 많은 관심을 기울여 내재되어 있는 유해한 편향성을 줄이는 것이 필요합니다.

2.3.2.1 문화적 맥락

콘텐츠 관리는 중립적인 과정이 아니며, 콘텐츠의 맥락을 이해하는 것이 매우 중요합니다. 따라서 문화적 맥락이 다르면 대규모 콘텐츠 검토에 상당한 장벽이 발생할 수 있습니다. 예를 들어, 특정 언어나 방언에 대한 학습 데이터나 관리가 충분하지 않아 과잉 또는 과소 관리가 발생할 위험이 있습니다. 영어와 같이 널리 사용되는 언어의 경우 학습 데이터에서 많은 텍스트 말뭉치를 보유하여 더 정확한 정보를 제공하기 때문에, 결과적으로 더 나은 관리로 이어질 수 있습니다.

다른 언어와 문화는 학습 데이터에서 과소 대표되는 경우가 많으며, 이는 콘텐츠 관리가 덜 정확하거나 존재하지 않을 가능성이 높음을 의미합니다. 소수 집단은 모델을 개발하고 학습하는 사람들 사이에서 심각하게 과소 대표되는 경향도 있습니다.[125] 또한 문화적 맥락 및 국내법과 관련된 중요한 문제가 있는데, 특정 장소와 상황에서는 사회적으로 허용되고 합법인데 다른 곳에서는 금기이거나 불법일 수 있습니다.

콘텐츠관리와 관련된 상황적 복잡성 때문에 이는 자동화에 적합하지 않을 수 있는 작업이기도 합니다. 출력물을 조정하고 학습 데이터에 주석을 다는 작업은 어떤 경우 자동화되어 있지만 수작업이 수반되는 경우도 많습니다. 데이터 정제, 콘텐츠 분류, 콘텐츠관리와 같은 과정은 인간의 정신적인 노동력을 필요로 하는 경우가 많습니다. 이에 대해서는 이하 노동 착취 절에서 자세히 설명합니다.

2.3.2.2 오픈 소스 모델과 콘텐츠 필터의 한계

실제로 콘텐츠관리는 중앙 집중식 폐쇄 소스 모델에서만 작동합니다. 스테이블 디퓨전과 같은 오픈 소스 모델에서는 모델이 생성할 수 있는 콘텐츠를 통제하는 것이 사실상 불 가능합니다. 개인을 포함한 다운스트림 개발자는 합법성과 무관하게 모든 종류의 이미지를 만들 수 있는 모델을 학습시키고 공유할 수 있습 니다. 이 모델은 인터넷 연결이나 클라우드 서버에 대한 접근 없이 로컬에서 실행되므로 모델을 배포 한 회사가 모델이 생성할 수 있는 내용을 검열하거나 제한할 수 없습니다.

2.4 프라이버시와 개인정보 보호

프라이버시권은 민주 사회의 핵심 가치 중 하나입니다. 프라이버시는 타인과의 통신에 대한 프라이버시, 정체성과 생각에 대한 프라이버시, 자신에 대한 데이터와 정보에 대한 프라이버시 등 다양한 측면을 포괄합니다. 개인정보 보호는 특히 온라인 서비스의 맥락에서 프라이버시의 하위 개념이자 중요 부분이지만, 프라이버시는 개인에 대한 보호를 훨씬 더 광범위하게 포괄합니다.

개인정보는 오랫동안 산업계에서 매우 가치있는 것 으로 간주되어 왔으며, 특히 개인과 집단을 대상으로 광고를 표적화하고 참여도를 측정하거나 회사 서비스를 개선하는 등의 목적으로 사용될 수 있습니다. 생성형 AI 모델이 인터넷에서 스크랩한 자료로 학습할 때, 이 학습 데이터는 일반적으로 대량의 개인정보를 포함하고 있습니다. 생성형 AI가 개발 및 배치됨에 따라 개인정보와 관련된 이러한 문제는 상당한 프라이버시 침해를 낳을 수 있습니다.

2.4.1 모델 학습에 사용되는 데이터세트와 관련된 프라이버시 문제

이미지 생성기는 일반적으로 실제 사람의 이미지가 포함된 방대한 데이터세트에 대해 학습합니다. 이러한 이미지는 예를 들어 소셜 미디어와 검색 엔진에서 가져올 수 있으며, 법적 근거나 당사자 인지 없이도 사진 속 인물을 사용할 수 있습니다. 마찬가지로 텍스트 생성기는 개인에 대한 개인정보 또는 개인 간의 대화를 포함할 수 있는 데이터세트를 학습합니다.

맥락에서 벗어난 개인정보로 생성형 AI 모델을 학습시키는 경우, 이는 개별 소비자의 문맥 무결성을 침해할 수 있습니다. 예를 들어 소셜 미디어와 같은 온라인에 자신의 사진을 업로드할 때, 소비자는 이 사진이 AI 모델을 학습시키는 데 사용될 것이라고 예상할 수 없었습니다. 해당 개인은 이러한 일이 발생할 것이라고 고지받은 바 없고, 자신의 초상권 사용에 동의한 적이 없으며, 자신의 프라이버시 및 개인정보에 대한 권리가 침해되었다는 사실도 알지 못할 가능성이 높습니다.

생성형 AI 모델의 학습 방식에 대한 대중의 인식이 높아짐에 따라 학습에 개인정보를 사용하는 것이 위축 효과를 초래할 수 있습니다. 당국이 생성형 AI 모델 을 배치하는 기업에 대해 GDPR과 같은 현행 법률을 집행하고 사람 이미지 사용에 대한 보호 장치와 제한 사항이 마련되지 않는 한, 자신의 이미지가 학습 데이터로 사용되는 것을 원하지 않는 소비자가 선택할 수 있는 유일한 방법은 온라인에 사진을 게시하지 않는 것뿐입니다. 이러한 해결책은 분명 불충분합니다.

2.4.2 생성된 콘텐츠와 관련된 프라이버시 문제

생성형 AI 모델이 딥페이크처럼 개인에 대한 새로운 이미지를 생성할 수 있는 경우 특히 문제가 됩니다. 이는 개인이 통제할 수 없는 방식으로 개인에 대한 '새로운' 개인정보를 생성하는 것에 해당합니다.
이는 생성된 콘텐츠에 묘사된 개인의 무결성을 침해하며, 잠재적으로 매우 침입적이거 나 유해한 방식으로 이루어질 수 있습니다.

때로는 생성형 AI 모델이 사진을 정확하게 재현할 수 있습니다. 이는 모델이 특정 데이터에 대해 '과학습'된 경우 일어납니다. 예를 들어, 모나리자는 매우 유명한 예술 작품이기 때문에 예술 작품이 포함된 학습 데이터세트에서 과대 표현될 가능성이 높습니다. 이 경우 모델이 모나리자의 얼굴에 대해 과도하게 학습하여 그림을 아주 정확하게 그릴 수 있습니다. 특정 인물의 사진을 과도하게 학습하면 동일한 효과를 얻을 수 있습니다. 이는 생성형 AI가 일반 인터넷 사용자보다 유명 연예인의 사진을 재현할 가능성이 더 높다는 의미입니다. 그러나 스테이블 디퓨전과 같은 오픈 소스 모델을 사용하면 개인을 포함한 모든 다운스트림 개발자가 누구의 얼굴로든 모델을 학습시켜 딥페이크 제작에 사용할 수 있습니다.

사진 생성 및 이로 인해 소비자에게 미칠 수 있는 부정적인 영향 외에도 개인에 대한 텍스트 생성의 문제도 있습니다. 여기에는 사람들에 대하여 허위 또는 비방하는 주장을 생성하는 텍스트 생성기의 문제가 있습니다. 예를 들어, 챗GPT는 어떤 교수가 성희롱 스캔들에 연루되었다는 허위의 주장이나 어떤 시장이 감옥에서 복역했다는 허위의 주장 등, 관련된 사람들에게 위험한 결과를 초래할 수 있는 내용의텍스트를 생성한 바 있습니다.[126]

2.5 보안 취약성 및 사기

악의적인 행위자가 생성형 AI 모델을 악용하여 범죄적 행동을 확대하거나 강화할 수 있습니다. 다른 분야와 마찬가지로 생성형 AI는 사기, 피싱 및 기타 행동을 보다 효율화하기 위해 사용될 수 있습니다. 이 모델은 기존 보안 시스템에 문제를 일으킬 수도 있습니다. 생성형 AI를 사용하여 수행할 수 있는 사이버 범죄 유형이 새로운 것은 아니지만, 기술의 편재성과 사용 편의성으로 인해 이러한 공격 유형이 더욱 확대될 수 있습니다.

사기꾼은 대량의 언어 모델을 사용하여 피해자를 속이기 위해 그럴듯해 보이는 텍스트를 대량으로 생성할 수 있습니다. 마찬가지로, 사기범이 정기적인 접촉을 통해 피해자와 신뢰를 형성하는 캣피싱 사기 역시 첨단 챗봇을 사용하면 그럴듯하게 자동화할 수 있습니다. 즉, 범죄자가 더 적은 시간과 자원을 들여 더 많은 피해자를 대상으로 효과적으로 사기행각을 할 수 있다는 의미입니다.

딥페이킹은 보안 수단을 우회하는 데에도 사용될 수 있습니다. 사진과 목소리를 그럴듯하게 위조할 수 있게 되면 새로운 방식으로 사기 행각을 할 수 있게 됩니다. 예를 들어, 한 기자는 자신의 목소리 클립을 위조하여 은행 계좌의 음성 인식 기능을 우회할 수 있었습니다.[127] 마찬가지로, 음성 생성기가 범죄 목적으로 가족을 사칭하는 데 사용된 것으로 알려졌습니다.[128]

LLM은 필터와 보안 조치를 우회하는 악용('탈옥'), 학습 데이터를 고의적으로 조작하는 악용(' 데이터 중독'), 이메일 등에 숨겨진 명령을 통해 모델이 특정 작업을 수행하도록 유도하는 악용('프롬프트 인젝션')에 취약합니다.[129]

"오픈AI와 같은 기업이 데이터를 사용하는 방식에 대한 투명성이 부족하기 때문에 기밀 정보가 남용될 수 있다는 우려도 커지고 있습니다. "

이러한 보안 취약성은 기업들이 충분한 보안 테스트 없이 생성형 AI를 다양한 서비스에 빠르게 통합하여 경쟁에서 앞서 나가려고 할 때 심각한 문제가 될 수 있습니다.

사이버 보안 전문가들은 텍스트 생성기가 멀웨어와 같은 악성 코드를 작성하는 데 사용되어 무기화될 수 있다고 경고했 습니다.[130] 이는 사이버 범죄자들이 전통적으로 이러한 활동과 관련되었던 기술적 숙련도 없이도 바이러스 및 기타 유해한 코드를 생성할 수 있다는 것을 의미합니다. 마찬가지로, 약물 발견을 위해 구축된 AI 모델은 잠재적으로 생물학적 무기를 설계하는 데에도 사용될 수 있습니다.[131] 유로폴은 LLM이 다양한 유형의 사이버 범죄에 사용될 가능성에 대해서도 경고했습니다. 이 기관에 따르면 이러한 제한을 우회하거나 모델을 탈옥할 수 있는 방법이 많기 때문에 콘텐츠 관리로는 충분하지 않을 수 있습니다.[132]

오픈AI와 같은 기업이 데이터를 사용하는 방식에 대한 투명성이 부족하기 때문에 기밀 정보가 남용될 수 있다는 우려도 제기되고 있습니다. 몇몇 유명 기업들은 직원들에게 챗GPT에 회사 정보를 입력하는 데 대하여 경고하거나 금지했습니다.[133] 아마존은 텍스트 생성기가 회사 내부 문서와 거의 일치하는 텍스트를 생성하는 것을 발견했습니다.[134] 이는 생성형 AI 모델을 통해 기밀 정보가 유출될 위험이 있음을 의미합니다.

2.6 소비자 대면 애플리케이션에서 인간을 전체적 또는 부분적으로 생성형 AI로 대체하기

생성형 AI 모델이 처음 대중 앞에 나타났을 때는 주로 최종 사용자가 콘텐츠를 생성할 수 있는 독립형 시스템이었습니다. 이들 시스템에 대한 관심이 높아지면서 시스템 소유자는 API를 통해 이를 다른 애플리케이션 및 시스템에 통합할 수 있도록 하였습니다. 이는 오락용 애플리케이션 및 시스템에 도입되는 부가 기능일 수 있지만, 의사 결정 시스템을 부분적으로 또는 완전히 자동화하거나 소비자 대면 서비스에서 인간 상호 작용을 대체할 것이 예상되기도 합니다.

이는 광범위한 영향을 미칠 수 있습니다. 예를 들어, 오픈AI 설립자인 샘 알트먼은 미래에는 생성형 AI 모델이 너무 가난해서 의료 서비스를 받을 수 없는 사람들을 위해 의료 조언을 할 수 있을 것이라고 주장했습니다.[135] 2023년 5월, 미국의 한 비영리 섭식장애환자 지원단체는 상담센터의 직원과 자원봉사자를 해고하고 AI 챗봇으로 대체했습니다.[136] 이 단체 대변인은 챗봇이 상담센터를 직접 대체하는 것이 아니라고 주장했지만, 그럼에도 불구하고 상담센터 폐쇄와 더불어 대화를 나눌 수 있는 실제 사람을 남겨두지 않았습니다.

이러한 업무를 자동화하면 학습 데이터나 모델 자체에 문제가 있는 경우 치명적인 실수가 발생할 위험이 배가될 수 있습니다.

수년 동안 기업들은 챗봇을 통해 고객 서비스를 자동화하는 등 소비자와의 상호작용을 자동화하려는 시도를 해왔습니다. 많은 기업이 소비자가 사람과 접촉하는 것을 어렵게 만들었으며, 이는 FAQ 및 유사 문서에서 다루는 표준에서 벗어난 문제를 가진 소 비자에게 부정적인 영향을 미칩니다. 생성형 AI의 등장으로 기업이 소비자가 실제 인간과 접촉하는 것을 더욱 어 렵게 만들 위험이 있습니다.

2.6.1 인간과 자동화된 의사 결정을 결합하는 문제

자동화 시스템은 윤리적 성찰, 공감 또는 이해 능력이 없습니다. 일반적으로 사람들은 경미한 위반으로 인해 고통받지는 않지만, 자동화 시스템은 경미한 위반과 중대한 위반을 구분할 수 없습니다.

소비자가 대금 결제를 하루 놓친 경우, 인간 담당자는 엄격한 규정 준수보다 고객 관계를 우선시할지 여부를 검토해서 추가 요금 없는 연체를 허용할 수 있습니다. 자동화 시스템에서는 이러한 고려를 할 수 없습니다. 따라서 절차를 자동화하는 과정에서 공정성에 대한 공감과 원칙이 사라질 수 있습니다.

완전히 자동화된 시스템은 일반적으로 이 의사 결정으로 인한 추가 위험을 고려하기 위해 추가적인 법적 조항과 보호로 규제됩니다. 이에 경우에 따라 인간의 개입을 요구하는 요구사항을 두거 나,[137] 기업이 법적 조사를 방지하기 위해 절차에 인간을 투입하기로(human in the loop) 결정할 수 있습니다.[138] 그러나 인간을 투입하는 것은 복잡한 조치라서 몇 가지 곤란한 점이 있습니다.

인간은 자동화된 시스템의 결과에 과도하게 의존하거나 과소 의존할 수 있으며,[139] 결정을 설명하거나 해석할 수 없는 자동화된 컴퓨터 시스템에서 특히 문제가 두드러지게 나타납니다. 개인이 자신의 결정에 지나치게 의존하는 것과 정반대로, 자동화된 시스템의 출력에 지나치게 의존하는 것은 가장 새로운 유형의 문제를 등장시켰으며, 이는 전적으로 수동적인 의사 결정 절차와 유사한 문제를 나타냅니다.

전체 또는 부분적으로 자동화된 시스템에서 지나친 의존은 서로 다른 사람에게 각각 영향을 미칠 수 있습니다. '절차에 투입된 사람'은 신중해야 하는 경우에도 시스템에 이의를 제기하지 않을 수 있고, 결정의 영향을 받는 사람은 이의를 제기 하거나 결정에 대한 인적 검토를 요구하지 않을 수 있습니다. 두 경우 모두 결정의 영향을 받는 사람의 이익을 위험에 처하게 합니다.

이전 절에서 설명한 바와 같이 챗GPT와 같은 텍스트 생성기의 출력은 매우 그럴듯해 보이는 것으로 밝혀졌습니다. 소비자에게 영향을 미치는 의사 결정 절차에 텍스트 생성기를 사용할 경우, 그 결과에 과도하게 의존할 위험이 증가할 수 있습니다. 이러한 효과는 최종 사용자가 확률적 텍스트 생성기가 아닌 지각이 있는 지적인 존재와 상호작용하고 있다고 믿을 때 더욱 악화될 수 있습니다.

담당자가 어떤 결정을 이해하고 이를 신뢰할 수 없다고 판단하면 이를 거부할 것을 검토할 때에도 추가적인 장벽이 있을 수 있습니다. 사업적 관점에서 보면 절차의 전체 또는 일부를 자동화함으로써 얻을 수 있는 효율성이 있습니다. 기계의 결정이 일반적으로 유지되는 경우, 결정을 거부하려면 결정을 수용할 때보다 더 힘겨운 논박이 필요할 수 있습니다. 반복적으로 결정을 거부해서 효율성을 저해하는 담당자는 문제를 일으키는 사람으로 보일 수 있습니다.

담당자 책임(responsibility) 또는 법적 책임(liability) 문제가 있어 개인인 인간 담당자가 결정을 번복하는 것이 어려울 수 있습니다. 컴퓨터 시스템의 잘못된 결정은 해당 시스템에 책임을 물을 수 있지만, 그 결정을 거부하는 것은 담당자가 그 결정에 대하여 책임을 지게 되므로 담당자의 위기의식이 상당히 높아질 수 있습니다. 법적 책임 제도는 담당자의 실제적 또는 인지적 위기의식을 고조시킬 수 있습니 다.

"소비자에게 영향을 미치는 의사 결정 절차에 텍스트 생성기를 사용할 경우, 그 결과에 과도하게 의존할 위험이 증가할 수 있습니다."

2.7 환경 영향

학계 및 과학계에서 점점 더 많은 사람들이 생성형 AI 모델 개발이 환경에 미치는 영향에 대하여 문제제기하고 있습니다. 기후 변화와 천연자원 부족이 전세계 모두의 문제인 상황에서 생성형 AI가 기후 변화를 해결할 수 있다는 주장과 이러한 기술이 실제 환경에 미치는 영향 사이에 딜레마가 발생하고 있습니다.

이 절에서는 이러한 주장 중 일부를 자세히 살펴보고 생성형 AI가 현재와 근미래 모두에 환경에 미치는 현실적인 영향에 대하여 비판적으로 검토합니다.

이러한 영향 중 대부분이 광범위한 기술 분야에 적용되는 문제이지만, 생성형 AI를 둘러싼 과대 광고 속에서 이러한 관점을 잃지 않는 것이 중요합니다.

2.7.1 기후 영향

생성형 AI 분야의 일부 주체들은 이 기술이 기후 변화의 위기에서 우리를 구할 수 있는 잠재력을 가지고 있다고 주장합니다.[140] 그러나 현재 접할 수 있는 데이터에 따르면, 대형 기술 회사들이 지금까지 운영되어 온 것과 동일한 방식으로, 생성형 AI의 배치는 기후 변화, 물 부족, 높은

에너지 소비와 같은 문제들을 해결하기 보다 문제를 보태고 있습니다.

기술 산업은 이미 상당한 양의 탄소를 배출하고 있습니다. UNEP에 따르면 2021년 기술 업계의 탄소 배출량은 전 세계 탄소 배출량의 2~3%를 차지했습니다.[141] 2022년 11월 MIT는 "클라우드의 탄소 발자국이 항공 산업 전체보다 더 크다"고 보고했습니다. 생성형 AI도 이러한 부정적인 추세에서 예외가 아닙니다.

2023년 5월, AI는 "다른 형태의 컴퓨팅보다 더 많은 에너지를 사용하며, 단일 모델을 학습하는 데 100개의 미국 가구가 1년 동안 사용하는 것보다 더 많은 전력을 소비할 수 있다"고 합니다.[142] 데이터센터는 엄청난 양의 에너지를 사용하는 것으로 알려져 있으며, 이미 5년 전에 전 세계 컴퓨팅의 에너지 수요가 10년 내에 전 세계 총 전력 생산량을 초과할 것으로 예측된 바 있습니다[143]. 이는 생성형 AI가 빠르게 개발되어 배치되기 전의 수치입니다.[144] 생성형 AI 모델이 기하급수적으로 성장하고 이러한 성장을 뒷받침하는 인프라에 대한 투자가 증가함에 따라 에너지 사용량과 탄소 배출량이 급증할 것으로 예상됩니다.

포브스는 최근 "생성형 AI가 데이터센터를 파괴하고 있다"고 보도했습니다. [145] 실제로 티리아스 리서치의 조사에 따르면 AI 개발로 인해 데이터센터 인프라 및 운영 비용이 2028년까지 760억 달러 이상으로 증가할 것으로 예상됩니다. 티리아스 리서치는 "이는 현재 전 세계 클라우드 인프라 서비스 시장의 3분의 1을 점유하고 있는 아마존의 클라우드 서비스 AWS의 연간 예상 운영 비용의 2배가 넘는 비용"이라고 추정합니다.[146] 이러한 기하급수적인 성장에는 환경에 대한 대가가 따릅니다. 정확한 비용은 아직 계산되지 않았지만, 하나의 지표로서, LLM을 검색 엔진에 통합하려는 계획이 시행될 경우 개별 검색 쿼리당 에너지 사용량이 4배 증가 할 수 있습니다.[147]

즉, AI 기술은 높은 탄소 발자국을 동반하며,[148] 생성형 AI 모델을 설계, 학습, 개발, 배치 및 사용하는 모든 단계에 에너지가 필요하다는 점이 분명합니다.[149] 문제는 생성형 AI 개발에 필요한 에너지 양에 대한 데이터가 아직 부족하다는 점입니다.[150] 이 글을 쓰는 현재 시점까지 생성형 AI 모델의 수명주기 동안 얼마나 많은 에너지가 필요한지 그 수치를 공개한 기업은 없었습니 다.

"데이터에 따르면, 대형 기술 회사들이 지금까지 운영되어 온 것과 동일한 방식으로, 생성형 AI의 배치는 기후 변화, 물 부족, 높은 에너지 소비와 같은 문제들을 해결하기 보다 문제를 보태고 있습니다.

생성형 AI의 에너지 소비는 기하급수적으로 증가하고 있으며, 향후 5~10년 동안 더 많은 연구에 반영되어서 소비자는 정보에 접근할 수 있어야 하고 정책 입안자들은 이 산업이 얼마나 많은 에너지를 배출해야 하는지 규제할 수 있어야 할 것입니다. 예를 들어, 딥러닝 모델 학습에 사용되는 컴퓨팅 파워의 양은 2012년부터 2018년까지 6년 동안 30만 배 증가했습니다.[151]

현재 AI 모델의 탄소 배출량을 측정할 수 있는 표준화된 방법은 없으며, AI 중심 기술 기업이 필요한 정보를 공개하려는 호의를 보이지도 않습니다. 메타, 구글, 마이크로소프트 와 같은 기존 기술 기업은 매년 지속가능성 보고서를 발간하여 에너지 및 물 사용량과 탄소 배출량을 자체적으로 보고하는 반면, 오픈AI와 같 은 AI 기업은 환경에 미치는 영향과 이를 완화하는 방법에 대한 어떠한 정보도 공개하지 않고 있습니다.

참고로, 뮌헨 공과대학교의 2021년 연구에 따르면 대형 기술 기업들이 보고서에 노력을 기울이고 있더라도 그들 자신의 탄소 배출량을 과소 보고하고 있을 가능성이 있습니다.

"설문조사에 참여한 56개 주요 기술 기업의 표본에서 2019년 배출량 중 절반 이상이 자체 보고서에서 제외되었습니다. 누락된 배출량은 약 390메가톤의 이산화탄소 환산량으로, 호주의 탄소 발자국과 비슷한 수준입니다.".[152]

기술 기업은 생성형 AI가 발생시키는 탄소 배출량을 계산하는 데 대한 관심이 적은데, 이는 이들 기업이 대량의 컴퓨팅 능력으로[153] 다른 모든 고려 사항을 희생해서라도[154] 보다 정확한 결과를 얻는데 관심을 두기 때문입니다. AI 업계에서는 모델과 데이터세트의 기하급수적인 크기를 다른 무엇 보다도 중요하게 여기는 "클수록 좋다"는 접근 방식이 지속되어 온 것 같습니다.[155] 안타깝게도 이러한 접근 방식은 지속 가능하지 않습니다. 연구자들은 이러한 현상을 " 레드 AI"라고 부릅니다.[156] 이 현상은 컴퓨팅 비용과 탄소

비용을 급격히 증가시키는 결과를 초래합니다. 연구진은 모델의 효율성과 환경 영향 감소에 초점을 맞춰 친환경 AI를 구현할 수 있다고 주장합니다.

생성형 AI의 환경 영향에 대한 보다 투명한 접근 방식도 가 능합니다. 보다 윤리적이고 투명한 AI 산업을 위해 노력하 는 스타트업 허깅 페이스(Hugging Face)는 자체 LLM인 BLOOM의 탄소 배출양에 대한 데이터를 공개하였습니다.[157] BLOOM은 이산화탄소를 배출하지 않는 원자력 에너지로 구동되는 프랑스 슈퍼컴퓨터로 학습되었기 때문에 비슷한 크기의 LLM에 비해 배출량이 현저히 낮습니다. 하지만 아직 배치되지 않은 상태에서 모델 학습을 마친 BLOOM은 이미 뉴욕과 런던 간 60회 비행에 해당하는 양의 이산화탄소를 배출했습니다[158].

일부에서는 기술 업계가 AI 개발의 탄소 배출량 측정을 거부하고 있다고 주장하고, 다른 한편에서는 활동 위치에 따라 에너지 사용량이 다르기 때문에 탄소 배출량 측정이 매우 어렵다고 말합니다.[159] 그러나 생성형 AI가 기후 변화로부터 우리 를 구할 수 있다고 믿는다면, 자체 탄소 배출량을 계산할 수 있는 능력이 있다고 기대하는 것이 합리적일 것입니다.

생성형 AI가 환경에 미치는 중대한 영향을 해결하기 위해 기업은 에너지 사용량, 에너지 조달 방법, 특히 학습, 개발, 배치 및 사용을 포함한 전체 수명 주기 동안 모델이 배출 하는 탄소량을 공개해야 합니다. 정책 입안자들이 제3자 전 문가가 이상적으로 측정하고 통제하는 이 데이터에 접근 할 수 없다면, 업계에 책임을 묻고 기후와 환경에 대한 통제 되지 않고 불균형적인 영향을 제한하는 것은 불가능합니다.

AI가 기후 변화로부터 우리를 구할 수 있을지 의문을 제기하는 것은 분명 가치있는 일입니다. 액센츄어의 기술 지속 능력 혁신 부문 상무이사이자 글로벌 책임자인 산제이 포더는 "데이터의 기하급수적인 증가와 이로 인한 에너지 수요 증가는 오히려 기후 변화에 대한 전 세계적인 진전을 방해하고 지연시킬 수 있다"고 말합니다.[160] 저자 나오미 클라인은 기후 변화를 막는 데 필요한 데이터는 부족하지 않지만, 국가와 탄소 배출량이 많은 기업의 구체적인 행동과 배출량 감축이 필요하다고 지적합니다.[161]

2.7.2 물 발자국

물은 기후 위기의 중심 문제입니다. 기후 변화에 관한 정부 간 협의체(IPCC)에 따르면 전 세계 인구의 약 절반이 일 년 중 적어도 일부 기간 동안 심각한 물 부족을 경험하고 있습니다.[162] 세계기상연구소에 따르면 이 수치는 기후 변화로 인해 더욱 증가할 것으로 예상됩니다.[163]

또한 산업 성장으로 인해 2000년부터 2050년까지 전 세계 물 수요가 55% 증가할 것이라는 전망도 있습니다.[164] 생성형 AI 의 개발 및 배치를 포함한 기술 산업은 물 수요 증가를 보태는 분야입니다. 물은 주로 데이터 센터를 냉각하는 데 사용됩 니다. 예를 들어, 마이크로소프트는[165] 2022년에 전년보다 170만m3 더 많은 640만m3의 물을 소비할 것이라고 보고했습니다.

AI의 개발, 학습, 배치, 사용으로 인해 물 수요가 더욱 커지고 있습니다. 최근 연구에 따르면 오픈AI의 LLM인 GPT-3을 학습시키기 위해 원자로 냉각탑을 채울 수 있는 양의 물이 필요했습니다.[166] 이 연구에 따르면 챗GPT는 최종 사용자와 기초적인 소통을 완료하는 데만 0.5리터의 물을 소비했습니다.[167] 이 사례는 마이크로소프트의 최첨단 미국 데이터 센터에서 이루어지는 물 소비량을 측정한 것이지만, 연구자들은 에너지 효율이 낮은 데이터 센터의 경우 물 소비량이 3배 더 많을 것으로 추정했습니다. GPT-4와 같은 최신 모델에서는 물 수요가 더 증가할 것으로 예상됩니다.[168]

메타, 구글, 마이크로소프트를 비롯한 일부 기업은 2030년까지 '물 포지티브'를 목표로 하고 있다고 주장하지만, 오픈AI와 같은 기업은 자기 활동의 물 사용 유형에 대하여 보고한 바 없습니다. AI 개발의 물 발자국은 아직 제대로 측정되고 있지 못한 것입니다.[169]

"생성형 AI가 환경에 미치는 중대한 영향을 해결하기 위하여, 기업은 얼마나 많은 에너지를 사용하는지, 어떻게 조달하는지, 특히 학습, 개발, 배치 및 사용을 포함한 전체 수명 주기 동안 모델이 얼마나 많은 탄소를 배출하는지 공개해야 합니다."

2.7.3 그린 워싱과 그린 AI에 대한 희망

대형 기술 기업들은 활동에 필요한 물과 에너지의 기하급수적인 수요를 완화하기 위해 상쇄 방안(물 보충 프로젝트 및 탄소 상쇄 방안)에 의존하고 있습니다. 또한 "물 포지티브화" 또는 "탄소 중립화", 심지어 마이크로소프트가 2030년까지 목표로 하는 "탄소 네거티브"와 같은 논란의 여지가 있는 주장을 사용하기도 합니다.[170] AI 기업이 일반적으로 배출량 감축 또는 상쇄 계획에 대해 공개하지 않기 때문에 현재까지 AI가 탄소 중립적이라는 주장은 존재하지 않습니다.

기술 기업의 탄소 중립 주장은 항상 탄소 상쇄 투자에[171] 의존하면서 자체 공급망이나 사업 활동에서 이산화탄소를 제거하기 보다는 다른 기업(주로 개발도상국)이 탄소 배출을 하지 않도록 비용을 지불하는 방식입니다. 이러한 탄소 상쇄 제도는 널리 비판을 받고 있으며, 오해의 소지가 있으며, '탄소 중립'과 같은 것이라고 볼 수 없습니다.[172]

탄소 상쇄는 더 효율적인 컴퓨팅 작업을 통해 더 작은 모델을 만드는 것보다 탄소 배출량을 손쉽게 상쇄할 수 있습니다. 게다가 이러한 탄소 중립 주장은 비표준화된 방법론에 의존하고 현재 배출되는 탄소와 장기 탄소 포집 계획을 맞추려 하기 때문에 모든 산업 분야에서 국제적인 비판을 받고 있습니 다.[173]. 따라서 탄소 상쇄는 종종 배출량을 줄이지 않아도 원하거나 필요한 만큼 배출하고 비용을 지불하면 되는 공짜 카드로 취급됩니다.

이러한 주장이 종종 그린워싱에 해당하기 때문에 EU는 탄소 중립 주장에 대해 금지하거나 최소한 훨씬 더 엄격한 규칙을 만드는 것을 고려하고 있습니다.[174]

기술 기업은 상쇄 방안을 넘어 에너지 소비가 적은 AI 모델을 설계하고 생성형 AI 모델을 개발하는 네 단계[175].

> "생성형 AI를 개발하고 활용하는 기업이 어떤 출처에서 얼마나 많은 에너지를 사용하는지, 얼마나 많은 에너지를 사용할 계획인지 투명해지지 않는 한, 이들에게 책임을 묻고 실질적인 감축을 약속하게 하는 것은 불가능할 것입니다."

모두에서 배출량을 줄이고 자원을 절약할 방법을 모색해야 합니다. 이는 또한 기하급수적으로 더 큰 모델을 필요로 하거나 원하는 경우 그 성능이 가져올 선형적 이득을 재고해야 함을 의미합니다.[176]

AI 부문의 지속가능성을 높이기 위한 시도는 투명성을 높이는 것으로부터 시작해야 합니다. 생성형 AI를 개발하고 활용하는 기업이 어느 출처에서 얼마나 많은 에너지를 사용하는지, 얼마나 많은 에너지를 사용할 계획인지 투명해지지 않는 한, 이들에게 책임을 묻고 실질적인 감축을 약속하게 하는 것은 불가능할 것입니다. 또한 소비자도 이러한 데이터에 접근하여 기후와 환경에 미치는 부정적인 영향이 적은 AI 시스템을 선택하거나 아예 시스템 사용을 자제할 수 있어야 합니다.

천연 자원이 점점 더 부족해지고 전기와 식수에 접근하는 시장이 점점 더 불안정해지는 기후 변화의 맥락에서, 무엇을 우선할 것인지 정치적 결정이 필요합니다. 배출량을 측정하거나 보고하지 않는 기업인지, 더 효율적일 수 있는 대형 모델에 전기를 사용할 것인지, 가정 난방과 같은 다른 용도로 에너지를 사용할 것인지 말입니다.[177]

2.8 노동에 미치는 영향

생성형 AI가 기후 변화로부터 인류를 구할 것이라는 신화 외에도 이 기술이 빈곤을 해결할 수 있다는 신화도 만연해 있습니다.[178] 거대 기술 기업들은 빈곤과 억압에 맞서 싸우기 보다 기존의 권력 구조를 강화하고 이용하고 있으며, 빈곤을 해결하기는커녕 오히려 강화할 수 있습니다.

2.8.1 노동 착취와 유령 노동

기술 기업들은 적어도 두 가지 방식으로 AI와 관련한 노동력을 착취합니다. 첫째, 세계 남반구 저임금 노동자들에게 어렵고 일시적이며 종종 정신적 충격을 주는 작업을 아웃소싱함으로써 노동력을 착취합니다. 둘째, 기술 기업들은 인간의 개입이 필요없고 스스로 기능할 수 있다는 착각에 빠져 생성형 AI를 개발하는 기업들은 이러한 노동자들을

보이지 않게 만들고 빈곤 및 트라우마와 맞선 이들의 투쟁을 거의 잊혀지게 만듭니다. 자동화의 인적 비용에 대한 이러한 모호함은 '유령 작업'이라고 불립니다.[179]

이에 적합한 예시로는 챗GPT의 유해성을 줄이려는 오픈AI의 시도를 들 수 있습니다. 이는 모델이 성폭력, 근친상간, 야만적 행위 등 폭력적인 행위와 언어를 인식하도록 함으로써 가능해졌습니다.[180] 이를 위해 회사는 유해 콘텐츠에 라벨을 붙이는 사람의 개입이 필요했고, 이 작업을 미국에 본사를 둔 회사 사마에 아웃소싱했습니다. 이 회사는 스스로가 5만 명을 빈곤에서 벗어나게 한 "윤리적 AI 접근 방식"을 갖춘 회사라고 홍보합니다.[181]

오픈AI와 사마 간의 성공적인 거래에도 불구하고 케냐의 노동자들은 시간당 미화 2달러 미만으로 지급받고 하루 9시간씩 유해 악성 데이터에 라벨을 부착해야 하는 중압감에 시달리면서도 심리적 지원은 거의 받지 못했습니다. 근로자들은 계 약이 끝나면 해고되었습니다.

> "거대 기술 기업들은 빈곤과 억압에 맞서 싸우기 보다 기존의 권력 구조를 강화하고 이용하고 있으며, 빈곤을 해결하기는커녕 오히려 강화할 수 있습니다."

오픈AI는 회사 이름을 공개 하지 않았지만, AI 기업의 공급망 전반에서 윤리 가이드라인을 준수하도록 하는 투명성 요구사항이 적용되어야 합니다. 타임지의 조사로 케냐 노동자들의 실체가 드러났지만, 이보다 LLM에 관여하는 더 많은 유령 노동자들이 대중 앞에 드러나야 합니다.

잡지 Sustain에 따르면, 오픈AI, 틱톡 등에서 일하는 관리자와 플랫폼 노동자들이 저임금에 시달리고 업무에 필요한 심리적 지원을 받지 못하고 있으며, 노동조합을 결성하지 못한다는 보고가 끊이지 않고 있습니다.[182] 이들이 처한 곤경 문제는 AI 논쟁에서 종종 간과됩니다.[183]

2.8.2 노동의 자동화와 일자리 위협

생성형 AI의 사용이 증가함에 따라 이 기술이 어떻게 노동자 업무를 보강할 수 있는지, 특정 일자리를 남아돌게 만들 것인지, 영향을받는 직업에 어떤 영향을 미칠 것인지에 대한 논의가 제기되었습니다.[184] 이는 여러 유형의 기술과 관련되어 온 주제이지만 생성형 AI의 급속한 성장으로 인해 노동의 자동화 문제가 전면에 등장했습니다.[185]

회사가 AI 모델에 텍스트나 이미지를 생성하도록 간단히 지시할 수 있다면, 이는 창의적 산업이나 보도 등의 분야에서 직원을 해고할 동기나 구실이 될 수 있습니다.

예를 들어, 이미지 생성기는 컨셉 드로잉이나 스톡사진 분야 직업을 불필요하게 만들 위험이 있는데 회사로서는 이미지를 제작하기 위해 아티스트나 사진작가에게 비용을 지불하는 것보다 이미지 생성기를 사용하는 것이 더 저렴할 것이기 때문입니다. 위에서 설명한 바와 같이 콘텐츠 제작의 자동화는 실제 사람 인력의 가치를 떨어뜨리는 동시에 활용되는 콘텐츠의 일반적인 품질도 떨어뜨릴 것입니다.

직원이 자동화된 시스템으로 대체되는 경우, 고객 지원과 같은 분야에서 서비스 품질이 저하될 수 있습니다. 이는 섭식 장애 상담센터에서 직원을 해고하고 챗봇으로 대체한 경우에서처럼, 최종 사용자가 인간 서비스를 접촉하는 데 의존하는 분야에서 특히 심각한 결과를 초래할 수 있습니다.[186]

2.9 지적 재산권

생성형 AI 모델은 이미 존재하는 콘텐츠를 기반으로 새로운 콘텐츠를 생성하기 때문에 학습 데이터의 원본과 생성된 결과물의 지적 재산 모두에 대해 여러 의문이제기됩니다.

많은 생성형 AI 모델의 학습 데이터에는 지적 재산권법으로 보호되는 방대한 양의 콘텐츠가 있습니다.

현재로서는 아티스트/작가/사진가/피사체의 동의 없이 생성형 AI 모델 을 학습시키는 것이 합법적인지 여부가 불분명합니다. 예를 들어, 지적 재산인 콘텐츠로 학습된 이미지 생성기의 개발과 사용에 대해 예술계에서 대대적인 항의가 일었습니다.[187] 특히 AI 모델이 특정 아티스트의 스타일이나 고유한 특징을 모방하여 새로운 이미지를 생성할 수 있을 때 논란이 더욱 커집니다.[188]

2023년 1월, 세 명의 아티스트가 스테이블 디퓨전에 대하여 스테이블 AI와 미드저니를 상대로 소송을 제기하면서, 이 도구가 수백만 아티스트의 저작권이 있는 이미지를 학습 데이터로 사용한다는 이유를 들었습니다.[189] 스테이블 AI는 아티스트가 자신의 작품이 스테이블 디퓨전 학습에 사용되는 것을 거부(옵트아웃, opt out)할 수 있는 시스템을 도입했지만, 이는 애초에학습 데이터세트에 포함되겠다고 요청한 바 없는 개별 아티스트에게 시간을 소모하는 부담을 줍니다.[190] 나아가 아티스트들은 원본 작품과 유사하게 생성된 합성 콘텐츠가 "기괴한 조롱거리"이며 아티스트의 역할을 평가 절하한다고 주장하고 있습니다.[191]

생성형 AI를 사용하여 생성된 저작물의 저작권을 누가 소유하는지에 대해서도 몇 가지 해결되지 않은 법적 문제가 있습 니다.[192] 컴퓨터는 지적 재산권을 가질 수 없으며, 모델의 최종 사용자가 생성형 AI를 사용하여 생성한 저작물에 대한 저작권을 어느 정도까지 취득할 수 있는지 불분명합니다.

3. 규제

기존의 법 제도는 항상 새로운 기술의 등장으로 도전받으며, 생성형 AI도 예외는 아닙니다. 관련된 상황에 생성형 기술이 사용되는 모든 경우에 법률이 생성형 AI에도 기술 중립적으로 적용될 수 있습니다. 그러나 참고할 선례나 판례가 없기 때문에 생성형 AI의 학습, 배치, 설계, 사용의 합법과 불법 사이의 경계를 정하는 데 규제 기관이 중요한 역할을 합니다. 이를 통해 기존 법 제도에 생성형 AI와 관련한 허점이 있는지, 있다면 어디에 있는지 명확히 할 수 있습니다. 또한 규제기관은 생성형 AI를 개발 및 배치하는 기업이 입법자들이 이미 정한 경계를 준수하도록 하는 데 중요한 역할을 합니다. 이러한 방식으로 생성형 AI의 개발과 학습이 안전하고 공정하며 책임감 있게 이루어질 수 있습니다.

기존 법률이 신기술의 위험을 충분히 해소하지 못한다면 법률을 개정하거나 새로운 법률을 도입해야 할 수도 있습니다.

전 세계적으로 AI 규제하기 위한 방안에 많은 진전이 있었으며, 이 중 일부는 생성형 AI와 관련성이 높습니다. 유럽에서는 소비자 권리를 개선하고 기술로 인한 피해를 최소화하기 위해 기존 제도를 개선하거나 새로운 규제를 마련하기 위하여 몇 가지 절차가 진행 중입니 다. EU 의원들은 이러한 기회를 잘 활용해야 합니다.

다음 장은 개인정보 보호, 소비자법, 제품 안전법 등 본 보고서 2장에서 설명하였던 생성형 AI 의 소비자 문제를 해결하기 위해 가장 중요하고 관련성이 높은 몇몇 법적 분야에 대한 것입니다. 이 장에서는 유럽의 법 제도를 중심으로 설명하되, 특히 관련성이 높은 경우 미국의 사례를 참고합니다.[193] 유럽 AI법안 초안, AI 책임 지침, 제조물 책임 지침 개정안과 같은 새로운 법률도 생성형 AI와 관련된 한도 내에서 설명합니다. 아래 표에서 가장 중요한 몇 가지 사항을 요약하였습니다.

	현행 법률인가? 장래 법률인가?	생성형 AI에 적용 가능한가?	생성형 AI에 미치는 영향은?	무엇이 필요한가?
일반 개인정보 보호 규정 (GDPR)	현행.	특히 생성형 AI 시스템의 학습 데이터, 입력 및 출력을 포함하여 개인정보와 관련된 생성형 AI의 모든 부분에 적용 가능함.	컨트롤러(controller)는 개인정보 처리에서 GDPR을 준수해야 함. 여기에는 정정 및 삭제권 등 정보 주체의 몇 가지 권리가 포함됨.	규제 기관은 생성형 AI 시스템을 조사하여 현행 법 제도 준수를 보장하여야 함. 일부 개인정보 보호 감독기관(DPA)는 이미 특정 생성형 AI 시스템을 조사하고 있음.
불공정 상거래 관행 지침 (UCPD)	현행. 지속적인 적합성 점검으로 인해 지침이 변경될 수도 있음.	상거래 상황에서 생성형 AI 시스템에 적용 가능함.	거래자는 UCPD에 따라 호도하거나 공격적인 상거래 행위 또는 거래업체의 성실 의무를 위반하는 관행에 해당하는 방식으로 생성형 AI를 사용해서는 안 됨.	소비자 당국은 생성형 AI 시스템을 조사하여 UCPD 준수를 보장하여야 함. EU 집행위원회는 지속적인 적합성 점검을 통해 UCPD의 적용 범위를 충분히 넓히고 효과적인 시정 메커니즘을 확보해야 함.
일반 제품 안전 지침 (GPSD)	현행.	잠재적으로 적용 가능하지만 GPSD의 적용 범위 및 피해 정의와 관련하여 몇 가지 불확실성이 존재함.	생산자는 안전하지 않은 제품을 시장에 출시해서는 안 됨.	제품 안전 승인기관은 GPSD에 따라 가능한 범위 내에서 생성형 AI로 인한 피해를 해결하기 위한 예방적 조치를 취해야 함.
일반 제품 안전 규정 (GPSR)	2024년 말에 발효됨.	적용 가능함.	생산자는 안전하지 않은 제품을 시장에 출시해서는 안 됨.	제품 안전 승인 기관은 GPSR이 시행될 때를 대비하여 이를 생성형 AI에 적용하고 시장에 안전하지 않은 제품이 출시되지 않도록 준비해야 함.

	현행 법률인가? 장래 법률인가?	생성형 AI에 적용 가능한가?	생성형 AI에 미치는 영향은?	무엇이 필요한가?
디지털 서비스법 (DSA) **콘텐츠 관리 관련**	2024년 2월에 해당 범위의 모든 사업체에 대하여 전면 적용될 예정이며, 초대형 온라인 플랫폼(VLOP) 및 초대형 온라인 검색 엔진(VLOSEs)을 2023년 여름 말까지 지정할 계획임.	생성형 AI 시스템에 직접 적용되지 않을 것으로 보임. DSA가 적용되는 디지털 서비스에 내장된 생성형 AI 시스템 또는 생성된 콘텐츠의 다운스트림 사용에 적용될 가능성이 높음.	생성된 텍스트에 대한 콘텐츠 관리 요구 사항의 적용.	
EU 경쟁법	현행.	적용 가능함.	생성형 AI를 개발 또는 배치 중인 기업이 시장지배적 지위를 남용할 수 있음.	경쟁 당국은 생성형 AI 시장을 모니터링하여 반경쟁적 관행이 없도록 보장하여야 함.
AI 법 (AIA)	현재 협상 중이며, 2023년 3자 협상이 개시될 예정임. 2024년 1월까지 3자 합의가 도출될 경우, 빠르면 2026년 4~5월에 완전히 적용될 것으로 예상됨.	적용 가능성은 높지만, 생성형 AI 시스템을 파운데이션 모델로 별도 규제할 것인지(의회 입장), 고위험 시스템/금지된 관행/챗봇 및 딥페이크의 맥 락에서 규제할 것인지 (집행위원회 초안), 범용 목적 AI 시스템 (이사회 입장)으로 규제할 것인지 불확실함	여전히 매우 불확실함.	EU 입법자들은 AIA가 이 보고서 2장에서 설명한 피해를 고려하도록 보장해야 하며, 전체 생성형 AI 행위자망에서 소비자 권리와 필요 의무를 보장하여야 함.
제조물 책임 지침 (PLD)	현행.	적용되지 않을 가능성이 높음.		

	현행 법률인가? 장래 법률인가?	생성형 AI에 적용 가능한가?	생성형 AI에 미치는 영향은?	무엇이 필요한가?
개정 제조물 책임 지침 (개정 PLD)	예정. 현재 협상 중.	불확실함. 현행 PLD에 대한 최근의 판례로 인함.	소비자가 보상을 청구할 수 있으나 비물질적 피해인 경우는 해당되지 않음. 이는 생성형 AI의 맥락에서 상당한 한계임.	EU 의원들은 제안을 수정하여 개정 PLD에서 비물질적 피해에 대해서도 소비자에게 보상 청구 권리를 부여하여야 함.
AI 책임 지침 (AILD)	예정, 현재 협상 중.	적용될 수 있음. AIA에 따름.	소비자가 보상을 요구할 수도 있음. 그러나 현재 상당한 한계를 내포하고 있음.	AID는 아직 정치적으로 초기 단계에 있지만, EU 의원들은 제안을 수정하여 소비자들이 생성형 AI로 인한 피해에 대하여 보상을 청구할 수 있는 효과적인 방안을 제시하여야 함.

이 보고서에서 다룬 법 제도 목록은 모든 경우를 포괄한 것이 아니며, EU 법률만을 다루고 있습니다. 인권법, 차별 금지법, 노동법 등 다른많은 EU 법률도 다양한 상황에서 생성형 AI에 적용될 수 있으며, 이들 중 상당수는 이러한 평가에 포함될 수 있었으나 용량 제약으로 인해 제외되었습니다. 마찬가지로, 생성형 AI에 대한 다양한 법 제도의 적용 가능성에 대한 검토들도 전체를 포괄한 것이 아닙니다.

따라서 이 보고서에서 제시된 개요가 생성형 AI로 인한 피해를 구제하는 방안을 논의하는 데에는 기여할 수 있지만, 이들 제도가 생성형 AI에 미치는 영향을 판단하기 위해서는 광범위한 법적 분석이 필요할 것입니다.

3.1개인정보 보호법

유럽 연합에 설립된 회사 또는 유럽 연합 외부에 설립된 회사가 유럽 연합(EU) 또는 유럽 경제 지역(EEA)에 있는 정보 주체의 개인정보를 처리하는 경우, 일반 개인정보 보호 규정(GDPR)[194]이 적용됩니다.[197]

GDPR의 의무는 주로 개인정보 처리의 목적과 수단을 결정하는 주체인 "컨트롤러(controller, 개인정보처리자)"에게 적용됩니다.[198] 일부 의무는 컨트롤러를 대신하여 개인정보를 처리하는 주체인 "프로세서(processor, 수탁자)"[199]에게도 적용됩니다. 1.1.2장에서 언급한 바와 같이, 생성형 AI의 개발 및 배치에는 여러 처리 단계에서 여러 주체가 관여합니다. 생성형 AI 행위자망에 포함된 여러 행위자가 각자의 역할을 명확하게 정의하여 전체 처리 절차에서 GDPR을 준수하는 것이 중요합니다.

기업이 생성형 AI 모델을 개발하고 배치할 때, GDPR은 생성형 AI 모델을 개발하는 데 사용되는 학습 데이터, 생성형 AI 모델의 결과물, 생성형 AI 모델 자체 등, 시스템의 최소 세 가지 측면에 적용될 수 있습니다.

이 보고서 전반에 걸쳐 설명한 바와 같이, 생성형 AI 모델은 일반적으로 인터넷에서 스크랩한 대량의 데이터를 분석합니 다. 이러한 데이터 포인트 중 일부는 명백한 개인정보이므로 GDPR이 그 처리에 적용될 수 있습니다. 마찬가지로 GDPR은 개인 계정 등을 통해 개인의 프롬프트에서 생성형 AI 모델로 이어지는 개인정보 처리에도 적용됩니다.

생성형 AI를 사용하여 식별 가능한 자연인과 관련된 이미지, 텍스트, 비디오 및 오디오를 생성할 수 있습니다. 따라서 GDPR은 입력 뿐 아니라 일부 출력에도 분명히 적용될 것입니다.[200] 이는 생성된 정보가 정확한지 여부와 관계가 없는 사실이며, 식별 가능한 개인과 관련된 딥페이크 사진이나 허위 진술도 여전히 개인정보에 해당한다는 의미입니다.

생성형 AI 모델이 작동하는 방식에 따라 모델에 개인정보가 필히 포함되지 않았는데도(모델 자체에는 개인의 실제 사진이 포함되지 않습니다), 결과물은 실제 사람의 식별 가능한 이미지일 수 있습니다. 그러나 모델에 개인정보가 직접 포함되지 않았더라도 연구자들은 LLM에서 학습 데이터를 추출할 수 있었습니다. 일반적으로 LLM은 소규모 언어 모델보다 이러한 추출에 더 취약합니다.[201] 위에서 언급한 바와 같이 학습 데이터에 개인정보도 포함되므로 생성형 AI 모델에서 개인정보가 추출될 수 있습니다. 일부 저자는 모델에서 개인정보를 추출할 수 있다는 것은 모델 자체가 개인정보로 간주될 수 있음을 의미한다고 주장하기도 합니다.[202] 따라서 GDPR은 모델의 입력 및 출력뿐 아니라 모델 자체에도 적용될 수 있습니다.

GDPR에 따라 개인정보를 처리하려면 법적 근거가 필요합니다.[203] 특수 범주 개인정보 처리는 일반적으로 금지되어 있는데, 이는 인종 또는 민족적 기원, 정치적 의견, 건강 및 생체 인식 정보를 드러내는 개인정보 범주를 포함합니다.[204] 생성형 AI 모델의 학습 데이터 및 결과물에 특수 범주의 개인정보가 포함된 경우, 컨트롤 러는 이 금지 조항을 면제할 수 있는 법적 근거가 있어야 합니다.

2023년 5월 현재, 일부 개발자들이 생성형 AI 모델 개발을 위한 개인정보 처리의 법적 근거라고 주장하는 부분에 대하여 일부 조망이 이루어졌습니다.

이탈리아 개인정보 보호당국의 조사 후에,[205] 오픈AI는 해외 사용자를 위한 개인정보 처리방침에 계약의 이행을 위한 법적 근거와 서비스 개발, 개선 및 향상 등 정당한 이익을 광범위하게 주장하는 법적 근거를 추가했습니다,[206]

반면, 구글은 아 직 챗봇 바드를 EU에 출시하지 않았습니다.[207] 이는 GDPR 때문일 수 있다는 추측이 제기되고 있으며,[208] 이 글을 쓰는 시점에서는 바드에서 개인정보를 처리하는 법적 근거에 대하여 언급된 바 없습니다.[209]

개인정보 처리를 위한 법적 근거를 요구하는 것 외에도, 생성형 AI 모델의 학습, 개발, 배치 및 사용을 위한 개인정보 처리와 관련한 다양한 관련 법적 요구사항들이 있습니다. 머신러닝 모델 학습과 관련하여 개인정보보호 중심설계(data protection by deisgn)와 중심설정(by default)[210], 데이터 최소화[211] 및 목적 제한[212] 원칙에 대한 논의는 새로운 것이 아니며, 이들 원칙은 개인정보가 포함된 경우 생성형 AI 모델 학습에도 적용됩니다.[213]

데이터 최소화 원칙은 명시된 처리 목적을 위해 가능한 한 적은 양의 개인정보를 수집하고 처리하는 것을 포함합니다. 목적 제한 원칙에는 수집 시점에 명시된 목적 외의 다른 목적으로 개인정보를 사용하지 않고, 이러한 목적을 달성하 는 데 필요한 기간보다 더 오래 개인정보를 저장하지 않는 것이 포함됩니다. 생성형 AI 모델 학습는 대량의 데이터를 필요로 하고 범용적으로 개발되는 경우가 많기 때문에 이들 원칙들이 많은 생성형 AI 모델 개발자가 취하는 접근 방식과 충돌할 수 있습니다.

그러나, 오픈AI는 개인정보 처리방침에서 "모델 작동 방식의 기술적 복잡성으로 인해 부정확한 개인정보를 정정 하지 못할 수도 있습니다."라고 밝혔습니다.[217] 즉, 오픈AI가 정보 주체의 권리를 보장하고 GDPR을 준수하는 것이 기술적으로 가능한지 여부가 매우 의문입니다.

오픈AI가 구현한 것과 같은 옵트아웃 시스템이 GDPR을 준수할 수 있는지도 의문입니다. 옵트아웃 시스템이 효과적이려면 생성형 AI 모델이 자신의 개인정보에 대해 학습하였다는 사실을 개인이 인지해야 합니다. 이는 생성형 AI 모델 을 자주 사용하지 않는 한 소비자에게 명확하게 드러나지 않으며, 자주 사용하는 경우라 하더라도 개인이 개인정보 처리 한도를 이해할 가능성이 낮습니다.[218]

학습 데이터에서 개인정보를 삭제하는 것과 관련해서는 생성형 AI 모델을 학습하는 데 사용되는 데이터세트의 엄청난 크기가 중요한 장벽입니다.[219] 관련 작업은 다음과 같습니다.

데이터세트의 수집, 정제 및 준비는 모델 개발에 비하여 AI 실무자 들에게 일반적으로 우선순위가 높지 않습니다.[220] 결과적으로 기업이 데이터세트에 대한 감독과 문서화가 미비하여 개인정보 보호법에 위배되는 개인의 데이터 흔적을 찾고 삭제할 수 있는 능력이 손상될 수 있습니다.

3.1.1 정보 주체 권리

개인정보가 처리되는 사람(정보주체)은 GDPR에 따라 몇 가 지 권리를 갖습니다. 여기에는 삭제(개인정보를 삭제할 권 리)[214], 정정(개인정보를 수정할 권리)[215] 및 반대(개인정보 처리에 대해 항변할 권리)가 포함됩니다.[216]

생성형 AI 모델을 개발하고 배치하는 기업이 실제로 정보주체 권리 행사 요청을 어떻게 이행할 수 있을지는 아직 불분명합니다. 이탈리아 개인정보 보호당국이 챗GPT를 면밀히 조사한 후, 오픈AI는 개인정보에 대한 옵트아웃 메커니즘을도입해서 개인정보를 학습 데이터에서 제거할 수 있도록 하였고, 부정확한 개인정보의 정정 가능성도 열어두었습니다.

3.1.2 이탈리아 DPA의 챗GPT 결정

이미 생성형 AI 모델에 GDPR을 적용하려는 노력이 있었습니다. 2023년 3월 31일, 이탈리아 개인정보 보호당국(DPA)은 이탈리아 시민의 개인정보 처리와 관련하여 챗GPT의 소유자인 오픈AI에 일시적인 제한을 부과했습니다. 동시에 DPA는 이 사례의 사실관계 조사를 시작했습니다.[221] 이는 챗GPT 서비스의 최종 사용자에 대한 개인정보 처리, 모델 학습과 관련된 개인정보 처리, 콘텐츠 생성 중 개인정보 처리와 관련된 문제 등 몇 가지 잠정적인 GDPR 위반이 있었기 때문입니다. 그 결과 오픈AI는 이탈리아에 거주하는 개인에 대한 챗GPT 접근을 일시적으로 차단했습니다. 이를 통해 DPA가 설명한 일부 개인정보 보호 문제가 해결되었지만, 이탈리아 바깥 시민들이 이탈리아 시민에 관한 개인정보를 생성하는 일은 여전히 가능했습니다.

이탈리아 DPA가 지적한 잠정적 위반 사항 중 일부는 다른 위반 사항보다 더 광범위한 결과를 초래할 수 있습니다. 오픈AI는 모델을 크게 변경하지 않고도 개인정보 유출 및 연령 확인 메커니즘과 같은 문제를 해결하기 위한 조치를 구현할 수 있습니다. 부정확한 개인정보를 생성하는 문제는 오픈AI가 모델의 정확도를 높이려고 전반적으로 노력하고 있음에도 해결하기가 더 어려워 보입니다. 위에서 언급한 바와 같이, 오픈AI는 이미 부정확한 개인정보를 수정하지 못할 수도 있다고 명시하고 있으며,[222] 2.3.2절에서 설명한 바와 같이 회사가 모델 자체적으로 또는 콘텐츠관리를 통해 개인에 대한 정확한 개인정보를 보장할 가능성은 매우 희박해 보입니다.

이탈리아 DPA가 제기한 마지막 가장 치명적인 문제는 오픈AI가 모델 학습을 위해 이탈리아 시민의 개인정보를 처리할 법적 근거가 없는 것으로 보인다는 점입니다. GDPR은 EU에서 통합 입법(harmonized)이기 때문에, 이는 오픈AI가 EU 또는 EEA에 있는 어떠한 정보 주체의 개인 정보로도 생성형 AI 모델을 학습시킬 수 있는 법적 근거가 사실상 없다는 것을 의미합니다. 오픈AI는 학습 데이터에 대한 정보를 공유하지 않았지만, 인터넷에서 스크랩한 EU 및 EEA의 정보주체에 대한 개인정보가 여기에 포함 되어 있다고 가정해도 무방합니다.

후속 GPT 모델의 학습을 위해 새로운 데이터세트를 준비하고 EU 및 EEA의 정보 주체에 대한 개인정보를 제거하는 데이터세트 정제가 기술적으로 가능하긴 하지만, 이는 시간과 리소스가 매우 많이 소요되며 개발을 상당히 중단시킬 수 있습니다. 어쨌든 이 문제는 범용 생성형 AI 모델[223]과 GDPR이 현재 형태대로 공존할 수 있는지 의문을 제기합니다.

2023년 4월 28일, 오픈AI가 다양한 개인정보 보호 기능을 도입한 후 이탈리아에서 챗GPT가 복구되었습니다. 위에서 설명한 옵트아웃 메커니즘, 개인정보 삭제권 행사 메커니즘, 개인정보 처리에 사용된 법적 근거를 비롯한 신규 처리방침 고지, 연령 지정 요건 등의 조치가 도입되었습니다.[224] 오픈AI는 문서상으로는 추가적인 개인정보 보호 조치를 도입했지만, 소비자가 실제로 자신의 개인정보가 생성형 AI 모델 학습에 사용되는 것을 옵트아웃할 권리 등 자신의 권리를 효과적으로 활용할 수 있는 방안은 불분명합니다.

오픈AI가 정보 주체의 기본권보다 우선하는 정당한 이익을 가지고 있다고 하더라도,[225] 오픈AI는 생성형 AI 모델을 대중에게 공개하기 전에 이러한 조정을 잘 수행하지 않은 것으로 보입니다. 따라서 오픈AI가 처리의 적법성 또는 책임원칙[226]을준수하는지는 기껏해야 모호한 수준입니다. 이탈리아 DPA가 이러한 주장을 명백히 수용한 것은 문제가 될 수 있는데, GDPR를 준수하는 것은 오픈AI의 임시방편 그 이상이기 때문 입니다

GDPR이 개인을 보호하기 위해서는 광범위한 분석과 신속한 집행이 필요하다는 사실이 분명해 보입니다. GDPR이 개인을 보호하는 데 기여할 수 있음에도 불구하고, 이 규정은 특히 국경을 넘는 사건에서 집행이 느리고 복잡하다는 비판을 받아왔습니다.[227] 그러나 유럽 정보보호 이사회(EDPB)가 챗GPT에 대한 조사 및 집행을 조정하기 위해 범EU 차원의 "태스크포스"를 발족시켰기 때문에 분명 더 많은 법적 발전이 있을 것입니다.[228] 프랑스 DPA는 챗GPT와 관련된 여러 진정을 접수하고 조치 계획을 발표했습니다.[229] 독일과 스페인 DPA도 모두 조치를 고려하고 있습니다.[230]

지금까지는 이 분야에서 오픈AI와 챗GPT가 GDPR 집행의 주요 초점이었지만, 여러 모델이 널리 사용되면 다른 경우들도 뒤따를 가능성이 높습니다. 집행에 대한 욕구 가 분명히 존재하므로 GDPR은 모든 정보주체의 개인정보 보호 권리를 존중하는 생성형 AI를 구현하는 데 중요한 법 제도가 될 것입니다.

3.2 소비자법

불공정 상거래 관행 지침(UCPD)[231]은 EU 및 EEA의 기업 대 소비자 관행에 적용되는 법적 조항을 명시하고 있습니다. 이 지침은 기술 중립적이며 모든 기업-소비자 간 거래에 적 용되고, 이는 상거래 환경에서 소비자의 의사 결정을 보호하기 위해 포괄적으로 작용한다는 의미입니다. 이 지침은 주로 거래업체의 [232] 관행에 대한 개방 및 공개 요구사항을 통해 이들 상거래 관행[233]이 불공정하지 않도록 보장합니다.

특정 상거래 관행은 UCPD 부속서 1의 불공정 상거래 관행 금지목록을 통해 전면적으로 금지됩니다. 부속서 1의 금지 관행 외에도 UCPD에는 몇 가지 광범위한 재량적 법률 조항이 있습니다. 상거래 관행이 호도하거나[234] 공격적이어서[235], 보통의 소비자가 다른 경우에는 하지 않았을 거래 결정을 내리도록 유도하는 원인이 되는(또는 원인이 될 수 있는) 경우, 이는 불공정합니다.

거래 결정은 광범위하게 정의되어 왔으며 가상 장바구니에 상품을 추가하거나 상점에 입장하는 등의 소비자 결정을 포함합니다. 2021년 12월에 발표된 가장 최근의 UCPD 가이드라인에서 EU 집행위원회는 검색 또는 스크롤을 통해 서비스를 계속 사용하는 등의 사례도 포함시켰으며,[236] 이로써 UCPD의 거래 결정 테스트 범위가 분명히 확대되어 주목 경제에서 핵심적인 사업 관행을 포괄하였습니다. 가이드라인은 법적 구속력이 없기 때문에 소비자 당국과 법원에서 실제로 어떻게 해석할지는 아직 명확하지 않습니다.

또한 어떤 관행이 직업적 성실 요구사항을 위반하여 보통의 소비자의 경제 행동을 왜곡하거나 왜곡할 가능성이 있는 경우 불공정 행위로 간주하는 일반 조항이 있습니다. 이는 금지 목록에 포섭되지 않는 불공정 행위에 대한 안전망의 역할을 하고 있고, 호도할 여지가 있는 관행이나 공격적인 관행은 여전히 UCPD의 공정성 평가의 대상이 될 수 있습니다.[237] 직업적 성실 요구사항은 차별금지법과 같은 다른 법률 체계와의 가교 역할을 할 수 있으며, 소비자법 맥락에서 이러한 법률의 법리를 통합할 수 있게 합니다.[238]

독립형 모델로서 또는 다른 소비자 대상 서비스에 내장된 생성형 AI는 UCPD에 의해 잠정적으로 여러가지 방식으로 다루어질 수 있으며, 전통적인 금전적 관점이나 소비자를 서비스에 계속 참여시키는 관점 등에서 논의될 수 있습니다. 어느 경우로든 UCPD의 적용 가능성은 생성형 AI 모델이 상거래 관행에서 사용되는 상황에 따라 달라집니다.

빙은 현재 생성형 AI 검색에 광고를 사용하고 있으며,[239] 상거래 관행이 호도될 여지가 없도록 보장하기 위해서는 별도의 라벨링이 필요합니다 예를 들어 지속적인 커뮤니케이션으로 소비자가 서비스에 계속 참여하도록 설득하기 위해 텍스트 생 성기를 사용하는 경우, 특히 파악된 소비자의 약점을 겨냥하는 것은(연애 관계를 생성하고 시뮬레이션하도록 프로그래밍된 챗봇의 경우처럼) 또한 공격적인 관행에 해당할 수 있습니 다. 2.1.4.3장에서 언급한 바와 같이, 기업들은 이미 생성형 AI를 쇼핑 경험에 통합하려고 시도하고 있으며, 이는 부정확한 정보를

제공하여 소비자가 제품을 구매하도록 호도할 가능성이 있습 니다.

UCPD를 통해 생성형 AI로 인한 소비자 문제를 해결 해야 한다는 요구가 있었습니다. 유럽 소비자 단체인 BEUC는[240] 2023년 4월 21일에 생성형 AI, 특히 텍스트 생성기에 관한 서한을 DG JUST와 소비자 보호 협력 네트워크에 보냈습니다.[241] 이 서한에서 BEUC는 UCPD를 위반하는 방식으로 소비자 행동에 영향을 미칠 수 있는 생성형 AI의 다양한 배치 방식에 주목했으며, 아동과 같은 취약 계층도 고려하였습니다.

점점 더 많은 서비스에서 생성형 AI가 사용됨에 따라 소비자 당국이 상업적 맥락에서 생성형 AI의 불법적인 사용을 검토하고 해결해야 할 필요성이 분명히 있습니다. UCPD는 상거래 관행의 맥락에서 생성형 AI와 관련된 특정 문제를 해결하는 데 사용될 수 있습니다. 특히 관련성이 높은 예시로는 생성형 AI 모델이 소비자에게 제품에 대하여 허위 또는 오해의 소지가 있는 정보를 제공하는 방식으로 사용되거나 시스템 소유자가 서비스 약관에 정보를 누락하는 경우 등이 있습니다.

동시에 생성형 AI에 대한 UCPD의 적용 가능성에는 몇 가지 잠정적인 한계도 있습니다. 아마도 생성형 AI는 분명히 한계가 있음에도 불구하고 지각 있는 대화 상대처럼 보일 수 있는 AI의 기능을 활용함으로써, 거래업체의 서비스에 대한 소비자 상호 작용을 늘리고 참여를 높이기 위해 사용될 수 있습니다. 예를 들어 챗봇이 소비자를 대상으로 연애 감정을 시뮬레이션하도록 설계된 경우 소비자들은 모델이 만들어낸 환상에 빠져 평소보다 서비스에 훨씬 더 많은 시간과 주의를 기울이도록 호도될 수 있습니다.

그러나 소비자의 관심과 참여를 부당하게 추출하는 관행에 대한 UCPD의 유용성은 아직 확실하지 않습니다. 이를 위해서는 무엇이 "거래 결정"에 해당하는지에 대한 폭넓은 이해가 필요한데, 지금까지는 법 문구가 아닌 유럽 집행위원회의 (구속력 없는) 가이드라인에만 근거해 왔습니다. 따라서 많은 관련 관행들이 UCPD에서 명확하게 다루어지고 있지 않습니다.[242]

또한 공개 요구사항에 집중할 뿐, 디지털 서비스의 필수적인 직업적 성실로서 '설계상 공정성'을 요구하지 않는 방식으로 적용되는 UCPD의 구제 메커니즘이 충분한지에 대한 의문도 있습니다.

EU 집행위원회는 이러한 문제를 해결하기 위해 현행 EU 소비자법이 높은 수준의 보호를 보장하는 데 적합한지를 평가하는 지속적인 디지털 적합성 점검을 활용하여[243] 이러한 문제를 해결해야 합니다.

3.2.1 미국 환경에서의 소비자 법률

미국에서는 연방거래위원회(FTC)가 FTC법을 시행하고 있으며, 이 법은 FTC에 불공정하고 기만적인 거래 관행 위반에 대한 집행권을 부여했습니 다.[244] 유럽 소비자 당국과 달리 FTC는 FTC법의 범위 내에서 특정 기만적 관행 또는 불공정 경쟁 방법을 대상으로 하는 규칙을 제정할 임무가 있습니다.[245] 2023년 6월 현재 FTC는 항상 불공정하고 기만적인 특정 관행을 금지하는 새로운 규칙을 제정하는 폭넓은 절차를 진행 중입니다. 이는 FTC가 시장의 새로운 기술에 대해 보다 강력하고 구체적으로 대응할 수 있게 되었음을 의미합니다.

FTC는 이미 2021년에 투명성, 편향적이지 않은 결과물, 책임성 등을 요구하는 AI 관련 가이드라인을 발표한 바 있습니다.[246] 2023년에는 생성형 AI 제품 및 서비스가 급증하는 가운데 FTC는 AI 광고가 신뢰할 수 있고 책임감 있는 방식으로 이루어져야 할 필요성을 기업들에게 상기시키는 성명을 발표하기도 했습니다.[247]

FTC가 아직 생성형 AI를 배치하거나 학습하는 기업에 대하여 집행 조치를 취한 바는 없지만, 알고리즘 분리(Algorithmic Disgorgement)를 비롯해 강력한 공평성 구제 조치를 취한 사례가 있습니다.[248] 이 구제수단은 수집 과정에서 소비자의 권리가 침해된 경우 해당 기업이 데이터와 해당 데이터를 기반으로 구축된 모델/알고리즘을 삭제할 것을 요구합니다. 이 강력한 구제수단은 모델이 삭제되는 것을 저어하는 기업의 관행을 개선할 수 있습니다.

AI 디지털 정책 센터(Center for AI and Digital Policy, 미국 비영리연구소)는 2023년 3월 3일 GPT4 이후 챗GPT의 추가 상용 버전 출시를 유예하고 특히 생성형 AI와 관련된 규칙 제정을 요청하는 민원을 FTC에 제기했습니다.[249] FTC는 개별 소비자에 대한 사건을 공개하지는 않지만 이러한 진정 및 신고를 바탕으로 조사를 진행할 수 있습니다. 따라서 FTC가 생성형 AI의 대량 도입으로 인한 소비자 피해를 최소화하기 위해 이 특별한 진정 사건에 대응할 가능성이 높습니다.

2023년 4월 25일, FTC, 소비자금융보호국(CFPB), 평등고용기회위원회(EEOC), 법무부 민권국은 '자동화된 시스템의 차별과 편향성'에 대해 단속할 계획이라고 발표하였고[250], 백악관은 5월 5일 공공 부문의 위험 완화 노력을 포함한 몇 가지 이니셔티브를 발표했습니다.[251]

3.3 일반 제품 안전법

제품 안전법은 시장에 출시되는 제품을 소비자가 안전하게 사용할 수 있도록 보장하기 위한 법입니다. 현재 유럽의 일반 제품 안전법은 일반 제품 안전 지침(GPSD)을 기반으로 합니다.[252] 2024년 말에는 일반 제품 안전 규정(GSPR)[253]이 GPSD를 대체할 예정입니다. 생성형 AI의 상황에 두 개의 법 제도가 모두 관련이 있습니다.

3.3.1 일반 제품 안전 지침

GPSD는 부문별 법률을 보완하고 다른 법률에서 다루지 않는 제품의 모든 위험에 적용됩니다.[254] 실제로 GSPD는 유럽 시장의 모든 제품에 대하여 안전 요구사항을 보장하는 안전망의 역할을 수행합니다.

이 법은 생산자가 안전한 제품만 시장에 출시할 것을 요구합니다.[255] 지침의 제품 정의는 이론적으로 제품과 연관된 소프트웨어로 인한 피해도 포함할 수 있을 정도로 광범위하지 만,[256] 그 범위는 소프트웨어를 명시적으로 포함하거나 제외하지 않고 있습니다. 따라서 GPT 모델 및 기타 순수 소프트웨어 기반 생성형 AI 모델에 대한 적용 가능성이 아직 불확실합니다.

제품이 정상적이고 예측 가능한 사용 환경에서 소비자의 안전과 건강에 위험을 초래하지 않거나 최소한의 위험만 초래 할 경우 그 제품은 안전한 것으로 간주됩니다.[257] 이는 전통적으로 신체적 상해나 재산적 피해와 같이 사람에게 물리적으로 미치는 영향을 다루어 왔으며 정신 건강은 GSPD에 명시적으로 언 급되어 있지 않습니다. 일부에서는 제품으로 인한 내재적 정신 건강 위험도 GPSD에 포함될 수 있다고 주장하지만,[258] 명시적인 언급이 없기 때문에 정신 건강 위험에 대한 적용 여부는 더욱 불확실합니다.

관할 당국은 제품이 시장에 출시된 이후라도[59] 실제로 안전한지 여부를 검토해야 합 니다. 이러한 검토는 사전 예방 원칙을 고려해야 하며, 이는 제품의 잠재적 위해 및 유해한 영향 여부에 대한 과학적 확신이 없는 경우 제품이 안전하지 않은 것으로 추정될 수 있다는 의미입니다.

이 보고서 전반에 걸쳐 설명한 바와 같이, 생성형 AI가 실제로 소비자에게 상당한 위험, 특히 정신 건강 위험을 초래 할 수 있다는 점이 분명해 보입니다. 이러한 위험은 예를 들어 부정확한 개인정보나 딥페이크의 생성과 후속적인 유포, 고도로 조작적이고 인간화한 생성형 AI 모델의 배치, 또는 소비자가 정신 건강 또는 의료 자문 목적으로 생성형 AI 모델을 사용하는 상황에서 발생할 수 있습니다.

안전 당국에 생성형 AI의 안전 위험을 조사하도록 요구하는 등 GPSD를 생성형 AI에 적용하려는 움직임이 있었습니다. 유럽 소비자 단체 BEUC는 이와 관련하여 2023년 4월 12일 소비자 안전 네트워크에 서한을 보냈습니다.[260] 이 서한은 특히 소비자의 정신 건강에 대한 위험에 주목했습니다.

3.3.2 일반 제품 안전 규정

새로운 일반 제품 안전 규정(GPSR)이 EU의 승인을 받았으며, 2024년 말에 발효될 예정입니다. 이 새로운 규정은 GPSD를 폐지하고 소프트웨어를 포함시켜 제품의 범위를 넓히고, 정신 건강을 명시적으로 언급할 것입니다.[261] 또한 모든 생산자는 제품의 위험을 평가할 때 제품의 진화, 학습 및 예측 기능을 고려하여야 하며,[262] 이는 생성형 AI 제품의 상황과 분명히 관련이 있습니다.

생성형 AI 모델에 대한 GPSD의 적용 여부는 불확실할 수 있지만, 생성형 AI 모델에 GPSR이 적용될 것임은 분명해 보입니다. 안전 당국은 현행 법 제도 하에서 가능한 범위 내에서 생성형 AI로 인한 피해를 해결하기 위해 예방 조치를 취해야 합니다.

이는 또한 안전 당국이 GPSR이 발효되고 생성형 AI 모델에 적용되기 시작하는 즉시 이를 집행할 수 있도록 철저히 준비하는 데 도움이 될 것입니다.

3.4 경쟁법

EU 경쟁법의 핵심은 반경쟁적 관행을 근절하여 경쟁할 수 있는 시장 환경을 유지하고 소비자가 더 낮은 가격, 더 나은 제품과 서비스 품질, 더 많은 선택권과 혁신의 혜택을 누리도록 하는 데 있습니다.

비록 시행은 다른 규정으로 이루어지고 있지만, EU 경쟁법의 핵심은 유 럽 연합 기능에 관한 조약(TFEU)에서 찾을 수 있습니다. 첫째, 기업이 반경쟁적 합의를 하는 것은 금지되어 있습니다.[263] 둘째, 기업은 지배적 지위를 남용할 수 없습니다.[264]

'지배적 지위'의 개념은 고도로 맥락 기반적이며 '관련 시장'이 어떻게 정의되는지에 따라 달라집니다. 이는 예를 들어 대체 제품의 가용성 및 이러한 대체 제품으로 전환하려는 소비자의 의지를 포함합니다.[265] 앞서 2.1.3절에서 논의한 바와 같이, 생성형 AI의 배치는 소수의 행위자에게 권력을 집중시킬 위험이 있습니다. 이로 인해 특정 기업이 해당 시장에서 지배적인 위치를 차지하게 될 수 있습니다(예: 생성형 AI 기반 검색 엔진, 쇼핑 도우미 등).

디지털 부문은 흔히 빅 테크라고 불리는 극소수의 거대 기업들이 차지하고 있는 것으로 악명이 높습니다. 생성형 AI와 관련된 신흥 시장이 조기에 독점 금지 기관의 조사를 받도록 하여 이 시장에서 자신의 지위를 남용하려는 유혹을 받을 수 있는 지배적인 기업이 출현하는 것을 방지하는 것이 매우 중요합니 다. 몇몇 빅 테크 기업이 생성형 AI에 막대한 투자를 하고 있다는 점에 주목할 만합니다.

경쟁 규제 기관은 이 보고서에서 설명한 피해 일부를 해결하기 위해 분명히 해야 할 역할이 있습니다. 영국 경쟁 및 시장 당국(CMA)은 생성형 AI 개발 및 사용 시 경쟁 및 소비자 보호 고려사항에 대한 초기 검토에 착수했습니다. 다른 경쟁 당국은 이 분야의 향후 발전 상황을 면밀히 모니터링하고, 경쟁을 저해할 수 있는 행위나 관행을 발견하면 조기에 개입할 준비를 해야 합니다. 이는 새롭게 떠오르는 AI 분야가 공정하고 경쟁적으로 유지되도록 하는 데 도움이 될 것입니다.

3.5 콘텐츠관리

디지털 서비스법(DSA)[266]은 불법 콘텐츠 제거 메커니즘을 개선하고 표현의 자유 및 높은 수준의 소비자 보호 등 개인의 권리를 보호하는 것을 목표로 하는 새로운 EU 규정입니 다. 이 법은 온라인 서비스에 새로운 콘텐츠관리 기능을 도입하는 중요한 수단이 될 것입니다.

DSA는 온라인 중개 서비스, 즉 도관 서비스, 캐싱 서비스 및 호스팅 서비스에 적용됩니다.[267] 실제로 이는 온라인 마켓플레이스, 소셜 미디어 플랫폼, 클라우드 호스팅 서비스, 인터넷 접속서비스 제공업체 등 소비자를 상품, 서비스, 콘텐츠에 연결하는 서비스에 DSA가 적용된다는 의미입니다.

이 규정은 2024년 2월부터 해당 범위에 속하는 모든 사업체에 전면 적용되며, 이미 유럽 집행위원회에서 일부를 지정하기 시작한 초대형 온라인 플랫폼(VLOP)과 초대형 온라인 검색 엔진 (VLOSE)[268]은 2023년 여름 말부터 새로운 규정을 준수해야 합니다.

생성형 AI 모델이 DSA의 적용 대상인지는 명확하지 않습니다. DSA의 적용을 받는 가장 관련성이 높은 서비스 유형은 '호스팅 서비스'입니다.[269] 이 경우에도 생성형 AI 모델이 제공하는 콘텐츠는 대부분 소비자나 기타 제3자가 아니라 모델 자체에 의해 생성되기 때문에 DSA의 적용 여부가 명확하지 않습니다.

독립형 서비스로서의 생성형 AI 모델은 DSA의 적용을 받지 않을 수 있지만, 플랫폼과 서비스에 생성형 AI 모델을 포함하고자 하는 기업에는 DSA가 적용될 수 있습니다.

예를 들어, 초대형 온라인 플랫폼으로 지정된 검색 엔진 빙이 챗GPT를 통합하는 것은 생성된 콘텐츠에 대하여 DSA의 콘텐츠 관리 요구사항을 사실상 촉발시킬 수 있습니다.

생성형 AI에 의해 생성되고 이후 소비자가 DSA의 적용을 받는 서비스에서 공유하거나 저장하는 모든 콘텐츠도 마찬가지로 콘텐츠 관리 요구사항이 적용됩니다. 두 경우 모두 DSA는 콘텐츠 생성 모델 자체보다 는 생성된 콘텐츠의 다운스트림 제공 및 활용에 적용되는 것으로 보입니다.

3.6 AI법 초안

2021년 4월, EU 집행위원회는 "내부 시장에서 AI의개발, 사용 및 활용을 촉진하기 위해" EU와 EEA 전반에 걸친 통일된 규칙으로 AI법(AIA)에 대한 제안을 발표하였습니다.[270]

AI 규제를 목표로 하는 EU의 법 제도는 생성형 AI도 규제할 것으로 예상해야 합니다. 그러나 집행위원회의 AIA 제안 은 2022/2023년 겨울에 생성형 AI가 널리 도입되기 전에 발표되었습니다. 그 여파로 EU 의원들 사이에서 AIA의 일부로서 생성형 AI를 적절히 규제하는 방법에 대한 논의가 활발하게 이루어지고 있습니다.

AIA는 아직 완성되지 않았기 때문에 실제로 생성형 AI에 어떻게 적용될지는 아직 확실하지 않습니다. 이하에서는 2023년 5월 현재 AIA에 대한 집행위원회 초안과 이사회 및 의회의 입장을 간략하게 설명 합니다. 이 개괄은 AIA에 대한 집행위원회 제안의 관련된 부분에 대한 것으로부터 시작하겠습니다.

3.6.1 EU 집행위원회의 제안

집행위원회 AIA 제안(이하 "AIA 초안")은 AI 시스템을 시장에 출시하는 모든 제공업체에 적용됩니다.[271] AI 시스템은 머신러닝 접근법, 논리 및 지식 기반 접근법 또는 통계적 접근법을 기반으로 결과물을 생성하는 시스템으로 광범위하게 정의되었습니다.[272] 즉, AIA 초안의 범위는 광범위하여 여러 유형의 시스템을 포괄합니다.

AIA는 위험 기반 접근 방식을 채택하여 개인 또는 사회에 미치는 위험에 따라 다양한 유형의 AI 시스템을 규제합니다. 명시적으로 열거된 특정 관행은 금지되며,[273] 유럽 시장에 출시될 수 없습니다. 또한 예를 들어 부속서III[274]에 열거된 고위험 시스템에 속하는 AI 시스템의 경우 고위험으로 분류될 수 있습니다.

AIA 초안의 대부분은 고위험 AI 시스템을 규제하고 이러한 시스템의 AI 운영자에 대한 법적 요구사항을 규정하는 데 중점을 두고 있습니다.[275] 여기에는 위험 관리 시스템을 포함한 품질 관리 시스템 구축,[276] 데이터 품질 기준 충족,[277] 정확성, 견고성 및 사이버 보안 조치는 물론,[278] 기술 문서 작성 등과 같은 법적 요구사항이 포함됩니다.[280]

특히 AIA 초안에서는 고위험 시스템의 범위를 벗어나는 시스템 제공자에게는 규정하는 요구사항이 거의 없거나 아예 없습니다. 챗봇과 같은 애플리케이션과 딥페이크 자료에 대하여 일부 제한적인 투명성 요구사항이 있을 뿐입니다.[281] 고위험이 아닌 시스템의 모든 제공자도 고위험 시스템에 대한 요건을 자발적으로 준수할 수는 있지만,[282] 그렇게 해야 할 법적 의무는 없 습니다.

따라서 AIA 초안은 매우 광범위한 AI 시스템을 대상으로 하면서도 극소수의 시스템에만 의무를 부과하고 회원국이 추가적인 의무를 부과하는 것을 금지하고 있습니다.[283] 따라서 고위험 AI 시스템의 범위가 특히 중요합니다. 또한 AIA 초안에는 소비자에 대한 권리가 매우 제한적으로만 포함되어 있습 니다.

생성형 AI 시스템이 2021년부터 시행되는 집행위원회 AIA 초안에 어떻게 포함되는지는 불분명합니다. AIA 초안의 일반적인 적용 범위에 포함될 가능성이 높습니다. 보다 구체적인 요구사항이 요구되려면 생성형 AI 시스템이 부속서 III의 고위험 범주 중 하나와 관련되어야 하고, 제한된 투명성 요건이 적용되려면 챗봇 또는 딥페이크 상황에서 사용되어야 하며, 또는 금지된 관행과 관련되면 금지됩니다.[284]

3.6.2 AIA에 대한 EU 이사회 입장

AI법에 대한 이사회 입장(이하 "이사회 입장")에서는 [285] 범용 AI에 대하여 다루고 있습니다. 그 정의는 다음과 같습니다:

> 오픈 소스 소프트웨어를 포함하여 시장에 출시되거나 서비스에 투입되는 방식과 무관하게 이미지 또는 음성 인식, 오디오 및 비디오 생성, 패턴 탐지, 질의 답변, 번역 등 일반적으로 적용 가능한 기능을 수행하는 것을 제공자가 의도한 AI 시스템. 범용 AI 시스템은 복수의 상황에서 사용될 수 있고 복수의다른 AI 시스템에 통합될 수 있다.[286]

이는 이 보고서에서 다루는 모든 유형의 생성형 AI에 적용될 수 있습니다. 이사회 입장에서는 범용 AI가 "고위험 AI 시스템 또는 고위험 AI 시스템의 구성 요소로 사용될 수 있는 경우" 고위험 의무의 적용을 받게 됩니다.[287] AI 제공자가 생성형 AI와 관련된 지침 또는 정보에서 모든 고위험 사용을 명시적으로 배제하는 경우 범용 AI 시스템은 면제됩니다.[288] 이 면제는 위의 배제가 신의성실에 따라 이루어진 경우에만 적용될 수 있습니다.[289] 실제로 제공업체가 이사회 입장 부록 3에 정의된 고위험 환경에서 해당 시스템이 절대 사용될 수 없도록 보장하는 것은 매우 어려울 것입니다. 따라서 '신의성실' 요건의 한도가 매우 중요하며, 사실상 모든 생성형 AI 시스템에 고위험 의무를 적용하도록 요구하는 경우나 기준이 너무 낮은 것으로 판명되는 경우 모두가 개발자에 대하여 무의미하게 면책하는 근거가 될 수 있습니다. 어떤 경우든, 이사회의 입장이 생성형 AI 시스템에 미칠 영향은 불확실합니다.

3.6.3 AIA에 대한 EU 의회 입장

2023년 5월 현재, EU 의회의 입장은 여전히 협상 중입니다. EU 의회 LIBE 및 IMCO 위원회는 5월 11일에 절충안을 승인했습니다.[290] 이 절충안은 6월 중순에 본회의에서 표결에 부쳐질 예정입니다. 따라서 다음은 이 문서 작성 시점에 의회 입장으로 추정되는 내용(이하 "의회 입장")에 대한 것입니다.

전반적으로 유럽 의회의 입장은 유럽 집행위원회 제안을 크게 개선했습니다. 소비자에게는 고위험 AI 시스템의 결정 대상이 될 때 고지받을 권리,[291] AI 시스템에 대해 당국에 진정을 제기할 권리,[292] 감독 당국이 조치를 취하지 않을 경우 법원에 제소할 권리 등 새로운 권리가 부여되었습니다.[293] 또한 소비자에게 AI 시스템이 소비자 집단에 피해를 입힌 경우 집단 구제를 청구할 수 있는 권리도 부여되었습니다.[294]

생성형 AI와 관련하여 유럽 의회는 이사회 입장처럼 범용 AI에 초점을 맞추지 않고 ' 파운데이션 모델'이라는 새로운 개념을 도입했습니다. 의회 입장에서는 ' 파운데이션 모델'을 "광범위한 데이터에 대해 대규모로 학습되고, 결과물의 일반성을 위해 설계되었으며, 다양한 고유 작업에 적용할 수 있는" AI 모델로 정의합니다.[295] 파운데이션 모델의 모든 제공자는 모델이 독립형 모델로 제공되는지 시스템에 내장되어 있는지, 모델이 오픈 소스인지 폐쇄 소스인지에 무관하게 추가적인 의무를 갖습니다.[296] 이러한 의무에는 건강, 안전 및 법치 등에 대한 위험을 식별하고 완화하기 위한 요구사항, 데이터 거버넌스 조치, 예를 들어

적절한 수준의 성능, 예측 가능성, 수명 주기 전반에 걸친 해석 가능성, 에너지 효율 측정 및 기술 문서 작성 등이 포함됩니다.[297]

생성형 AI 시스템의 기초로 사용되는 모델은 파운데이션 모델의 정의에 포함되는 것이 분명합니다. 의회 입장은 파운데이션 모델 제공자에게 의무를 부과하는 조항에서 이러한 시스템을 명시적으로 언급하고 있습니다. 생성형 AI 시스템에 사용되는 파운데이션 모델에는 투명성, 불법 콘텐츠 생성에 대한 적절한 보호 장치, "저작권법에 따라 보호되는 학습 데이터의 사용에 대한 충분히 상세한 요약"을 게시할 의무가 추가로 부과됩니다.[298]

3.6.4 AI법은 소비자를 보호해야

생성형 AI 모델은 특정 상황에 맞춰 구축된 것이 아니며 광범위하게 사용할 수 있기 때문에 일부 저자는 생성형 AI 모델이 AIA 초안의 위험 기반 시스템에 적합하지 않다고 주장합니다.[299] 그 대신 이들은 시스템 위험 모니터링과 같은 표적화된 조치를 고려할 것을 주장합니다. 동시에 이 보고서 전반에 걸쳐 설명 한 바와 같이, 생성형 AI 시스템은 시스템이 시장에출시되는 시점이나 출시 이후가 아니라 시스템 개발 단계에서 완화해야 하는 중대한 위험을 야기합니다.[300]

AIA가 생성형 AI에 어떻게 적용될지는 아직 확실하지 않습니 다. 그러나 AIA를 통해 유럽 입법자들은 생성형 AI의 위험으로부터 소비자를 보호할 수 있는 강제력 있는 보호 장치를 도입 할 특별한 기회를 얻었습니다. AIA가 확정되기 전까지 앞으로 몇 달 동안 이 기회를 효과적으로 활용하여 소비자 권리를 도입하고 생성형 AI 전체 행위자망에 의무를 적용함으로써 이 보고서에서 설명한 피해 문제를 해결해야 합니다.

EU 입법자들은 업계 로비로 인해 AIA에서 소비자의 의무와 권리가 사라지지 않도록 보장해야 합니다. Corporate Europe Observer의 보고서에 따르면, 범 용 AI 시스템을 규제 대상에서 제외하도록 촉구하는 등 업계의 로비 활동으로 인해 집행위원회가 제안한 규제안에서 관련 조항이 크게 약화되었습니 다.[301] EU 입법자들은 협상의 마지막 단계에서 더욱 강화될 로비 전략에 넘어가지 않도록 경계해야 합니다.

입법자들이 AIA를 확정하는 동안 다른 주요 규제 기관들도 소비자의 안전과 권리를 보장해야 합니다. AIA는 몇 년 간 완전히 적용되지는 못할 것이며,[302] 그 동안 다른 법 제도를 소관하는 규제 기관은 위에서 설명한 대로 생성형 AI의 피해로부터 소비자를 보호할 필요가 있습니다.

"그러나 AIA를 통해 유럽 입법자들은 생성형 AI의 위험으로부터 소비자를 보호할 수 있는 강제력 있는 보호 장치를 도입할 특별한 기회를 얻었습니다."

3.7 책임법

EU에는 결함이 있는 제품으로 인해 소비자가 피해를 입었을 때 공정한 보상을 받을 수 있도록 하기 위해 여러가지 관련 책임법이 있습니다. 일부 법 제도는 이미 시행 중이며, 다른 제도는 아직 협상 중입니다.

3.7.1 제조물 책임 지침

제조물 책임 지침은 소비자가 결함이 있는 제품으로 인한 피해에 대해 보상을 청구할 수 있도록 합니다. 현행 EU의 책임 규칙인 제조물 책임 지침(PLD)은 1985년에 채택되었으며, 이 규정이 생성형 AI에 적용되는지 여부는 명확하지 않습니다.

첫째, PLD가 생성형 AI와 같은 디지털 서비스 및 소프트웨어에 적용되는지 여부에 대한 합의가 이루어지지 않았습니다.

둘째, 생성형 AI와 같은 디지털 서비스에 PLD가 적용되더라도 유럽연합 사법재판소의 판결에 따르면 제품이 제공하는 정보는 PLD의 적용을 받지 않습니다.[303] 생성형 AI의 출력물은 본질적으로 음성, 텍스트 또는 이미지 형태의 정보이므로, 정보가 PLD의 적용을 받지 않는다는 것은 생성형 AI가 PLD의 적용을 받지 않을 가능성이 크다는 것을 의미합니 다.

3.7.2 개정 제조물 책임 지침

유럽 집행위원회는 제조물 책임 지침(개정 PLD)에 대한 개정안을 발표했습니다. 개정 지침은 AI 시스템을 포함한 소프트웨어에도 적용될 예정입니다.[304]

개정 PLD 제안은 제조물 책임 지침과 유사하게 무과실 책임주의에 기반한 책임 체계로 운영됩니다. 소비자는 제품 운영자나 생산자의 과실을 입증할 필요는 없지만, 제품의 관련 결함, 소비자 피해, 결함과 피해 사이의 인과관계는 소비자가 입증해야 합니다.
현행 PLD에 대한 최근 사법재판소의 판결이 개정 PLD의 맥락에서 어떻게 적용될지, 이에 따라 제품에서 제공하는 정보가 개정 PLD의 적용을 받을 수 있을지 여부는아직 지켜봐야 합니 다.

어쨌든 생성형 AI 시스템의 경우 2장에서 설명한 것처럼 잠재적인 소비자 피해 대부분이 비물질적입니다. 이러한 피해는 물질적 피해만 보장하는 PLD에서 제외됩니다.

전체적으로 볼 때, 현행 PLD나 개정 PLD 모두 생성형 AI 시스템으로 인한 소비자 피해에 대한 보상을 인정하는 데 적합하지 않은 것으로 보입니다. 그러나 PLD 제안의 최종 성안 문구를 검토해봐야 할 것입니다.

3.7.3 AI 책임 지침

유럽 집행위원회는 소비자가 AI 시스템으로 인한 피해에 대하여 보상을 청구할 수 있도록 AIA와 병행하는 AI 책임 지침(AILD)을 PLD에 대한 추가 지침으로 제안했습니다. 집행위원회가 초안을 마련하였지만, AIA가 채택될 때까지는 이 지침도 확정되지 않을 가능성이 높습니다.

AILD 제안은 소비자가 모든 물질적 피해에 대하여 보상을 청구할 수 있는 가능성을 제공하며, 각국의 법 제도에서 허용하는 경우 비물질적 피해에 대해서도 보상을 청구할 수 있도록 하였습니다. 그러나 이 제안에는 소비자 피해 보상에 대한 효과를 크게 떨어뜨릴 수 있는 심각한 한계가 있습니다.[305]

AI 시스템으로 인한 피해에 대하여 보상을 청구하려는 소비자는 AI 시스템 운영자의 과실을 입증해야 합니다. AILD의 맥락에서 과실을 입증한다는 것은 소비자가 AI 시스템 운영자가 AIA를 비롯한 EU 규칙에 따라 운영되지 않는다는 사실을 입증해야 함을 의미합니다.
이러한 규정 미준수 를 입증하려면 고도의 기술 및 법률 지식이 필요하며, 일반 소비자는 이러한 지식을 보유하고 있지 않거나 보유하고 있을 것이라고 예상하기 어렵습니다. 과실 입증은 결함과 피해로 이어지는 결과 사이의 인과관계 추정과같은 AILD의 다른 메커니즘을 위한 전제 조건이므로 이러한 제한은 상당한 것입니다. AILD가 소비자를 효과적으로 보호하려면 소비자 청구 에 무과실 책임을 인정하고 입증 책임을 전환해야 합니다.[306]

생성형 AI의 경우, AIA가 생성형 AI를 'AI 시스템'으로 분류할지 여부가 아직 불확실합니다. 만약 AIA에 의해 AI 시스템으로 분류된다면, AILD의 AI 시스템 정의가 AIA의 정의를 참조하기 때문에 AILD 역시 생성형 AI에 적용될 수 있습니다. 그러나 부정확한 정보(텍스트, 이미지 또는 오디오 형태)로 인해 발생한 손해에 대한 보상 청구는 AILD 제안에서 통합되지 못했습니다. 따라서 생성형 AI와 관련된 소비자에 대한 보상 청구는 국가 차원에서 사례별로 평가되어야 할 것입니다.

AILD는 아직 정치 과정의 초기 단계에 있으며, 각 국가별 규칙과 무관하게 EU 입법자들이 AI로 인한 피해에 대하여 소비자에게 보상할 수 있는 실질적인 선택권을 제공하는 방식으로 이 제안을 수정할 수 있는 여지가 남아 있습니다. 이는 2장 에서 설명한 피해에 직면해 있는 소비자 보호를 강화하기 위해 필요합니다.

3.8 업계 표준 및 가이드라인

업계 관계자들은 이미 생성형 AI 모델의 개발과 사용에서 투명성을 높이기 위한 가이드라인을 개발하고 있습니다.[307] 새로운 생성형 AI 모델의 개발을 중단해야 한다는 업계의 요구도 있었습니다.[308] 개발 중단 요구는 일반적으로 매우 첨단인 모델의 위험에 초점을 맞추고 있으며, 오픈AI에서 자발적으 로 GPT5 개발을 중단한 것과 궤를 같이 합니다.[309] 그러나 이 보고서 2장에서 확인 및 논의한 바처럼 GPT4 기반 시스템 등 현재의 생성형 AI 모델의 여러 위험에 대해서는 충분히 다루어지지 않고 있습니다.

또한 생성형 AI의 개발자와 배치자를 위한 자발적인 행동 지침을 만들어야 한다는 요구도 점점 더 커지고 있습니다.[310] 특히 EU의 경우, 집행위원회는 AI 법의 새로운 규칙들에 앞서 기업들과 협약을 맺는 것을 목표로 하고 있습니다. EU의 정책 입안자들은 "수개월 내"에 행동 지침을 공동 작성할 계획인 것으로 알려졌는데, 이는 AIA의 3자 협상과 동시에 행동 지침이 작성되거나 협상될 수 있음을 의미합니다. 이런 방식에서는 구글과 같은 업계 대표자들이 로비 활동을 강화할 수 있는 완벽한 위치에 놓이게 됩니다.

이러한 절차는 두 가지 위험을 초래합니다. 첫째, 유럽 집행위원회는 3자 협상에서 역할을 수행해야 하는데, 같은 주제에 대해 업계 및 제3국과 행동 지침이나 다른 자율규제 규칙을 동시에 협상하는 경우 그 역할을 공평하게 수행하기 어렵습니다. 둘째, EU에서 이러한 행위자에 대한 법적 요구사항이 아직 정의되지 않은 상태에서 자발적 협약에 어떤 요구사항이 포함될 수 있을지 불분명합니다. 자발적 약속이 최종 법률 구문과 일치하지 않을 명백한 위험이 있습니다. 자발적 행동 지침은 소비자 권리 및 인권이 아니라 제품의 사업성과 수익 창출 가능성에 대한 업계 관계자들의 견해에 크게 영향을 받을 수 있습니다. 마지막으로, AIA는 업계 관계자의 과도한 영향을 받을 수도 있습니다. 이는 용납할 수 없으며 허용되어서는 안 됩니다.

업계 표준과 지침은 이미 시장에 출시된 GPT 모델의 배치로 인한 위험을 해결하기에 부적합합니다. 이들은 최저 공통 분모 역할을 하는 경향이 있기 때문에 충분한 집행 메커니즘과 독립적인 감독이 부족한 상태입니다. AIA가 아직 적용되지 않는 동안 업계의 자발적인 약속에 의존하는 대신에, 당국은 소비자 보호법, 개인정보 보호법 또는 제품 안전 관련 법률과 같은 기존 법률의 집행에 집중해야 합니다. 정책 입안자와 입법자들은 자율 규제 체제를 피하기 위해 노력을 다해야 합니다.

4. 앞으로 나아갈 길

이 보고서에서는 생성형 AI의 개발, 학습, 배치, 사용과 관련된 심각한 피해와 문제들에 대해 설명했습니다. 이는 미래의 디스토피아에 대한 가상적 위험이 아니라 오늘날의 사람과 인구에 영향을 미치는 실질적인 피해입니다.

이들 문제가 우려스럽기는 하지만 극복할 수 없는 것은 아니라고 생각합니다. 생성형 AI와 관련된 많은 문제는 다른 분야에서 잘 알려져 있는 문제들의 반향이고, 생성형 AI 모델이 빠르게 개발되고 도입되고 있다는 사실은 즉 그 피해를 해결하기 위한 조치를 취하는 것이 타당하다는 것을 의미합니다. 기술이 우리 삶과 사회 구조에 깊숙이 자리 잡을 때까지 기다렸다가 그 개발과 사용 방향을 바꾸려고 하면 너무 늦습니 다.

기술은 길들일 수 없는 짐승이 아니라 민주 사회의 규칙과 가치에 적응하고 그에 따라 형성되어야 합니다. 생성형 AI가 소비자 권리 및 인권에 따라 개발되고 사용되도록 하려면 기업이 스스로 규제하도록 하는 것만으로는 충분 하지 않습니다. 기술이 어떻게 학습되고, 개발되고, 배치되고, 사용되어야 하는지 경계를 설정하는 것이 정책 입안자와 규제 기관의 책임입니다. 따라서 입법자들은 업계 관계자들이 기술적으로 실현 가능하다고 주장하는 것에 기반하여 법안을 통과시킬 것이 아니라, 향후 몇 년 동안 안전하고 소비자 중심적인 기술을 제공하는 데 필요한 것이 무엇인지에 기반하여야 합니다.

이하에는 사회가 생성형 AI에 접근하는 방식에서 핵심이 되어야 한다고 생각되는 몇 가지 기본 원칙이 제시되어 있습니다. 이어서 규제 기관, 정책 입안자, 입법자들을 위한 몇 가지 실행 사항이 제시됩니다. 이들 사항이 기술에 대한 인간 중심적 접근을 위한 청사진이 되기를 바랍니다.

4.1안전하고 책임성 있는 AI를 위한 핵심 소비자 권리 원칙

안전하고, 신뢰할 수 있으며, 공정하고, 공평하고, 책임성 있는 AI을 구현하기 위해서는 소비자 권리를 다루는 포괄적인 원칙이 필요합니다. 아래에 제시된 원칙은 정책 입안자와 규제 기관이 생성형 AI의 혜택과 위험 문제에 어떻게 접근해야 하는지에 대한 초석이 될 수 있습니다.

많은 원칙이 이미 현행 소비자법에 정의되어 있지만, 이러한 원칙을 실제로 생성형 AI의 개발과 배치를 위한 기반으로 보장할 것을 정책 입안자와 규제 기관에 촉구합니다. 이는 현재와 향후 몇 년 동안 소비자의 기본적 권리를 존중하는 기술 환경을 보장하기 위한 핵심입니다.

- **소비자의 권리는 존중되어야 합니다**. 정보 공개 및 투명성, 공정성 및 차별금지, 안전 및 보안, 프라이버시 및 개인정보 보호, 권리 구제 등 이미 확립된 소비자 권리 및 인권이 생성형 AI의 등장으로 인해 훼손되거나 대체되어서는 안 됩니다.
- 소비자에게 중대한 영향을 미치는 결정을 내리는 데 생성형 AI 모델이 사용될 때마다 소비자는 **이의를 제기하고 설명을 들을 권리**를 가질 수 있어야 합니다.
- 소비자는 생성형 AI에서 개인정보를 **삭제**할 수 있는 '**잊힐 권리**'를 가져야 하며, 자신에 대하여 생성된 잘못된 정보로 인한 피해를 **정정**할 권리를 가질 수 있어야 합니다.
- 소비자는 고객 서비스와 같은 상황과 관련하여 **생성형 AI 대신 인간과 상호작용할 권리**를 가질 수 있어야 합니다. 이는 소비자에게 추가 비용을 발생시켜서는 안 되며, 소비자가 지불 능력에 따라 다르거나 불공정한 대우를 받지 않도록 해야 합니다.
- 소비자는 생성형 AI 사용으로 인해 입은 **피해에 대한 구제 및 보상을 받을 권리**를 가질 수 있어야 합니다.
- 소비자는 **집단 구제를 받을 권리**를 가질 수 있어야 하며, 소비자 단체 및 기타 시민 사회 단체의 대리를 받아 권리를 행사할 수 있어야 합니다.
- 소비자는 생성형 AI 모델의 사용이 법률을 위반하는 경우 **감독 당국에 진정을 제기하거나 법원에 법적 소송을 제기할 권리**를 가질 수 있어야 합니다.
- 생성형 AI 모델의 개발자와 배치자는 이러한 권리를 실제로 소비자가 **사용할 수 있도록 보장하는 체계를 구축해야** 합니다.

4.2 정책 권고 사항

기술을 사회에 통합하는 방법에 대한 결정은 본질적으로 정치적 문제입니다. 선출직 공무원과 정부는 기술이 소수의 기업의 처분이 아니라 사람을 위해 봉사하도록 보장할 책임이 있습니다. 소비자 중심의 기술 정책은 사람과 사회가 실험적인 기술을 위한 테스트 실험실로 이용되어서는 안 된다는 것을 의미합니다. 시민의 권리와 사회에 미치는 영향을 충분히 고려하지 않은 채 디지털 사회로 광범위하게 전환하면서 얻은 교훈은 정부가 생성형 AI에 접근하는 방식에 참고가 되어야 합니다.

사회적 측면에서 책임감 있고 공정하며 책임감 있는 혁신을 보장하기 위해서는 과대광고에 휩쓸린 후 그 여파에 따라 방향을 수정하는 것이 아니라 미래를 대비한 견실한 정책이 필요합니다. 이하에서는 정부와 정책 입안자들이 생성형 AI 및 유사 기술에 어떻게 접근해야 하는지에 대한 몇 가지 실행 사항을 제시합니다.

4.2.1 규제 기관의 행동 및 권한 부여 촉구

생성형 AI와 같은 신기술은 때때로 규제의 황무지로 묘사되기도 하지만, 이미 포괄적인 법 제도가 마련되어 있습니다. 이들 규제 중 상당수는 이 보고서 2장에서 설명한 여러 문제를 해결하는 데 이미 적합하다고 생각합니다. 그러나 착취, 차별, 기타 권력 남용으로부터 사람들을 효과적으로 보호하려면 이러한 법률이 집행되어야 합니다.

효과적인 집행을 위해서는 규제 기관이 필요한 권한과 전문성, 재원을 갖추고 있어야 하며, 규정을 준수할 수 없거나 준수하기를 거부하는 기업을 충분히 처리할 수 있어야 합니다. 이 절에서는 규제 기관이 소비자 친화적인 방식으로 기술을 형성하기 위하여 기존의 수단을 사용하는 데 필요한 몇 가지 접근 방식과 전제 조건을 제시합니다.

- **규제 기관은 다가오는 규제를 기다리지 말아야 합니다**. 그 보다, 생성형 AI 시스템을 **즉시 조사**하고 개인정보 보호, 경쟁, 제품 안전, 소비자법 등 각각의 법 제도 **관련 법률 조항을 적용**해야 합니다.
- 생성형 AI로 인한 위험을 관리하기 위하여 여러 규제 기관이 동일한 조사에 참여하는 **부문 간 공동 조사**가 필요할 수 있습니다. 이러한 공동 조사의 진행을 보장하기 위해 알고리즘 집행에 대한 조정자를 지정해야 할 수도 있습니다.
- **규제 기관은** 생성형 **AI 모델에 대한 시판 후 감시를 수행할 수 있는 권한을 부여받고** 관련 법률을 준수하지 않는 **알고리즘 시스템 또는 그 일부에 대해 제품 리콜 또는 폐기**를 명령할 수 있어야 합니다. 이러한 명령에는 잘못된 관행을 억제하기 위해 상당한 금전적 벌금이 수반되어야 합니다.
- **규제 기관은** 개인적, 기술적 역량과 필요한 기술 수단을 비롯하여 각각의 법 제도의 **위반에 대한 단속에 필요한 모든 자원을 갖추어야 합니다**. AI로 생성된 콘텐츠가 범람함에 따라 시장 감시 및 단속을 확대해야 할 것입니다.
- 규제 기관의 집행 노력을 지원하기 위해 다국적 및 국가별 기술 전문가 기구를 설립해야 합니다.
- 기술을 사용하여 집행을 강화하는 방법에 대한 연구가 수행되어야 합니다.

4.2.2 의사 결정권자 - 전략적 조치

- **정부는 국가 AI 전략에서 생성형 AI에 대한 비판적 관점을 고려해야 합니다**. 안전하고 인간 중심적인 생성형 AI를 촉진하기 위한 가장 중요한 원칙은 추가 방식이 아니라 처음부터 내포되어 있어야 합니다. 추가 방식은 비용이 소요되고 실수로 인해 신뢰가 훼손하기 때문입니다.
- **정부는 공공 부문에서 생성형 AI를 사용할 때 비판적이고 예방적인 접근 방식을 취해야 합니다.** 공공 부문은 합법적이고 신뢰할 수 있는 방식으로 생성형 AI를 사용해야 할 특별한 책임이 있으며, 공공 조달을 활용하여 독립형 생성형 AI 소프트웨어 또는 생성형 AI가 내장된 시스템을 제공하는 업체에 적극적인 영향을 미쳐야 합니다. 특히 공공 부문은 공공 부문의 상황에 기술을 사용하기 전에 투명성을 확보하여 해당 기술을 이해할 수 있어야 합니다.
- 정부는 기관을 설립하거나 기존 기관에 권한을 부여하여, 기술이 공익을 위해 개발, 배치, 사용될 수 있도록 지속적으로 감독하고 공개적으로 논의하며 의무 원칙을 정의할 것을 강력히 고려해야 합니다.
- 정부는 데이터 관행과 생성형 AI로 인한 소비자 및 사회적 피해에 대한 연구에 공공 자금을 지원해야 합니다.
- **국제 무역 협정으로 인해** 생성형 AI에 대한 **투명성 의무나** 소비자 권리를 보장하는 데 필요한 기타 의무가 **무력해져서는 안 됩니다.**
- 생성형 AI 시스템을 개발 및 배치하는 기업의 **이해관계자 및 투자자**, 특히 공공 부문의 이해관계자 및 투자자는 **착취적 관행, 환경 영향 등을 방지 및 완화하기 위한 조치를 취하도록 요구해야 합니다.** 기업은 윤리 지침을 마련하고 취한 조치에 대해 공개해야 합니다.

4.2.3 새로운 입법 조치

이미 시행중인 많은 법 제도가 생성형 AI의 피해에 대응하는 데 적합할 수 있지만, 법적 공백과 허점이 있는 영역이 분명히 존재할 것입니다. 기존 법률이 충분하지 않은 경우에는 소비자를 피해로부터 보호하기 위한 새로운 법 제도를 만들어야 합니다. 이전 장에서 설명한 바와 같이 이미 여러 입법 과정이 진행 중이며, 이러한 과정 을 통해 소비자 권리 및 인권에 기반하여 강력하게 미래를 보장하는 규칙이 만들어지는 것이 중요합니다.

우리는 정책 입안자들과 입법자들이 소비자 보호와 인권 보장에 대하여 강력한 입장을 취할 것을 촉구합니다. 생성형 AI 시스템의 개발자와 배치자가 투명하고 책임성 있는 방식으로 운영하도록 엄격한 의무를 부과하고, 이들 권리와 근본적으로 양립할 수 없는 시스템의 개발, 배치 및 사용을 제한하는 등 강력한 법적 조치가 필요합니다.

4.2.3.1 추가 조사가 필요한 특정 형태의 생성형 AI

- **생성형 AI 시스템에서 특정 형태의 조작 기술은 금지되어야 합니다**. 예를 들어, 1인칭 언어 사용, 이모티콘 및 유사한 기호 사용, 인간의 감정 및 유사한 속성 시뮬레이션 등 의인화된 모델에 대한 상당한 제한이 포함될 수 있습니다. 이러한 제한은 사용 상황과 목적에 따라 달라질 수 있습니다. 아동 등 취약 계층이 사용할 때는 허용되는 기술 및 애플리케이션의 한도가 더 높아야 합니다.
- 생성형 AI 시스템을 특정 용도로 사용할 경우, 배치 전에 관련 규제 기관의 **사전 승인**을 받아야 할 수도 있습니다. 소비자, 특히 아동 등 취약한 소비자의 착취나 차별을 초래할 수 있는 모델이 그 예가 될 수 있습니다.
- **정책 입안자들은 미래에도 적용 가능한 법안을 마련하여**, 당국이 급속한 기술 발전에 뒤처지지 않도록 해야 합니다. 여기에는 기술 중립적인 원칙과 규정이 포함됩니다.

4.2.3.2 생성형 AI 개발자 및 배치자의 의무

책임성 있는 생성형 AI의 개발, 배치 및 사용은 시스템이 작동하는 방식을 통제하고, 학습 데이터를 검사하고, 사회적 및 환경적 영향을 감독하는 등의 일이 가능하다는 것을 전제로 합니다. 투명성 자체가 만병통치약은 아니지만, 이는 기술이 소비자 권리 및 인권을 훼손하지 않도록 보장하기 위한 전제 조건입니다. 이는 기업의 책임으로만 맡겨둘 수는 없습니다.

생성형 AI 시스템에 대한 독립적인 감독, 연구 및 감사가 시급히 필요합니다. 이는 문제가 발생할 경우 기업이 책임을 지도록 하고, 편향성과 부정확성을 식별 및 근절하며, 법률 준수를 보장하고 피해를 완화하기 위한 것입니다. 따라서 우리는 생성형 AI 시스템의 개발자와 배치자에게 부과해야 할 몇 가지 조치를 제시합니다.

투명성

- 생성형 AI 시스템의 개발자와 배치자는 **위험 평가**, 위험 완화 계획, 콘텐츠관리 방법, 표준화된 성능 지표 등에 관한 **문서를 공개하고 발표할 의무가 있습니다**. 이는 두 가지 수준, 즉 일반 소비자를 위한 짧고 덜 기술적인 설명과 시민 사회, 학계 및 기타 제3자를 위한 심층적인 설명 모두가 이루어져야 합니다.
- 생성형 AI 시스템을 개발 및 배치하는 모든 기업은 생성형 AI 모델의 전체 수명주기 동안 **에너지 사용, 물 사용, 탄소 배출에 대한 모든 정보를 발표**하고 일상적인 사용에 대한 향후 배출량 예측을 제시할 **의무가 있습니다**. 여기에는 하드웨어 생산, 모델 학습, 개발, 배치 및 사용에 필요한 자원에 대한 정보가 포함됩니다. 기업이 자체 계산 시스템을 개발하기보다는 모든 기업이 사용할 수 있는 배출량, 물 사용량 에너지 사용량 계산에 대한 표준화된 모델이 구축되어야 합니다.
- 개발자와 배치자는 모든 공급업체의 이름을 공개하고, 폭력적이고 유해한 콘텐츠의 관리자를 위한 생활임금 및 심리적 지원, 특정 작업에 일시적으로만 필요한 노동자를 위한 계획을 비롯하여 **전체 공급망의 노동 조건에 대해 투명한 방식으로** 공개해야 합니다.

- 생성형 AI의 기본 모델 개발자는 관련 생성형 AI 모델을 감독할 수 있도록 중앙 집중식 공개 시스템에 모델을 등록할 의무가 있습니다.
- 소비자 대면 인터페이스 및 서비스에 생성형 AI를 배치하는 자는 **생성된 콘텐츠가 개발자, 배치자 또는 제3자의 상업적 이해관계에 의해 어떻게 영향을 받는지 공개할 의무**가 있습니다. 이는 특히 검색 질의 또는 이와 유사한 상황에서 생성되는 콘텐츠처럼 생성된 콘텐츠가 소비자의 선택을 위한 정보를 제공하는 경우와 관련이 있습니다.
- 생성형 AI 시스템의 배치자는 **소비자가 생성형 AI 시스템과 상호 작용할 때마다**, 그리고 소비자 대면 시스템이 의사 결정 결과에 영향을 미치는 AI을 사용하는지 여부를 **공개할 의무**가 있습니다.
- 공공 및 민간 기관은 **생성형 AI에 의해 생성된 콘텐츠**가 소비자, 보다 광범위한 소비자 권리 또는 민주적 절차에 영향을 미치는 결정에 영향을 미칠 수 있는 경우 이를 **공개할 의무**가 있습니다.

위험 완화

- 생성형 AI 시스템 배치자는 생성형 AI 시스템이 배치될 상황을 신중하게 검토해야 할 의무가 있습니다. 생성형 AI 시스템의 배치자는 **신중한 위험 평가 없이 생성형 AI 시스템을 사용해서는 안 되며**, 이는 시스템이 해결하고자 하는 문제에 대한 매핑, 시스템이 관련 법률을 준수하는지 확인, 소비자 및 소비자 권리에 대한 위험, 인권에 대한 위험, 취약 계층에 영향을 미칠 수 있는 위험, 환경에 대한 부정적 영향, 예측 가능한 사회적 및 집단적 피해, 개인정보 피해 등을 포함하여야 합니다.
- 생성형 AI 시스템의 배치자는 시스템을 배치하기 전에 위험 평가에서 **발견된 위험을 완화하기 위한 효과적인 조치를 구현**하여 수용 가능한 잔여 위험 수준에 도달할 의무가 있습니다. 위험을 완화할 수 없거나 시스템이 해결하고자 하는 문제를 해결하지 못하는 경우 해당 상황에 시스템을 배치해서는 안 됩니다.
- 생성형 AI 시스템의 개발자와 배치자는 **해당 기술의 영향을 받을 수 있는 집단**, 특히 소외되고 취약한 집단과 커뮤니티의 **대표를 참여시킬 의무**가 있습니다. 이는 민주적 참여와 분야 간 참여를 촉진합니다. 이해관계자의 참여는 생성형 AI 모델의 개발 및 학습, 그리고 다양한 문화적 맥락, 언어 등을 고려해야 하는 위험 평가, 위험 완화 계획, 콘텐츠관리와 같은 관련 주제 활동에 필요합니다.
- 생성형 AI 시스템의 배치자는 시스템 배치 후 **시스템이 소비자에게 미치는 영향을 모니터링하고 해결해야 하는 의무가 있으며**, 특히 소외되고 취약한 집단과 커뮤니티에 미치는 영향을 고려하여 수용 가능한 잔여 위험 수준에 도달하기 위해 지속적인 위험 평가 및 완화를 수행해야 합니다.

책임성

- 프라이버시, 안전, 소비자 권리, 기본권 일반에 대한 피해 등 **생성형 AI 시스템의 유해한 영향에 대한 책임과 의무에 대한 명확한 규칙**이 있어야 합니다. 이러한 규칙은 공급망에서 어떤 회사가 책임을 지는지 명확하게 표시하거나 생성형 AI 개발자와 배치자가 상호 간의 책임과 의무를 명확하게 설정할 것을 요구해야 합니다.
- 모든 책임 제도는 소비자, 규제 기관 및 법원이 소비자 피해에 대한 책임을 기업에 쉽게 물을 수 있도록 하여야 합니다.
- 생성형 AI 시스템 개발자는 **사용하는 데이터**, 데이터세트의 대표성, 데이터 정제 및 라벨링 관행, 시스템의 모든 다운스트림 사용에 영향을 미칠 기타 설계 선택에 대해 **책임을 져야 합니다**. 이러한 선택 사항은 신중하게 문서화하여 다운스트림 개발자와 배치자가 생성형 AI 시스템의 위험과 적합성을 검토할 수 있도록 해야 합니다.
- **기술 표준과 인증 제도가 개발되고** 사용되어 생성형 AI 시스템의 개발자와 배치자가 책임성 있고 합법적인 방식으로 시스템을 개발, 학습, 배치 및 사용할 수 있도록 지원하여야 합니다. 그러나 정책 입안자들은 인권, 정치적, 법적 문제를 표준 기관에 위탁해서는 안 됩니다. 정부는 이러한 기구 에 시민사회의 참여를 보장해야 하며, 시민사회의 참여가 부족한 기구에 의존해서는 안 됩니다.
- 생성형 AI 시스템과 모델은 **독립적인 연구자, 규제 기관 및 기타 제3자가 감사할 수 있어야** 합니다. 이는 편향성과 차별의 위험을 완화하고, 학습 데이터의 책임성 있는 사용을 보장하며, 관련 법적 요구사항을 준수하는 데 필수적입니다.
- 감사는 최소한 학습 데이터, 데이터 수집 관행, 데이터 라벨링 관행, 콘텐츠관리 관행, 지속가능성 보고서 및 알고리즘 모델에 대해 이루어져야 합니다. 감사는 책임성과 재현성을 보장하기 위해 신중하게 문서화되어야 하며, 표준화된 감사 요구사항에 기반해야 합니다.
- 기업은 생성형 AI를 개발하고 배치할 때 발생하는 탄소 배출량, 에너지 및 물 사용량에 대한 계산을 기반으로 그 **소비를 줄이기 위한 정량적이고 시간제한적인 약속을 이행해야 할 의무**가 있습니다. 또한 이러한 진행 상황은 공개 발표를 통해 외부의 독립적인 기관에게 감사를 받아야 합니다. 이는 대규모 모델이 축소되고 덜 규모화할 수 있음을 의미합니다. 순탄소 배출 제로 활동과 탄소 상쇄 제도에 대한 주장이 기업으로 하여금 배출량을 '보상'하게 하는 기본 모델이 되어서는 안 되며, 기업은 자체 활동에서 배출량을 **줄여야** 합니다.

더 자세한 정보는:

핀 뤼초우-홀름 미르스타드, 디지털 정책 담당 이사

노르웨이 소비자위원회

이메일: finn.myrstad@forbrukerradet.no

www.forbrukerradet.no/ai

FORBRUKERRÅDET