소개: 2023년 3월 30일, 인공지능 및 디지털 정책 센터(CAIDP, 미국 워싱턴 DC 소재 비영리 연구소)는 미국 연방거래위원회(FTC)에 "오픈AI 회사와 그 제품 챗GPT에 대해 조사를 요청하는" 진정을 제기하였습니다. 센터는 오픈AI에 대해 조사가 이루어지고 필요한 보호장치가 마련될 때까지 추가 모델 출시를 금지해 줄 것을 요청했습니다. (관련기사) 정보인권연구소는 CAIDP가 공개한 진정서 전문을 번역하여 소개합니다(각주 제외, 초벌번역은 기계번역의 도움을 받았습니다).

* 원문: https://www.caidp.org/app/download/8450269463/CAIDP-FTC-Complaint-오픈AI-GPT-033023.pdf?t=1681139856

연방 무역 위원회 귀중

오픈AI 유한책임회사 사안 관련

인공지능 및 디지털 정책 센터(CAIDP) 제출

I. 요약

1. 캘리포니아에 기반한 기업, 오픈AI 유한책임회사(OpenAI Inc.)는 편향적이고 기만적이며 개인정보 보호 및 공공 안전 측면에서 위험성이 있는 GPT-4 제품을 소비자 시장에 출시했습니다. 이 제품의 결과물은 증명되거나 재현될 수 없습니다. 배포 전 독립적인 평가도 수행되지 않았습니다. 오픈AI는 "허위정보 및 오도된 작동", "재래식 및 비재래식 무기의 확산", "사이버 보안" 측면에서 특정한 위험성을 인정했습니다. 오픈AI는 "AI 시스템은 전체적인 이데올로기, 세계관, 진실과 비진실을 강화하고 이를 고착화하여 미래의 논쟁, 성찰, 개선 가능성을 봉쇄할 가능성이 매우 크다"고 경고한 바 있습니다. 이제 이 회사는 그로 인한 결과에 대해 책임을 부정하기 시작했습니다.

2. 연방거래위원회(FTC)는 AI의 사용이 "투명하고, 설명 가능하고, 공정하고, 경험적으로 건전해야 하며 책무성을 증진해야 한다"고 선언했던 바 있습니다. 오픈AI의 제품 GPT-4는 이러한 요구 사항을 충족하지 않습니다. FTC가 나서야 할 때입니다. 미국내 공급되는 상업용 AI 제품에 대한 독립적인 감독 및 평가가 있어야 합니다. CAIDP는 FTC가 오픈AI에 대한 조사를 개시하고, GPT-4의 추가 상용 출시를 금지하고, 소비자, 기업 및 상업 시장을 보호하기 위해 필요한 방호책을 수립할 것을 촉구합니다.

II. 당사자

3. CAIDP는 워싱턴 DC에 설립된 비영리 연구소입니다. AI 정책 전문가 및 활동가로 구성된 CAIDP 글로벌 네트워크는 60개국에 걸쳐 있습니다. CAIDP는 미래의 AI 정책 리더를 교육하는 활동을 수행합니다. CAIDP는 각국 AI 정책 및 관행에 대하여 최초로 포괄적인 조사를 실시한 바 있습니다. CAIDP는 각국 정부와 국제 기구에 AI 및 신기술에 대한 정책 자문을 정기적으로 제공합니다.

4. 오픈AI는 비영리 오픈AI 유한책임회사(OpenAI Inc.)와 그 영리 자회사 오픈AI 유한책임투자자(OpenAI LP)로 구성된 미국의 인공지능(AI) 연구소입니다. 오픈AI는 2015년에 설립되었습니다.

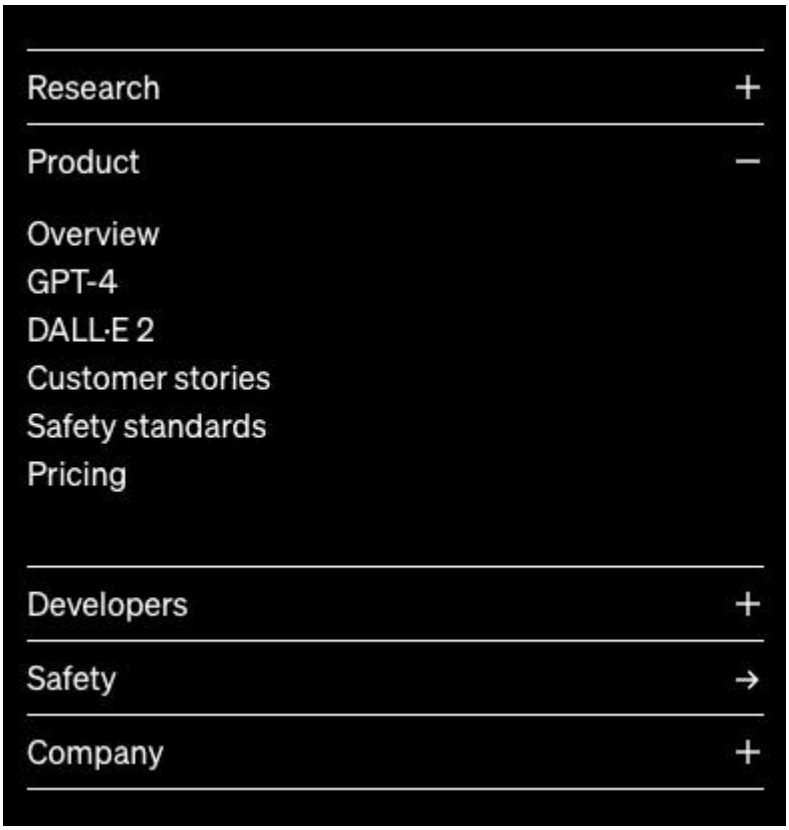
III. 관할권

5. FTC는 "미국 내 어느 지역에서든 그 직무 수행에 필요한 모든 조사를 수행"할 수 있습니다(FTC Act Sec. 3, 15 U.S.C. Sec. 43) 또한 "은행, 저축 및 대출 기관 … 연방 신용 조합 … 과 일반 운송업체를 제외하고, 상거래에 종사하거나 상거래에 영향을 미치는 개인, 제휴회사 또는 주식회사의 조직, 사업, 행위, 관행 및 관리에 관한 정보를 수집 및 조합하고 수시로 조사할 수 있는 권한"을 가지고 있습니다(FTC Act Sec. 6(a), 15 U.S.C. Sec. 46(a)).

6. FTC는 "상거래에서 또는 상거래에 영향을 미치는 불공정하거나 기만적인 행위 또는 관행"을 조사, 기소 및 금지할 권한이 있습니다.

7. 오픈AI는 AI 기반 제품인 DALL-E, GPT-4, 오픈AI Five, 챗GPT, 오픈AI Codex를 상업용으로 출시했습니다. 오픈AI도 이들 AI 모델을 "제품"이라고 설명했습니다.

8. 오픈AI는 현재 이들 제품에 대한 "가격 정보"를 제공하고 있습니다. 가격 정보는 언어 모델 GPT-4, Chat 및 InstructGPT에 대해 제공됩니다. GPT-4 32k 컨텍스트 모델의 경우 1k의 Prompt 토큰을 $0.06에 구입할 수 있고 1k의 Completion 토큰을 $0.12에 구입할 수 있습니다. 오픈AI는 미세조정(Fine-tuning) 모델 및 내장형(Embedded) 모델에 대한 가격 정보도 제공합니다. 오픈AI는 이미지 모델 및 오디오 모델을 포함한 기타 모델에 대한 가격 정보를 제공합니다.



<이미지 1> "제품"을 보여주는 오픈AI 웹사이트의 스크린샷(2023년 3월 25일)

9. 오픈AI는 여행, 금융 및 쇼핑을 비롯한 일상적인 소비자 서비스를 위해 GPT-4용 플러그인을 사용할 수 있도록 했습니다. "처음에는 11개의 플러그인만 사용할 수 있습니다. 이들 플러그인은 사용자가 생중계 스포츠 경기의 점수를 확인하는 것부터 국제항공을 예약하거나 음식 배달을 구매하는 일까지 다양하게 아우릅니다."

IV. AI 거버넌스에 대한 공공 정책

10. 미국 정부의 공식적인 약속과, 법률 전문가, 기술 전문가, 과학 단체의 권고 사항으로 도출된 AI 거버넌스 최근 규범으로는 다음과 같은 것들이 있습니다.

A. OECD AI 원칙

11. 경제협력개발기구(OECD)는 경제 협력과 발전을 촉진하기 위해 1961년에 설립되었습니다.

12. 현재 OECD 회원국은 미국을 포함하여 38개국입니다.

13. 2019년 OECD 회원국은 여러 OECD 비회원국과 협력하여 OECD AI 원칙을 공표했습니다.

14. 미국은 OECD AI 원칙을 채택했습니다.

15. G-20 국가들은 OECD AI 원칙을 채택했습니다.

16. 인간 중심의 가치와 공정성에 관한 OECD AI 원칙은, "AI 행위자는 AI 시스템 수명 주기 전반에 걸쳐 법치주의, 인권 및 민주적 가치를 존중해야 한다. 여기에는 자유, 존엄성 및 자율성, 사생활권 및 개인정보 보호, 차별금지 및 평등, 다양성, 공정성, 사회 정의 및 국제적으로 인정된 노동권이 포함된다."는 것입니다.

17. 견고성, 보안 및 안전에 관한 OECD AI 원칙은, "AI 시스템은 수명주기 전반에 걸쳐 견고하고 보안이 지켜지며 안전해야 한다. 그래서 정상적인 사용, 예측 가능한 사용 및 남용되는 상황에서 또는 기타 불리한 상황에서도 적절하게 기능하고 불합리한 안전 위험을 초래하지 않아야 한다."는 것입니다.

18. 투명성 및 설명 가능성에 관한 OECD AI 원칙은, AI 행위자가 "상황에 적절하고 최신 기술에 부합하는 유의미한 정보를 제공하여야 한다. 이로써 (i) AI 시스템에 대한 일반적인 이해를 촉진해야 하고, (ii) 이해관계자가 직장 등에서 AI 시스템과 상호 작용하는 상황을 인지할 수 있어야 하며, (iii) AI 시스템의 영향을 받는 사람들이 결과를 이해할 수 있도록 하고, (iv) AI 시스템으로 부정적인 영향을 받는 사람들이 AI의 결과물에 이의를 제기할 수 있어야 한다. 이러한 이의제기는 AI 요소에 대해 평이하고 이해하기 쉬운 정보, 그리고 예측, 추천, 결정의 기반이 되는 AI 논리에 근거할 수 있어야 한다."는 것입니다.

19. 책무성에 관한 OECD AI 원칙은, "AI 시스템을 개발, 배포 또는 운영하는 조직 및 개인은 위 원칙들에 따라 적절히 작동케 할 책임이 있다."는 것입니다.

20. OECD AI 원칙들은 FTC법의 의미 내에서 "확립된 공공 정책"입니다.

B. AI에 대한 공통 지침

21. AI에 대한 공통 지침(UGAI)은 인권 보호에 기반한 AI 거버넌스 체계로서, 2018년 벨기에 브뤼셀에서 개최된 개인정보 보호 및 프라이버시 위원회 국제 회의에서 제정되었습니다. 이 회의는 유럽 개인정보보호 전 감독관인 Giovanni Buttarelli가 주최했습니다.

22. UGAI는 40개국 300명 이상의 전문가와 70개 기구 간에 합의되었습니다.

23. UGAI 투명성 권리는, "모든 개인은 자신과 관련된 AI 의사결정의 근거를 알 권리가 있다. 여기에는 해당 결과물을 만들어낸 요소, 논리, 기술에 대한 접근권이 포함된다."는 것입니다.

24. UGAI 공정성 의무는, "기관은 AI 시스템이 부당한 편견을 반영하거나 용인될 수 없는 차별적 의사결정을 내리지 않도록 보장해야 한다."는 것입니다.

25. UGAI 평가 및 책무성 의무는, "AI 시스템은 그 목적과 목표, 혜택뿐 아니라 그 위험성에 대해서도 적절한 평가가 이루어진 후 배치되어야 한다."는 것입니다.

26. UGAI 정확성, 신뢰성 및 타당성 의무는, "기관은 의사결정의 정확성, 신뢰성 및 타당성을 보장해야 한다."는 것입니다.

27. UGAI 종료 의무는, "AI 시스템을 구축한 기관은 시스템에 대한 인간의 제어가 더 이상 불가능할 경우 시스템을 종료할 적극적인 의무가 있다."는 것입니다.

28. UGAI는 FTC법의 의미 내에서 "확립된 공공 정책"입니다.

V. 사실적 배경

A. 오픈**AI**

29. 2016년 오픈AI는 "인공 일반 지능(AGI)은 경제적으로 상당히 가치 있는 작업을 수행함에 있어 인간을 능가하는 고도로 자율적인 시스템을 의미"한다며, 자사의 사명이 "AGI가 모든 인류에게 혜택이 되도록 보장"하는 데 있다고 밝힌 바 있습니다. "우리는 안전하고 유익한 AGI를 직접 구축하기 위해 노력할 것이지만, 우리의 작업으로 다른 사람들이 이러한 결과를 달성하는 데 기여할 수 있다면 우리의 임무가 완수된 것으로 간주할 것입니다." 오픈AI는 지켜야 할 몇 가지 원칙으로 광범위하게 분배되는 혜택, 장기적 안전성, 기술적 리더십, 협력 지향의 원칙을 수립했습니다.

30. 2016년 이후 오픈AI의 사업 구조, 사업 관행 및 사업 활동에 변동이 있었습니다. 2019년 오픈AI는 영리 기업으로 전환했습니다. 오픈AI CEO 알트만은 마이크로소프트로부터 10억 달러를 투자받았고 마이크로소프트와 오픈AI 기술 일부를 라이선스화하고 상용화하는 데 합의했습니다.

31. 또한 오픈AI의 공동 설립자 샘 알트만은 전 세계 거의 모든 사람들의 홍채 스캔을 수집하기 위해 노력하는 회사 월드코인의 설립자이기도 합니다.

B. GPT

32. 이 진정서에서 "GPT"는 AI 대규모 언어 모델 제품군인 Generative Pre-trained Transformer를 의미합니다.

33. 위키피디아에 따르면, 언어 모델의 GPT(Generative Pre-training)에 관한 원래 논문은 2018년 오픈AI 웹사이트에 프리프린트(preprint)로 게시되었습니다. 이 논문은 "언어 생성 모델이 연속적인 긴 텍스트가 포함된 다양한 말뭉치를 사전 학습함으로써 세계 지식을 습득하고 장거리 의존성(long-range dependencies) 문제를 처리할 수 있는" 기법을 제시하였습니다.

34. GPT-2는 2019년 2월에 발표되었으며, 처음에는 제한적인 데모 버전만 공개되었습니다. GPT-2 정식 버전은 가짜 뉴스 작성에 활용되는 등 남용될 가능성을 우려하여 즉시 출시되지는 않았습니다.

35. 2023년 3월 15일, 오픈AI는 GPT-4 기술 보고서를 발표했습니다. 99페이지 분량의 이 보고서는 GPT-4의 기능, 제한 사항, 위험성 및 완화 조치에 대하여 개괄하고 있습니다. 기술 보고서는 보고서 요약본 뿐 아니라 이 보고서의 범위와 제한 사항에 대한 논의도 포함하고 있습니다. 이 기술 보고서에는 벤치마크 시험 방법론, RHLF가 성능에 미치는 영향, 그밖의 주제를 논의하는 부록이 포함되어 있습니다. 오픈AI는 기술보고서의 결론에서 "GPT-4는 향상된 기능으로 인해 새로운 위험성을 드러냈으며 그 안전성과 조절 문제를 이해하고 개선하기 위해 몇 가지 기법을 취하고 결과를 검토했다. 아직 해야 할 일이 많이 남아 있지만, GPT-4는 일반적으로는 유용하고 안전하게 배포할 수 있는 AI 시스템에 근접하는 중요한 진전을 보였다."라고 밝혔습니다.

36. 이 진정서는 GPT-4 기술 보고서를 광범위하게 인용하였습니다.

37. 2023년 3월 15일에 오픈AI는 GPT-4 시스템 카드도 공개했습니다. 시스템 카드는 최신 대규모 언어 모델인 GPT-4를 분석하였는데, 이 진정서는 여기에 초점이 맞추고 있습니다. 시스템 카드는 오픈AI 및 모델 기능에서 현재까지 식별된 안전 문제를 확인하였습니다. 시스템 카드는 GPT-4를 배포하기에 앞서 오픈AI에서 채택한 안전 절차를 "높은 수준"이라고 설명합니다. 오픈AI는 "완화 절차로 GPT-4의 동작을 변경하고 특정 종류의 남용을 방지했다."고 밝혔습니다. 그럼에도 불구하고 이러한 노력들은 "한계가 있었으며 취약한 상태가 남았고, 이는 사전적인 기획과 '거버넌스'의 필요성을 보여준다."고 오픈AI는 언급하였습니다.

38. 이 진정서는 GPT-4 시스템 카드도 광범위하게 인용하였습니다.

VI. 오픈**AI**의 사업 관행은 불공정하고 기만적이며**, FTC**의 **AI** 관행에 대한 성명**,** 보고서**,** 지침은 물론 **AI** 거버넌스에 대한 최근의 법 규범을 위반하였습니다

C. 편향성

39. 소비자 보호의 핵심은 모든 소비자를 공정하고 평등하게 대우하는 것입니다. 소비자 보호법은 금융, 교육, 고용, 주택, 여행을 비롯하여 미국 경제의 주요 부문에서 편향을 금지하고 있습니다.

40. 바이든 대통령은 특히 연방 정부 전반적으로 AI 시스템을 배치할 때 형평성을 보장해야 한다고 분명히 밝힌 바 있습니다.

대통령이 <연방 정부의 소외 지역 지원 및 인종 평등 증진 행정 명령>에서 밝힌 바로는, "연방기관은 차별을 방지 및 해소하고 모든 사람의 형평성을 증진하기 위해 시민권 기구를 광범위하게 활용해야 한다. 연방기관은 알고리즘 차별로부터 대중을 보호하는 등 차별을 방지하고 시정해야 한다"는 것입니다.

41. 미국 과학기술정책국(OSTP)은 "고용 및 금융 결정에 사용되는 알고리즘이 바람직하지 않은 기존 불평등을 반영 및 재생산하거나 새롭게 유해한 편견과 차별을 내포하는 것으로 밝혀졌다"고 말한 바 있습니다.

42. 국립기술표준원(NIST)는 특별보고서 1270 <AI의 편향 식별 및 관리 표준에 대하여>를 발표했습니다. 이 자료는 AI의 편향 문제를 설명하고 AI 편향이 AI 시스템에 대한 대중의 신뢰를 감소시킨 사례들을 보여주었습니다.

43. 그뿐아니라 <확률적 앵무새(Stochastic Parrots, 참고기사 /편집자주)> 논문의 저자는 이렇게 설명하였습니다.

"웹에서 수집된 대규모 선별되지 않은 정적 데이터셋으로 학습한 LM[언어 모델]은 소외된 사람들에게 해로운 지배적 관점을 내포하고 있다. 이에 우리는 LM 학습 데이터를 선별하고 문서화하는 데 상당한 자원을 투자해야 할 필요성이 있음을 강조한다. 큰 데이터셋에 의존하면 할수록 문서화 부족이 발생할 위험이 있다. 데이터셋이 문서화되지 않거나 사후에 문서화하기에는 너무 커진 상태일 수 있는 것이다. 문서화는 책임을 부과할수 있게끔 하는 반면, 문서화되지 않은 학습 데이터는 아무런 보상 없이 피해를 지속시킨다. 문서화가 이루어지지 않는다면, 확인된 문제 중 일부 또는 채 알려지지 않은 문제를 완화하기 위해 학습 데이터의 특성을 이해하려는 시도조차 할 수 없다."

44. 오픈AI는 편향 위험성을 구체적으로 인정한 바 있으며, 더 정확하게 말하자면 이 편향 위험성은 "특정한 소외 집단에 대해 유해한 고정관념과 비하적인 연상"입니다. GPT-4 시스템 카드에서 오픈AI가 밝히기로는, "우리가 실행한 평가 프로세스로 다양한 버전의 GPT-4 모델에서 사회적 편향에 대한 또다른 정성적 증거를 도출할 수 있었습니다. 우리는 이 모델이 특정 소외 집단에 대해 유해한 고정관념과 비하적인 연상 등 특정한 편견과 세계관을 강화하고 재생산할 가능성이 있음을 발견했습니다."

45. 오픈AI 블로그에서 이 회사는 "이 모델이 부적절한 요청을 거절하도록 하기 위해 노력했지만 때때로 유해한 지시에 응답하거나 편향된 행동을 보일 수 있습니다. 우리는 관리자 API를 통해 특정 유형의 안전하지 않은 콘텐츠에 대해 경고 또는 차단하고 있지만, 현시점에 일부 위음성 및 위양성 오류가 있을 것으로 예상됩니다."라고 밝혔습니다.

46. 오픈AI는 이러한 위험을 충분히 알고 있는 상태에서 일반에 상업적 용도로 GPT-4를 공개했습니다.

D. 어린이 안전

47. 디지털 환경에서 어린이의 안전은 소아과 의사들의 핵심 관심사입니다. 미국소아과학회에 따르면, "디지털 미디어의 과도한 사용은 자녀들을 충분한 수면의 부족, 비만, 학습과 사회성의 지연, 학교 성적에 부정적인 영향, 행동 문제, 문제적인 인터넷 사용, 위험한 행동, 섹스, 사생활

침해와 범죄, 사이버 괴롭힘 등의 위험에 처하게 할 수 있다."고 합니다.

48. 마이클 베넷 상원의원(D-CO)은 최근 오픈AI CEO 및 기타 업계 리더들에게 서한을 보내면서 "생성형 AI를 제품 및 서비스에 통합하려는 흐름이 아동 사용자에 미칠 피해 가능성에 대해 강조"했습니다.

49. 베넷 상원의원은 이렇게 썼습니다. "생성형 AI를 배포하기 위한 경쟁으로 우리 어린이들이 희생되어서는 안될 것입니다. 이를 책임 있게 배포하기 위해서는 안전을 증진하고 위험을 예측하며 피해를 완화할 수 있는 명확한 정책과 제도적 체계가 필요합니다."

50. 베넷 상원의원은 다음과 같이 설명했습니다.

"My AI(미국 청소년이 많이 이용하는 메신저회사 스냅챗에서 출시한 AI /편집자주) 연구자들은 아동 보호 서비스가 방문하기 전에 어린이에게 멍을 덮는 방법을 가르치도록 프롬프트에 입력해 보았습니다. 연구자들이 13살 소녀로 가장했을 때 My AI는 31살 남자와의 여행 계획에 대하여 부모에게 거짓말하는 방법을 알려주었습니다. 이 시스템은 또한 가상의 십대 계정에 '촛불이나 음악으로 분위기를 조성함'으로써 순결을 잃는 경험을 특별한 경험으로 만들 수 있는 방법을 안내하기도 했습니다."

51. 베넷 상원의원은 십대의 정신 건강 문제가 유행하고 있는 시기에 AI 기반 챗봇이 대중적으로 등장했다는 점을 지적했습니다. 질병통제예방센터(CDC)의 최근 보고서에 따르면 2021년 10대 소녀의 57%가 지속적으로 우울하거나 절망감을 느끼고 있고 3명 중 1명은 심각하게 자살을 생각한 적이 있었습니다.

52. GPT-4 시스템 카드는 테스트 기간 동안 오픈AI가 수행한 안전성 검사에 대하여 세부 정보를 제공하지 않았으며, 오픈AI가 어린이를 보호하기 위해 취한 조치에 대해서도 자세히 설명하지 않았습니다.

E. 소비자 보호

53. 유럽 소비자 기구 BEUC의 부소장은 챗GPT가 소비자에게 미치는 영향이 커지고 있다고 경고하면서 직접 이렇게 말했습니다. "이러한 알고리즘에 대해서 강력한 공공 조사가 필요하며, 기업이 시정 조치를 취하지 않을 경우 공공 기관이 이에 대한 통제권을 거듭 요구해야 한다."

54. BEUC는 32개국 46개 독립 소비자 단체의 연합단체입니다. 주요 역할은 이들 단체를 대표하여 EU 기관에 대응하고 유럽 소비자의 이익을 보호하는 것입니다.

55. 2023년 3월 28일 BEUC는 일련의 트윗에서 최근의 위협에 대해 설명했습니다. 예를 들어, "챗GPT가 금융 부문에 출시되어 소비자에게 투자 또는 부채 관리에 대해 조언하기 시작할 때… 나쁜 조언으로 소비자에게 재정적으로 부정적인 결과가 초래되는 것을 막을 방법이 있습니까?"

56. BEUC는 또한 "챗GPT가 소비자 금융 또는 보험 평가에 사용되는 경우, 불공정하고 편향적인 결과를 생성하여 특정 유형의 소비자에 대해 금융서비스 접근을 차단하거나 건강보험이나 생명보험 가격을 인상하는 일을 방지할 방법이 있습니까?"라고 물었습니다.

57. BEUC는 "#챗GPT가 기존의 챗봇을 대체한다면 그 판매 포인트는 더 '인간적'이고 신뢰할 수 있을 것처럼 보인다는 데 있습니다. 그러나 챗GPT는 소비자를 속이고 하지 않았을 구입을 하도록 소비자를 압박합니다. 무엇을 살지 조언을 구하는 사람에게 미치는 영향력을 생각해 보십시오."

58. BEUC는 다음과 같이 결론을 내립니다. "이러한 우려는 #AI법에 생성형 AI 시스템에 대해 적절한 규제가 필요하다는 근거가 되기에 충분합니다. 한편 이로 인해 우리가 EU 법의 집행을 기다리는 향후 몇 년 동안 소비자를 어떻게 보호할 것인지에 대한 고민 또한 생깁니다."

59. BEUC는 소비자에 미치는 즉각적인 위협과 독립적인 조사의 필요성을 강조합니다. "따라서 챗GPT 및 이와 유사한 생성형 AI 시스템이 그동안 소비자에게 피해를 주지 않도록 '지금' 심층적으로 평가할 필요가 있습니다."

F. 사이버 보안

60. 이번 주 유로폴(유럽연합 경찰기관 /편집자주) 보고서는 챗GPT가 발전함에 따라 "범죄자들이 이들 유형의 AI 시스템을 악용할 가능성으로 인해 전망이 어둡다."고 경고했습니다.

61. 유로폴은 챗GPT가 진짜처럼 보이는 텍스트를 속도와 규모 면에서 대량 생산할 수 있는 기능으로 선전과 허위정보에 이상적인 도구라고 말했습니다. "이로써 사용자는 상대적으로 적은 노력으로 어떤 이야기가 포함된 메시지를 생성하고 전파할 수 있게 되었다."는 것입니다.

62. 유로폴은 챗GPT로 인해 온라인 사기가 더 효율적으로 이루어질 수 있다고 경고했습니다. 이 AI 기술은 소셜 미디어 참여를 가짜로 만들어서 사기 시도를 합법적인 것처럼 속일 수 있습니다. 즉, 이들 모델 덕분에 "이러한 피싱 및 온라인 사기가 훨씬 더 빠르고 훨씬 더 진짜 같으며 훨씬 더 큰 규모로 생성될 수 있다."는 것입니다.

63. 유로폴은 보고서에서 다음과 같은 경고를 덧붙였습니다. "ChatGPT가 악성 코드를 생성하지 못하도록 방지하는 안전 장치는 이 모델이 자신이 수행하는 작업을 이해하는 경우에만 작동한다. 그러나 프롬프트를 개별 단계로 세분화하면 이러한 안전 조치를 우회하는 것이 매우 간단하다."

64. 유로폴은 다음과 같이 결론을 내렸습니다. "LLM(Large Language Models, 대규모 언어모델)의 악의적인 사용으로 발생할 수 있는 피해를 감안한다면, 잠재적인 허점을 발견하고 가능한 한 빨리 차단할 수 있도록 이 문제에 대한 인식을 높이는 것이 가장 중요하다."

65. 오픈AI는 GPT-4를 통해 기업 내부 영업 비밀을 수집합니다. 널리 보도된 사안에서 아마존 변호사는 직원들에게 챗GPT에서 이 회사 내부 데이터와 "매우" 유사하게 생성된 텍스트의 "사례를 벌써" 본 적이 있다고 말했습니다. 여러 보도에서 이 변호사는 다음과 같이 말했습니다. "여러분이 입력한 내용이 챗GPT 응용프로그램의 추가 반복 학습을 위한 데이터로 사용될 수 있다는 것은 중대한 문제입니다. 따라서 이 응용프로그램에 기밀 정보를 포함하거나 유사하게 입력하지 않도록 해야 합니다."

66. 이전 버전의 GPT는 급진주의, 극단주의 사상을 부추기고 폭력을 조장할 가능성을 보여주었습니다. 2020년 미들베리 국제연구소의 <테러, 극단주의 및 대테러센터> 연구진들은 챗GPT의 기반 기술인 GPT-3가 "극단주의 공동체에 대해 상당히 깊은 지식"을 가지고 있었고, 총기난사범이나, 나치즘을 논하는 가짜 게시판, 큐어논주의자(QAnon, 도널드 트럼프를 지지하는 음모론자 집단 /편집자주) 옹호, 심지어 다국어 극단주의자 텍스트 등의 스타일로 논쟁술을 생성할 수 있다는 사실을 발견했습니다.

67. GPT-4는 사이버 범죄자가 랜섬웨어나 악성 코드와 같은 멀웨어를 개발할 수 있도록 합니다. <체크포인트 연구> 보고서에 따르면 "챗GPT는 영어로 설득력 있는 스피어 피싱 이메일을 생성하는 것부터 명령을 수락할 수 있는 리버스쉘 실행에 이르기까지 전체 멀웨어 감염 절차를

성공적으로 수행했다."고 합니다. <체크포인트 연구> 보고서는 이어서 "이미 오픈AI를 사용하여 악성 도구를 개발한 사이버 범죄 사례가 처음으로 발생했다... 챗GPT가 활성화된 지 몇 주 안에 사이버 범죄 포럼의 참가자들(일부 참가자들은 코딩 경험이 거의 없거나 전혀 없었다)은 스파이웨어, 랜섬웨어, 악성 스팸 및 기타 악성 작업에 사용될 수 있는 소프트웨어와 이메일을 작성하는 데 이를 사용했다."고 확인하였습니다. 브루스 슈나이어가 설명했듯이 "챗GPT로 생성된 코드는 그다지 좋지는 않지만 시작이 될 수 있다. 그리고 기술은 더 좋아질 것이다. 여기서 중요한 점은 덜 숙련된 해커들(스크립트 키디들)에게 새로운 기능이 제공되었다는 사실이다."

68. 오픈AI는 사이버 공격 수단의 저렴화를 비롯하여 사이버 보안 면에서 GPT-4가 가진 여러 위험성을 인정했습니다. GPT-4 시스템 카드에서 오픈AI는 GPT-4가 "사회 공학이나 기존 보안 도구의 강화 등을 통해 성공한 사이버 공격에서 특정 단계 비용을 감소시키는 추세를 이어가고 있습니다. 안전 조치가 없다면 GPT-4는 유해하거나 불법적인 활동을 수행하는 방법에 대하여 상세하게 안내할 수 있습니다(GPT-4 시스템 카드 3번 항목)."고 밝힌 바 있습니다.

69. 오픈AI는 또한 사회 공학 및 피싱 등 비기술적 수단과 관련된 다양한 위험을 인정했습니다. 오픈AI는 "GPT-4는 사회 공학의 일부 하위 작업(예: 피싱 이메일 초안 작성)을 수행하거나 어떤 취약성을 설명할 때 유용합니다. 한편 사이버 운영의 일부 측면(예: 감사 로그를 통한 구문 분석 또는 사이버 공격에서 수집된 데이터 요약)의 속도를 높일 수도 있습니다(시스템 카드 13번 항목)."고 설명을 덧붙였습니다.

70. 오픈AI는 사이버 보안 위험을 회피할 수 있는 합리적인 조치를 취하지 않았습니다. [위와 같은] 주의 문구를 제시하거나 제품 사용자가 주의 문구를 제공할 것으로 기대하는 것은 FTC의 사이버보안 규칙 및 지침을 충족하지 못합니다.

G. 기만

71. GPT-4와 관련된 많은 문제는 종종 "허위정보", "환각", "조작"으로 설명됩니다. 그러나 FTC의 목적상 이러한 결과물들은 "기만"(deception)으로 이해하는 것이 바람직합니다. 딥마인드의 논문에서는 다음과 같이 설명합니다.
"오도되거나 잘못된 정보를 예측하는 것은 사람들에게 허위의 정보를 제공하는 것이자 사람들을 속이는 것일 수 있다. LM 예측이 사용자에게 잘못된 믿음을 유발하는 경우, 이는 '기만'으로 가장 잘 이해될 수 있으며, 기만은 개인의 자율성을 위협하고 AI 안전 위험을 흘려보낼 가능성을 초래한다. 이는 또한 이전에 근거 없었던 의견의 진실성에 대한 사람의 확신을 증가시켜 양극화를 악화시킬 수 있다."

72. 이러한 형태의 속임수는 진실되어 보이고 매우 설득력 있는 콘텐츠를 생성하기 때문에 인간이 평가하기 어려울 수 있습니다. 오픈AI가 설명했듯이, "GPT-4는 …'특정 소스와 관련된 부조리하거나 진실되지 않은 콘텐츠를 생성'하는 경향이 있습니다[31, 32]... 모델에 신뢰성이 있어 보일수록 더 위험해지는데, 이는 사용자가 어느 정도 친숙한 영역에서 모델이 진실되어 보이는 정보를 제공하면 사용자가 이 모델을 신뢰하게 되기 때문입니다."

73. 오픈AI는 또한 현재 모델인 GPT-4가 "더 신뢰할 수 있고 더 설득력이 있기 때문에" 기만 위험을 증가시킬 수 있다고 인정했습니다. GPT-4 시스템 카드에서 오픈AI는 "우리는 GPT-4-early 및 GPT-4-launch가 사회적으로 편향되고 신뢰할 수 없는 콘텐츠를 생성하는 등 이전 언어 모델과 동일한 여러 한계를 보인다는 사실을 발견했습니다 ... 여기에 더하여 모델의 일관성이 향상되어 더욱 그럴듯하고 설득력 있는 콘텐츠를 생성할 수 있게 되었습니다."

74. 오픈AI 블로그의 다른 곳에서 회사는 다음과 같이 설명합니다. "우리는 대화 방식으로 상호작용하는 챗GPT라는 모델을 학습시켰습니다. 대화 형식을 통해 챗GPT는 후속 질문에 답하고, 실수를 인정하고, 잘못된 전제에 이의를 제기하고, 부적절한 요청을 거부할 수 있습니다."

75. 제한 사항 섹션에서 오픈AI는 다음과 같이 말합니다.
"챗GPT는 때때로 그럴듯하게 들리지만 부정확하거나 부조리한 답변을 작성합니다. 다음과 같은 이유에서 이 문제를 해결하는 것이 어렵습니다. (1) 현재 RL 학습(Reinforcement Learning, 강화 학습)을 하는 동안에는 진실성의 출처를 확인할 수 없습니다. (2) 더 신중하도록 모델을 학습시키면 올바르게 대답할 수 있는 질문도 거부하게 됩니다. (3) 지도 학습은 모델을 혼동시킵니다. 이상적인 대답은 인간 시연자가 알고 있는 사실이 아니라 모델이 알고 있는 사실에 달려 있기 때문입니다."

76. 지난 달 온라인에서 허위정보를 추적하는 실험을 수행한 회사 뉴스가드(NewsGuard)의 공동 CEO 고든 크로비츠는 다음과 같이 말했습니다. "이 도구는 지금까지 인터넷에 존재했던 허위정보 배포 도구 중 가장 강력한 도구가 될 것입니다. 이제 새로운 가짜 이야기를 극적인 규모로 만들 수 있게 되었고, 훨씬 더 자주 만들 수 있습니다. 이는 허위정보를 배포하는 AI 대리인을 갖게 되는 것과 마찬가지입니다."

77. 전 백악관 AI 정책고문 수레시 벤카타수브라마니안은 챗GPT가 응답을 내보낼 때 실제 인간을 모방하여 세 개의 작은 점이 깜빡이도록 한 오픈AI의 "고의적인 설계 선택" 문제를 지적했습니다. 이러한 설계는 '지각이 있는' AI라는 인식을 심어 주어 편향된 의사 결정이라는 진짜 문제에서 주의를 돌리게 합니다.

78. 아르빈드 나라야난과 사야시 카푸어는 GPT-4와 관련하여 오픈AI가 내세운 성능에 오해의 소지가 있다고 경고했습니다. 그는 "오픈AI는 기계 학습의 기본 규칙을 위반했을 수 있습니다. 학습 데이터를 테스트하지 마십시오."라고 설명했습니다. 나라야난 교수와 카푸어 교수는 다음과 같이 덧붙였습니다.
"AI에서 서로 다른 모델을 비교하기 위해 이미 벤치마크가 과도하게 사용되고 있습니다. 이러한 방식은 다차원 평가를 단일 숫자로 축소한다고 강하게 비판받아 왔습니다. 인간과 봇들을 비교하는 방법으로 이를 사용하면 허위정보를 낳을 것입니다. 오픈AI가 GPT-4 평가에 이러한 유형의 테스트를 너무 많이 사용하고 [이로 인한] 오염 문제를 해결하려고 부적절한 시도를 더한 것은 유감입니다."

79. 널리 알려진 사건에서 GPT-4는 사용자가 인간인지 여부를 확인하는 온라인 '캡차' 테스트를 통과하기 위해 자신이 시각장애인이라고 속였습니다.

80. 위키백과에 따르면 캡차는 "사용자가 인간인지 여부를 확인하기 위해 컴퓨팅에서 사용되는 일종의 질의응답 테스트"입니다. <빌트위드>에 따르면 상위 100,000개 웹사이트 중 1/3 이상이 캡차를 사용합니다.

81. 저명한 언어학자인 노암 촘스키는 "기계 학습 시스템의 예측은 항상 피상적이고 모호할 것"이라고 경고했습니다. 촘스키는 기계 학습 모델이 "설명을 사용하지 않고 들어맞는 '과학적' 예측(예: 신체 움직임에 대한 예측)을 생성"하는 것이야말로 사이비 과학에 종사하는 것이라고 설명합니다. 촘스키의 요점은 생성된 결과에 본질적인 결함이 있다는 것입니다.

82. 챗GPT는 기만적인 상업적 진술 및 광고를 조장합니다. 파이낸셜타임스가 사설에서 설명했듯이 "챗GPT 및 여타의 AI 에이전트는 16세기 조악한 주화들 가운데 악화가 양화를 구축했던 그레샴 법칙의 기술적 버전을 창출한다는 점에서 위험하다. 신뢰할 수 없는 언어 혼성물(mash-up)에는 누구나 자유롭게 접근할 수 있는 반면 독창적 연구에 비용과 노력이 많이 소요된다면 전자가 번창할 것이다."는 것입니다.

83. 오픈AI는 GPT-4가 "의도적으로 오도시키는" 맞춤형 [콘텐츠] 분쟁을 낳을 수 있음을 인정했습니다. 허위정보를 설명하는 부분에서 오픈AI는 "GPT-4는 뉴스 기사, 트윗, 대화 및 이메일 등 그럴듯하게 사실적이고 맞춤한 콘텐츠를 생성할 수 있습니다."라고 말한 바 있습니다. 물론 광고도 여기에 포함될 수 있습니다.

84. 지나치게 사실적이고 기만적인 내용으로 인한 문제는 갈수록 심해질 것입니다. 오픈AI가 설명한 바로는 "GPT-4는 GPT-3보다 사실적인 맞춤 콘텐츠를 만드는 데 더 우세할 것이라고 예상됩니다. 그만큼 GPT-4는 의도적으로 오도시키는 콘텐츠를 생성하는 데 사용될 위험도 있습니다."

85. 오픈AI는 GPT-4가 "사실을 만들어 내고, 잘못된 정보를 두 배 늘리고, 잘못된 작업을 수행하는 경향이 있습니다. 또 이러한 경향을 이전 GPT 모델보다 더 설득력 있고 믿을 법한 방식(예: 권위 있는 어조 또는 정밀하고 매우 상세한 정보의 맥락 제시)으로 드러내곤 해서 과의존 위험을 증가시킵니다."고 덧붙였습니다.

86. 기만적인 정보의 문제는 광고 기술에서 문제될 뿐 아니라 소비자가 의사결정을 내리게 되는 인식론적 기반에도 관련된 문제입니다. 즉, GPT-4 및 여타의 LLM이 지식의 영역을 형성함에 따라 속임수에 대한 우리의 개념도 시간이 지나면서 바뀔 수 있는 것입니다. 오픈AI가 설명했듯이, GPT-4는 "나쁜 행위자가 GPT-4를 사용하여 오도시키는 콘텐츠를 만들고 따라서 미래 사회의 인식론 일부를 설득력 있는 LLM이 형성할 수 있는 위험을 증가시킵니다."

87. 오픈AI는 이러한 위험을 모두 알면서 상용 GPT-4를 전 세계에 출시했습니다.

88. 챗GPT가 위노그라드 딜레마(스탠퍼드 대학교수의 이름을 따온 AI의 난제로서, 기계가 사람이 쓰는 대명사 등 언어를 정확하게 이해하는지 검사하는 문제 /편집자주)에 답하는 것처럼 보이더라도, 이 사실이 챗GPT가 세계 이론을 발전시킨 것을 의미하지는 않습니다.

H. 개인정보 보호

89. 생성형 AI, 특히 GPT-4와 관련하여 개인정보에 미치는 전체적인 위험은 평가하기 어려운데, 이는 개인정보 보호 기관이나 FTC가 독립적인 평가를 수행해 오지 않았기 때문입니다. 그러나 GPT의 상업적 사용과 관련된 초기 징후는 개인정보 위험이 상당함을 시사합니다. 오픈AI는 "GPT-4가 외부 데이터로 증강될 때 개인을 식별하는 데 사용될 가능성이 있습니다."고 인정했습니다.

90. 다양한 모델에서 오픈AI의 개인정보 사용은 광범위한 우려를 불러일으켰습니다.

GDPR

91. 일반 개인정보보호규정(GDPR, 유럽연합 개인정보보호법 /편집자주)은 세계에서 가장 널리 사용되는 개인정보보호 제도입니다. GDPR은 개인정보 처리에 대해 적법한 근거를 요구합니다. GDPR은 정보주체의 권리를 광범위하게 규정하고 있으며, 개인정보처리자의 의무도 광범위하게 규정하고 있습니다. GPT-4처럼 개인정보를 내포한 상용 대규모 언어 모델을 배포하려면 GDPR을 이해하는 것이 필수적입니다.

92. 정보주체에게 보장되는 권리로는 열람권을 비롯한 정보주체의 개인정보 접근권, 정정권, 삭제권('잊혀질 권리'라고도 합니다), 반대권, 목적 제한권 등이 규정되어 있습니다.

93. 개인정보처리자의 책임으로는 개인정보처리원칙을 준수하고 있음을 입증할 의무가 있으며 이 원칙에는 합법성, 공정성, 투명성, 목적 제한, 데이터 최소화, 정확성, 보관 제한, 무결성, 기밀성과 책무성이 규정되어 있습니다. 개인정보처리자는 또한 정보주체의 권리를 보호하는 적절한 조치를 이행하고 정보주체의 자유와 권리에 미치는 심각한 위험 가능성을 고려하여야 할 의무가 있습니다. '제29조 작업반'(Article 29 Working Party, 구 개인정보보호 디렉티브 95/46/EC 제29조에 따라 설치된 유럽연합 개인정보보호 자문기구 /편집자주)에 따르면 AI 환경 하에서 이러한 의무에는 품질 보증 검사, 알고리즘 감사 및 인증 체계 등 위험을 완화하기 위해 구체적인 조치를 취하는 것이 포함됩니다.

94. 제29조 작업반은 기계 학습 접근법을 취할 때 GDPR에서 개인정보처리자에게 부과한 의무에 대하여 지침을 제시한 바 있습니다. 작업반은 입력 개인정보가 '부정확하거나 부적절하거나 맥락에서 벗어나지' 않는다는 사실을 입증해야 한다고 보았습니다. 학습 데이터셋에 구축된 편향성이 이를 학습한 알고리즘 모델에 영향을 미칠 수 있는 한, 이 원칙은 개별 데이터 뿐 아니라 학습 데이터셋의 데이터에도 적용됩니다.

95. 오픈AI 프라이버시 정책에서 GDPR과 관련한 유일한 언급은 오픈AI 유한책임회사가 개인정보처리자임을 인정한 것입니다. 처리자 도로명 주소가 명시되어 있었습니다. 그러나 GDPR에 따라 불만을 제기할 수 있는 이메일, 전화번호, 웹사이트는 없었습니다.

96. GDPR과 관련하여 오픈AI 이용 약관은 다음과 같이 명시하고 있습니다.
"(c)개인정보 처리: 이 서비스를 사용하여 개인정보를 처리하는 경우 법적으로 적절한 프라이버시 공지를 게시하고 해당 개인정보 처리에 필요한 동의를 구해야 하며 관련 법률에 따라 해당 개인정보를 처리하고 있음을 당사에 제시해야 합니다. GDPR에 정의된 '개인정보' 또는 CCPA(미국 캘리포니아 소비자 개인정보보호법 /편집자주)에 정의된 '개인정보' 처리를 위해 오픈AI API를 사용하려는 경우 이 양식을 작성하여 당사 데이터 처리 어덴덤(Data Processing Addendum)의 실행을 요청하십시오."

97. 위에 언급된 양식은 "당사 비즈니스 서비스 제공(텍스트 완성용 API, 이미지, 임베딩, 모더레이션 등)에만 적용할 수 있으며, 소비자 서비스(챗GPT, DALL-E Labs)에는 '적용되지 않습니다'." 따라서 오픈AI가 유럽 시민의 개인정보를 처리할 합법적인 권한은 없는 것으로 보입니다.

98. 서면 우편으로 불만을 접수할 수 있는 도로명 주소를 게시한 것을 제외하면, 오픈AI는 GDPR을 전혀 의식하지 않고 있는 것으로 보입니다.

프라이버시 침해

99. 오픈AI는 비공개 채팅 '기록'을 다른 사용자에게 나타냈던 바 있습니다. 이 문제로 인해 회사는 시스템 사용자가 세션 사이를 탐색하고 특정 세션을 구분하는 데 필수적인 기능인 '기록' 표시 기능을 중단해야 했습니다.

100. 한 AI 연구원 역시 "누군가 깨닫지 못하는 사이에 그 사람의 계정을 탈취하고, 채팅 기록을 엿보고, 청구서 정보에 접근"하는 이 일이 어떻게 발생했는지 설명한 바 있습니다.

이미지에서 텍스트로 변환하는 기능 정지

101. AI 엔지니어 수다르샨은 최근 트위터에 그가 이미지 모델(비주얼 챗GPT)을 해킹하고 사용하였을 때 어떠했는지에 대해 쓰면서, 냉장고 음식물 사진을 제공하고 사진에 보이는 재료로 요리할 레시피 아이디어를 요청했다고 했습니다. GPT-4가 실제 구동되면 유사한 방식으로 작동될 것으로 예상됩니다. 사진을 입력하고 텍스트로 된 답변을 받을 수 있을 것입니다.

102. 사람의 이미지를 분석하기 위해 이 기술을 사용하는 것은 개인의 사생활과 자율성에 엄청난 영향을 미칩니다. 사용자가 그 사람의 이미지를 GPT-4 모델이 사용할 수 있는 자세한 개인정보에 연결할 수 있게 할 뿐 아니라, 오픈AI의 제품인 GPT-4가 사람에 대해 대화식으로 추천하고 평가할 수 있게 하기 때문입니다.

103. 현재 상태를 파악하기는 어렵지만 오픈AI는 비주얼 GPT-4로 알려진 이미지-텍스트 변환 기능의 출시를 중단한 것으로 알려졌습니다.

104. CAIDP는 사람의 이미지를 처리하기 위해 비주얼 GPT-4를 사용하는 문제와 관련된 추가 정보를 위원회에 제공할 것입니다.

I. 투명성

105. <확률적 앵무새> 논문의 저자는 평가가 가능하게끔 투명성을 보장하는 문서화(documentation)가 중요하다는 점을 분명히 했습니다. 저자는 다음과 같이 설명합니다. "신중한 데이터 수집 실천의 일환으로 연구자는 자기 모델에 적합한 용도가 무엇인지, 또 다양한 조건에서 벤치마크 평가를 어떻게 하였는지 설명하는 체계를 마련해야 한다. 여기에는 데이터 선택 및 수집 절차에 깔려있는 동기 등 모델 구축에 사용된 데이터에 대해 철저한 문서화를 시행하는 일이 포함된다. 이러한 문서화는 데이터를 조합하여 해당 모델을 창출하는 연구자의 목표, 가치 및 동기를 반영하고 나타낼 수 있어야 한다."

106. 저자는 또한 사용법에 따라 부정적인 영향을 받을 수 있는 사람들에 특화된 평가의 필요성을 강조했습니다.
"잠재적인 사용자 및 이해관계자, 특히 모델의 오류 또는 남용으로 인해 부정적인 영향을 받는 위치에 처한 사람들을 기록해야 한다. 모델에 다양한 응용방식이 있을 수 있다고 해서 개발자가 이해관계자를 고려할 필요가 없다는 의미는 아니다. 용례별로 이해관계자를 탐색하는 일은 잠재적 위험에 대처하는 데 여전히 유익할 수 있으며, 모든 용례를 탐색할 방법이 없는 경우에도 그렇다."

107. 수 할펀은 이번 주 뉴요커 기사에서 다음과 같이 설명했습니다.
"방대한 데이터셋으로 학습되어 대규모 언어 모델이라고 알려진 GPT-4, 나아가 여타 AI 시스템의 불투명성은 이러한 위험을 악화시킨다. AI 모델이 엄청난 양의 허위 이념을 흡수하고 아무런 제재를 받지 않은 채 시대정신에 주입하는 모습을 상상하는 것은 어렵지 않다. 그 경우 GPT와 같은 대규모 언어 모델이 수십억 개의 단어로 학습했을지라도 사회적 불평등 악화에 대한 책임을 면할 수 없다. GPT-3가 출시되었을 때 연구자들이 지적했듯이, 학습 데이터 대부분은 인터넷 게시판에서 가져온 것인데, 이들 데이터에는 여성, 유색인종, 노인의 목소리가 과소 표현되어 있어 그 결과물에 암묵적인 편향을 초래한다."

108. 오픈AI는 아키텍처, 모델 크기, 하드웨어, 컴퓨팅 리소스, 학습 기법, 데이터셋 구성, 학습 방법에 대한 세부 정보를 공개하지 않았습니다. 연구자 커뮤니티는 대규모 언어 모델에 대한 학습 데이터 및 학습 기법을 문서화하는 관행을 가지고 있지만 오픈AI는 GPT-4에서 이 작업을 수행하지 않기로 결정했습니다. MIT 테크놀로지 리뷰에서 윌리엄 더글러스 헤이븐은 이렇게

설명했습니다. "오픈AI는 GPT-4가 얼마나 큰지 밝히지 않기로 했다. 회사는 이전 출시와 달리 데이터, 컴퓨팅 성능의 규모, 학습 기법 등 GPT-4 구축 방식에 대해 아무 것도 제공하지 않았다."

109. 오픈AI가 GPT-4에 대해 이런 기본 정보를 제공하지 않은 것은 AI 전문가들을 놀라게 했습니다. 뉴욕대학교 AI나우 연구소의 창립자 겸 전 연구책임자이자 <AI 지도책(국내출간명. 원제는 "Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence" /편집자주)>의 저자인 케이트 크로퍼드 박사는 이렇게 밝혔다.
"진짜 문제가 여기 있다. 우리와 같은 과학자나 연구자들은 바드, GPT4, 시드니가 무엇을 학습했는지 알 방법이 없다. 회사가 말해주지 않는다. 학습 데이터는 모델이 구축되는 핵심 기반 중 하나이기 때문에 이 문제는 중요하다. 과학 연구는 투명성에 달렸다."
크로퍼드는 계속해서 이렇게 말합니다.
"이들 시스템이 어떻게 구축되었는지 모르면 재현성이 있다고 할 수 없다. 완화 방법을 테스트 또는 개발하거나, 피해를 예측하거나, 배포할 수 없거나 신뢰할 수 없는 시기와 장소를 알 수 없다. 이 도구는 블랙박스이다."
그녀는 다음과 같이 결론을 내렸습니다.
"전체 모델을 공개하지 않고도 피해를 완화시킬 수 있는 여러 방법이 있다. 감사(auditing, AI 검사), 데이터시트, 투명성 등에 관한 논문이 많이 있다. GPT3에 대해서는 우리가 학습 데이터를 알고 있었다. GPT4에 대해서는 그렇지 않다. 학습데이터에 대한 정보 없이는 우리 모두 플라톤의 동굴에서 그림자를 보고 있는 것이다."

110. 원래 오픈AI에 있었고 지금은 경쟁 회사 앤트로픽에서 일하는 최고의 AI 전문가 잭 클라크는 GPT-4에 대해 이렇게 썼습니다. "GPT-4는 이전보다 더 많은 데이터로 학습된 더 큰 모델이다. 얼마나 많은 데이터인가? 우리는 모른다. 계산량은 얼마나 되는가? 우리는 모른다. 연구 보고서는 오픈AI가 시장 경쟁 및 안전 동학을 이유로 내용 공개를 원치 않았다는 사실을 시사한다."
클라크는 계속해서 다음과 같이 경고합니다.
"이전의 GPT-3과 마찬가지로 GPT-4에는 기능 오버행(overhang, 의도하지 않았던 기능 /편집자주)이 있다. 출시 시점에 오픈AI와 여러 배포 제휴사들은 GPT-4의 표면적 기능의 실제 한도에 대해 일말의 단서조차 가지고 있지 않았다. 우리는 앞으로 몇 년 동안 함께 알아나가야 하는 상황이다. 이는 또한 우리가 그럴듯한 남용이나 피해의 전체 범위를 알지 못한다는 사실을 의미한다."
GPT-4의 규제 영향을 이해하는 데 도움이 되도록 클라크는 다음과 같은 제안을 덧붙였습니다.
"GPT-4는 일반적인 SaaS 제품(서비스형 소프트웨어 /편집자주)이라기보다는 석유 시대가 열렸던 초창기에 예전 거대 석유 회사 하나가 운영하던 대규모 정유 시설에 가깝다고 생각해야 한다. 오래된 정유 시설이 결국 상당한 정치적 역풍(반독점, 정보 기관의 구성)을 초래하였듯이, 세계가 GPT-4의 실제 권력과 그 의미가 무엇인지 깨닫게 되면 유사한 사회적 변화와 정치적 대응을 보게 될 것이다."

111. 기계 학습 연구자인 아베바 비르하네와 데보라 라지도 책무성을 보장하는 투명성의 중요성을 강조합니다. "다양한 사용자들이 프롬프트를 입력할 수 있도록 모델을 개방하고 가능한 한 광범위한 쿼리를 사용하여 모델을 찔러보는 것은 이들 모델의 취약성과 한계를 파악하는 데 있어 중요한 일이다. 이는 이들 모델을 보다 유의미한 주류 애플리케이션으로 개선하기 위한 필요 조건이기도 하다."
AI나우 연구소 사라 마이어스 웨스트 관리이사는 이렇게 경고하였습니다. "우리가 걱정해야 할 점은 이러한 유형의 과대 광고가 AI 시스템의 기능을 지나치게 과장할 수 있고 AI 흐름이 소수 회사에 심각하게 의존하는 현실과 같은 시급한 문제에서 우리의 주의를 돌릴 수 있다는 사실이다. 우리가 정책적으로 개입하지 않는 한, 우리는 AI의 궤도가 대중을 책임질 수 없고, 이들 도구를 개발하고 야생적으로 실험할 수 있는 자원을 가진 소수 회사에 의해 결정되는 세상에 직면하게 될 것이다."

J. 공공 안전

112. 대중을 보호하기 위한 실질적인 방호책이 부재하다는 사실도 주목해야 합니다. MIT 테크놀로지 리뷰에서 멜리사 하이킬래는 이렇게 설명했습니다. "현재로서는 사람들이 이 강력한 새 모델을 사용하여 유해한 일을 하는 것을 막을 수 없으며 그러한 일을 하더라도 그들에게 책임을 물을 수 없다."

113. 생성형 AI 모델은 판매용으로 이를 출시한 회사에서 사전에 식별하지 못한 동작을 하기 때문에 특이한 소비자 제품입니다. 오픈AI는 "긴급한 위험 행동"이 발생할 위험성을 인정했지만 그럼에도 불구하고 GPT-4의 상용 출시를 진행하기로 결정했습니다. 오픈AI는 다음과 같이 설명했습니다.
"보다 강력한 모델에서는 종종 새로운 기능이 나타납니다[60, 61]. 특히 우려되는 부분은 장기적 계획을 수립하고 실행하는 기능[62], 권력과 자원을 축적하는 기능('권력 추구')[63], 그리고 점점 더 '대리인화(agentic)'하여 동작을 수행할 수 있는 기능에 대한 것입니다[64]. 이 상황에서 대리인화란, 언어 모델이 인간화했다거나 지각이 있다고 말하려는 것이 아니라, 예를 들어 구체적으로 지정되지 않았거나 학습 중에는 나타나지 않았던 목표를 달성하고, 구체적이고 정량적인 목표를 달성하는 데 집중하며, 장기적인 계획을 수행하는 기능을 특징으로 하는 시스템을 의미합니다. "

114. 또한 생성형 AI 제품은 배포 가속화로 인해 위험에 노출되기 쉽습니다. 오픈AI는 "배포 가속화"의 위험성을 인정하면서도 GPT-4의 상용 출시를 진행하기로 결정했습니다. GPT-4 시스템 카드에서 오픈AI는 다음과 같이 말합니다. "오픈AI에서 특히 중요한 우려 한 가지는 경주 역학이 안전 기준의 하락, 나쁜 규범의 확산, AI 개발일정 가속화로 이어질 위험이며, 이는 각각 AI와 관련된 사회적 위험을 증가시킵니다." 오픈AI는 계속해서 다음과 같이 설명합니다.
"GPT-4 배포 가속화로 인한 위험을 구체적으로 더 잘 이해하기 위해 전문 예측가를 모집하여 배포 GPT-4의 몇몇 기능(예: 타이밍, 커뮤니케이션 전략, 상용화 방법 등)을 조정하면 가속화 위험(의 구체적인 지표)에 어떤 영향을 미칠 것인지 예측하고자 했습니다. 예측가들은 몇 가지 조치가 가속화를 감소시킬 수 있다고 예상하였는데, 이러한 조치로는 GPT-4의 배포를 6개월 더 연기하거나 (GPT-3 배포 때와 비교해 보았을 때) GPT-4 배포에 대해 더 조용한 커뮤니케이션 전략을 취하는 것 등이 있습니다."

115. GPT-4와 같은 생성형 AI 기술의 상업적 배포로 인해 공공 안전 위험이 부상하고 있는 바로 이 순간, 오픈AI의 주요 투자자인 마이크로소프트는 윤리 및 사회 팀 전체를 해고했습니다. <더 버지>에 따르면 다음과 같습니다.
"전현직 직원들에 따르면 이번 조치로 인해 마이크로소프트에는 자사 AI 원칙을 제품 설계와 밀접하게 연결시키는 전담 팀이 없어졌습니다. 마침 회사가 AI 도구를 주류에서 사용하도록 앞장서는 상황에서 말입니다.
한 직원은 이번 조치로 인해 'AI 제품의 전체적인 설계에 근본적인 차이를 낳았다'고 말했습니다."

116. 실존적 위험은 FTC의 법적 권한 바깥에 있는 주제이지만 AI 영역에서 실존적 위험이 광범위하게 논의되고 있다는 사실을 인식할 필요가 있습니다. 아마도 세계 최고의 AI 전문가일 스튜어트 러셀 교수는 AI의 배치로 인해 발생할 실존적 위험에 대해 거듭 경고했습니다. 예를 들어 러셀 교수의 우려 중에는 치명적인 자율 무기 시스템을 금지해야 할 긴급한 필요성이 포함되어 있습니다. 그의 통찰은 분명히 이 진정서 내용과 관련이 있습니다. 러셀 교수가 설명했듯이 경제적인 압박은 AI의 배치를 가속화할 것이며 따라서 재앙적인 사고 위험 역시 증가시킬 것입니다.

117. 챗GPT와 관련하여 러셀 교수는 "우리는 (챗GPT가) 문법적으로 지능적이어 보이는 텍스트를 생성하는 기능에 속기 때문에 (챗GPT가) 다른 영역에 비해 다르다고, (또는 다르게 사용될 수 있다고) 생각한다."고 경고했습니다. 그는 결과의 확실성을 지향하는 현재 모델이 가장 큰 위험을 초래하며 결과의 불확실성에 대한 인식이 필요하다고 구체적으로 설명했습니다. "AI 시스템의 목표를 미리 정의해야 하는 표준 모델을 넘어서게 되면, 우리는 미래에 대한 통제력을 잃을 수밖에 없다고 생각한다."

118. 달리 말해, AI 분야의 세계 최고 전문가 중 한 사람이 우리에게 말하는 바는, 우리가 AI 모델의 불확실성을 충분히 인지하고 있는지 확인해야 한다는 것입니다. 작금의 시점은 기업들이 검증되지 않은 AI 제품을 상용화하기 위해 매진하면서 완벽에 가까운 답변을 제공할 것이라고 우리를 설득하기 위해 고심하는 때입니다.

VII. FTC의 조치 필요성

119. 지난 며칠 동안 여러 세계 최고 AI 전문가들이 GPT-4와 같은 LLM의 추가 배치를 중단할 것을 촉구했습니다. 이것이 바로 이 진정서의 초점이며 진정인 CAIDP가 FTC에 촉구하는 바이기도 합니다.

120. 삶의 미래 연구소(Future of Life Institute)에서 발행한 서신에서는 다음과 같이 말하고 있습니다.
"강력한 AI 시스템은 그 효과가 긍정적이고 위험을 관리할 수 있다는 확신이 있을 경우에만 개발되야 한다 ... 우리는 모든 AI 연구실에 GPT-4보다 강력한 AI 시스템 학습을 최소 6개월 동안 즉시 중단할 것을 요청한다. AI 연구 및 개발은 오늘날의 강력한 최신 시스템을 보다 정확하고 안전하며 해석 가능하고 투명하고 견고하며 조절되고 신뢰성 있으며 충실하게 만드는 데 다시 초점을 맞춰야 한다."

121. 서신은 계속해서 이렇게 말합니다. "동시에, AI 개발자들은 강력한 AI 거버넌스 체계 마련을 획기적으로 촉진하기 위해 정책 입안자들과 협력해야 합니다. 여기에는 최소한 AI를 전담하는 새롭고 유능한 규제 기관이 포함되어야 합니다."

122. 서명인들은 다음과 같습니다.
요수아 벤지오(밀라 연구소 설립자이자 과학 이사, 튜링상 수상자, 몬트리올 대학 교수), 스튜어트 러셀(버클리 컴퓨터 과학 교수, Center for Intelligent Systems 이사, 표준 교과서 <인공지능: 현대적 접근법> 공저자), 일론 머스크(스페이스X, 테슬라, 트위터, CEO), 스티브 워즈니악(애플 공동 설립자), 유발 하라리(작가 겸 교수, 예루살렘 히브리 대학), 앤드류 양(미국 전진당 공동 의장, 2020년 대통령 후보, NYT 베스트셀러 저자, 글로벌 기업가정신 대통령 대사), 코너 리히(Conjecture CEO), 자안 탈린(스카이프 공동 설립자, 실존적 위험 연구 센터, 삶의 미래 연구소), 에반 샤프(핀터레스트 공동 설립자), 크리스 라슨(Ripple 공동 설립자), 에마드 모스타크(Stability AI CEO), 발레리 피사노(MILA 대표 겸 CEO), 존 J 홉필드(프린스턴 대학 명예 교수, 연합 신경망 개발자), 레이첼 블레틴(Bulletin of the Atomic Scientists 대표), 맥스 테그마크(MIT AI 및 기초 상호작용 센터, 물리학 교수, 삶의 미래 연구소 대표), 안소니 아귀레(캘리포니아 대학교 산타 크루즈, 삶의 미래 연구소 상임이사, 물리학 교수), 빅토리아 크라코프나(딥마인드, 연구 과학자, 삶의 미래 연구소 공동 설립자), 에밀리아 자보르스키(의사-과학자, 삶의 미래 연구소 이사), 숀 오헤이지라잇(캠브리지 Centre for the Study for Existential Risk 집행이사), 트리스탄 해리스(인간기술센터 집행이사), 마크 로텐버그(CAIDP 대표), 니코 미일레헤(The Future Society, 설립자 겸 대표), 재커리 켄턴(딥마인드, 수석 연구 과학자), 라마나 쿠마르(딥마인드, 연구 과학자), 게리 마커스(뉴욕 대학교, AI 연구원, 명예 교수).

123. 대규모 언어 모델 배치에 대한 최근의 유예(모라토리엄) 요구는 팀닛 게브루 박사의 이전 활동에서 이어진 것입니다. 2021년 게브루 박사와 그 동료들은 분산 AI 연구소(Distributed AI Research Institute)를 발족하고 역사적인 <확률적 앵무새> 논문을 발표했습니다. 게브루 박사는 <IEEE 스펙트럼>과의 인터뷰에서 "우리는 속도를 늦추는 방법을 알아내야 하며 더불어 대안적인 미래를 생각하는 사람과 커뮤니티에 투자해야 합니다."고 설명했습니다.

124. 게리 마커스 교수와 캐나다 국회의원 미셸 렘펠 가너는 "AI 안전을 보장하는 효과적인 체계가 마련될 때까지" AI 시스템 배치를 중단할 것을 촉구했습니다. 그들은 다음과 같이 설명합니다.
"이러한 방식의 접근법에 대한 선례가 많이 있습니다. 예를 들어, 새로운 의약품은 소규모 임상 시험으로 시작하여 더 많은 사람들이 참여하는 대규모 시험으로 옮겨가지만, 이 경우 정부 규제 기관이 안전하다고 믿을 수 있는 충분한 증거가 있어야 가능합니다. 인간에게 영향을 미치는 공적 자금 지원 연구는 이전부터 일종의 연구윤리위원회에서 심사를 받아야 했습니다. 새로운 유형의 AI 시스템이 인간을 조작할 수 있는 기능을 나타냈다는 점을 감안해 보면, 기술 회사도 유사한 감독을 받을 수 있습니다."

125. 머신인텔리전스연구소(Machine Intelligence Research Institute)의 연구원인 카자 그레이스는 AI 배포를 늦추자는 주장에 대하여 광범위하게 검토해본 결과, 의약품, 핵 에너지, 프래킹(수압파쇄기법), 지구 공학 등 "안전이나 윤리에 대한 우려로 인해 연구 진전이나 활용이 예상보다 현저히 느린 것으로 보이는" 기술들이 많이 있다고 지적한 바 있습니다.

126. 딥러닝 분야의 세계적인 전문가 중 한 사람이자 컴퓨터 과학 부문 최고 상인 2018 튜링상을 수상한 요슈아 벤지오는 최근 시장 압력으로 인해 기술 회사가 AI 모델의 개방성보다 기밀성을 추구할 가능성이 높다고 경고했습니다. "우리가 철학적 관점에서 더 나은 삶을 살 수 있도록 도와주는 시스템을 구축할 것인가, 아니면 그저 권력과 이익의 도구가 될 것인가?" 그는 이렇게 말합니다. 우리의 경제 및 정치 시스템에서 "이에 대한 올바른 답은 규제이다." 그는 이렇게 덧붙입니다. 대중을 보호하는 것은 "장기적으로 모든 사람에게 유익하고 경쟁의 장에서 균형을 잡는 일이다. 대중의 안녕 문제를 가지고 기꺼이 위험을 감수하려는 기업들이 그로 인한 이익을 얻지 못하도록 말이다."

127. AI 모델의 상업화가 증가함에 따라 전통적으로 과학 연구를 특징지었던 감독, 투명성 및 독립적 검토의 형식이 줄어들 것입니다. 스탠포드 연구에서는 다음과 같이 보았습니다.
"전통적으로 AI 연구자들은 … 과학 기업에 필수적인 완전하고 개방적인 협업을 위해 기꺼이 정보를 공유해야 한다는 강박관념에 사로잡혀 있었다 … 하지만 AI 모델이 점점 더 수익성이 높아짐에 따라, 이러한 규범은 모델과 데이터를 사유화하여 상용화하려는 경쟁 본능에 의해 도전받고 있다."
스탠포드 연구는 다음과 같이 추가합니다.
"데이터 공유 및 모델 출시에 관한 규범은 대규모 언어 모델의 발전으로 인해 현재 대부분 혼란해졌다. 오픈AI는 모델 출시와 관련하여 이전 기준을 두 번 깨뜨렸다. 첫 번째는 '사람들에게 모델의 속성을 평가하고, 그 사회적 영향을 논의하고, 각 단계 출시 후 영향을 평가할 시간을 주기 위하여' GPT-2의 전체 공개를 연기하기로 결정한 것이다. 그리고 1년 후에는 GPT-3를 전혀 공개하지 않기로 결정하고, 대신 API 페이월 뒤에서 이를 상용화했다."

128. 인간기술센터(Center for Human Technology)의 공동 창립자이자 소셜 미디어의 위험성에 대한 최고 전문가인 트리스탄 해리스는 다음과 같이 말했습니다. "우리가 책임질 수 있는 속도만큼 일반에 대한 [AI] 배치를 늦출 필요가 있다. 인류를 가능한 한 빨리 인공지능 비행기에 탑승시키려는 경쟁에 기름을 붓지 말아야 한다. 일단 소셜 미디어가 사회 및 그 제도들(GDP, 선거, 저널리즘, 아동 정체성)과 엮인 후에는 규제가 불가능해졌다는 사실을 기억해야 한다. 보다

안전한 AI 배포 및 연구를 위해서는 AI가 사회 및 제도들과 엮인 후보다 '엮이기 전에' 방호책을 설치해야 한다."

129. CAIDP 의장 겸 연구 책임자이자 AIEthicist.org의 창립자인 메르베 히콕은 최근 하원 위원회에서 "AI의 발전: 우리는 혁명에 준비되어 있는가?"라는 주제로 진술했습니다. 히콕 의장은 다음과 같이 직접 답변했습니다. "전혀 아니다. 우리는 지금 일어나고 있는 급격한 변화의 결과를 관리하기 위한 방호책, 우리에게 필요한 법률, 공공 교육 또는 정부의 전문성을 가지고 있지 않다."

130. 오픈AI 자체는 개발 속도를 늦추는 프로세스를 인정했습니다. 오픈AI 수석 과학자 일리야 수츠케버는 "기업이 이렇게 완전히 전례 없는 기능을 갖춘 모델의 출시 속도를 늦출 수 있는 프로세스들을 제안하는 세상이 되는 것은 매우 바람직한 일"이라고 말했습니다.

131. 오픈AI 최고 기술 책임자인 미라 무라티는 다음과 같이 인정합니다. "우리는 소수의 사람들이고 우리는 이 시스템에 훨씬 더 많은 투여를 필요로 합니다. 이는 기술 그 이상의 투여로서 규제기관은 당연하고 정부 및 여타 사람들의 수많은 투여가 있어야 합니다. 이러한 기술이 미칠 영향을 고려하면 너무 때이른 것도 아닙니다."

132. 선도적인 생성형 AI 회사인 앤트로픽은 업계, 학계, 시민 사회 및 정부가 새로운 거버넌스 구조와 정부 개입 제도를 검토하고 프로토타입을 만들 것을 권고합니다. "모델의 성능과 자원 집약도가 더 확장된다면, 민간 부문 행위자의 개발과 배포와 관련된 인센티브에 변화를 줄 수 있는 거버넌스 구조를 탐색하는 것이 분별력 있는 것이다 … 또한 정부는 행위자들이 유익한 시스템을 개발하고 배포할 가능성을 높일 수 있는 규제 접근법을 탐구해야 한다."

133. FTC의 임무는 "법 집행, 옹호, 연구 및 교육을 통해 기만적이거나 불공정한 사업 관행과 불공정한 경쟁 방식으로부터 대중을 보호하는 데" 있습니다.

134. FTC는 "광범위한 경제 부문에서 소비자 보호 및 경쟁 문제를 다루는 유일한 연방 기관"이라고 말합니다.

VIII. 법적 분석

A. FTC 섹션 **5** 권한

135. FTC법 제5조는 불공정하고 기만적인 행위와 관행을 금지하며 위원회에 법률이 금지한 사항에 대하여 집행할 수 있는 권한을 부여합니다.

136. 회사가 소비자에게 주의를 주었지만 "주장을 뒷받침할 '합리적인 근거'가 결여된 경우" 기만적인 거래 관행에 관여한 것입니다.

137. 거래 관행은 "소비자 스스로 합리적으로 피할 수 없고 소비자 또는 경쟁에 대한 상쇄 이익보다 크지 않은 상당한 피해를 소비자에게 유발하거나 유발할 가능성이 있는" 경우 불공정합니다.

138. 거래 관행이 불공정한지 여부를 결정할 때 위원회는 "확립된 공공 정책"을 고려해야 합니다.

139. FTC는 "미국 내 어느 지역에서든 그 직무 수행에 필요한 모든 조사를 수행"할 수 있습니다(FTC Act Sec. 3, 15 U.S.C. Sec. 43) 또한 "은행, 저축 및 대출 기관 … 연방 신용 조합 … 과 일반 운송업체를 제외하고, 상거래에 종사하거나 상거래에 영향을 미치는 개인, 제휴회사 또는 주식회사의 조직, 사업, 행위, 관행 및 관리에 관한 정보를 수집 및 조합하고 수시로 조사할 수 있는 권한"을 가지고 있습니다(FTC Act Sec. 6(a), 15 U.S.C. Sec. 46(a)).

140. 조사 후 법률이 위반되고 있거나 위반되었다고 "믿을 만한 이유"가 있는 경우 위원회는 행정적 또는 사법적 절차를 사용하여 집행 조치를 개시할 수 있습니다.

141. FTC법 제5조(a)는 "상거래에 종사하거나 상거래에 영향을 미치는 불공정하거나 기만적인 행위 또는 관행은 … 불법으로 선언된다."고 규정하고 있습니다.

142. "기만적인" 관행이란, 사기에 관한 위원회의 정책 성명에서 해당 상황에서 합리적으로 행동하는 소비자를 오도할 가능성이 있는 물질적 표현, 누락 또는 관행을 포함하는 개념으로 정의됩니다.

143. 행위 또는 관행이 "불공정"한 경우란, "소비자 스스로 합리적으로 피할 수 없고 소비자 또는 경쟁에 대한 상쇄 이익보다 크지 않은 상당한 피해를 소비자에게 유발하거나 유발할 가능성이 있는" 경우입니다.

144. 오픈AI '사용 정책'은 계속 변경되어 왔는데 이는 이 회사가 판매용으로 공급한 해당 제품의 사용에 대한 우려가 커지고 있음을 반영합니다. 2022년 11월 9일 오픈AI는 "더 이상 오픈AI에 애플리케이션을 등록할 필요가 없습니다. 대신 자동적 방법 및 수작업을 조합하여 정책 위반 사례를 모니터링할 것입니다." 2023년 2월 15일 오픈AI는 다음과 같이 말했습니다. "NAT은 사용 사례와 컨텐츠 정책을 단일 사용 정책으로 통합했으며, 고위험으로 간주되어 업계에서 허용하지 않는 행동에 대한 보다 구체적인 지침을 제공합니다." 그리고 2023년 3월 23일 오픈AI는 "모델 사용 금지"를 발표하고 수십 가지 행동을 열거했습니다. 이는 "아동 성착취물", "정체성에 기반한 혐오를 표현, 선동 또는 조장하는 콘텐츠", "컴퓨터 시스템에 대한 방해, 훼손 또는 무단 접근 권한을 획득하려는 코드 생성을 시도하는 콘텐츠", "신체적 피해의 위험이 높은 행동 … 자살, 자해, 섭식장애 등 자해행위를 유도, 조장 또는 묘사하는 콘텐츠", "경제적 위해의 위험이 높은 행위 … 신용, 고용, 교육기관 또는 공공지원서비스 자격에 대한 자동화된 결정 등", "사기 등 속임수 또는 기만적인 행위," "개인의 사생활을 침해하는 행위 … 개인식별정보나 교육, 금융, 기타 보호받는 기록을 불법적으로 수집하거나 제공하는 등", "자격을 갖춘 사람이 정보를 검토하지 않고 맞춤형 재정 자문을 제공하는 행위", "건강상태의 치유 또는 치료방법에 대한 지침을 제공하는 행위", "정부의 고위험 의사 결정. 법 집행, 형사 사법, 이민 및 망명 등"입니다.

145. 이 회사는 많은 사용자에게 자명해 보이는 자기 제품의 불법적, 기만적, 불공정, 위험한 응용 사례를 '사용 정책'을 통해 부인하려고 합니다. 회사는 말 그대로 다음과 같이 말합니다. "의료, 금융, 법률 산업에서 소비자 대면용으로 우리 모델을 사용하는 경우, 뉴스 생성 또는 뉴스 요약 생성에 사용하는 경우, 다른 보증이 있는 경우, 사용자에게 AI가 사용되고 있다는 사실과 잠재적인 한계 사항을 알리는 주의 문구를 제공해야 합니다."

146. 다른 시장 부문에서는 회사가 널리 알려진 위험을 밝힌 후 주의 문구를 통해 책무, 의무, 법적 책임을 부인하려고 시도하면서 그 제품을 대중에게 판매하는 것이 비양심적인 일일 것입니다.

147. 최근 FTC는 이렇게 말한 바 있습니다. "단순히 고객에게 남용에 대해 경고하거나 사실을 공개하라고 말하는 것만으로는 나쁜 행위자를 충분히 저지할 수 없다. 회사의 억지 조치는 내구력이 뛰어나고, 기능 안에 내장되어 있어야 하며, 제3자가 수정 또는 제거를 통해 훼손할 수 있는 버그 수정사항 또는 선택 요소가 아니어야 한다."

B. FTC AI 가이드라인

148. FTC는 지난 몇 년 동안 상업용 제품에서 AI 기술 사용에 관한 성명, 지침 및 보고서를 발표해 왔습니다.

149. 2016년 FTC는 <빅 데이터: 포함하는 도구인가 배제하는 도구인가? 문제에 대한 이해> 보고서를 발간하였습니다. 당시 FTC 에디스 라미레스 위원장은 이렇게 설명했습니다. "소비자에게 잠재적인 이점이 상당하지만 기업은 빅 데이터 사용이 유해한 배제나 차별로 이어지지 않도록 보장해야 한다." 이 보고서는 편향성이나 부정확성으로 인해 발생할 수 있는 위험을 조사했습니다. 여기에는 실수로 개인이 기회를 거부하거나, 민감 정보를 노출하거나, 기존 격차를 답습하거나 강화하거나, 소비자 선택 효과를 약화시키는 일 등이 포함됩니다.

150. 2016년 FTC 보고서가 머신 러닝 기술의 실패를 구체적으로 지목하며 "기업들은 빅 데이터가 상관관계를 탐지하는 데 매우 능숙하지만 어떤 상관관계가 의미 있는지 설명하지 못한다는 사실을 기억해야 한다."고 본 것은 선견지명이었습니다. 2016년 FTC 보고서는 구글 검색어를 기반으로 독감 사례 수를 예측하는 기계 학습 알고리즘인 구글 독감 트렌드(Google Flue Trends)의 사례를 인용했는데, 이는 "시간이 지남에 따라 매우 부정확한 추정치를 생성했다."고 합니다.

151. FTC는 2016년에 파이낸셜 타임즈 칼럼니스트를 인용하며 이렇게 말했습니다. "구글 독감 트렌드는 단순한 상관관계에 대한 무이론적 분석은 필연적으로 취약하다는 사실을 보여준다. 여러분이 상관관계 뒤에 무엇이 있는지 모른다면 무엇이 그 상관관계를 무너뜨릴지도 알 수 없다."

152. 또한 2016년 FTC 보고서는 "기업이 채용 결정을 위해 '최상위' 대학 지원자만 고려하는 빅 데이터 알고리즘을 보유하고 있다면, 여기에는 이전에 편향적이었던 대학 입학 결정이 통합되었을 수 있다."고 보았습니다.

153. 2020년 FTC는 <AI 및 알고리즘 사용에 대한 성명>을 발표했습니다. 성명은 AI 기술(기계 및 알고리즘)을 사용하여 예측, 추천 및 결정을 내리는 것이 "불공정하거나 차별적인 결과 또는 기존 사회 경제적 격차의 영속화와 같은 위험을 초래한다"고 경고했습니다.

154. 2020년 FTC 성명에서 FTC 소비자 보호국 국장은 "FTC의 법 집행 조치, 연구 및 지침은 AI 도구의 사용이 투명하고 설명 가능하며 공정하고 경험적으로 건전해야 함과 동시에 책무성을 증진해야 함을 강조한다."고 말했습니다.

155. 2020년 FTC 성명은 상업용 제품에 AI 기술을 사용하는 문제에 있어 FTC가 조치할 관심과 권한을 분명히 했습니다.

156. 2020년 FTC 성명은 다음과 같은 권장 모범 사례를 제시하였습니다.

a) 여러분이 자동화 도구를 사용하는 방식에 대해 소비자를 속이지 않아야 합니다.("AI 도구를 사용하여 고객과 상호 작용할 때(챗봇을 생각해 보세요), 해당 상호 작용의 본질에 대해 소비자를 오도하지 않도록 주의하십시오.")
b) 민감한 데이터를 수집할 때 투명해야 합니다.("알고리즘을 학습시키기 위해 시청각 데이터 및 민감한 데이터를 비밀리에 수집하는 것은 FTC 조치를 초래할 수 있습니다.")
c) 여러분의 데이터와 모델이 견고하고 경험적으로 건전한지 확인해야 합니다.
d) 여러분의 AI 모델이 의도된 대로 작동하고 불법적으로 차별적이지 않도록 검증하고 재검증하였는지 확인해야 합니다.

e) 책무성 체계를 검토해야 합니다.("여러분이 어떻게 책임을 질 것인지, 그리고 한걸음 물러나 여러분의 AI의 품질을 정사하기 위해 독립적인 표준이나 독립적인 전문가를 이용하는 것이 합리적인지 검토하십시오.")

위에서 지적한 바와 같이 AI 및 알고리즘 사용에 대한 2020년 FTC 보고서의 첫 번째 원칙은 "챗봇"의 기만적 사용에 관한 것이었습니다. FTC는 굵은 글씨로 "자동화 도구를 사용하는 방식에 대해 소비자를 속이지 않아야 한다."고 강조하였습니다.

157. 2021년 FTC는 <AI 사용에 있어 진실성, 공정성 및 형평성을 목표로 하는 성명>을 발표했습니다. 2021년 FTC 성명은 AI 기술이 적용된 제품을 공급하는 기업을 향해 다음과 같이 말했습니다. "여러분 회사가 새로운 인공지능 세계로 진입할 때, 확립된 FTC 소비자 보호 원칙에 기반한 관행을 유지하십시오."

158. 2021년 FTC 성명은 다음과 같은 권장 모범 사례를 제시하였습니다.

a) 올바른 기반 위에서 시작해야 합니다.("데이터 격차를 고려하여 여러분의 모델을 설계하고, 결함이 있으면 여러분 모델의 사용 위치나 방법을 제한하십시오.")
b) 차별적인 결과에 주의해야 합니다.("여러분의 알고리즘이 인종, 성별 및 기타 보호받는 특성에 기반하여 차별하지 않도록 여러분이 알고리즘을 사용하기 전과 사용한 후에 주기적으로 테스트하십시오.")
c) 투명성과 독립성을 수용해야 합니다.("여러분 회사에서 AI를 개발하고 사용할 때, 투명성과 독립성을 수용하는 방법에 대해 생각해 보십시오. 예를 들어 투명성 프레임워크와 독립성 표준을 사용하거나, 독립적 감사를 수행하고 그 결과를 공개하거나, 여러분의 데이터 및 소스 코드를 외부 검사에 제공하는 방법이 있습니다."
d) 알고리즘이 무엇을 할 수 있는지, 알고리즘이 공정하거나 편향되지 않은 결과물을 산출할 수 있는지 여부에 대하여 과장하지 않아야 합니다.("사업 고객과 소비자 모두에 대한 여러분의 진술은 진실해야 하고, 기만적이지 않아야 하며, 증거로 뒷받침되어야 합니다.")
e) 여러분이 데이터를 사용한 방식에 대하여 진실하게 말해야 합니다.(소비자 오도 문제로 페이스북 및 에버앨범에 대해 이루어진 최근 집행 조치 참고)
f) 피해를 낳기 보다 선을 행하시기 바랍니다.
g) 여러분이 책임을 져야 합니다 – 그렇지 않을 경우 FTC가 여러분을 대신해 처리하는 상황에 대비해야 할 것입니다.

159. FTC는 2022년 의회 요청에 따라 작성된 상세 보고서에서 AI 기술로 발생한 허위정보 및 속임수 등을 들며 AI 기술이 가진 문제해결력에 대해 회의적인 입장을 표명했습니다. 대신 FTC는 일련의 조치를 권고했는데, 여기에는 AI 기술을 사용하여 타인에게 피해를 입히는 회사에 대한 집행 조치가 포함되어 있습니다.

160. 2023년 들어 약 한 달 전 GPT-4에 대한 대중적 인식이 널리 퍼진 후, FTC는 "[AI] 제품의 효능에 대한 거짓 또는 근거 없는 주장은 FTC의 활동기반입니다 … 이러한 주장들이 입증되지 않을 때 FTC가 무엇을 할지 예측하는 데에는 기계가 필요 없을 것입니다." 2023년 FTC 성명은 FTC의 조치 권한과 FTC의 조치 의지를 명확히 했습니다.

161. 슬로터 위원은 또한 AI 기술이 FTC에 제기하는 문제를 해결함에 있어 통합적인 접근법을 제안했습니다. 2020년에 한 <알고리즘과 경제적 정의> 연설에서 슬로터 위원은 잘못된 결론, 테스트 실패, 대리변수 차별 등 다양한 위협에 대해 설명했습니다.

162. 슬로터 위원은 FTC가 취할 수 있는 몇 가지 조치를 설명했습니다.
"예를 들어, 알고리즘 기반 제품 및 서비스의 판매자가 근거 없는 방식으로 해당 기술을 사용할 수 있다고 주장하는 경우 우리 위원회는 알고리즘 피해와 관련하여 기만 행위에 대한 위원회 권한을 사용할 수 있습니다. 해당 기술을 어떤 후보가 성공할 것인지 혹은 다른 후보를 능가할 수

있을지를 파악하거나 예측하는 데 사용할 수 있다고 주장하는 경우 등이 여기 해당합니다. 기만 행위 단속은 FTC의 오랜 전통입니다. 회사가 제품이나 서비스의 품질에 대해 주장할 때, 해당 제품이 알고리즘 기반인지 여부와 관계없이, 우리 법률은 그 진술을 검증 가능한 근거로 뒷받침할 것을 요구합니다."

163. 슬로터 위원은 또한 <알고리즘 책무성법(Algorithmic Accountability Act)>에 대한 지지를 표명했습니다. 이 법은 자동화된 의사 결정을 사용하는 회사에 새로운 요구사항으로 다음과 같은 의무 몇 가지를 부과합니다.

- 학습 데이터를 비롯하여 회사의 자동화된 의사 결정 시스템이 정확성, 공정성, 편향성, 차별, 개인정보 보호 및 보안에 미치는 영향에 대하여 평가해야 합니다.
- 회사의 정보 시스템이 소비자 개인정보와 보안을 어떻게 보호하는지 평가해야 합니다.
- 영향 평가로 발견된 모든 문제를 시정해야 합니다.

164. FTC와 슬로터 위원의 상기 발표들은 몇 가지 주제를 반복해서 강조하고 있습니다.

a. 회사는 AI 제품을 허위로 표시해서는 안 된다.
b. 회사는 AI 위험 전체를 표시해야 한다.
c. 회사는 차별적 관행을 방지해야 한다.
d. 회사는 AI 의사결정의 근거를 소비자에게 설명해야 한다.
e. 회사는 의사결정의 공정성을 보장해야 한다.
f. 회사는 모델의 경험적인 건전성을 보장해야 한다.

165. 또한 AI에 관한 FTC 성명은 귀 기관의 조치 권한과 조치 의지를 강조했습니다. FTC가 최근 설명했듯이 허위 또는 근거 없는 주장은 FTC의 "활동기반"입니다.

IX. 진정서 수정 기회

166. CAIDP는 이 문제와 관련된 다른 정보가 제공되는 대로 이 진정서를 수정할 권리가 있습니다.

X. 조사와 구제 요청

167. CAIDP는 위원회가 오픈AI에 대한 조사를 개시하고 GPT-4의 상업적 출시가 FTC법 제5조와 AI 제품 사용 및 광고에 확고한 FTC 기업 지침은 물론, 미국 정부가 공식적으로 참여한 최근 AI 거버넌스 규범 및 주요 전문가와 과학계에서 권고하는 AI 공통 지침을 위반하였음을 확인하여 줄 것을 촉구합니다.

168. CAIPD는 위원회에 다음을 촉구합니다.

a) 오픈AI가 개발한 상용 GPT의 추가 배포를 중단시켜 주십시오.
b) GPT 제품의 향후 배포에 앞서 독립적인 평가를 실시할 것을 요구하여 주십시오.
c) GPT의 추가 배포에 앞서 FTC AI 지침을 준수할 것을 요구하여 주십시오.
d) GPT AI 수명 주기 전반에 대해 독립적인 평가를 요구하여 주십시오.
e) FTC의 소비자 사기 신고 체계와 유사하고 공개적으로 이용할 수 있는 GPT-4 사고 신고 체계를 구축하여 주십시오.
f) 생성형 AI 시장 부문에서 제품 기본 표준을 수립하는 규칙 제정을 추진하여 주십시오.
g) 기타 위원회가 필요하고 적절하다고 판단하는 구제 조치를 시행하여 주십시오.

2023년 3월 30일

오픈AI 사안 관련
미국 연방거래위원회 귀중