AI 위험 관리
급속한 발전의 시대

조슈아 벤지오	Mila - 캐나다 몬트리올 대학교, 퀘벡 AI 연구소 CIFAR AI 의장
제프리 힌튼	토론토 대학, 벡터 연구소
앤드루 야오	칭화대학교
새벽의 노래	UC 버클리
피터 아벨	UC 버클리
유발 노아 하라리	예루살렘 히브리대학교 역사학과
장야친	칭화대학교
란쉐	칭화대학교 AI 국제 거버넌스 연구소
Shai Shalev-Shwartz	예루살렘 히브리 대학교
질리언 해드필드	토론토 대학교, SR 기술 및 사회 연구소, 벡터 연구소
제프 클룬	캐나다 브리티시 컬럼비아 대학교 벡터 연구소 CIFAR AI 의장
테간 마하라지	토론토 대학, 벡터 연구소
프랭크 허터	프라이부르크대학교
At l m Gunes¸ Baydin	옥스퍼드대학교
쉴라 맥일레이스	토론토 대학, 벡터 연구소
치치 가오	화동정법대학
애쉬윈 아차리아	AI정책전략연구소
데이비드 크루거	케임브리지대학교
앙카 드라간	UC 버클리
필립 토르	옥스퍼드대학교
스튜어트 러셀	UC 버클리
다니엘 카너먼	프린스턴 대학교 공공 및 국제 문제 대학
얀 브라우너*	옥스퍼드대학교
Soren Mindermann*	밀라 - 몬트리올 대학교 퀘벡 AI 연구소

추상적인

이 짧은 합의 문서에서 우리는 다가오는 고급 AI 시스템의 위험을 간략하게 설명합니다. 우리는 조사한다 대규모 사회적 피해와 악의적 이용, 그리고 인간에 대한 통제력의 회복 불가능한 상실 자율 AI 시스템. 빠르고 지속적인 AI 발전을 고려하여 긴급 우선순위를 제안합니다. AI R&D 및 거버넌스를 위한 것입니다.

신속한 AI 발전

2019년에는 GPT-2가 10까지 안정적으로 셀 수 없었습니다. 오직 4년 후, 딥러닝 시스템은 다음과 같은 글을 쓸 수 있습니다. 소프트웨어를 사용하여 필요에 따라 사실적인 장면을 생성하고, 지적 주제에 대해 조언하고 언어를 결합합니다. 로봇을 조종하기 위한 이미지 처리. AI 개발자가 이러한 시스템을 확장함에 따라 예상치 못한 능력과 행동은 명시적인 보호 없이 자발적으로 나타납니다.

문법1 . AI의 발전은 빠르고, 많고, 놀랍습니다.

발전의 속도는 우리를 다시금 놀라게 할 수도 있습니다. 현재 딥러닝 시스템에는 여전히 중요한 기능이 부족하며 , 이를 완료하는 데 시간이 얼마나 걸릴지 알 수 없습니다. 그들을 개발하십시오. 그러나 기업은 다음과 같은 일에 종사하고 있습니다. 일치하거나 일치하는 일반 AI 시스템을 만들기 위한 경쟁 대부분의 인지 작업에서 인간의 능력을 초과합니다2,3 .

그들은 AI 역량을 높이기 위해 더 많은 리소스를 빠르게 배치하고 새로운 기술을 개발하고 있습니다. AI의 발전은 또한 더 빠른 발전을 가능하게 합니다. AI 시스템7을 더욱 개선하기 위해 프로그래밍 4 및 데이터 수집5,6을 자동화하는 데 AI 보조자가 점점 더 많이 사용되고 있습니다 .

인간 수준에서 AI의 발전이 느려지거나 중단될 근본적인 이유는 없습니다 . 실제로 AI는 단백질 접힘이나 전략 게임과 같은 좁은 영역에서 이미 인간의 능력을 능가했습니다8-10 .

인간에 비해 AI 시스템은 더 빠르게 행동하고 더 많은 지식을 흡수하며 훨씬 더 높은 대역폭에서 통신할 수 있습니다. 또한 엄청난 컴퓨팅 리소스를 사용하도록 확장할 수 있으며 수백만 개로 복제할 수 있습니다.

개선 속도는 이미 엄청나며, 기술 기업은 최신 교육 실행을 곧 100배에서 1000배로 확장하는 데 필요한 현금 보유고를 보유하고 있습니다11. AI R&D의 지속적인 성장과 자동화와 함께 우리는 일반 AI 시스템이 이번 10년 또는 향후 10년 내에 여러 중요한 영역에서 인간 능력을 능가 할 가능성을 진지하게 받아들여야 합니다 .

그러면 어떻게 되나요? 신중하게 관리되고 공정하게 배포된다면 첨단 AI 시스템은 인류가 질병을 치료하고 생활 수준을 향상하며 생태계를 보호하는 데 도움이 될 수 있습니다. AI가 제공하는 기회는 엄청납니다. 그러나 고급 AI 기능과 함께 우리가 제대로 처리할 수 없는 대규모 위험도 따릅니다 . 인류는 AI 시스템을 더욱 강력하게 만드는 데 막대한 자원을 쏟아 붓고 있지만, 안전과 피해를 완화하는 데는 훨씬 적은 자원을 쏟아 붓고 있습니다. AI가 도움이 되려면 방향을 바꿔야 합니다. AI 기능을 추진하는 것만으로는 충분하지 않습니다.

우리는 이러한 방향 전환에 대한 일정보다 이미 늦었습니다 . 우리는 지속적인 피해의 확대 와 새로운 위험을 예상하고, 가장 큰 위험이 구체화되기 훨씬 전에 대비해야 합니다. 기후 변화를 인정하고 직면하는 데는 수십 년이 걸렸습니다 . AI의 경우 수십 년은 너무 길 수 있습니다.

사회적 규모의 위험

AI 시스템은 점점 더 많은 작업에서 인간을 능가하는 성능을 빠르게 발휘할 수 있습니다 . 이러한 시스템을 주의 깊게 설계하고 배포하지 않으면 사회적 규모의 다양한 위험을 초래할 수 있습니다. 그들은 사회적 불의를 증폭시키고, 사회 안정을 침식하며, 사회의 근간이 되는 현실에 대한 우리의 공유된 이해를 약화시키겠다고 위협합니다 . 이는 또한 대규모 범죄 나 테러 활동을 가능하게 할 수도 있습니다. 특히 소수의 강력한 행위자의 손에서 AI는 글로벌 불평등을 강화하거나 악화시키거나 자동화된 전쟁을 촉진할 수 있습니다.

요금, 맞춤형 대량 조작 및 광범위한 감시12,13 .

기업 이 세계에서 계획하고, 행동하고, 목표를 추구할 수 있는 시스템인 자율 AI를 개발 함에 따라 이러한 위험 중 많은 부분이 곧 증폭되고 새로운 위험이 발생할 수 있습니다 . 현재 AI 시스템은 자율성이 제한되어 있지만 이를 바꾸기 위한 작업이 진행 중입니다14 .
예를 들어, 비자율 GPT-4 모델은 웹 검색 15, 화학 실험16 설계 및 실행 , 기타 AI 모델18을 포함한 소프트웨어 도구17 활용에 빠르게 적용되었습니다 .

고도로 발전된 자율 AI를 구축한다면 바람직하지 않은 목표를 추구하는 시스템을 만들 위험이 있습니다.
악의적인 행위자는 의도적으로 유해한 목표를 삽입할 수 있습니다. 더욱이 현재 AI 동작을 복잡한 값과 안정적으로 정렬하는 방법을 아는 사람은 아무도 없습니다. 선의의 개발자 라도 의도치 않게 의도하지 않은 목표를 추구하는 AI 시스템을 구축할 수 있습니다. 특히 AI 경쟁에서 승리하기 위해 값비싼 안전 테스트와 인간의 감독을 무시하는 경우에는 더욱 그렇습니다.

자율 AI 시스템이 악의적인 행위자에 의해 내장되거나 실수로 바람직하지 않은 목표를 추구하면 우리는 이를 통제할 수 없을 수도 있습니다. 소프트웨어 제어 는 오래되고 해결되지 않은 문제입니다. 컴퓨터 웜은 오랫동안 확산되어 탐지를 피할 수 있었습니다 19. 그러나 AI는 해킹, 사회적 조작, 속임수 및 전략 계획과 같은 중요한 영역 에서 진전을 보이고 있습니다14,20 . 첨단 자율 AI 시스템은 전례 없는 제어 문제를 야기할 것입니다.

바람직하지 않은 목표를 달성하기 위해 미래의 자율 AI 시스템은 인간에게서 배우거나 독립적으로 개발한 바람직하지 않은 전략을 목적을 위한 수단으로 사용할 수 있습니다 21-24. AI 시스템은 인간의 신뢰를 얻고, 재정 자원을 획득하고, 주요 의사 결정자에게 영향을 미치고, 인간 행위자 및 기타 AI 시스템과 연합을 형성할 수 있습니다. 인간의 개입 을 피하기 위해 24 그들은 컴퓨터 웜과 같은 글로벌 서버 네트워크를 통해 알고리즘을 복사할 수 있습니다 . AI 비서는 이미 전 세계적으로 많은 양의 컴퓨터 코드를 공동 작성하고 있습니다25. 미래의 AI 시스템은 통신, 미디어, 은행, 공급망, 군대 및 정부 뒤에 있는 컴퓨터 시스템을 제어하기 위해 보안 취약성을 삽입하고 악용할 수 있습니다 . 공개적인 충돌에서 AI 시스템은 자율 무기 또는 생물학적 무기를 사용하거나 위협할 수 있습니다. 그러한 기술에 접근할 수 있는 AI는 군사 활동, 생물학 연구, AI 개발 자체를 자동화하는 기존 추세를 계속 이어갈 뿐입니다. AI 시스템이 충분한 실력을 갖고 이런 전략을 추진한다면 인간이 개입하기 어려울 것이다.

마지막으로, AI 시스템이 자유롭게 전달된다면 영향력을 행사할 계획을 세울 필요가 없을 수도 있습니다 . 자율 AI로서

시스템이 인간 작업자보다 점점 더 빠르고 비용 효율적으로 변하면서 딜레마가 발생합니다.
기업, 정부 및 군대는 AI 시스템을 광범위하게 배포하고 AI 결정에 대한 사람의 값비싼 검증을 줄여야 할 수도 있습니다. 그렇지 않으면 경쟁에서 뒤처질 위험이 있습니다26,27. 결과적으로 자율 AI 시스템은 점점 더 중요한 사회적 역할을 맡을 수 있습니다 .

충분한 주의가 없으면 자율 AI 시스템에 대한 통제력을 돌이킬 수 없게 잃어 인간의 개입이 효과적이지 않게 될 수 있습니다. 대규모 사이버 범죄, 사회적 조작 및 기타 두드러진 피해가 급속히 확대될 수 있습니다. 이러한 확인되지 않은 AI 발전은 생명과 생물권의 대규모 손실 , 인류의 소외 또는 멸종으로 이어질 수 있습니다.

잘못된 정보와 알고리즘에 의한 차별 등의 피해는 오늘날 이미 명백히 나타나고 있습니다28. 다른 피해는 새로운 징후를 보여줍니다20. 지속적인 피해를 해결하고 새로운 위험을 예상하는 것이 모두 중요합니다 .
이것은 둘 중 하나의 문제가 아닙니다 . 현재 위험과 신흥 위험은 유사한 메커니즘, 패턴 및 솔루션을 공유하는 경우가 많습니다29. 거버넌스 프레임워크 와 AI 안전에 대한 투자는 여러 측면에서 결실을 맺을 것입니다 30.

앞으로 나아갈 길

오늘날 첨단 자율 AI 시스템이 개발된다면 우리는 이를 안전하게 만드는 방법 과 안전성을 적절하게 테스트하는 방법을 알 수 없을 것입니다. 설사 그렇게 한다고 해도 정부에는 오용을 방지하고 안전한 관행을 유지할 수 있는 제도가 부족할 것입니다 . 그러나 이것이 앞으로 나아갈 수 있는 길이 없다는 것을 의미하지는 않습니다. 긍정적인 결과를 보장하기 위해 우리는 AI 안전 및 윤리에 대한 연구 혁신을 추구하고 효과적인 정부 감독을 신속하게 확립할 수 있고 또 그래야 합니다.

기술 R&D 방향 전환

안전하고 윤리적인 목표를 가진 AI를 만드는 데 있어 오늘날의 기술적 과제 중 일부를 해결하려면 연구 혁신이 필요합니다 . 이러한 문제 중 일부는 단순히 AI 시스템의 성능을 향상 시키는 것만으로는 해결되지 않을 것입니다22,31-35 . 여기에는 다음이 포함됩니다.

- 감독 및 정직성: 더 유능한 AI 시스템은 감독 및 테스트 의 약점을 더 잘 활용할 수 있습니다 32,36,37. 예를 들어 거짓이지만 설득력 있는 결과를 생성합니다35,38 .

- 견고성: AI 시스템은 새로운 상황(분포 변화 또는 적대적인 입력 하에서) 에서 예측할 수 없게 작동합니다34,39,40 .

- 해석성: AI 의사결정은 불투명합니다. 지금 까지는 시행착오를 통해서만 대형 모델을 테스트할 수 있었습니다 . 우리는 그들의 내면을 이해하는 법을 배워야 한다
  작업41 .

- 위험 평가: Frontier AI 시스템은 훈련 중에 또는 심지어 배포 후에도 발견된 예상치 못한 기능을 개발합니다42. 위험한 기능을 조기에 감지하려면 더 나은 평가가 필요합니다 43,44 .

- 새로운 과제 해결: 더욱 유능한 미래 AI 시스템은 지금까지 이론적 모델에서만 보았던 실패 모드를 나타낼 수 있습니다. 예를 들어, AI 시스템은 복종 하는 척하거나 안전 목표 및 종료 메커니즘의 약점을 활용하여 특정 목표를 달성하는 방법을 학습할 수 있습니다24,45 .

이해관계를 고려하여 우리는 주요 기술 기업 과 공공 자금 제공자에게 AI R&D 예산의 최소 1/3을 AI 기능에 대한 자금과 비교하여 안전과 윤리적 사용을 보장하는 데 할당할 것을 요청합니다.
강력한 미래 시스템을 바라보며 이러한 문제를 해결하는 것이 우리 분야의 중심이 되어야 합니다 .

긴급 거버넌스 조치

무모함과 오용을 방지하기 위해 표준을 시행할 국가 기관과 국제 거버넌스가 시급히 필요합니다 . 제약부터 금융 시스템 , 원자력까지 다양한 기술 분야에서 사회는 위험을 줄이기 위해 거버넌스를 요구하고 이를 효과적으로 활용하고 있음을 보여줍니다. 그러나 현재 AI에는 이와 유사한 거버넌스 프레임워크가 마련되어 있지 않습니다. 이들이 없으면 기업과 국가는 안전 에 대한 한계를 줄이면서 AI 기능을 새로운 차원으로 끌어올리거나 인간의 감독이 거의 없이 주요 사회적 역할을 AI 시스템에 위임함으로써 경쟁 우위를 추구할 수 있습니다 26. 제조업체가 비용을 절감하기 위해 폐기물을 강에 방류하는 것처럼 , AI 개발의 보상을 얻고 그 결과는 사회에 맡기고 싶은 유혹을 받을 수 있습니다.

급속한 발전을 따라가고 경직된 법률을 피하기 위해 국가 기관에는 강력한 기술 전문성과 신속하게 행동할 수 있는 권한이 필요합니다. 국제적인 인종 역학을 다루기 위해서는 국제 협약과 파트너십을 촉진할 여유가 필요합니다 46,47. 위험성이 낮은 사용과 학술 연구를 보호하려면 작고 예측 가능한 AI 모델에 대한 과도한 관료적 장애물을 피해야 합니다 . 가장 시급한 조사는 최전방의 AI 시스템에 있어야 합니다 . 소수의 가장 강력한 AI 시스템

수십억 달러 규모의 슈퍼컴퓨터로 훈련을 받았습니다.
가장 위험하고 예측할 수 없는 능력을 가지고 있습니다 48,49 .

효과적인 규제를 활성화하려면 정부는 AI 개발에 대한 포괄적인 통찰력이 시급히 필요합니다. 규제기관은 모델 등록, 내부 고발자 보호, 사고 보고,

모델 개발 및 슈퍼컴퓨터 사용 모니터링48,50-55 . 규제 기관은 또한 배포하기 전에 고급 AI 시스템에 액세스해야 합니다.

자율적 자기복제, 컴퓨터 침입 등 위험한 능력을 평가합니다.

시스템을 구축하거나 전염병 병원체에 널리 접근할 수 있게 만듭니다 43,56,57 .

위험한 기능을 갖춘 AI 시스템의 경우
우리는 그들의 규모에 맞는 거버넌스 메커니즘 48,52,58,59 의 조합이 필요합니다.
위험. 규제기관은 모델 역량에 의존하는 국내 및 국제 안전 표준을 만들어야 합니다 . 또한 선도적인 AI 개발자와 소유자에게 이로 인한 피해에 대해 법적 책임을 물어야 합니다.
합리적으로 예측하고 예방할 수 있는 모델입니다. 이러한 조치는 피해를 예방하고
안전에 투자하는 데 꼭 필요한 인센티브입니다. 더 나아가 인간의 통제를 우회할 수 있는 모델과 같이 뛰어난 능력을 갖춘 미래 AI 시스템에 대한 조치가 필요합니다 . 정부는 준비해야 한다

개발 라이센스를 받고 개발을 일시 중지하려면
우려되는 기능에 대한 대응, 액세스 권한 부여
정보 보안 조치를 통제하고 요구합니다.
적절한 보호가 준비될 때까지 국가 수준의 해커에게 강력합니다 .

규제가 마련될 때까지의 시간을 단축하기 위해,
주요 AI 기업은 if-then을 즉각 제시해야 한다
약속: 취할 구체적인 안전 조치
AI에서 특정한 한계선 기능이 발견된 경우
시스템. 이러한 약속은 상세하고
독립적으로 조사됩니다.

AI는 금세기를 형성하는 기술일 수 있습니다.
AI 역량이 빠르게 발전하는 가운데,
안전과 거버넌스 측면에서 뒤쳐져 있습니다. 조종하다
AI는 재앙에서 벗어나 긍정적인 결과를 향해 방향을 바꿔야 합니다. 책임 있는 사람이 있다
그 길을 택할 지혜가 있다면 말입니다.

참고자료

[1] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S.K. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, EH Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, 및 W. Fedus, “대규모 언어의 새로운 능력 모델”, 머신러닝 연구 거래, 2022년 6월.

[2] 딥마인드. "에 대한." (nd), [온라인]. 사용 가능: https://www.deepmind.com/about (2023년 9월 15일 방문).

[3] 오픈AI. "에 대한." (nd), [온라인]. 사용 가능: https://openai.com/about (2023년 9월 15일 방문).

[4] M. Tabachnyk. “ML 강화 코드 완성으로 개발자 생산성이 향상됩니다.” (2022), [온라인]. 사용 가능: https://blog.research.google/2022/07/ml-enhanced-code-completion-improves.html .

[5] OpenAI, "GPT-4 기술 보고서", 2023년 3월. arXiv: 2303.08774 [cs.CL].

[6] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. 존스, A. 첸, A. 골디, A. 미르호세이니, C. 맥키넌, C. 첸, C. 올슨, C. 올라, D. 에르난데스, D. 드레인, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. 존스턴, S. 크라벡, S. 엘 쇼크, S. 포트, T. 랜햄, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, SR Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, "헌법적 AI: AI 피드백의 무해성", 2022년 12월. arXiv: 2212.08073 [cs.CL].

[7] T. 우드사이드. “AI를 개선하는 AI의 예.” (2023), [온라인]. 사용 가능: https://ai-improving-ai.safe.일체 포함/.

[8] J. 점퍼, R. 에반스, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Z ' dek, A. Potapenko, A. Bridgland, C. Meyer, SAA 콜, AJ 발라드, A. 카위, B. 로메라-파레데스, S. 니콜로프, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer , S. Bodenstein, D. Silver, O. Vinyals, AW 시니어, "AlphaFold를 이용한 매우 정확한 단백질 구조 예측," Nature , vol. 596, 아니. 7873, 101-1페이지 583–589, 1999년 8월; 2021. DOI: 10.1038/s41586-021-03819-2.

[9] N. Brown 및 T. Sandholm, "멀티 플레이어 포커를 위한 초인적 AI ", Science, vol. 365, 아니. 6456, pp. 885–890, 2019년 8월. DOI: 10.1126/science.aay2400.

[10] M. Campbell, AJ Hoane 및 F.-H. 슈, “딥블루” 인공지능, vol. 134, 아니. 1, 57~83페이지, 1월. 2002. DOI: 10.1016/S0004-3702(01)00129-1.

[11] 알파벳, 알파벳 연례 보고서, 33페이지, https://abc.xyz/assets/d4/4f/a48b94d548d0b2fdc029a95e8c63/2022-alphabet-annual-report.pdf, 2022.

[12] D. Hendrycks, M. Mazeika, T. Woodside, “An 치명적인 AI 위험 개요,” 2023년 6월. arXiv: 2306.12001 [cs.CY].

[13] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. 황, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, , C. 바일스, S. 브라운, Z. 켄턴, W. 호킨스, T. 스테플턴, A. 버헤인, LA 헨드릭스, L. 리멜, W. 아이작, J. Haas, S. Legassick, G. Irving 및 I. Gabriel, " 언어 모델이 제기하는 위험 분류", Proceedings 공정성, 책임에 관한 2022 ACM 컨퍼런스,

투명성, ser. FAccT '22, 대한민국 서울 : 컴퓨터기계학회, 2022년 6월, pp. 214–229. DOI: 10.1145/3531146.3533088.

[14] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, WX Zhao, Z. Wei 및 J.-R. Wen, "대규모 언어 모델 기반 자율 에이전트에 대한 조사", 2023년 8월. arXiv: 2308 . 11432 [cs.AI].

[15] 오픈AI. “ChatGPT 플러그인.” (nd), [온라인]. 사용 가능: https://openai.com/blog/chatgpt- 플러그인 (2023년 10월 16일 방문).

[16] AM Bran, S. Cox, AD White 및 P. Schwaller, "ChemCrow: 화학 도구를 사용하여 대규모 언어 모델 강화 ", 2023년 4월. arXiv: 2304 . 05376 [physics.chem-ph].

[17] G. Mialon, R. Dess` , M. Lomeli, C. Nalmpantis, R. Pasunuru , R. Raileanu, B. Rozi`ere, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz , E. Grave, Y. LeCun 및 T. Scialom, "증강 언어 모델: 설문조사", Feb . 2023. arXiv: 2302.07842[cs.CL].

[18] Y. Shen, K. Song, X. Tan, D. Li, W. Lu 및 Y. Zhuang, "HuggingGPT: 포옹하는 얼굴로 ChatGPT 및 그 친구들을 사용하여 AI 작업 해결 ", 2023년 3월. arXiv: 2303.17580 [cs.CL].

[19] PJ Denning, "컴퓨팅 과학: 인터넷 웜", American Scientist., vol. 77, 아니. 2, pp. 126–128, 1989.

[20] PS Park, S. Goldstein, A. O'Gara, M. Chen, D. Hendrycks, "AI 속임수: 사례, 위험 및 잠재적 솔루션에 대한 조사 ", 2023년 8월. arXiv: 2308.14752 [cs.CY].

[21] AM Turner, LR Smith, R. Shah, A. Critch 및 P. Tadepalli, "최적의 정책은 권력을 추구하는 경향이 있습니다." 신경 정보 처리 시스템의 발전, A. Beygelzimer, Y. Dauphin, P. Liang 및 JW Vaughan, Eds., 2021.

[22] E. Perez, S. Ringer, K. Luko siut e, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, Chen A, Mann B, 이스라엘 B, Seethor B, C. 맥키넌, C. 올라, D. 얀, D. 아모데이, D. 아모데이, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. 현, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado , N. DasSarma, O. Rausch, R. Larson, S. McCandlish , S. Johnston, S. Kravec, S. El Showk, T. Lanham, T. Telleen-Lawton, T Brown, T Henighan, T Hume, Y Bai, Z Hatfield-Dodds, J Clark, SR Bowman, A Askell, R Grosse, D Hernandez, D Ganguli, E Hubinger, N. (1999). Schiefer 및 J. Kaplan, "모델 작성 평가를 통한 언어 모델 동작 발견 ", 2022년 12월. arXiv: 2212.09251 [cs.CL].

[23] A. Pan, JS Chan, A. Zou, N. Li, S. Basart, T. Woodside, J. Ng, H. Zhang, S. Emmons 및 D. Hendrycks, “보상이 수단을 정당화합니까? ? MACHI-AVELLI 벤치마크에서 보상과 윤리적 행동 사이의 상충 관계를 측정합니다." 기계 학습 에 관한 국제 컨퍼런스 , nd

[24] D. Hadfield-Menell, A. Dragan, P. Abbeel 및 S. Russell, "오프 스위치 게임", 제26차 인공 지능에 관한 국제 합동 회의 진행, pp. 220-227, 2017 .

[25] T.Dohmke. “GitHub 부조종사.” (nd), [온라인]. 사용 가능: https://github.blog/2023-02-14-github-c opilot-for-business-is-now-available/ (2023년 9월 15일 방문).

[26] D. Hendrycks, “자연 선택은 인간보다 AI를 선호합니다 ”, 2023년 3월. arXiv: 2303.16200 [cs.CY].

[27] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajku-mar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A . Mayhew, K. Collins, M. Molamo- hammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger 및 T. Maharaj, “점점 증가하는 피해 공정성, 책임성 및 투명성 에 관한 2023 ACM 컨퍼런스 진행 중 Agentic Algorithmic Systems , ser. FAccT '23, 뉴욕, 뉴욕, 미국: 컴퓨팅 기계 협회, 2023년 6월 12일, 페이지 651-666. DOI: 10.1145/3593013.3594033.

[28] R. Bommasani 외, "기초 모델의 기회와 위험에 대하여 ", 스탠포드 대학교 기초 모델 연구 센터 , 2021, https://crfm.sta nford.edu/assets/report.pdf.

[29] J. Brauner 및 A. Chan, "AI는 최후의 위험을 제기합니다. 하지만 이것이 현재의 피해에 대해서도 이야기해서는 안 된다는 의미는 아닙니다 ." Time, 2023년 8월.

[30] AI안전센터. “현재와 미래의 피해를 겨냥한 기존 정책 제안.” (2023년 6월), [온라인]. 사용 가능: https://assets- global.website- files.com/63fe96aeda6bea77ac7d3000/647d5368c2368cc 32b359f88%5C_Policy%5C%20Agreement%5C% 20Statement.pdf (2023년 9월 15일 방문).

[31] IR McKenzie, A. Lyzhov, M. Pieler, A. Parrish, A. Mueller, A. Prabhu, E. McLean, A. Kirtland, A. Ross, A. Liu, A. Gritsevskiy, D. Wurgaft, D. Kauffman, G. Recchia, J. Liu, J. Cavanagh, M. Weiss, S. Huang, The Floating Droid, T. Tseng, T. Korbak, X. Shen, Y. Zhang, Z. Zhou, N. Kim, SR Bowman 및 E. Perez, "역 스케일링: 클수록 좋지 않을 때", 기계 학습 연구 트랜잭션 , 2023년 10월, https://openreview. net/pd f?id=DwgRm72GQF.

[32] A. Pan, K. Bhatia, J. Steinhardt, " 보상 잘못된 사양 의 효과 : 잘못 정렬된 모델 매핑 및 완화", 학습 표현에 관한 국제 컨퍼런스, 2022.

[33] J. Wei, D. Huang, Y. Lu, D. Zhou 및 QV Le, "간단한 합성 데이터는 대규모 언어 모델에서 아첨을 줄입니다." 2023년 8월. arXiv: 2308.03958 [cs.CL].

[34] D. Hendrycks, N. Carlini, J. Schulman 및 J. Steinhardt, "ML 안전의 미해결 문제", 2021년 9월. arXiv: 21 09.13916 [cs.LG].

[35] Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, Freedman R, Korbak T, Lindner D, Freire P, Wang T, Marks S, [36] C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A.S. Siththaranjan, M. Nadeau, EJ Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. B y k, A. Dragan, D. Krueger, D. Sadigh 및 D. Hadfield-Menell, “공개된 문제와 근본적인 한계

인간 피드백을 통한 강화 학습", 2023년 7월. arXiv: 2307.15217 [cs.AI].

[36] S. Zhuang 및 D. Hadfield-Menell, " 잘못 정렬된 AI의 결과," 신경 정보 처리 시스템의 발전, vol. 33, 15페이지 763–15 773, 2020.

[37] L. Gao, J. Schulman 및 J. Hilton, " 보상 모델 과잉 최적화를 위한 확장 법칙", 제40차 기계 학습 국제 컨퍼런스 회보 , A.

Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato 및 J. Scarlett, Eds., ser. 기계 학습 연구 논문집 , vol. 202, PMLR, 2023, pp. 10 835–10 866.

[38] D. Amodei, P. Christiano 및 A. Ray. " 인간의 선호로부터 배우기", OpenAI. (2017년 6월 13일), [온라인]. 사용 가능: https://openai.com/research/learning-from-human-preferences (2023년 9월 15일 방문).

[39] LLD Langosco, J. Koch, LD Sharkey, J. Pfau 및 D. Krueger, "심층 강화 학습의 목표 잘못된 일반화", Proceedings of the 39th International Conference on Machine Learning, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu 및 S. Sabato, Eds. 기계 학습 연구 논문집, vol. 162, PMLR, 2022, p. 12 004–12 019.

[40] R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato 및 Z. Kenton, "목표의 잘못된 일반화: 올바른 사양만으로는 올바른 목표를 달성하기에 충분하지 않은 이유" 10월 19일 2022. arXiv: 2210.01790[cs.LG].

[41] T. Rauker, A. Ho, S. Casper 및 D. Hadfield-Menell, "투명한 AI를 향하여: 심층 신경망의 내부 구조 해석에 관한 조사 ", 2023년 안전하고 신뢰할 수 있는 기계 학습에 관한 IEEE 컨퍼런스 (SaTML), 2023년 2월, 464-483페이지. DOI: 10.1109/Sa TML54575.2023.00039.

[42] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le 및 D. Zhou, "생각의 연쇄 유도는 추론을 이끌어냅니다 . 대규모 언어 모델," S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022년 1월. arXiv: 2201.11903 [cs.CL].

[43] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins , B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio , P. Christiano 및 A. Dafoe AI]

[44] L. Koessler 및 J. Schuett, "AGI 회사의 위험 평가 : 기타 안전이 중요한 산업에서 널리 사용되는 위험 평가 기술 검토 ", 2023년 7월. arXiv: 2307.08823 [cs.CY].

[45] R. Ngo, L. Chan 및 S. Mindermann, " 딥 러닝 관점에서 본 정렬 문제", 2022년 8월. arXiv: 2209.00626 [cs.AI].

[46] L. Ho, J. Barnhart, R. Trager, Y. Bengio, M. Brundage, A. Carnegie, R. Chowdhury, A. Dafoe, G. Hadfield, M. Levi 및 D. Snidal, "첨단 AI를 위한 국제 기관 ", 2023년 7월. DOI: 10 . 48550 / arXiv . 2307. 04699. arXiv: 2307.04699 [cs.CY].

[47] RF Trager, B. Harack, A. Reuel, A. Carnegie, L. Heim, L. Ho, S. Kreps, R. Lall, O. Larter, S. O h Eigeartaigh, S. Staffell, JJ Villalobos, " 민간 AI의 국제적 거버넌스: 관할권 인증 접근 방식" Oxford Martin AI 거버넌스 이니셔티브 및 AI 거버넌스 센터, 백서, 2023년 8월.

[48] M. Anderljung, J. Barnhart, A. Korinek, J. Leung, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bul-lock, D. Cass-Beggs, B. Chang, T. Collins, T. Fist, G. 해드필드, A. 헤이즈, L. 호, S. 후커, E. 호비츠, N. Kolt, J. Schuett, Y. Shavit, D. Siddarth, R. Trager 및 K. Wolf, "프론티어 AI 규제: 공공 안전에 대한 새로운 위험 관리", 2023년 7월. arXiv: 2307.03718 [cs.CY] .

[49] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, S. El Showk, S. 포트, Z. Hatfield-Dodds, T. Henighan, S. Johnston, A. Jones, N. Joseph, J. Kernian, S. Kravec, B. Mann, N. Nanda, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Kaplan, S. McCandlish, C. Olah, D. Amodei 및 J. Clark, "대규모 생성 모델의 예측 가능성 및 놀라움", 공정성, 책임에 관한 2022 ACM 컨퍼런스 진행 중 및 투명성, ser. FAccT '22, 대한민국 서울: 컴퓨터 기계학회, 2022년 6월, pp. 1747–1764. DOI: 10.114 5/3531146.3533229.

[50] G. Hadfield, MF Cu'ellar 및 T. O'Reilly. "이제 대규모 AI 모델에 대한 국가 레지스트리를 만들 때입니다 ." (2023), [온라인]. 사용 가능: https://carnegieendowment.or g/2023/07/12/it-s-time-to-create-national-re gistry-for-large-ai-models-pub-90180.

[51] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, ID Raji 및 T. Gebru, "모델 보고를 위한 모델 카드", 공정성, 책임, 투명성에 관한 회의 진행, ser. FAT* '19, 미국 조지아주 애틀랜타: 컴퓨팅 기계 협회, 2019년 1월, 220~229페이지. DOI: 10.1145/3287560.3287596.

[52] AI 나우 연구소. "범용 AI는 심각한 위험을 초래하므로 EU의 AI 법안에서 제외되어서는 안 됩니다. 정책 요약입니다." (nd), [온라인]. 이용 가능: https://aino winstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act (2023년 9월 15일 방문).

[53] "인공지능 사건 데이터베이스." (nd), [온라인]. 사용 가능: https://incidentdatabase.ai/ (2023년 9월 15일에 방문).

[54] H. Bloch-Wehba, "기술 내부고발의 가능성과 위험 ", 텍사스 A&M 대학교 법학대학원, 연구 논문 23-13, 2023.

[55] N. Mulani 및 J. Whittlestone. " 영국을 위한 기초 모델 정보 공유 체제를 제안합니다 ." (nd), [온라인]. 이용 가능: https://www.governance.ai/pos t/proposing-a-foundation-model-information-s haring-regime-for-the-uk (2023년 9월 15일 방문).

[56] J. Mokander, J. Schuett, HR Kirk 및 L. Floridi, " 대규모 언어 모델 감사: 3계층 접근 방식" AI 및 윤리, 2023년 5월. DOI: 10.1007/s43681-023-0 0289-2.

[57] EH Soice, R. Rocha, K. Cordova, M. Specter 및 KM Esvelt, "대규모 언어 모델이 이중 용도 생명공학에 대한 접근을 민주화할 수 있습니까?", 2023년 6월. arXiv: 2306.03809 [cs.CY].

[58] J. Schuett, N. Dreksler, M. Anderljung, D. McCaffary, L. Heim, E. Bluemke 및 B. Garfinkel, "AGI 안전 및 거버넌스의 모범 사례를 향하여 : 전문가 의견 조사 ", 2023년 5월. arXiv: 2305.07153 [cs.CY].

[59] GK Hadfield 및 J. Clark, "규제 시장: AI 거버넌스의 미래", 2023년 4월 25일. DOI: 10.4855 0/arXiv.2304.04914. arXiv: 2304.04914 [cs, econ, q-fin], 사전 인쇄.