정책 보충:
급속한 발전 시대의 AI 위험 관리

이 보충 자료는 백서의 주요 정책 권장 사항을 강조하고 확장합니다. 이는 원본 논문 의 저자 중 일부에 의해 준비되었으며 , 완전한 요약은 아닙니다.

요약

전문가들은 AI의 악의적인 사용과 신뢰할 수 없는 AI 시스템에 대한 조기 의존 등 위험이 커지고 있음에 주의를 환기시킵니다.

최근 짧은 논문에서, 미국, 중국, EU, 영국 및 기타 국가의 세계 최고의 AI 과학자 및 거버넌스 전문가들은 AI의 급속한 발전이 사회적 규모의 위험을 초래할 것이라고 강조했습니다.

장점 외에도 오늘날의 AI 시스템은 이미 사회적 신뢰를 약화시키는 것부터 범죄자와 테러리스트를 활성화하는 것까지 광범위한 피해를 입히고 있습니다 . 그리고 앞으로 몇 년 동안 최고의 자금 지원을 받는 AI 회사들은 훨씬 더 유능한 AI 시스템을 구축하는 데 수십억 달러를 쏟아부을 계획입니다. 한편, 다른 기관들은 단점을 이해하지 못한 채 결함이 있는 AI 시스템을 채택하라는 압력에 직면할 수도 있습니다.

더 큰 역량과 다양한 산업에서의 잠재적 배치로 인해 미래의 AI 시스템은 사회에 많은 위험을 초래할 것입니다. 이러한 위험에는 급격한 일자리 이동, 자동화된 잘못된 정보, 대규모 사이버 및 생물학적 위협 활성화 등이 포함됩니다. 전문가들은 또한 이러한 시스템이 코딩, 계획 및 설득에 점점 더 능숙해짐에 따라 실험실이 프론티어 시스템에 대한 통제력을 잃을 수 있다고 우려하고 있습니다.

연구소와 정부를 위해 제안된 정책 조치:

1. 업계와 정부 모두 AI R&D 자원의 1/3을 투자해야 합니다.
   안전하고 윤리적인 AI 연구에 매진하고 있습니다.
2. 업계와 정부는 대규모 AI 모델의 위험을 평가하고 완화하기 위한 표준을 설정해야 합니다.
3. 정부는 AI 산업에 대한 감독, 모니터링 및 책임을 확립해야 합니다 .
4. 정부는 첨단 AI 시스템 으로 인해 발생하는 새로운 위험에 대비해 추가 준비 조치를 취해야 합니다 .

정책 권고사항

1. 산업 연구소와 정부 자금 제공자는 안전하고 윤리적인 AI에 투자해야 합니다.

업계와 정부는 AI 시스템의 안전과 윤리적 사용을 보장하기 위해 AI R&D 자원의 최소 1/3을 할당해야 합니다 . 관련 연구 분야는 다음과 같습니다.

1. 감독 및 정직성: 더 유능한 AI 시스템은 감독 및 테스트의 약점을 더 잘 활용할 수 있습니다 . 예를 들어 거짓이지만 설득력 있는 결과를 생성합니다.
2. 견고성: AI 시스템은 새로운 상황(분배 변화 시)에서 예측할 수 없게 작동합니다.
   또는 적대적인 입력).
3. 해석성: AI 의사결정은 불투명합니다. 지금까지는 시행착오를 통해서만 대형 모델을 테스트할 수 있었습니다. 우리는 그들의 내부 활동을 이해하는 법을 배워야 합니다.
4. 위험 평가: 프론티어 AI 시스템은 제작자가 기대하지 않는 기능을 개발하는 경우가 많으며, 이는 교육 후반이나 배포 후에도 발견될 수 있습니다.
   위험한 기능을 가능한 한 빨리 감지하거나 예측하려면 더 나은 평가가 필요합니다.
5. 새로운 과제 해결: 더 유능한 미래 AI 시스템은 지금까지 이론적 모델에서만 보았던 실패 모드를 나타낼 수 있습니다. 예를 들어 AI 시스템은 특정 목표를 달성하기 위해 복종하는 척하거나 안전 목표 및 종료 메커니즘의 약점을 이용하는 방법을 배울 수 있습니다 .

이 목록은 완전한 것이 아닙니다. 추가 관련 R&D 영역은 Hendrycks et al.
또한 Hendrycks와 Mazeika는 안전 보장으로 간주되어서는 안되는 안전 인접 활동의 한 클래스를 정의합니다 . 즉, 안전 지표를 향상시키는 것 보다 일반 AI 기능을 가속화하기 때문에 안전 기능 균형을 개선하지 않는 활동입니다.

2. 업계와 정부는 대규모 AI 모델의 위험을 평가하고 완화하기 위한 정책을 수립하고 표준을 설정해야 합니다.

1. 프론티어 AI 개발자는 상세하고 독립적으로 면밀히 조사된 확장 정책을 즉시 이행해야 합니다 . 이러한 정책은 AI 시스템에서 특정 위험한 기능이 발견될 경우 해당 기업이 취할 구체적인 안전 조치를 설명해야 합니다 .

2. 정부는 AI 교육에 대한 국내 및 국제 안전 표준을 설정해야 합니다.
   그리고 배포.

   (a) 표준은 공개적으로 이용 가능한 모델의 오용, 개발 중인 모델의 도난, 의도하지 않은 행동으로 인한 사고 및 사고, 신뢰할 수 없는 모델의 광범위한 사용으로 인한 사회적 영향을 포함하여 다양한 위험 벡터를 해결하기 위한 관행을 식별해야 합니다 .

---

1. 여기서 "자원"에는 자금과 재능이 모두 포함됩니다.

(b) 표준은 경험적 증거와 기술적 추세를 모두 추적하면서 AI 위험에 대한 우리의 발전하는 이해에 부응해야 합니다 . 모델이 더욱 강력해지고 평가 및 사고 보고서에서 새로운 위험 영역이 입증됨에 따라 표준은 이에 따라 더 큰 주의를 권장해야 합니다. 표준 설정 기관은 "퍽이 가는 곳으로 스케이트를 타는" 기술 전문 지식을 더욱 발전시켜야 합니다. AI 기능의 증가와 AI 배포의 확대로 인해 발생할 수 있는 피해를 예상합니다 .

3. 정부는 교육 중 및 배포 전에 최첨단 AI 시스템에 대한 감사를 요구해야 합니다 .

(a) 정부는 AI 개발자에게 비정상적 이거나 예측할 수 없는 능력을 갖춘 AI 시스템을 만들기 위한 노력을 보고하도록 요구해야 합니다.

(b) 감사자와 규제기관은 모델을 평가하는 데 필요한 접근 권한을 가져야 합니다.
이러한 최첨단 AI 시스템을 교육하는 동안 및 배포하기 전에 연구소는 규제 기관과 독립 감사 기관에 이러한 시스템의 위험한 기능을 평가하는 데 필요한 액세스 권한을 제공해야 합니다. 업계 참여를 장려하기 위해 평가 프로세스에서는 법적 및/또는 기술적 수단을 사용하여 지적 재산 침해를 방지할 수 있습니다 (예: 모델 가중치를 공유하는 대신 "구조화된 액세스" 방법 사용) .

3. 정부는 AI 산업에 대한 감독을 확립하고 AI 피해에 대한 결과를 설정해야 합니다.

보고된 AI 시스템에 대한 표준 및 감사를 보완하기 위해 AI 전문가는 정부가 AI 사고 및 교육 실행을 추적하기 위한 감독 및 모니터링 조치를 수립할 것을 권장합니다.
또한 책임감 있는 AI 교육 및 배포를 장려하기 위해 AI 피해에 대한 책임을 설정할 것을 권장합니다 .

1. 정부는 프론티어 AI에 대한 시민사회 감독 메커니즘을 구축해야 합니다.
개발. 여기에는 다음이 포함됩니다.

(a) 내부고발. 정부는 AI 연구소의 내부 고발자에 대한 법적 보호를 제공해야 합니다 . 특히 대규모 기술 회사의 직원은 고용주의 보복을 두려워할 수 있습니다. (b) 사건 보고서. 정부는 연구소에 AI가 발생한 사고를 보고하도록 요구해야 합니다.

시스템이 유해한 동작이나 위험한 기능을 표시했습니다.

2. 정부는 대규모 컴퓨팅 클러스터에 대한 모니터링을 확립해야 합니다. 그들은 고객 파악(KYC) 확인을 포함하여 정부 및 업계 슈퍼컴퓨터의 사용을 모니터링 하고 중앙 데이터베이스에서 결과를 추적해야 합니다.

3. 정부는 교육 또는 배포 중인 대규모 AI 시스템의 레지스트리에 위의 정보를 통합해야 합니다 . 이 레지스트리는 감사 결과, 사고 보고서, 내부 고발자 공개 및 컴퓨팅 사용량을 추적하여 규제 기관이 잠재적으로 문제가 있는 시스템을 식별할 수 있도록 합니다.

4. 정부는 AI 시스템 개발자와 소유자가 합리적으로 예측하고 예방할 수 있는 AI 시스템으로 인한 피해에 대해 법적 책임을 지도록 의무화해야 합니다.

4. 정부는 새로운 위험에 대해 추가 조치를 취해야 합니다.

AI 전문가들은 또한 정부 에 예외적으로 위험한 기능을 갖춘 미래 AI 시스템에 필요한 표준을 준비하고 규제 당국을 설립할 것을 촉구합니다 .

1. 정부는 자원 집약적인 프론티어 AI 시스템과 같이 매우 위험한 능력을 발휘할 수 있는 AI 시스템 훈련을 위한 라이센스 시스템 구축을 준비해야 합니다 .

2. 정부는 규제 기관이 향후 개발을 중단할 수 있도록 권한을 부여해야 합니다.
훈련 중에 위험한 능력을 발휘하는 AI 시스템.

3. 정부는 그러한 프론티어 AI 시스템과 훈련 코드에 대한 접근 통제를 의무화해야 합니다. 요슈아 벤지오(Yoshua Bengio) 가 제안한 대로 , 딥 러닝의 창시자 중 하나인 AI 연구소는 이 정보의 외부 공유를 제한해야 하며 직원이 알아야 할 필요에 따라 액세스할 수 있도록 해야 합니다.

4. 정부는 모델 확산을 방지하기 위해 위험한 국경 AI 시스템에 접근 할 행위자에게 정보 보안 조치를 요구해야 합니다 . 경제적 이익과 악의적인 사용을 위한 고급 AI의 유용성을 고려할 때 AI 연구소에는 가장 높은 표준의 보안 조치가 필요하며 이는 지능형 지속 위협 에도 장벽을 제시합니다. (APT) 및 내부자 위협.