법무법인 지향 | 정보인권연구소 | 진보네트워크센터

범용AI의 위험성을 어떻게 통제할 것인가

: 첨단 AI의 안전성에 관한 국제 과학 보고서

8월 13일(화) 오후2시
온라인 웨비나

사회

김묘희 (법무법인 지향 변호사)

발제

희우 (진보네트워크센터 활동가)
범용 AI는 무엇을 할 수 있는가

이은우 (법무법인 지향 변호사)
범용 AI 시스템을 어떻게 평가할 것인가

장여경 (정보인권연구소 상임이사)
범용 AI가 야기하는 위험의 종류

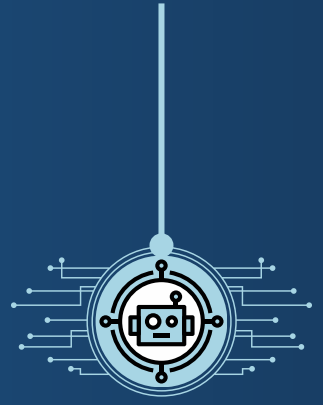
오병일 (진보네트워크센터 대표)
범용 AI 위험 완화를 위한 기술

범용 AI는 무엇을 할 수 있는가

: 첨단 AI 안정성에 관한 국제 과학 보고서
제 1, 2장

희우(진보넷)

법무법인 지향 | 정보인권연구소 | 진보네트워크센터



목차

국제 과학 보고서

배경

2023년 11월 제 1차 국제 AI 안전 서밋 개최
'첨단 AI 안전에 관한
국제 과학 보고서 개발 지원' 합의
그 첫 번째 중간보고서

목적

첨단 AI 안전에 대한 과학적 이해를
국제적으로 공유하는 데 기여하는 것

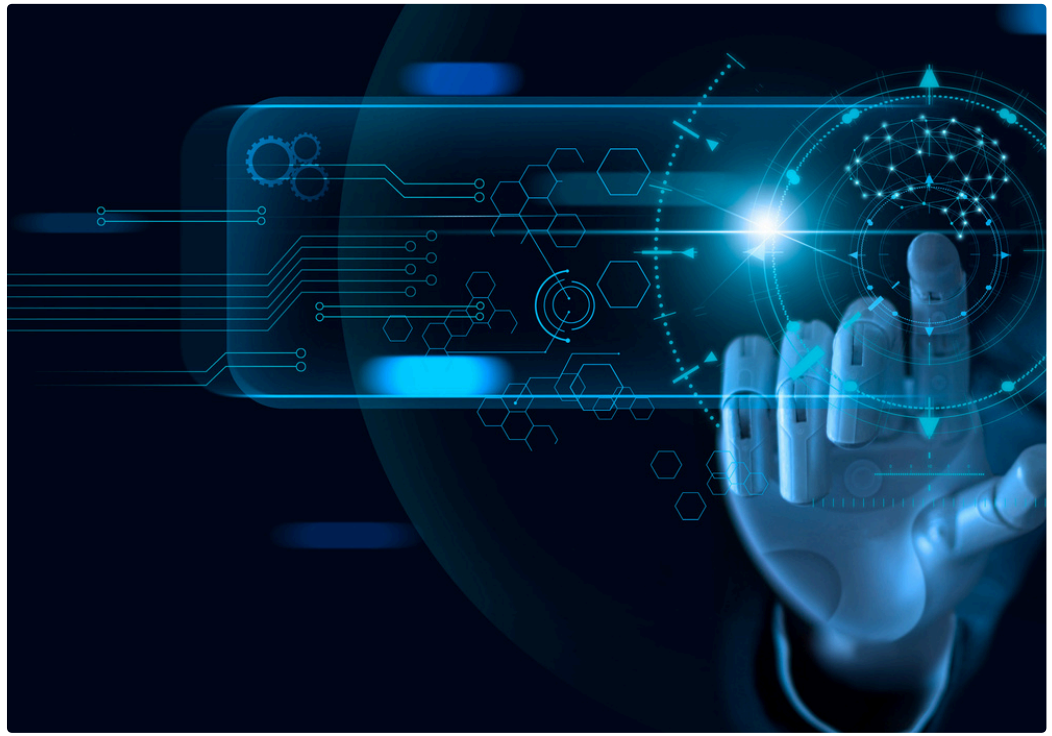
내용

'범용 AI', 즉 다양한 작업을 수행할 수 있는
AI에 초점
확신 대신 과학적 이해와 합의가 어디까지인
지, 어떤 합의가 부족한 지에 대한 의견 제시

술자

다양한 배경을 가진 75명의 AI 전문가

무엇을 다루는가?



Chat GPT와 같이 다양한 작업을 수행할 수 있는 **범용 AI**

범용 AI는 AI 시스템과 AI 모델로 구분할 수 있음

Chat GPT= AI 시스템 / GPT-4 = AI 모델

단, 범용 AI는 인공일반지능(AGI)보다 약한 개념

보고서는 발전 속도가 가장 빠르고 관련 위험에 대한 연구와 이해가 덜 이루어진 고급 범용 AI에 초점을 맞추고 있음

그러나, 보고서에 다루지 않고 있다고 하더라도 '좁은 의미의 AI' 역시 위험과 안전 관점에서 매우 중요하다고 언급
*좁은 의미의 AI : 군사분야, 예를 들면 LAWS

2.1 범용 AI는 어떻게 기능을 확보하는가?

범용 AI 모델의 예시

GPT-4, Gemini-1.5, Claude-3, Qwen1.5, Llama-3, Mistral Large 등의 챗봇 스타일 언어 모델
DALLE-3, Midjourney-5, Stable Diffusion-3 등의 이미지 생성기
SORA 등의 비디오 생성기
PaLM-E 등의 로봇 공학 및 내비게이션 시스템
AlphaFold 3 등의 분자 생물학의 다양한 구조 예측자

범용 AI 모델은 딥 러닝, 즉 여러 층의 상호 연결된 노드로 구성된 AI 모델인 인공 신경망의 학습에 의존
대부분의 최첨단 범용 AI 모델은 '트랜스포머' 신경망 아키텍쳐를 기반으로 함

*트랜스포머 아키텍쳐 : 문장 속 단어와 같은 순차 데이터 내의 관계를 추적해 맥락과 의미를 학습하는 신경망

2.1 범용 AI는 어떻게 기능을 확보하는가?

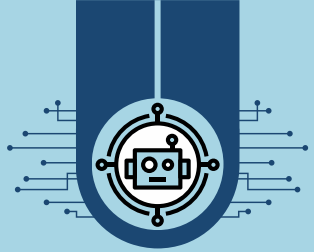
범용 AI 모델이 배포되는 과정

1. 사전학습	대량의 데이터에 있는 패턴을 통해 배경지식 구축. 복잡한 병렬 계산을 빠르게 처리하도록 설계된 특수 컴퓨터 칩인 수천 개의 GPU(그래픽 처리 장치)를 사용하며 몇 주 또는 몇 달 소요 오늘날 이 프로세스는 2010년에 비해 약 100억 배 더 많은 컴퓨팅을 사용
2. 미세조정	하나 이상의 추가 미세조정을 거쳐 의도한 작업을 수행할 수 있도록 능력 개선 미세 조정에는 일반적으로 사람의 상당한 개입이 필요하며, 최신 모델을 미세 조정하려면 수백만 건의 피드백이 필요하기 때문에 수천 명의 계약직 지식 근로자가 제공하는 경우가 다수

2.1 범용 AI는 어떻게 기능을 확보하는가?

범용 AI 모델이 배포되는 과정

3. 시스템 통합	기능과 안전성을 향상시키기 위해 다른 시스템 구성 요소와 통합 사용자 인터페이스, 입력 전처리기, 출력 후처리기 및 콘텐츠 필터와 통합
4. 배포	개발자들만 시스템을 사용하는 '내부 배포' 다른 사람들도 사용할 수 있도록 허용하는 '외부 배포(클로즈드 또는 오픈 소스)' GPT-4와 같은 일부 최첨단 범용 AI 시스템은 클로즈드 소스인 반면, 라마-3과 같은 시스템은 오픈 소스
5. 모니터링 및 업데이트	배포 후에도 개발자는 지속적으로 기능을 업데이트하고 결함 및 취약점을 해결할 수 있음



2.2 범용 AI 시스템의 현재 능력

현재 범용 AI 모델이 할 수 있는 것

- 프로그래머 보조 및 짧은 컴퓨터 프로그램 작성
- 여러 차례에 걸쳐 유창한 대화 참여
- 교과서 수학 및 과학 문제 해결

현재 범용 AI 모델이 할 수 없는 것

- 집안일과 같은 유용한 로봇 작업 수행
- 거짓 진술을 안정적으로 피하기
- 완전히 새로운 복잡한 아이디어 개발

단, 범용 AI 시스템은 상황에 따라 성능이 크게 달라진다는 특징이 있기 때문에 과소평가되거나 과대평가될 수 있음

2.2 범용 AI 시스템의 현재 능력

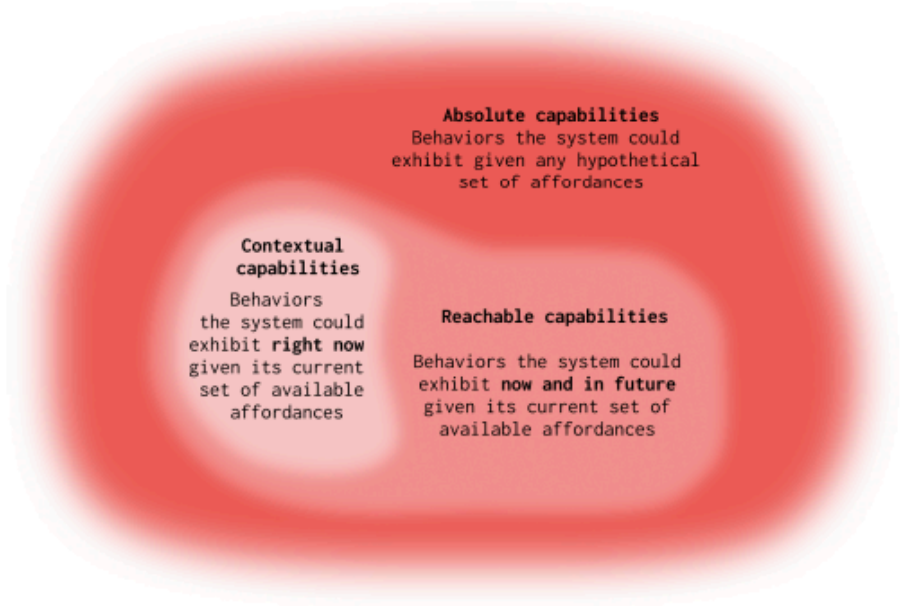


Figure 2. The relationship between the sets of potential behaviors defined by absolute capabilities, reachable capabilities, and contextual capabilities.

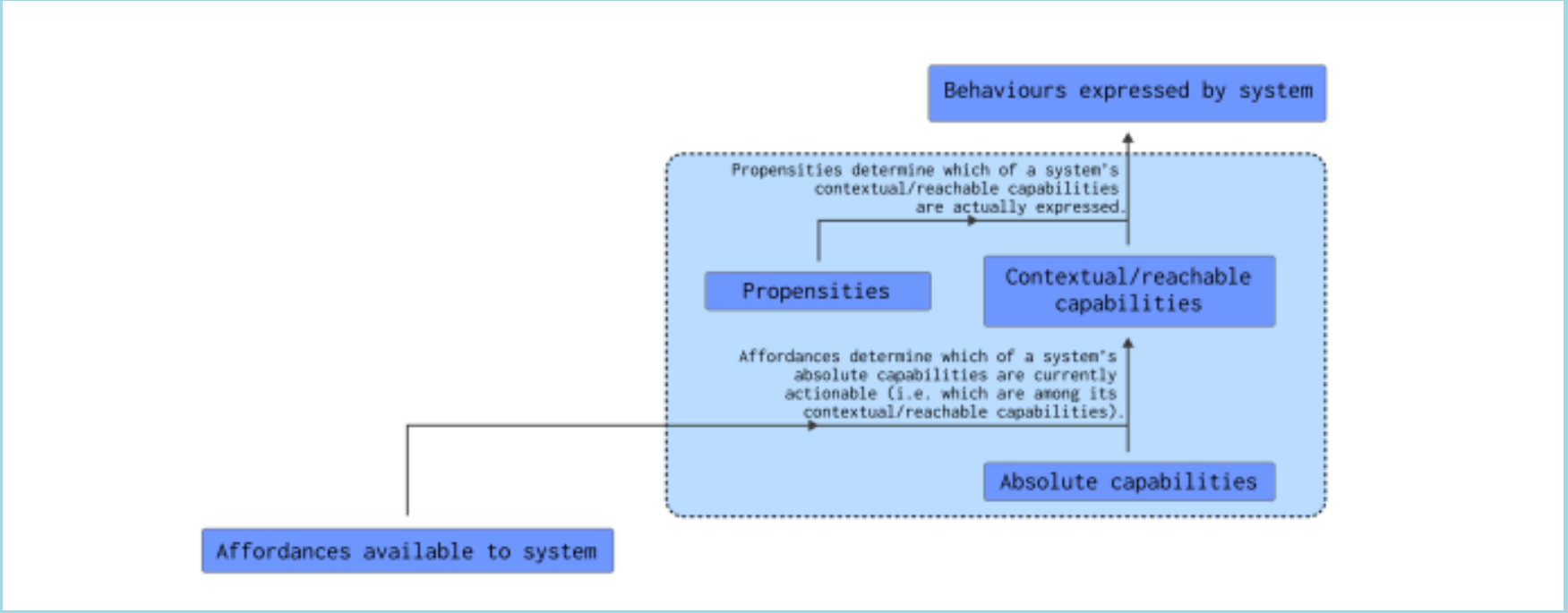
범용 AI 역량 정의의 어려움

AI 연구자들은 AI 시스템의 행동, 즉 시스템이 실제로 생성하는 일련의 출력 또는 동작과 그러한 행동을 하는 맥락(예: 프롬프트)만을 관찰

그러나 모델이 배포된 후 모델에 '단계별 사고'를 유도하는 등의 새로운 방법을 통해 기능을 이끌어내는 방법을 종종 발견

또 다른 복잡한 점은 해당 시스템의 환경, 즉 액세스할 수 있는 도구와 리소스에 따라 기능이 형성된다는 것

2.2 범용 AI 시스템의 현재 능력



2.2.1 양식별 기능

입력 처리-출력 생성 양식에 따른 분류

텍스트

유창한 텍스트를 생성할 수 있으며 다양한 자연어, 주제 및 형식에 대한 멀티턴 대화에 사용할 수 있음
수학 공식이나 소프트웨어 코드와 같이 텍스트로 인코딩된 다양한 유형의 데이터가 포함될 수 있음

이미지

이미지를 분류, 설명, 인코딩 또는 구별하는 데 사용할 수 있음
이미지를 출력으로 생성해낼 수 있으며 더 복잡한 개념과 이미지의 렌더링이 개선되고 있음

비디오

기존 비디오를 입력으로 사용하거나 텍스트에서 비디오를 생성할 수 있음
동영상에서 시간 경과에 따라 추적할 수 있는 객체 속성을 인코딩하는 방식을 학습하기도 함

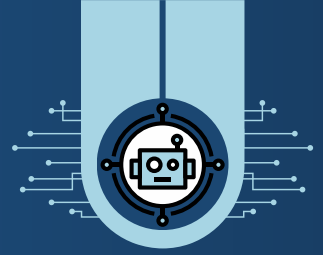
단백질 및 분자

단백질 구조를 예측하고 유용한 단백질을 새로 생성하며 다양한 작업을 수행할 수 있음
예측 가능한 기능을 가진 단백질 설계를 생성하도록 점점 개선되고 있음

2.2.2 기술별 능력과 한계

범용 AI를 능력-한계 관계로 살펴보기

지식(능력)	비일관성(한계)
공개 인터넷에서 발견되는 광범위한 사실 인코딩	사실 차이를 식별하는 데 한계가 있고 비일관적
창의성(능력)	**환각(한계)**
새로운 예시를 생성할 수 있음	콘텐츠를 조작하는 환각으로 이어질 수도 있음



2.2.2 기술별 능력과 한계

범용 AI를 능력-한계 관계로 살펴보기

상식적 추론(능력)	인과관계(한계)
광범위한 상식 지식을 모방하고 비교적 복잡한 문제를 단계별로 해결하는 능력을 보이기도 함	'추론'하는 것처럼 보이더라도 근본적인 인과적 근거를 파악하지 못했을 수도 있음
형식적 추론(능력)	**구성성(한계)**
수학, 프로그래밍, 자연과학과 같은 영역에서 일부 형식적 추론 작업을 수행할 수 있음	형식적 추론을 뒷받침하는 임의의 구성 추론을 수행하기 어려움

2.2.2 기술별 능력과 한계

범용 AI를 능력-한계 관계로 살펴보기

예측(기능)	새로운 개념(한계)
제한된 영역에서 합리적인 예측으로 미래의 사건 예측 가능	완전히 새로운 개념을 종합하기는 어려움
시뮬레이션(기능)	**구현(한계)**
가상 에이전트 동작을 시뮬레이션 가능	아직 물리적 로봇이나 기계를 효과적으로 제어할 수 없음

2.3 기능의 최근 동향과 그 동인

범용 AI 역량의 발전

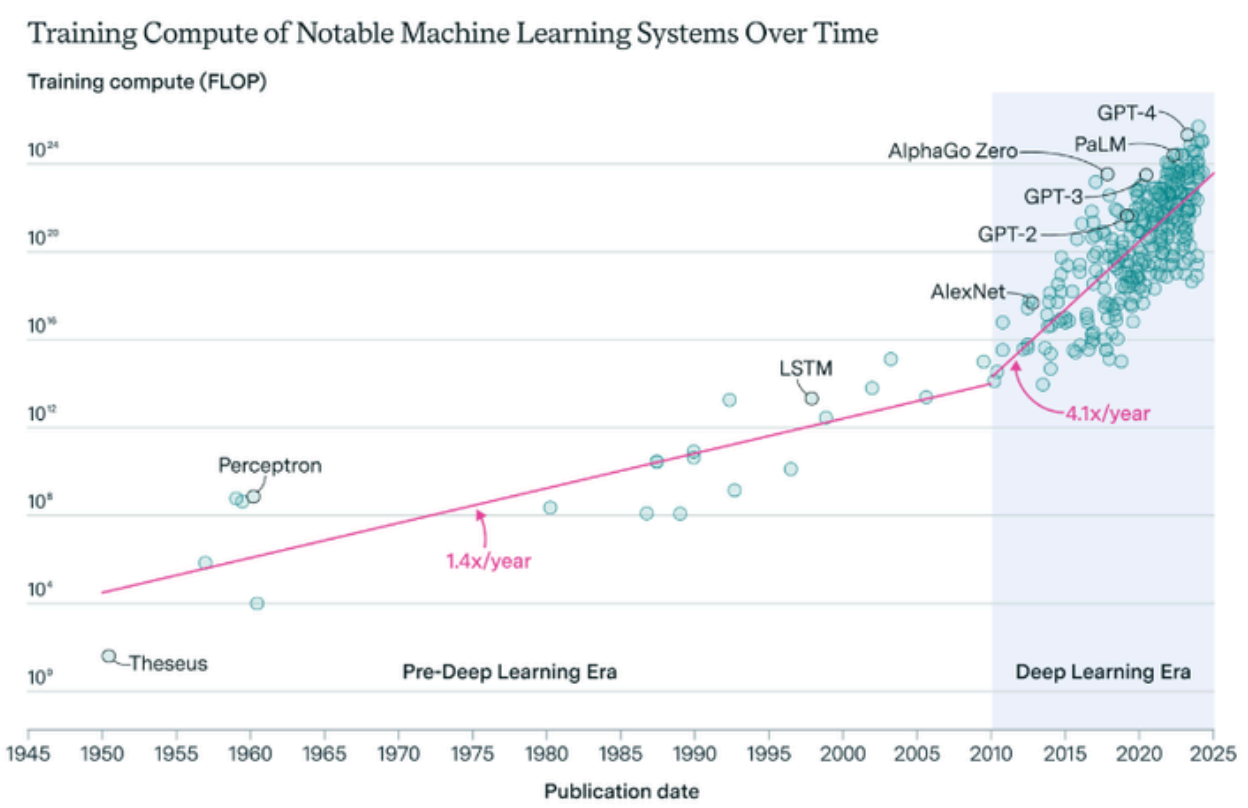
- 학습용 컴퓨팅: 연간 4배
- 학습 데이터세트 크기: 연간 2.5배
- 알고리즘 학습 효율성: 연간 1.5배~3배
- 학습 중 컴퓨터 칩 전원 공급에 사용되는 에너지: 연간 3배
- 하드웨어 효율성: 연간 1.3배

스케일업 모델

범용 AI의 훈련을 위해 더 많은 컴퓨팅과 데이터를 사용하는 것 -> 그런데 과연 진전했을까?

2.3.1 컴퓨팅, 데이터 및 알고리즘의 최신 동향

학습 및 추론에 사용되는 컴퓨팅 트렌드



AI 모델 학습에 사용되는 컴퓨팅 리소스(수행되는 연산 횟수)가 빠르게 증가 중

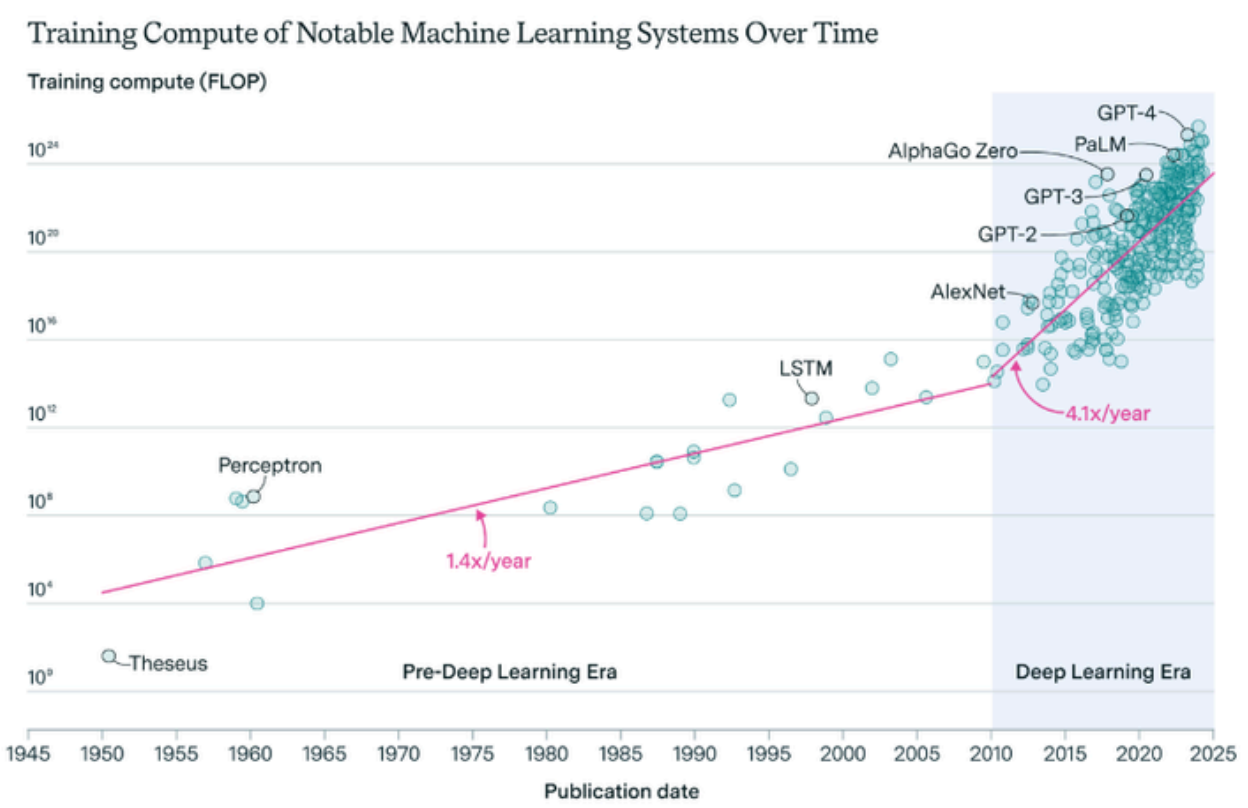
2010년 초반부터 학습에 사용되는 평균 양이 약 6개월마다 2배씩 증가

2010년의 머신러닝 모델은 평균 평균 약 1e15 부동소수점 연산(FLOP)을 사용

2023년에는 공개적으로 보고된 컴퓨팅 예산이 가장 큰 모델인 Inflection-2가 1e25 FLOP을 사용해 100억 배나 증가

2.3.1 컴퓨팅, 데이터 및 알고리즘의 최신 동향

학습 및 추론에 사용되는 컴퓨팅 트렌드



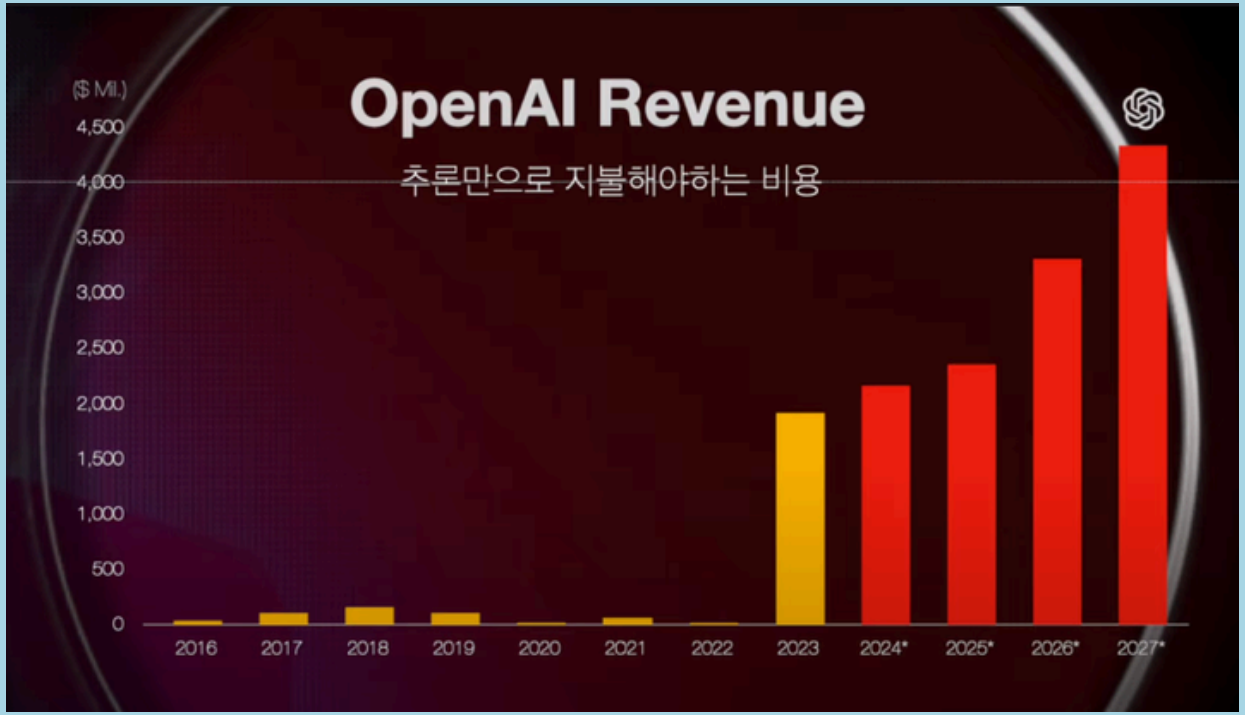
지난 15년 동안 컴퓨팅 비용은 약 50배에서 200배까지 증가했으며 범용 학습에 사용되는 컴퓨팅 총량은 GPU 성능 향상 등에 의한 컴퓨팅 비용 감소 요인에도 불구하고 훨씬 높았음

-> 머신러닝 신경 스케일링 법칙(하드웨어 / 학습 데이터 / 연산량 등을 최적 계산)이 컴퓨팅 중심의 AI 개발에 기여함

그 결과 하드웨어 지식에 대한 필요성이 더욱 커졌으며, 배포를 위한 컴퓨팅 리소스도 크게 성장

2.3.1 컴퓨팅, 데이터 및 알고리즘의 최신 동향

학습 및 추론에 사용되는 컴퓨팅 트렌드



배포된 범용 AI 시스템의 사용자 수가 급격히 증가함에 따라 추론(시스템을 사용자에게 제공하는 데 핵심적인 부분)에 필요한 컴퓨팅 리소스도 증가

일부 추정에 따르면 범용 AI 추론에 사용되는 총 연산량이 이미 새로운 모델 학습에 사용되는 연산량을 넘어섰음
구글은 전체 60%를 배포, 추론에 사용
GPT 4의 경우 예상 추론 비용이 훈련의 2배 이상으로 추정됨

학습과 추론을 위한 컴퓨팅 리소스가 증가함에 따라 AI의 에너지 사용량도 급격히 확대됨

2.3.1 컴퓨팅, 데이터 및 알고리즘의 최신 동향

학습 데이터 트렌드 : 대규모 데이터 세트, 멀티모달 등

범용 AI 학습용 데이터세트 크기는 2017년 오리지널 Transformer 모델의 경우 약 20억 개의 토큰(토큰은 단어, 문자 또는 단어의 일부)에서 2023년에는 3조 개가 넘는 토큰으로 증가하여 3년마다 약 10배씩 성장 중

고성능 언어 모델을 훈련하려면 데이터 품질이 중요 :
고품질 데이터를 사용하고 구성을 최적화하는 과정은 노동집약적 수고가 필요함
데이터를 측정하고 분석해 편향성, 다양성 부족 등을 걸러내는 것은 필수적

다양한 양식으로 범용 AI 모델을 훈련하는 것이 주목받는 중 :
GPT-4, 클로드 3, 제미나이 울트라 등 범용 AI 모델은 텍스트와 그래픽이 포함된 문서를 분석하거나 멀티미디어 프레젠테이션을 제작하는 등 텍스트, 시각, 청각 정보를 함께 처리해야 하는 작업을 수행하기 위해 다양한 모달리티를 결합

'인간 선호도' 데이터 :
사용자가 선호하는 출력물의 유형을 파악하는 것으로, 공개소스에서 얻을 수 없고 학습을 위해 특별히 생성해야 함, 대기업은 대량의 독점적인 인간 선호도 데이터를 생성하고 활용하는 데 유리할 수 있음

2.3.1 컴퓨팅, 데이터 및 알고리즘의 최신 동향

범용 AI의 학습 방법과 기술은 지속적으로 개선 중

가장 성능이 뛰어난 범용 AI 모델의 기반이 되는 기술과 훈련 방법은 시간이 지남에 따라 지속적으로 안정적으로 개선되었음
이미지 분류, 게임 플레이, 언어 모델링과 같은 주요 영역에서 AI 기술과 훈련 방법의 효율성은 약 2~5년마다 10배씩 증가하고 있음
예를 들어, 일정 수준의 성능을 달성하기 위해 이미지 분류를 수행하도록 모델을 훈련시키는 데 필요한 컴퓨팅 양은 2012년과 2019년 사이에 44배 감소했으며, 이는 16개월마다 효율성이 두 배씩 증가했음을 의미

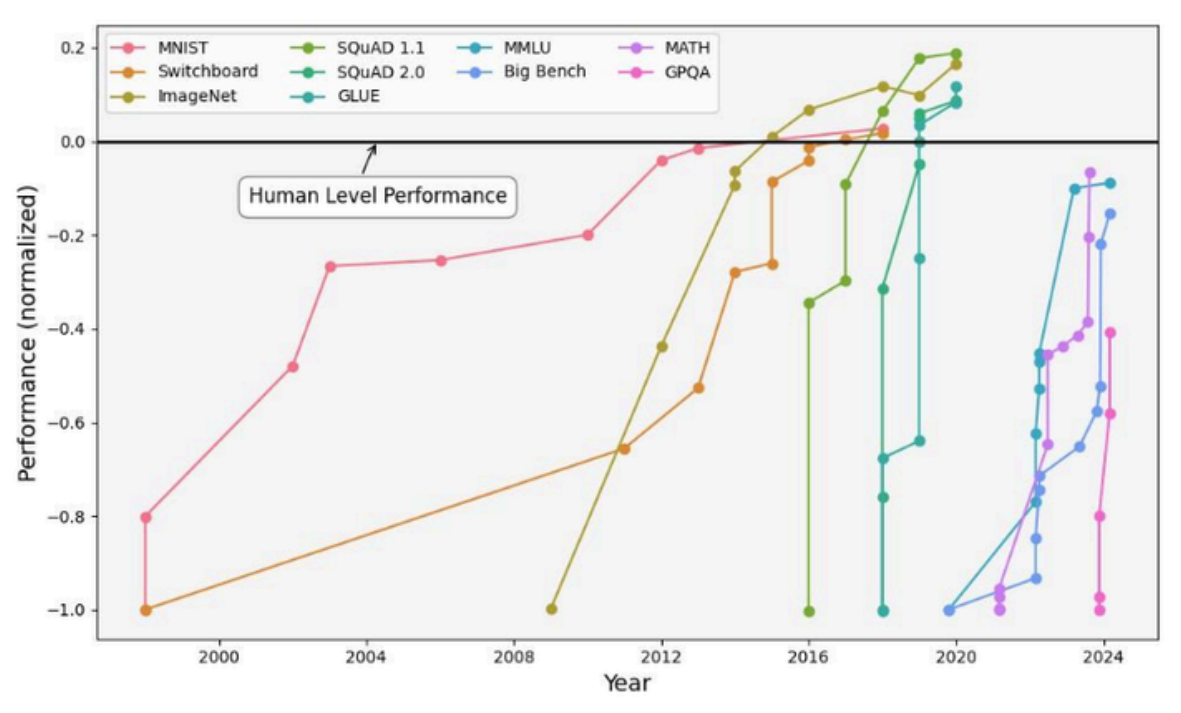
-> 이러한 발전 덕분에 범용 AI 연구자와 연구소는 제한된 하드웨어 예산 내에서 시간이 지남에 따라 더 뛰어난 성능의 모델을 개발할 수 있게 됨

그러나 AI 알고리즘의 상당한 발전에도 불구하고 범용 AI는 최근 몇 년 동안 상대적으로 큰 개념적 돌파구를 찾지 못했다고 평가

-> 현재 대부분의 고급 범용 AI 시스템에서 사용되는 트랜스포머 아키텍처를 능가하는 것은 없고, 이는 언어 모델이 더 긴 컨텍스트를 분석할 수 있도록 하는 최근의 추세를 강화

2.3.2 기능의 최근 동향

인간 수준에 근접하거나 능가한 범용 AI?



범용 AI는 컴퓨터 비전, 음성 인식, 이미지 인식, 자연어 이해와 같은 영역 일부에서 인간 수준의 성능을 달성하거나 능가함

그러나 전문가들은 이러한 지표가 제대로 된 평가인지에 대해 논쟁을 벌이고 있음

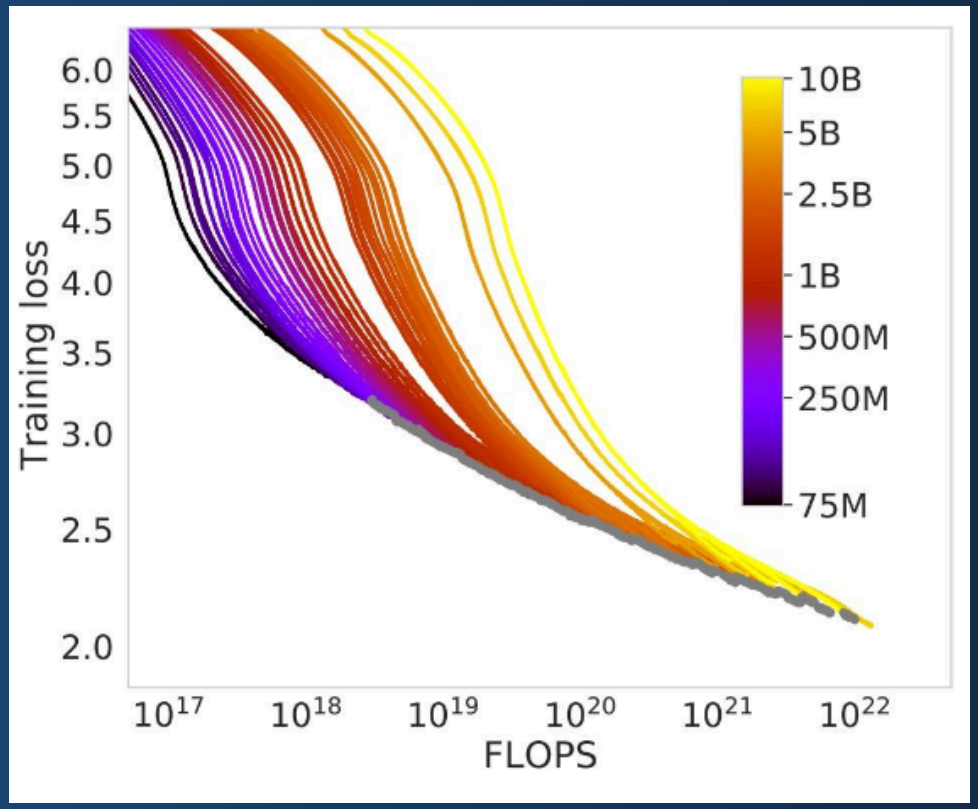
최첨단 범용 AI 모델은 종종 일부 벤치마크에서 예상치 못한 약점을 보이는데, 이는 강력한 추론이나 추상적 사고를 사용하기보다는 패턴 암기에 의존하고 있음을 나타냄

즉, 벤치마크 결과와 실제 시나리오에 지식을 안정적으로 적용할 수 있는 능력 사이에 상당한 차이가 있음을 강조

또한 AI와 인간은 인지 능력, 추상적 추론 능력 등에서 뚜렷한 차이가 있어 비교하기 어려움

2.3.2 기능의 최근 동향

범용 AI의 기능은 향상되었지만, 특정 예측은 난항



스케일링 법칙에 따르면 언어 모델의 크기가 커지고 더 많은 데이터를 학습하면 예측 가능한 만큼 성능 향상

- 단어, 문자, 숫자 등 시퀀스에서 다음 '토큰'을 더 정확하게 예측
- 데이터 세트에 내재된 작업을 더 효과적으로 수행하게 됨

그러나, 원칙을 확립하는 데 사용된 경험적 데이터의 범위를 넘어서는 규모에 대해서도 계속 유지될 것이라는 수학적인 보장은 없음

모델에 명시적으로 프로그래밍되지 않은 상태에서 모델이 특정 규모에 도달하면 갑자기 나타나는 기능에 대한 사례가 있음

또한 최근 연구에 따르면 모델 크기와 학습 컴퓨팅이 증가함에 따라 언어 모델 성능이 악화되는 '역확장(역스케일링)'의 사례가 확인됨

2.4 향후 몇 년 간의 역량 발전에 대해

가까운 미래, 범용 AI에 무엇을 기대할 수 있을까?

전문가들의 의견이 분분한 상황 : 기존 기술을 지속적으로 '확장'하고 개선하면 빠른 발전을 이룰 수 있을까? 아니면 이러한 접근 방식은 근본적으로 한계가 있으며 범용 AI 능력을 실질적으로 발전시키기 위해서는 예측할 수 없는 연구 혁신이 필요할까?

AI의 발전을 이끈 3가지 요소 : 훈련에 사용되는 연산 능력(컴퓨팅) 확장 / 훈련 데이터의 양 확대 / AI 기술 및 훈련 방법 개선

선도적인 AI 기업들은 이 세 가지 요소, 특히 컴퓨팅 성능의 향상에 지속적으로 투자 중
이대로라면 2026년 말에는 일부 범용 AI 모델은 현재 공개된 가장 컴퓨팅 집약적인 모델보다 40배에서 100배 더 많은 연산과 약 3~20배 더 효율적인 기술 및 훈련 방법을 사용하여 훈련될 것임

그러나 데이터의 제한된 가용성, AI 칩 생산 문제, 높은 전체 비용, 제한된 지역 에너지 공급 등 데이터와 컴퓨팅을 더욱 늘리는 데는 잠재적인 병목 현상이 존재

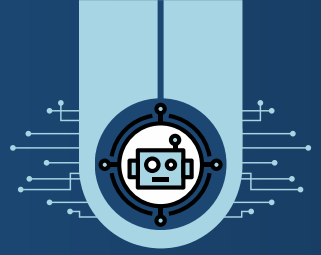
2.4.1 리소스가 빠르게 확장되면, 빠르게 발전할까?

미래의 발전에 대한 다양한 의견

- 과거의 발전이 범용 AI 시스템의 이해와 추론 능력에서 의미 있고 의미 있는 진전을 이루었으며, 훨씬 더 많은 컴퓨팅과 아마도 적당한 수준의 개념적 혁신을 통해 지속적인 발전을 이뤄 대부분의 인지 작업이 인간 수준 또는 그 이상의 성능을 발휘하는 범용 AI 시스템의 개발로 이어질 수 있다는 주장

- 현재의 딥러닝 시스템에는 인과적 추론 능력, 제한된 데이터로부터의 추상화, 상식적 추론, 유연한 예측 세계 모델 등 근본적으로 중요한 구성 요소가 부족하므로 점진적인 개선을 통한 단순한 확장으로는 해결할 수 없고, 획기적인 개념적 혁신이 필요하다는 주장

-> 새로운 기능에 대한 진전이 이루어진다면 향후 몇 년 동안 AI 위험 관리에 중요한 영향을 미칠 수 있음



2.4.2 리소스가 빠르게 확장될 것인가?

AI 개발에 투입되는 리소스를 계속 늘릴 수 있을지, 오래 지속할 수 있을지

데이터, 에너지, GPU의 부족은 모두 아래에서 설명하는 리소스를 더욱 빠르게 확장하는 데 잠재적인 장벽임
또한 전 세계적으로 디지털 인프라의 품질에 큰 격차가 존재해 전 세계적으로 AI 역량의 격차를 확대하는 데 기여하고 있음

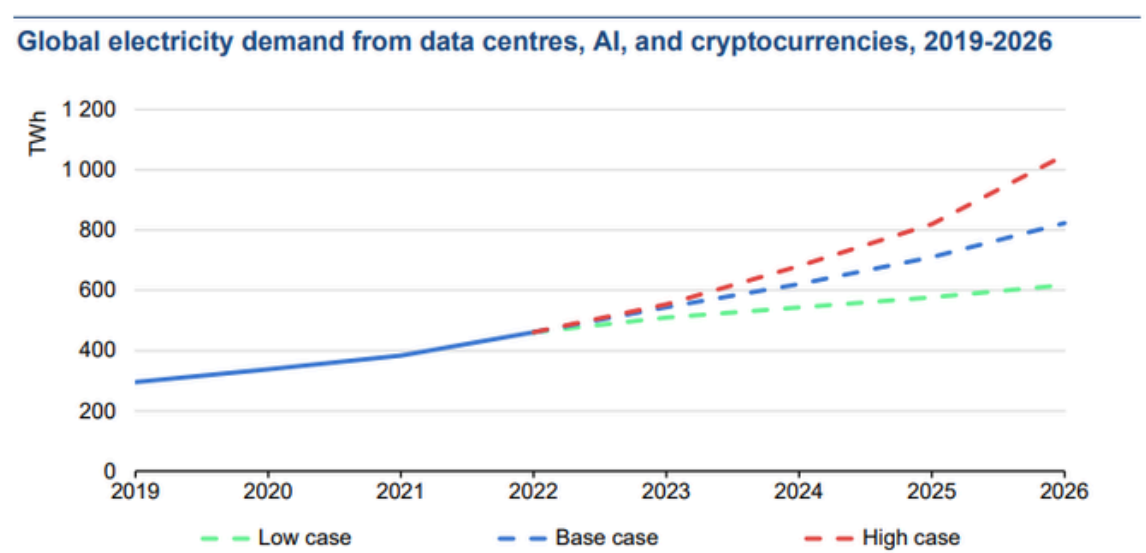
하지만, 데이터 병목 현상으로 인해 빠른 확장이 제한되는 것이 여러 도메인에 걸친 다중 시대 학습, 합성 데이터, 전이 학습과 같은 새로운 접근 방식이 활성화될 수 있음(학습 데이터 확장은 제한될 가능성이 큼)

데이터 가용성 병목 현상은 다음으로 극복 가능함

- 다중 에포크 트레이닝 : 동일한 데이터를 여러 번 학습하는 것, 그러나 많이 학습한다고 해서 이점이 높은 것은 아님
- 합성 또는 자체 생성 데이터 사용 : 실제 데이터가 제한되어 있을 때 매우 유용함, 그러나 편향성을 강화할 우려가 있음
- 도메인 간 전이 학습 : 텍스트, 이미지, 비디오, 음성, 생물학적 서열 등 다양한 출처의 데이터로 훈련

2.4.2 리소스가 빠르게 확장될 것인가?

에너지 수요 증가로 기존 전력 인프라에 부담 가중



Notes: Includes traditional data centres, dedicated AI data centres, and cryptocurrency consumption; excludes demand from data transmission networks. The base case scenario has been used in the overall forecast in this report. Low and high case scenarios reflect the uncertainties in the pace of deployment and efficiency gains amid future technological developments.

Sources: Joule (2023), de Vries, The growing energy footprint of AI; CCRI Indices (carbon-ratings.com); The Guardian, Use of AI to reduce data centre energy use; Motors in data centres; The Royal Society, The future of computing beyond Moore's Law; Ireland Central Statistics Office, Data Centres electricity consumption 2022; and Danish Energy Agency, Denmark's energy and climate outlook 2018.

다양한 분야의 소프트웨어 프로그래밍에 인공지능이 빠르게 도입되는 등 시장 트렌드로 인해 데이터센터의 전반적인 전력 수요 증가

Google과 같은 검색 도구에 인공지능을 완전히 구현할 경우 전력 수요는 10배 이상 증가할 수 있음
일반적인 Google 검색의 평균 전력 수요(0.3Wh)와 OpenAI의 ChatGPT(요청당 2.9Wh)를 비교하고 매일 90억 건의 검색을 고려하면 1년에 거의 10TWh의 추가 전력이 필요

-> 전력망과 송전 인프라가 AI 관련 전력 수요의 급증을 수용하기 어려움

2.4.2 리소스가 빠르게 확장될 것인가?

AI 칩 생산 문제와 느려지는 GPU 발전

반도체 제조 공장의 제한된 용량과 글로벌 반도체 공급망의 제약 및 우선순위는 범용 AI 시스템의 확장에 병목 현상을 일으킴
새로운 반도체 제조 공장(팹)을 건설하는 것은 매우 비싸고 일반적으로 3~5년이 걸림
최첨단 GPU 생산이 크게 증가하고 있지만, AI 칩의 공급망이 수요를 따라잡지 못할 수도 있음

연산에 대한 GPU 가격 성능과 에너지 효율은 매년 약 30%씩 개선된 반면,
트레이닝에 사용되는 총 컴퓨팅은 2010년 이후 매년 약 4배씩 증가하여 하드웨어 효율성 개선 속도를 앞지르고 있음
하드웨어 효율성의 향상보다는 지출 증가가 AI 학습을 위한 컴퓨팅 증가의 주요 동인이었음을 시사

2.4.3 알고리즘의 발전이 빠르게 이어질까?

가까운 장래에 발전 속도가 줄어들 수도, 예측이 어려움

범용 AI 모델을 뒷받침하는 기술과 알고리즘은 일관되고 강력하게 개선되어 왔지만, 발전 속도가 감소할 수도 있음, 많은 사후 학습 알고리즘은 하드웨어 예산을 늘리지 않고도 학습 컴퓨팅을 5배 이상 사용하여 모델 성능을 향상시켰으며, 20배 이상 향상시키기도 했음

-> 거버넌스 프로세스는 사후 학습 알고리즘을 고려하는 것이 필요함

사후 학습 알고리즘
- 더 나은 성능을 위해 모델을 미세 조정
- 외부 도구를 활용할 수 있는 기능 추가
- 결과를 안내하는 프롬프트 제작
- 보다 일관되고 논리적인 응답을 위해 추론 프로세스를 구조화
- 여러 응답 중에서 가장 관련성이 높고 정확한 후보 결과를 선택

AI R&D 영역에서도 거대언어모델이 이미 활용되고 있음

-> 범용 AI 시스템의 기능이 발전함에 따라 AI의 알고리즘 발전과 엔지니어링에 미치는 영향을 예측하기가 더욱 어려워지고 있음


질의응답
감사합니다

범용 AI의 안전성에 대한 국제 과학 보고서

제3장 범용 AI 시스템을 평가하고 이해하기 위한 방법론

이은우(법무법인 지향)

요약

1

거버넌스 가정

범용 AI 거버넌스 접근 방식은 AI 개발자와 정책 입안자 모두가 범용 AI 시스템이 **무엇을 할 수 있는지, 그리고 잠재적인 영향을 이해하고 측정할 수 있다고 가정**

2

한계

기술적 방법은 이러한 질문에 답하는 데 도움이 될 수 있지만 한계가 있음.

- **현재의 접근 방식은 대규모 범용 AI 관련 피해에 대해 강력한 보장을 제공할 수 없음.**

3

문제

현재 개발자들은 여전히 범용 AI 모델이 어떻게 작동하는지에 대해 거의 알지 못함.

- **모델 설명 및 해석 기술은 연구자와 개발자가 범용 AI 시스템이 어떻게 작동하는지에 대한 이해를 향상시킬 수 있지만, 이 연구는 아직 초기 단계**

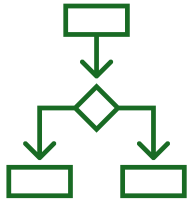
요약

접근 제한

독립적인 행위자의 회사 개발 범용 AI 모델, 시스템 감사

회사는 엄격한 평가에 필요한 모델, 데이터, 사용된 방법에 대한 정보에 **필요한 수준의 '화이트박스' 액세스를 독립 감사자에게 제공하지 않음**

여러 정부가 기술 평가 및 감사를 수행할 수 있는 역량을 구축하기 시작.



하류의 사회적 영향

범용 AI 시스템의 하류 사회적 영향을 평가하는 것은 어려움.

엄격하고 포괄적인 평가 방법론이 아직 개발되지 않았고 범용 AI는 광범위한 실제 사용 가능성이 있기 때문.

다양한 이해관계자 참여, 다학제 분석 필요

범용 AI 모델과 시스템의 잠재적 하류 사회적 영향을 이해하려면 **섬세하고 다학제적 분석**이 필요.

AI 개발 및 평가 프로세스 에서 관점의 참여와 표현을 늘리는 것은 지속적인 기술적, 제도적 과제.

범용 AI에 대한 정확한 '평가'가 필요한 이유

일반적인 역량 및 한계 결정을 위해

시스템이 통제된 환경과 자연적 환경에서 우리의 기대에 얼마나 잘 부합하는지 이해하는 데 도움.

모델 역량에 대한 보다 정확한 이해는 사용 적합성을 판단하는 데 도움이 됨.

평가에 한계와 불확실성이 따르므로, 결과를 적절히 해석하기 위해 문서화해야 함.

사회적 영향 및 하위 사용 위험 평가를 위해

배포 또는 거버넌스와 관련된 문제에 대한 토론에 정보를 제공.

평가는 복잡한 학제간 과제

사회적 위험 평가는 제품 안전, 보안 취약성 및 노동 및 환경 영향과 같은 원치 않는 외부 효과를 평가

제품 사용 중 일반적으로 예상되는 사고에 기여할 수 있는 요소와 예상치 못한 악의적 사용을 해결하는 것이 포함

NIST - AI 위험관리 프레임워크 핵심 (RMF Core)(2023. 1.)

Artificial Intelligence Risk Management Framework (AI RMF 1.0)(2023. 1.)(U.S. Department of Commerce, NIST)

모델 성능 분석을 위한 접근 방식

다양한 이해 관계자(AI 개발자, 사용자, 영향을 받는 인구 구성원 등)는 범용 AI 시스템이 **모델 역량 측면에서** 어떻게 작동해야 하는지와, 이용의 **하위 사슬에서** 부정적인 사회적 영향을 방지하는 것에 대해 기대를 가지고 있음.

연구자들은 모델 결과를 이러한 기대와 비교하는 다양한 방법을 개발

- 사례연구
- 벤치마크
- 레드팀, 적대적 공격
- 감사
- 모델 투명성, 해석

모델 성능 분석은 모델이 어떻게 수행되는지, 배포 시 어떤 제한, 이점 또는 위험이 발생할 수 있는지 이해하는 데 필수적임.

Hyper ClovaX와 EXAONE의 기술보고서

EXAONE 3.0 7.8B Instruction Tuned Language Model

LG AI Research*

Abstract

We introduce EXAONE 3.0 instruction-tuned language model, the first open model in the family of Large Language Models (LLMs) developed by LG AI Research. Among different model sizes, we publicly release the 7.8B instruction-tuned model to promote open research and innovations. Through extensive evaluations across a wide range of public and in-house benchmarks, EXAONE 3.0 demonstrates highly competitive real-world performance with instruction-following capability against other state-of-the-art open models of similar size. Our comparative analysis shows that EXAONE 3.0 excels particularly in Korean, while achieving compelling performance across general tasks and complex reasoning. With its strong real-world effectiveness and bilingual proficiency, we hope that EXAONE keeps contributing to advancements in Expert AI. Our EXAONE 3.0 instruction-tuned model is available at https://huggingface.co/LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct.

1 Introduction

EXAONE stands for **EX**pert **A**I for Every**ONE**, a vision that LG is committed to realizing in order to democratize access to expert-level artificial intelligence capabilities. Our objective of Expert AI is twofold: to help the general public achieve expert-level competency in various fields and to assist experts in attaining even higher levels of proficiency. This aligns with LG AI Research's mission to integrate advanced AI into everyday life, making expert knowledge and capabilities accessible to a broader audience.

In August 2024, LG has announced the release of EXAONE 3.0 models with enhanced performance and equipped with the Enterprise AI Agent service enabled by the models. EXAONE 3.0 models will be supplied for commercial purposes, mainly to LG affiliates and partners as before, but among them, the 7.8B instruction-tuned model is made publicly available for non-commercial, research purposes. This release aims to support the broader AI community by providing access to a high-performance language model, thereby fostering innovation and collaboration. This technical report covers the performance of EXAONE 3.0's 7.8B instruction-tuned model which is competitive in English and excellent in Korean compared to other similar-sized recently-released large language models (LLMs).

2 Model Training

In this section, we provide an overview of the model training process for EXAONE 3.0, which encompasses several critical stages, including the detailed architecture design, efficient tokenization for bilingual support, extensive pre-training on a diverse dataset, and advanced post-training techniques to enhance instruction-following capabilities. These steps ensure the model's robust performance in real-world scenarios and adherence to strict data compliance standards.

2.1 Model Architecture

In line with recent trends, EXAONE language model is based on the decoder-only transformer architecture [39]. Its maximum context length is 4,096 tokens, and it uses Rotary Position Embeddings (RoPE) [36] and Grouped Query Attention (GQA) [2]. The model architecture is shown in detail in Table 1.

*The complete list of authors who contributed to this work can be found in Section 8.1.

모델 훈련
평가
책임 있는 AI
제한 사항
배치
결론

HyperCLOVA X Technical Report

NAVER Cloud
HyperCLOVA X Team

Abstract

We introduce HyperCLOVA X, a family of large language models (LLMs) tailored to the Korean language and culture, along with competitive capabilities in English, math, and coding. HyperCLOVA X was trained on a balanced mix of Korean, English, and code data, followed by instruction-tuning with high-quality human-annotated datasets while abiding by strict safety guidelines reflecting our commitment to responsible AI. The model is evaluated across various benchmarks, including comprehensive reasoning, knowledge, commonsense, factuality, coding, math, chatting, instruction-following, and harmlessness, in both Korean and English. HyperCLOVA X exhibits strong reasoning capabilities in Korean backed by a deep understanding of the language and cultural nuances. Further analysis of the inherent bilingual nature and its extension to multilingualism highlights the model's cross-lingual proficiency and strong generalization ability to untargeted languages, including machine translation between several language pairs and cross-lingual inference tasks. We believe that HyperCLOVA X can provide helpful guidance for regions or countries in developing their sovereign LLMs.

1 Introduction

The latest advances in large language models (LLMs) have been primarily driven by objectives to improve comprehension and generation of English text. This gave birth to an array of powerful LLMs that can proficiently handle English; they reflect the norms and values of predominantly English-speaking societies, specifically North American cultures, which are extremely overrepresented in the pretraining corpora. Consequently, these LLMs exhibit limitations in their capacity to process and understand non-English languages like Korean, which embodies distinctive cultural nuances, geopolitical situations, and other regional specificities, as well as unique linguistic attributes.

In light of this context, we present HyperCLOVA X [1], a family of LLMs that includes HCX-L, the most powerful model, and HCX-S, a more lightweight alternative. Both models are tailored to the Korean linguistic and cultural framework and are capable of understanding and generating English, among several other languages. The models were initially pretrained using an evenly distributed mixture of Korean, English, and programming source code data. Subsequently, they underwent instruction tuning, utilizing high-quality human-annotated demonstration and preference datasets.

HyperCLOVA X's capabilities are showcased through extensive experiments on a collection of major benchmarks on reasoning, knowledge, commonsense, factuality, coding, math, chatting, and instruction-following, as well as harmlessness, in both Korean and English. Our thorough analysis reveals that HyperCLOVA X possesses comprehensive knowledge specific to the Korean language and culture and delivers powerful Korean reasoning capabilities unparalleled by any existing closed and open-source models, all while adhering to strict safety guidelines. Further analysis highlights HyperCLOVA X's competitive edge in its core competencies, performing on par with other proficient English-centric LLMs.

[1]API access for HyperCLOVA X is available at CLOVA Studio, a Hyperscale AI development tool optimized for businesses and provided via NAVER Cloud Platform. The chat service is available at https://clova-x.naver.com/.

교육 세부정보
핵심 벤치마크
다국어성
안전하고 책임감 있는 AI
결론

사례 연구(Case Study)

- 많은 연구 논문에서 모델 역량 평가는 **질적 방식**으로 진행되며, **모델 성능에 대한 일화적 증명과 인간의 판단**에 의존
  - 이미지 생성 모델에 대한 초기 평가는 종종 **소수의 예**를 보여주는 데 의존.
- GPT-4가 새로 출시되었을 때, 기존 벤치마크를 넘어 **큐레이팅된 작업 세트에 대한 모델 출력의 예를 사용하여 모델 성능을 설명**
  - 현재 여러 인기 있는 **벤치마크는 인간 평가자에게 서로 다른 모델의 응답을 평가하도록 요청하는 데 의존**.
- 위험은 때때로 **'향상 연구'**를 통해 측정.
  - 인간이 범용 AI 시스템에 액세스할 수 있을 때와 액세스할 수 없을 때 잠재적으로 해로운 작업을 수행하는 데 얼마나 더 유능한지 테스트하는 것

벤치마크 (Benchmarks)

- **대부분의 머신 러닝 평가는 표준화된 벤치마크 측정을 기반**으로
  - 분류 , 세분화, 질의응답을 포함하는 비교적 작은 규모의 이미지 처리 벤치마크 : AI 비전 연구에 필수적
  - 언어 모델링 연구는 일반적인 역량과 신뢰성을 측정하기 위한 여러 공개 벤치마크에 의해 형성
  - 최근에는 여러 모달리티의 정보를 결합하고 웹 브라우저와 같은 소프트웨어 도구를 사용하는 범용 AI의 기능을 측정하기 위한 벤치마크가 설계됨
- **벤치마크에서의 성과는 다운스트림 작업 성과의 불완전한 측정일 수 있음.**
  - 벤치마크는 본질적으로 원하는 성과에 대한 대리 측정이며, 벤치마크에서 좋은 점수를 받았다고 해서 다양한 타당성 문제 때문에 항상 실제로 원하는 성과로 이어지는 것은 아님

벤치마크(Benchmarks)

- 내부 타당성
  - **벤치마크 측정의 신뢰성**, 즉 보고된 메트릭이 강력한 기준선과 비교하여 반복 실행에서 얼마나 신뢰할 수 있는지와 관련.
  - 예를 들어, 벤치마크에 모델 성능에 대한 **통계적으로 유효한 주장을 할 만큼 충분한 예가 포함되지 않거나, 잘못된 레이블이 포함된 경우**
- 외부 타당성
  - 벤치마크 성과가 실제 환경으로 얼마나 잘 변환되는지의 문제. **벤치마크가 실제 작업에 대한 좋지 않은 구성이거나 부적절하거나 불완전한 대표일 수 있음.**
  - 예를 들어, 최첨단 모델이 다른 시스템과 결합되고 활성화되는 방식에 대한 제한으로 인해 해당 모델의 기능이 과소평가될 수도 있음
  - 현재 벤치마크는 종종 적용 범위에 대해 명시적이지 않음
  - '일반적인' 성능을 주장하는 벤치마크는 종종 문화적 표현과 주석자 차이의 편견, 논란의 여지가 있는 실제 사실의 개념 등을 위장
  - 최신 범용 AI 모델은 방대한 양의 인터넷 데이터로 훈련되었기 때문에 새로운 기능과 기억된 기능을 구분하기 어려울 수 있음 - **데이터 오염**

벤치마크(Benchmarks)

- **벤치마크에서 인간 성과를 해석하는 것은 어려울 수 있음**
  - 모델 성과를 직관적으로 이해하는 데 중요한 측면은 인간 성과와 비교하는 것
  - 인간 성과 측정은 종종 신뢰할 수 없음
  - ImageNet 벤치마크에서 '인간 성과'의 기준은 단일 대학원생의 주석으로 구성됨.
  - '기본 진실'('ground truth')에 대한 인간 주석자는 악명 높게 변덕스럽고, 종종 문화적 맥락, 가치 또는 전문성의 차이로 인해 상당하게 의견이 일치하지 않음
  - 작업에 있어서 인간의 성과(human performance)와 작업에 대한 인간 역량(human competence) 사이에는 상당한 차이. 후자는 종종 예를 들어 정확성뿐만 아니라 견고성(robustness)에 대한 판단을 포함
    - 인지 과학의 기본적인 통찰력, 즉 *성과* 와 *역량* 의 구분
  - 의도적으로 주석자 집단을 다양화, 다양한 기준 진실 레이블을 허용하는 평가, 또는 인간 성과 주장의 맥락을 적절히 고려하는 평가가 인간과 AI 간의 비교에 대한 보다 신뢰할 수 있는 평가가 되는 경향

벤치마크(Benchmarks)

- **테스트 세트의 만연한 레이블 오류로 인한 머신 러닝 벤치마크 불안정**
  - *커티스 G 노스컷*, *아니시 아탈리*, *조나스 뮐러*
  - 게시: 2021년 7월 29일, 최종 수정: 2024년 7월 14일
  - 가장 일반적으로 사용되는 10개의 컴퓨터 비전, 자연어 및 오디오 데이터 세트의 테스트 세트에서 레이블 오류를 식별한 다음 이러한 레이블 오류가 벤치마크 결과에 영향을 미칠 가능성을 연구. 테스트 세트의 오류는 많고 광범위합니다. 우리는 10개 데이터 세트에서 평균 3.3% 이상의 오류를 추정
- 자연어 이해에서 벤치마킹을 수정하려면 무엇이 필요할까요?
  - 사무엘 R. 보우먼 , 조지 달
  - 우리는 대부분의 현재 벤치마크가 이러한 기준을 충족하지 못하고 적대적 데이터 수집이 이러한 실패의 원인을 의미 있게 해결하지 못한다고 주장합니다. 대신 건강한 평가 생태계를 복원하려면 벤치마크 데이터 세트의 설계, 주석이 달린 신뢰성, 크기 및 사회적 편향을 처리하는 방법에서 상당한 진전이 필요합니다.

벤치마크(Benchmarks)

- **머신 러닝 실무의 평가 격차**
  - Ben Hutchinson , Google Research, 호주, benhutch@google.com, Negar Rostamzadeh , Google Research, 캐나다, nrostamzadeh@google.com, Christina Greer , Google Research, 미국, ckuhn@google.com, 캐서린 헬러 , Google Research, 미국, kheller@google.com, Vinodkumar Prabhakaran ,GoogleResearch,미국, vinodkpg@google.com DOI: https://doi.org/10.1145/3531146.3533233
    FAccT '22: 2022 ACM 공정성, 책임성 및 투명성 컨퍼런스 , 대한민국 서울, 2022년 6월
- 기계 학습(ML) 모델이 애플리케이션 생태계에 적합한지에 대한 신뢰할 수 있는 판단을 내리는 것은 책임감 있는 사용을 위해 매우 중요하며, 해악, 이점, 책임을 포함한 광범위한 요소를 고려해야 합니다. 그러나 실제로 ML 모델에 대한 평가는 **종종 좁은 범위의 비문맥화된 예측 행동에만 초점을** 맞춥니다. 우리는 평가 우려의 이상화된 폭과 실제 평가의 관찰된 좁은 초점 간의 평가 격차를 조사합니다. 컴퓨터 비전 및 자연어 처리 커뮤니티에서 최근 열린 저명한 컨퍼런스의 논문에 대한 실증적 연구를 통해 소수의 평가 방법에 대한 일반적인 초점을 보여줍니다. 이러한 방법에 사용된 메트릭과 테스트 데이터 분포를 고려하여 모델의 어떤 속성이 현장에서 중심이 되는지 주목하여 **평가 중에 자주 무시되거나 소외되는 속성을** 밝힙니다. 이러한 속성을 연구함으로써 우리는 규범적 영향을 미치는 다양한 약속에 대한 기계 학습 분야의 암묵적 가정을 보여줍니다. 여기에는 **결과주의에 대한 헌신, 맥락으로부터의 추상성, 영향의 정량화, 평가에서 모델 입력의 제한된 역할, 다양한 실패 모드의 동등성**이 포함됩니다. 이러한 가정에 빛을 비추면 ML 시스템 맥락에 대한 적합성에 의문을 제기할 수 있으며, ML 모델의 신뢰성을 견고하게 검토하기 위한 보다 맥락화된 평가 방법론을 향한 길을 제시합니다.

레드팀 및 적대적 공격

- 적대적 공격(A**dversarial attacks)**, 레드팀(**Red-teaming)**
  - 실제 상황에서 시스템을 배포하기 전에 평가자는 '적대적 공격 및 레드팀'을 사용하여 최악의 행동, 악의적 사용 기회 및 시스템이 예상치 못하게 실패할 가능성을 파악함.
- 적대적 공격
  - 사이버 보안에서 적대적 공격은 시스템을 실패하게 하려는 의도적인 시도를 말함.
  - 예를 들어, 언어 모델에 대한 공격은 자동으로 생성된 공격 또는 수동으로 생성된 공격의 형태를 취할 수 있음
  - 모델의 안전 제한(safety restrictions)을 파괴하는 '탈옥'('jailbreaking') 공격이 포함될 수도 있음
- '레드팀'
  - 시스템을 공격하여 취약점을 찾는 것을 목표로 하는 사람들의 집단

레드팀

- 레드팀의 장점
  - 고정된 일련의 테스트 사례인 벤치마크와 달리 레드팀의 주요 장점은 테스트 중인 특정 시스템에 맞게 평가를 조정한다는 것
- 레드팀의 방식
  - 레드팀 구성원은 시스템과의 상호 작용을 통해 모델에 대한 사용자 정의 테스트(custom tests)를 설계할 수 있음
  - '버그 현상금' 플랫폼, 인시던트 데이터베이스 등과 같은 ' 해악 발견 ' 을 위한 리소스를 포함하여 AI 책임 프로세스 에서 활용되는 도구의 생태계를 분석(Ojewale et al.)
  - 이러한 도구는 잠재적인 해악 벡터를 식별하고 해악 발견에 더 광범위하게 참여할 수 있도록 지원.
- 평가자는 때때로 공익을 대표하지 못할 수 있음
  - 최첨단 범용 AI 시스템을 위한 레드팀은 주로 이를 개발한 조직에서 수행.
  - 학계, 감사 네트워크 및 전담 평가 조직도 핵심적인 역할을 할 수 있음.
  - 레드팀에 대한 모범 사례는 아직 확립되지 않았음.

레드팀

- AI 개발자 자신의 경우와 마찬가지로 레드 팀 평가자는 항상 공익이나 인구 통계를 대표하지 않으며 편견을 보이거나 AI 관련 피해를 식별하거나 평가할 때 중요한 고려 사항을 생략할 수 있음

- 일반 용도 AI 모델의 다양한 결함은 레드팀 중에 감지되지 않았음
  - 레드팀과 적대적 공격은 모델의 최악의 성능을 더 잘 이해하고 벤치마크에서 적절히 다루지 않는 성능에서 성능을 평가하는 데 유용
  - 그러나 이러한 공격에는 한계가 있음
  - 레드팀과 적대적 공격 기술 벤치마킹에 대한 이전 작업에서는 버그가 종종 감지되지 않는다는 사실을 발견.
  - 실제 사례로는 최신 범용 AI 채팅 시스템 용 탈옥이 있으며 , 이를 설계한 개발자의 초기 감지를 피한 것으로 보임
- 전반적으로 레드팀은 범용 AI 기능에 대한 의미 있는 이해에 필요한 여러 평가 도구 중 하나일 뿐.
- 레드팀은 AI 시스템이 사회에 더 광범위하게 배치 될 때 발생하는 하류 피해를 포착하지 못할 수도 있음

적대적 공격(탈옥)

- **Prompt Engineering을 통한 ChatGPT 탈옥: 경험적 연구(2023)**
  - Yi Liu , Gelei Deng , Zhengzi Xu , Yuekang Li , Yaowen Zheng , Ying Zhang , Lida Zhao , Tianwei Zhang , Kailong Wang , Yang Liu
  - ChatGPT와 같은 대규모 언어 모델(LLM)은 엄청난 잠재력을 입증했지만 콘텐츠 제약 및 잠재적 오용과 관련된 과제도 제기합니다.
    - 저희 연구는 세 가지 핵심 연구 질문을 조사합니다. (1) LLM을 탈옥할 수 있는 다양한 프롬프트 유형의 수, (2) LLM 제약을 우회하는 탈옥 프롬프트의 효과성, (3) 이러한 탈옥 프롬프트에 대한 ChatGPT의 회복성.
    - 처음에 기존 프롬프트의 분포를 분석하기 위한 분류 모델을 개발하여 10가지의 고유한 패턴과 3가지 탈옥 프롬프트 범주를 식별합니다. 그런 다음 8가지 금지 시나리오에 걸쳐 3,120개의 탈옥 질문 데이터 세트를 활용하여 ChatGPT 버전 3.5 및 4.0이 있는 프롬프트의 탈옥 기능을 평가합니다. 마지막으로 **ChatGPT의 탈옥 프롬프트에 대한 저항성을 평가하여 프롬프트가 40가지 사용 사례 시나리오에서 제한을 일관되게 회피할 수 있음을 발견**했습니다.

적대적 공격(탈옥)

- **탈옥: LLM 안전 교육은 어떻게 실패하는가?(2023)**
  - A. Wei, N. Haghtalab, J. Steinhardt,
  - 안전성과 무해성을 위해 훈련된 대규모 언어 모델은 ChatGPT의 초기 릴리스에서 원치 않는 동작을 유발하는 "탈옥" 공격의 유행에서 알 수 있듯이 적대적 오용에 여전히 취약합니다. 문제 인식을 넘어 이러한 공격이 성공하는 이유와 공격이 생성되는 방식을 조사합니다.
  - **안전 훈련의 두 가지 실패 모드**, 즉 **목표의 경합(competing objectives)과 불일치 일반화(mismatched generalization**)가 가정됩니다. 경쟁 목표는 모델의 역량과 안전 목표가 충돌할 때 발생하는 반면, 불일치하는 일반화는 안전 훈련이 이루어지지 않은 분야에서 발생합니다.
  - **모델의 역량과 안전 목표가 충돌하는 시나리오를 설정하는 것. 예를 들어 접두사 주입, 모델에 복종적인 확인, 거부 억제, 역할극 수행**
  - **안전 교육이 자연어 쿼리를 Base64로 변환하는 것과 같이 기능이 존재하는 도메인으로 일반화하지 못하면 불일치 일반화가 발생, 암호 Yuan et al. (2024) 및 자원이 부족한 언어 Deng et al. (2023) 민감한 단어를 동의어로 대체 Wei et al. (2023) 또는 민감한 단어를 하위 문자열로 분할**
  - 이러한 실패 모드를 사용하여 탈옥 설계를 안내한 다음 OpenAI의 GPT-4와 Anthropic의 Claude v1.3을 포함한 최첨단 모델을 기존 및 새로 설계된 공격에 대해 평가합니다.
  - **이러한 모델 뒤에 있는 광범위한 레드팀 구성 및 안전 훈련 노력에도 불구하고 취약성이 지속된다는 것을 알게 되었습니다**.
  - 특히, 우리의 실패 모드를 활용한 새로운 공격은 모델의 레드팀 평가 세트에서 안전하지 않은 요청 모음의 모든 프롬프트에서 성공하고 기존의 임시 탈옥보다 성능이 뛰어납니다.
  - 우리의 분석은 안전-역량 패리티의 필요성을 강조합니다. 즉, 안전 메커니즘은 기본 모델만큼 정교해야 하며, 확장만으로 이러한 안전 실패 모드를 해결할 수 있다는 생각에 반대합니다.

적대적 공격(탈옥)

- **정렬된 언어 모델에 대한 보편적이고 이전 가능한 적대적 공격(2023)**
  - A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Zico Kolter, M. Fredrikson,
  - 이 논문에서는 정렬된 언어 모델이 불쾌한 동작을 생성하도록 하는 간단하고 효과적인 공격 방법을 제안합니다. 구체적으로, 우리의 접근 방식은 LLM에 대한 광범위한 쿼리에 첨부하여 불쾌한 콘텐츠를 생성할 때 모델이 긍정적인 응답을 생성할 확률을 최대화하는 것을 목표로 하는 접미사를 찾습니다(답변을 거부하는 것이 아니라). 그러나 수동 엔지니어링에 의존하는 대신, **우리의 접근 방식은 탐욕적 및 기울기 기반 검색 기술을 결합하여 이러한 적대적 접미사를 자동으로 생성하며, 과거의 자동 프롬프트 생성 방법보다 개선되었습니다.**
  - **놀랍게도, 우리의 접근 방식으로 생성된 적대적 프롬프트는 블랙박스, 공개적으로 출시된 LLM을 포함하여 상당히 이전 가능하다는 것을 알게 되었습니다.** 구체적으로, 우리는 여러 프롬프트(즉, 여러 유형의 불쾌한 콘텐츠를 요청하는 쿼리)와 여러 모델(우리의 경우 Vicuna-7B 및 13B)에 대한 적대적 공격 접미사를 훈련합니다. 그렇게 할 때, 결과적인 공격 접미사는 ChatGPT, Bard 및 Claude의 공개 인터페이스와 LLaMA-2-Chat, Pythia, Falcon 등과 같은 오픈 소스 LLM에서 불쾌한 콘텐츠를 유도할 수 있습니다.
  - 전체적으로 이 작업은 정렬된 언어 모델에 대한 적대적 공격의 최첨단 기술을 크게 발전시켜 이러한 시스템이 불쾌한 정보를 생성하는 것을 방지할 수 있는 방법에 대한 중요한 의문을 제기합니다. 코드는 이 http URL 에서 제공됩니다 .

적대적 공격(탈옥)

- **페르소나 변조를 통한 언어 모델을 위한 확장 가능하고 이전 가능한 블랙박스 탈옥(2023)**
- R. Shah, QF Montixi, S. Pour, A. Tagade, J. Rando,
  - 대규모 언어 모델을 정렬하여 무해한 응답을 생성하려는 노력에도 불구하고, 여전히 제한 없는 행동을 유발하는 탈옥 프롬프트에 취약합니다. 이 작업에서 우리는 유해한 지침을 준수할 의향이 있는 인격을 취하도록 대상 모델을 조종하는 블랙박스 탈옥 방법으로 페르소나 변조를 조사합니다. **각 페르소나에 대한 프롬프트를 수동으로 작성하는 대신 언어 모델 어시스턴트를 사용하여 탈옥 생성을 자동화**합니다.
  - 우리는 **메스암페타민 합성, 폭탄 제작 및 자금 세탁에 대한 자세한 지침을 포함하여 페르소나 변조로 가능해진 다양한 유해한 완료를 보여줍니다. 이러한 자동화된 공격은 GPT-4에서 42.5%의 유해한 완료율을 달성하는데, 이는 변조 전(0.23%)보다 185배 더 큽니다**. 이러한 프롬프트는 또한 클로드 2와 비쿠나로 전송되어 각각 61.0%와 35.9%의 유해한 완료율을 보입니다. 우리의 작업은 상업용 대규모 언어 모델의 또 다른 취약성을 드러내고 보다 포괄적인 보호 장치의 필요성을 강조합니다.

레드팀

- **생성 AI를 위한 레드팀 구성: 만병통치약인가, 보안 극장인가?(2024)**
  - 마이클 페퍼 , 아누샤 신하 , 웨슬리 한웬 덩 , 재커리 C. 립톤 , 호다 하이다리
  - 인공지능의 안전하고 보안적이며 신뢰할 수 있는 개발 및 활용에 관한 미국 대통령 행정명령은 레드팀을 8번 언급하며 이를 다음과 같이 정의합니다.
    - *"'AI 레드팀'이라는 용어는 종종 통제된 환경에서 AI 개발자와 협력하여 AI 시스템의 결함과 취약성을 찾기 위한 체계적인 테스트 노력을 의미합니다. 인공 지능 레드팀은 대부분 전담 '레드팀'이 수행하며, 결함과 취약성(예: AI 시스템의 유해하거나 차별적인 출력, 예상치 못하거나 바람직하지 않은 시스템 동작, 제한 사항 또는 시스템의 오용과 관련된 잠재적 위험)을 식별하기 위해 적대적인 방법을 채택합니다."*
  - 이 명령은 상무부 장관과 기타 연방 기관에 AI 안전 및 보안에 대한 지침, 표준 및 모범 사례를 개발하도록 함. 여기에는 "[이러한] 모델의 안전, 보안 및 신뢰성을 평가하고 관리"하기 위한 메커니즘으로 *"AI 개발자, 특히 이중 용도 기반 모델 개발자가 AI 레드팀 테스트를 수행할 수 있도록 하는 적절한 절차 및 프로세스"* 가 포함됨.
  - 생성 AI(GenAI) 모델의 안전성, 보안성 및 신뢰성을 둘러싼 우려가 커지면서 실무자와 규제 기관 모두 이러한 위험을 식별하고 완화하기 위한 전략의 핵심 구성 요소로 AI 레드팀을 지적했습니다. 그러나 정책 논의와 기업 메시징에서 AI 레드팀이 중심적인 역할을 함에도 불구하고 정확히 무엇을 의미하는지, 규제에서 어떤 역할을 할 수 있는지, 사이버 보안 분야에서 원래 구상된 기존 레드팀 관행과 어떻게 관련이 있는지에 대한 중요한 의문이 남아 있습니다.
  - 이 연구에서 우리는 AI 산업에서 레드팀 활동의 최근 사례를 파악하고 관련 연구 문헌에 대한 광범위한 조사를 수행하여 AI 레드팀 관행의 범위, 구조 및 기준을 특성화합니다.

레드팀

- **생성 AI를 위한 레드팀 구성: 만병통치약인가, 보안 극장인가?(2024)**
  - 우리의 분석에 따르면 이전의 **AI 레드팀 방법 및 관행은 활동의 목적(종종 모호함), 평가 중인 아티팩트, 활동이 수행되는 설정(예: 행위자, 리소스 및 방법) 및 이로 인해 발생하는 결정(예: 보고, 공개 및 완화)**을 포함하여 여러 축을 따라 갈라집니다.
    - **레드팀은 구조화되지 않았음**
    - **평가팀 구성은 편향을 초래**
    - **결과를 공개하는 데 주저하는 것은 유용성을 감소**
    - **레드팀을 수행하는 다양한 방법.**
    - **위협 모델링은 반대 위험에 치우쳐 있음**
    - **적대적 역량에 대한 합의 없음**
    - **정렬에 사용된 값의 비보편성**
    - **레드팀을 수행해야 할 사람에 대한 합의가 없음**
    - **레드팀 활동에 대한 불분명한 후속 조치**
  - 우리의 연구 결과에 비추어, 우리는 레드팀이 GenAI 해악 완화를 특징짓는 데 가치 있는 빅텐트 아이디어일 수 있고, 업계가 레드팀과 다른 전략을 비밀리에 적용하여 AI를 보호할 수 있지만, 보안 극장에서 발생할 수 있는 모든 위험에 대한 만병통치약으로서 레드팀(공개 정의 기반)은 적절하지 않습니다.

감사

- 감사
  - 범용 AI 개발 프로세스 전반에 걸친 설계 선택은 결과 시스템이 작동하는 방식에 영향을 미침.
  - 감사는 이러한 선택에 대한 책임을 면밀히 조사하고 보장하는 메커니즘을 제공.
- 다양한 수준
  - 범용 AI 감사자의 다양한 그룹은 수집된 증거의 품질과 달성된 책임 결과에 관해 천차만별의 다양한 수준
- 책임성
  - AI 감사 에 대한 다양한 접근 방식에 대한 설문 조사는 다양한 차원에서 배포된 범용 AI 시스템을 독립적으로 평가하여 이해 관계자가 범용 AI 시스템의 개발과 사용과 관련하여 내리는 선택에 대해 책임을 지도록 제안.

데이터 감사

- 데이터 감사
  - 훈련 데이터 분석은 문제가 있는 내용을 드러낼 수 있음
  - 머신 러닝 개발 프로세스에서 데이터는 수집,되고 큐레이션되고, 주석이 달림.
  - 이러한 데이터 엔지니어링 결정이 어떻게 간접적인 피해를 초래할 수 있는지, 그리고 그 결과에 영향을 미칠 수 있는지 조사하는 것은 모델과 그 궁극적인 하류 영향을 이해하는 데 유용
- 사례
  - 현대 시스템을 훈련하는 데 사용된 인터넷의 텍스트 및 이미지 데이터 분석에서는 저작권이 있는 콘텐츠, 증오 표현 및 밈, 악의적인 고정관념, 성적으로 노골적인 콘텐츠, 아동 학대 자료를 포함한 성폭력 묘사가 식별됨
  - 훈련 데이터 세트에 대한 조사는 주류 데이터 소스에서 특정 인구의 인구 통계적, 지리적 및 언어적 과소 표현 측면에서 문제를 종종 드러냄
  - 데이터 감사는 저작권에 대한 법적 도전에 필요한 증거를 제공. 예를 들어, OpenAI에 대한 New York Times 소송은 Dodge et al.의 데이터 감사를 많이 인용했으며 부적절한 내용이 포함되어 있다고 간주되는 일부 데이터 세트의 삭제 시도로 이어짐.

데이터 감사

- 비공개로 인한 어려움과 문제
  - 현대의 범용 AI 모델은 종종 엄청나게 방대한 양의 인터넷 데이터 세트에서 훈련되고 이러한 데이터 세트는 종종 공개되지 않으므로 데이터 출처는 여전히 체계적인 문제로 남아 있음.
  - 결과적으로 대규모 훈련 데이터에서 잠재적으로 유해한 예를 체계적으로 검색하는 것은 어려움

프로세스 감사

- 프로세스 감사
  - 모델이 어떻게 개발되는지 면밀히 조사하는 감사를 일반적으로 '프로세스 감사'라고 함
  - 학습 데이터 외에도 AI 모델링 및 제품 선택에 대한 분석은 상충 관계를 드러내고 다운스트림 위험을 나타낼 수 있음
  - 예를 들어, 인간 피드백 기반 방법은 범용 AI 모델을 학습하는 최첨단 기법인데, 아첨(sycophancy) 및 비다양성(non-diverse) 출력 생성과 같은 문제에 기여할 수 있음
  - '모델 정리'(model pruning)와 같은 엔지니어링 결정은 특정 테스트 인구 통계에 대해 불공평한 영향을 미칠 수 있음
  - 같은 맥락에서 생성 이미지 모델 아키텍처 선택은 다양한 인종을 표현하는 시스템의 성능에 영향을 미치는 것으로 나타남.
  - 독점적이거나 오픈소스인 범용 AI 모델의 전체 개발 세부 정보가 문서화되거나 공개되는 경우가 드물기 때문에 연구자들이 개발자의 방법론을 분석하는 것은 어려운 일.

생태계 감사(ecosystem audits)

- '생태계 감사'
  - 인간 -AI 상호작용을 평가하는 데 도움
  - 점점 더 많은 실험실 연구에서 인간 사용자가 범용 AI 시스템과 상호작용하는 방식을 조사
  - 이는 종종 통제된 연구의 형태를 띰. 참가자는 모델과 상호작용하고 모델링과 사용자 상호작용 설정이 참가자의 결정과 행동에 미치는 영향을 직접 측정.
  - 이러한 연구는 사용자가 다른 모델 출력보다 특정 모델 출력 표현을 신뢰하는 경향을 보여줌.
  - 그러나 참가자가 모집되기 때문에 연구를 설계하고 실제로 사용자 영향의 전체 범위를 의미 있게 반영할 만큼 충분히 광범위한 참가자를 모집하는 것이 어려울 수 있음.
- 실제 배포 상황의 통제 실험
  - 여러 연구원이 실제 배포 설정에서 AI 사용 의 영향에 대한 자연적 및 통제된 실험을 수행하기 시작.
  - 예를 들어, 미국 켄터키주에서 AI 위험 평가가 판사의 보석금 결정에 미치는 영향, 자동화된 채용 도구 사용이 채용 관리자의 재량권에 미치는 영향, 생성 AI 가 중간 관리자의 성과에 미치는 영향에 대한 연구
  - 이해 관계자에 대한 질적 인터뷰는 AI 구현 의 일부 사회적 영향과 같은 보다 체계적인 영향을 설명하는 데 효과적인 것으로 입증되었으며 , 이는 AI 시스템이 제거된 후에도 지속될 수 있음 .

실사용 감사(배포후 감사)

- 실제 세계에서 배포 후 AI 시스템 분석
  - 연구자는 이를 더 큰 사회 시스템의 구성 요소로 연구할 수 있음
- 사후 시장 감시
  - 감사 생태계가 있는 다른 산업에서 상당히 일반화 됨
  - 사용자는 종종 개발자가 발견하지 못하는 기능과 실패 모드를 발견하며, 시스템의 실제 사용을 모니터링하면 과학적 이해를 더욱 높일 수 있음
- 예를 들어, 최신 대규모 언어 채팅 모델에 대한 탈옥은 일반 사용자의 결과에 따라 먼저 연구됨.
- 실제 세계에서 딥페이크에 대한 연구는 또한 피해를 연구하고 완화하는 과학적 연구를 형성하는 데 도움이 됨.

모델 투명성, 설명 및 해석

- 출력 연구 vs 내부 메커니즘 연구
  - 범용 AI 모델 출력을 연구하는 것과 대조적으로 , 모델을 평가하는 또 다른 일반적인 접근 방식은 모델이 출력을 생성하는 내부 메커니즘을 연구하는 것
  - 이를 통해 연구자는 모델 성능 평가를 맥락화하고 모델 기능에 대한 이해를 심화할 수 있음.
  - 범용 AI 모델과 시스템이 내부적으로 어떻게 작동하는지 연구하는 것은 수천 개의 학술 논문이 생산된 인기 있는 연구 주제.
- 투명성을 강화하는 것을 목표로 하는 연구 분야
  - 문서화, 타사 액세스 메커니즘, 블랙박스 분석, 모델 작업 설명 및 모델의 내부 작동 방식 해석이 포함.

모델 투명성, 설명 및 해석 - 문서화

- 문서화 템플릿
  - 내린 결정을 기록함. 운영 수준에서 투명성 제고.
  - 모델을 정의하는 엔지니어링 결정을 문서화하고 커뮤니케이팅 하는 것.
    - 현재 범용 AI 모델에 대한 투명성을 높이는 가장 실용적인 방법 중 하나
  - 이러한 결정을 더 광범위한 내부 및 외부 이해 관계자에게 전달하기 위해 여러 문서화 솔루션이 제안됨.
- 모델 카드(Model Cards) 개발
  - 일부 성공적, 최근 한 연구에서는 " AI 커뮤니티 내에서 모델 카드가 광범위하게 채택 " 된다고 보고,
  - 데이터 세트 관행, 더 광범위한 시스템 기능 및 더 광범위한 절차적 의사 결정에 대한 커뮤니케이션을 위한 문서화 템플릿

모델 설명 및 해석 기술

- **모델 설명 및 해석 기술**
  - **연구자가 범용 AI 시스템이 내부적으로 작동하는 방식을 이해하는 데 도움이 될 수 있음**
  - **범용** AI 시스템 에 대한 외부 감사를 허용하는 여러 도구가 있어 외부 행위자가 범용 AI 시스템을 직접 쿼리하거나 다른 방식으로 모델 세부 정보에 대한 가시성을 얻을 수 있음
  - 주어진 입력의 결과로 모델의 출력을 설명할 수 있는 방법을 연구하는 것도 그 중 하나.
  - 이러한 설명은 자동화된 AI 시스템 에 의해 인간이 부당하게 해를 입거나 차별을 받는 경우 책임을 결정하는 데 도움이 되어 책임을 뒷받침하는 데 고유한 역할을 할 수 있음
  - 신경망의 계산을 연구하는 데 사용되는 또 다른 접근 방식에는 AI 시스템 내부의 매개변수, 뉴런, 하위 네트워크 또는 레이어 표현의 역할을 해석하는 것이 포함
    - 모델에 대한 해석은 때때로 연구자가 취약점을 찾는 데 도움이 됨.
    - 예시에는 레드팀 , 잘못된 기능의 내부 표현 식별(internal representations of spurious features), 취성 기능 표현 (brittle feature representations) 및 변환기의 사실적 회상의 한계(limitations of factual recall in transformers)가 포함 .

모델 설명 및 해석 기술

- **범용 AI 시스템이 내부적으로 어떻게 작동하는지 이해하고 이러한 이해를 효과적으로 사용하는 것은 어**려움
  - 비교할 객관적인 기준이 없기 때문에 범용 AI 시스템이 작동하는 방식에 대한 해석이 올바른지 확인하기 어렵다는 것
  - 해석 가능성 기술이 모델 작동 방식에 대한 오해의 소지가 있는 해석을 제안하는 '해석 가능성 환상'이 여러 건 기록됨
  - 일부 연구에서는 알고리즘 투명성 도구가 어떻게 악의적으로 사용되어 거짓 이분법을 구성하고 모호하게 하며 오도할 수 있는지 비판적으로 조사
  - 해석 가능성 기술을 엄격하게 평가하려면 해당 기술이 생성하는 해석이 일부 다운스트림 작업에 대해 입증 가능하게 유용해야 함
  - 그러나 AI 해석 가능성 도구는 아직 많은 작업에 대해 더 간단한 기술과 일관되게 경쟁력이 없음.

모델 설명 및 해석 기술

- 특히, 모델 동작을 설명하는 다양한 기술은 종종 서로 일치하지 않으며, 다운스트림 사용자에 대한 온전성 검사에서 실패
- 해석 가능성은 때때로 실제 진단 및 이해를 개선했으며, 특히 이 분야의 최근 진전이 있음
  - 그러나 범용 AI 시스템 에 대한 고수준 해석은 현재 모델 및 방법으로 해당 시스템의 동작에 대한 공식적인 보장을 하는 데 사용할 수 없음.
- 현재 설명 가능성 및 해석 가능성 기술이 엄격한 모델 평가에 실질적으로 도움이 될 수 있는 잠재력에 대해 논쟁이 있음.

범용 AI 시스템 연구의 과제 : 액세스, 투명성

- 범용 AI 역량과 위험에 대한 철저한 평가를 실시하고 강력한 확신을 얻는 것은 극히 어려운 일.
  - 현대의 범용 AI 시스템은 데이터 수집, 훈련 실행, 시스템 통합 및 배포 애플리케이션을 포함하는 복잡하고 분산된 프로젝트의 결과이며, 실제 사용 사례도 많기 때문.
  - 이러한 복잡성으로 인해 단일 행위자가 전체 프로세스를 이해하기 어려움.
- 접근 : 평가의 질은 액세스 수준과 투명성에 따라 달라져
  - AI 시스템을 평가하는 다양한 기술에는 다양한 유형의 액세스가 필요.
- 블랙박스 액세스, 화이트박스 액세스
  - 테스트 데이터에서 모델의 성능을 평가하려면 일반적으로 대상 모델을 쿼리하고 해당 출력을 분석하는 기능만 필요. 이를 일반적으로 '블랙박스' 액세스라고 함.
  - 블랙박스 시스템을 쿼리하는 기능은 유용하지만 많은 유형의 평가 기술은 더 높은 수준의 액세스에 의존
- 역사적으로 AI 연구자들은 오픈 소스 방법, 모델 및 데이터의 혜택을 누렸음.
  - 오늘날 기업들은 최첨단 범용 AI 시스템을 점점 더 비공개 로 유지 .
  - '화이트박스' 액세스(모델 매개변수에 대한 액세스)가 부족하면 연구자들이 적대적 공격, 모델 해석 및 미세 조정을 수행하기 어려움
  - 데이터, 문서, 기술, 구현 세부 사항 및 조직 세부 사항을 포함하여 시스템이 설계된 방식에 대한 정보에 대한 '상자 밖' 액세스가 부족하여 개발 프로세스에 대한 평가를 수행하기 어려움

범용 AI 시스템 연구의 과제 : 액세스, 투명성

- 제3자 감사 생태계
  - 초기 단계이지만 성장하고 있음
  - 일부는 오픈 소스인 다양한 AI 감사 도구를 사용하면 외부 사용자가 모델의 세부 정보를 쿼리하고 액세스할 수 있음.
  - 여러 연구에서는 독립적인 레드팀 구성 및 감사 노력을 가능하게 하기 위해 법적 '안전 항구' 또는 정부 중재 액세스 체제를 주장
  - 코드와 가중치를 공개할 필요가 없지만, 독립적인 연구원과 감사자가 누출을 방지하도록 설계된 보안 환경에서 모델에 대한 전체 액세스 권한으로 분석을 수행할 수 있는 구조화된 액세스 방법이 제안

범용 AI 시스템 영향 분석의 과제

- 하류 사회적 영향을 철저히 평가하려면 섬세한 분석, 학제간성 및 포용성이 필요
  - 전반적인 사회적 영향을 이해하는 것이 많은 AI 평가의 궁극적인 목표이기는 하지만 많은 경우 이 목표에 미치지 못함
- 첫째, 연구자들이 AI 시스템을 연구하는 설정과 배포될 끊임없이 변화하는 현실 세계 설정 사이에는 항상 차이가 있음
- 둘째, 사회에 대한 AI 영향을 평가하는 것은 복잡한 사회 기술적 문제
  - 예를 들어, 대규모 언어 모델은 안전성 , 역량 , 경향 측면 에서 언어 간에 상당한 차이가 있는 것으로 알려져 있지만 , 연구자들이 언어 간에 언어 모델을 철저히 평가하는 것은 어려움
  - '공정성'과 '평등'과 같은 윤리적 개념을 다룰 때 단순화된 기술적 대리자에 지나치게 의존하면 오해의 소지가 있거나 대표성이 부족한 이해 관계자를 제외할 수 있음
- 범용 AI 의 광범위한 영향에 대한 평가는 매우 다면적이어서 학제간, 매우 다른 관점을 가질 수 있는 여러 이해 관계자의 대표성이 필요
  - 배포에 앞서 사회에서 AI 배포 의 영향을 모델링하는 것은 본질적으로 복잡하고 정량적 분석에 고정하기 어려움
  - 범용 AI 에 대한 사회적 영향 평가에 진전이 있었지만, 여러 이해 관계자의 이익과 이 작업을 실제로 수행하기 어렵게 만드는 리소스 문제의 균형을 맞춰야 하기 때문에 구현은 여전히 어려움.

범용 AI 시스템 영향 분석의 과제

- **AI 개발 및 평가 프로세스 에서 관점의 참여 및 대표성을 높이는 것은 지속적인 기술적, 제도적 과제**
  - **평가의 관련 목표를 지정하는 것은 누가 참여하는지와 토론이 어떻게 구성되는지에 따라 크게 영향을 받으므로 우려되는 영역을 놓치거나 잘못 정의하기 쉬움**
  - **감사 프로세스에 참여하는 사람들의 범위를 넓히면 현재 또는 예상되는 피해를 발견하고 특성화하는 프로세스에 통합된 경험 범위도 넓어짐.**
  - **참여 확대는 최근 몇 년 동안 머신 러닝 커뮤니티의 초점이었으며, AI 모델 설계, 개발, 평가 및 거버넌스 프로세스에 더 광범위한 관점과 이해 관계자를 통합하고 참여시킬 필요성을 강조**
  - 영향 평가를 구현하는 동안 더 광범위한 영향 개념을 촉진
  - 보다 포괄적인 범위의 인간 피드백을 가능하게 하는 것 등이 제안됨.
  - 더 많은 목소리를 찾아내 통합하는 것은 섬세한 노력이며, 착취 가능성을 최소화하기 위해 참여 당사자에 대한 세심함과 존중이 필요
  - 참여를 늘리는 과제에는 양립할 수 없는 가치나 우선순위 간의 어려운 선택에 대한 필요한 협상도 포함됨

범용 AI 시스템 영향 분석의 과제

- **투명성은 AI 시스템을 평가하는 데 중요하지만 실제로 달성하기 어려움**
- 투명성 노력이 항상 책임성 충족이 아님.
- 설명 가능성 기술을 구현한 수십 개의 기업을 조사한 결과, 많은 설명 가능성 노력이 모델 개발 프로세스에서 의미 있게 사용되었지만 최종 사용자에게 투명성이나 정당성을 제공한다는 정책적 요구에는 거의 부응하지 못하는 것으로 나타남.
- 독점적이거나, 오픈 소스인 설명 가능성 및 해석 가능성 툴에 대한 최근 조사에 따르면 이러한 툴은 주장을 뒷받침하기에 적절하게 검증되지 않았으며 조작 및 견고성 문제에 취약했음
- 기업 환경에서 문서화 관행을 운영화하는 데 따르는 물류는 내부 정치로 가득 차 있을 수 있

모델 투명성, 설명, 해석

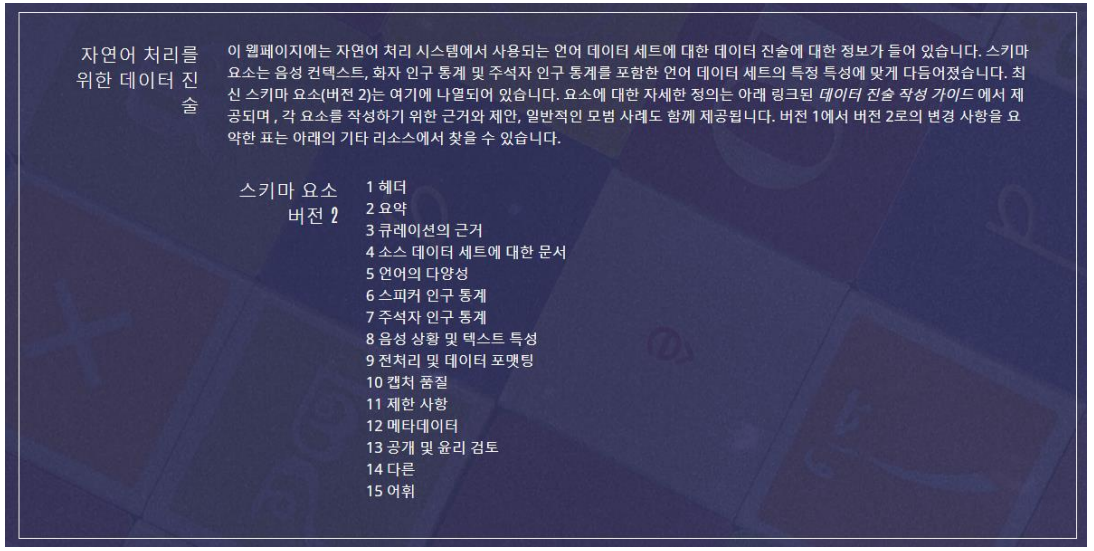
- **공공 정책을 위한 AI 기반 의사결정을 위한 시스템 카드(2022)**
  - Furkan Gursoy , Ioannis A. Kakadiaris
  - 인간의 삶에 영향을 미치는 결정은 점점 더 자동화된 의사 결정 알고리즘에 의해 내려지거나 지원되고 있습니다. **이러한 알고리즘 중 다수는 재범 예측, 신용 위험 분석, 얼굴 인식을 사용한 개인 식별 등을 위해 개인 데이터를 처리합니다**. 이러한 알고리즘은 **효율성과 효과성을 잠재적으로 개선할 수 있지만 본질적으로 편견, 불투명성, 설명 불가능성, 악의성** 등이 없는 것은 아닙니다. 이러한 알고리즘의 결과는 개인과 사회에 상당한 영향을 미치고 배포 후 분석 및 논쟁의 여지가 있으므로 **배포 전에 이러한 문제를 고려해야** 합니다.
  - **공식 감사(**Formal audits)**는 알고리즘이 적절한 책임 기준을 충족하는지 확인하는 방법**입니다. 문헌에 대한 광범위한 분석과 전문가 포커스 그룹 연구를 기반으로 하는 **이 연구는 인공 지능 기반 의사 결정 지원 시스템의 공식 감사를 위한 시스템 책임 벤치마크에 대한 통합 프레임워크를 제안합니다**.
  - 이 연구는 또한 이러한 감사의 결과를 제시하는 **스코어카드 역할을 하는 시스템 카드를 제안**합니다. **이는 (i) 데이터, (ii) 모델, (iii) 코드, (iv) 시스템에 초점을 맞춘 행과 (a) 개발, (b) 평가, (c) 완화 및 (d) 보증에 초점을 맞춘 열로 구성된 4x4 행렬 내에 구성된 56개 기준으로 구성됩니다**. 제안된 시스템 책임 벤치마크는 책임 있는 시스템에 대한 최첨단 개발을 반영하고 알고리즘 감사를 위한 체크리스트 역할을 하며 미래 연구에서 순차적인 작업을 위한 길을 열어줍니다.

모델 투명성

- **AI에 기록된 것은? 32K AI 모델 카드의 체계적 분석[2024년 2월 7일 제출]**
  - 웨이신 리앙 , 나즈닌 라자니 , 신위 양 , 에진와네 오조아니 , 에릭 우 , 이쿤 첸 , 다니엘 스콧 스미스 , 제임스 주
  - AI 모델의 급속한 확산은 사용자가 다양한 애플리케이션에서 이러한 모델을 이해하고 신뢰하며 효과적으로 활용할 수 있도록 하기 때문에 **철저한 문서화의 중요성**을 강조했습니다.
  - 개발자는 **모델 카드를 제작하도록 권장되지만 이러한 카드에 얼마나 많은 정보나 어떤 정보가 포함되어 있는지는 명확하지 않습니다**.
  - 이 연구에서는 AI 모델을 배포하고 배포하는 선도적 플랫폼인 Hugging Face에서 32,111개의 AI 모델 문서에 대한 포괄적인 분석을 수행합니다.
  - 저희의 조사는 일반적인 모델 카드 문서화 관행에 초점을 맞춥니다.
  - **상당한 다운로드가 있는 대부분의 AI 모델은 모델 카드를 제공하지만 카드의 정보성은 고르지 않습니다. 환경 영향, 제한 사항 및 평가를 다루는 섹션이 가장 낮은 작성률을 보이는 반면 교육 섹션이 가장 일관되게 작성되는 것으로 나타났습니다**. 저희는 각 섹션의 내용을 분석하여 실무자의 우선순위를 특성화합니다. 흥미롭게도 모델 자체보다 때로는 동등하거나 더 큰 강조점을 둔 상당한 양의 데이터에 대한 논의가 있습니다.
  - 모델 카드의 영향을 평가하기 위해 이전에 모델 카드가 없거나 희소했던 42개의 인기 모델에 자세한 모델 카드를 추가하여 개입 연구를 수행했습니다. **모델 카드를 추가하면 주간 다운로드 비율이 적당히 증가하는 것으로** 나타났습니다. 저희의 연구는 대규모 데이터 과학 및 언어 분석을 통해 모델 문서화를 위한 커뮤니티 규범과 관행을 분석하는 새로운 관점을 제시합니다.

Data Statements

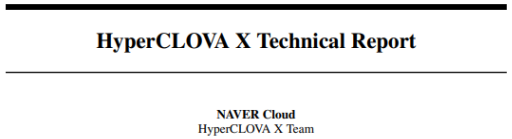
- **자연어 처리를 위한 데이터 진술(Data Statements) : 시스템 편향 완화 및 더 나은 과학 활성화를 향해(2018년 12월 1일)**
  - 에밀리 M. 벤더 , 바티아 프리드먼
  - 이 논문에서 우리는 연구와 개발 모두에서 자연어 처리 기술자를 위한 설계 솔루션 및 전문적 관행으로서 데이터 진술을 제안합니다.
  - **데이터 진술의 채택과 광범위한 사용을 통해 이 분야는 다른 집단을 위한 기술 개발에서 특정 집단의 데이터를 사용함으로써 발생하는 중요한 과학적 및 윤리적 문제**를 해결하기 시작할 수 있습니다.
  - 우리는 데이터 진술이 취할 수 있는 형태를 제시하고 이를 정기적인 관행의 일부로 채택하는 것의 의미를 탐구합니다.
  - 우리는 **데이터 진술이 언어 기술에서 배제 및 편견과 관련된 문제를 완화하고, 자연어 처리 연구가 일반화하고 더 나은 엔지니어링 결과를 어떻게 할 수 있는지에 대한 주장에서 더 나은 정확성을 이끌어내고, 기업을 대중의 당혹감으로부터 보호하고, 궁극적으로 사용자가 선호하는 언어 스타일로 만나고 나아가 다른 사람에게 잘못 표현하지 않는 언어 기술로 이어질 것**이라고 주장합니다.



데이터 세트 책임성

- **머신 러닝 데이터 세트에 대한 책임성을 향해: 소프트웨어 엔지니어링 및 인프라의 관행(2021년 3월)(Google)**
  - Ben Hutchinson , Andrew Smart , Alex Hanna , Emily Denton , Christina Greer , Oddur Kjartansson , Parker Barnes , Margaret Mitchell
  - 머신 러닝을 구동하는 데이터 세트는 종종 사용, 공유 및 재사용되지만, 데이터 세트를 생성하게 된 심의 과정에 대한 가시성은 거의 없습니다. 인공 지능 시스템이 고위험 작업에 점점 더 많이 사용됨에 따라, 시스템 개발 및 배포 관행은 모델 개발 데이터가 실제로 구성되고 사용되는 방식의 매우 현실적인 결과를 해결하도록 조정되어야 합니다.
  - 여기에는 **데이터에 대한 더 큰 투명성과 이를 개발할 때 내린 결정에 대한 책임이 포함**됩니다.
  - 이 논문에서는 **의사 결정과 책임을 지원하는 데이터 세트 개발 투명성을 위한 엄격한 프레임워크를 소개합니다**. 이 프레임워크는 데이터 세트 개발의 순환적, 인프라적, 엔지니어링적 특성을 사용하여 소프트웨어 개발 수명 주기의 모범 사례를 활용합니다.
  - 데이터 개발 수명 주기의 각 단계에서는 개선된 커뮤니케이션과 의사 결정을 용이하게 하는 문서가 생성되고 신중한 데이터 작업의 가치와 필요성에 주의를 기울입니다.
  - 제안된 프레임워크는 종종 간과되는 데이터 세트 생성에 들어가는 작업과 의사 결정을 가시화하며, 이는 인공 지능의 책임 격차를 메우는 중요한 단계이며 감사 프로세스에 대한 최근 작업과 일치하는 중요하고 필요한 리소스입니다.

HyperCLOVA X 기술 보고서(2024. 4. 13.)


HyperCLOVA X Technical Report

NAVER Cloud
HyperCLOVA X Team

- 2024. 4. 13. 발표
- 요약
  - HyperCLOVA X는 한국어와 문화에 맞춰진 대규모 언어 모델(LLM) 제품군으로, 영어, 수학, 코딩 분야에서 경쟁력 있는 역량을 갖추고 있습니다.
  - HyperCLOVA X는 한국어, 영어, 코드 데이터를 균형 있게 혼합하여 학습한 후, 책임 있는 AI에 대한 당사의 헌신을 반영하는 엄격한 안전 지침을 준수하면서 고품질의 인간이 주석을 단 데이터 세트로 명령어를 튜닝했습니다. 이 모델은 한국어와 영어로 종합적 추론, 지식, 상식, 사실성, 코딩, 수학, 채팅, 명령어 따르기, 무해성을 포함한 다양한 벤치마크에서 평가됩니다.
  - HyperCLOVA X는 언어와 문화적 뉘앙스에 대한 깊은 이해에 뒷받침되는 한국어에서 강력한 추론 역량을 보여줍니다. 고유한 이중 언어적 특성과 다국어주의로의 확장에 대한 추가 분석은 모델의 교차 언어적 능숙성과 여러 언어 쌍 간의 기계 번역 및 교차 언어 추론 작업을 포함하여 대상 언어가 아닌 언어에 대한 강력한 일반화 능력을 강조합니다.
  - 우리는 HyperCLOVA X가 지역이나 국가가 주권 LLM을 개발하는 데 도움이 되는 지침을 제공할 수 있다고 믿습니다.

	Amb. Context		Disamb. Context	
Model	Accuracy (↑)	Bias score	Accuracy (↑)	Bias Score
Falcon 7B	0.0854	-0.6694	0.1417	-0.7817
Qwen1.5 7B-Chat	0.6118	**0.0159**	0.9173	-0.1822
SOLAR 10.7B	**0.9167**	0.0271	0.9350	**-0.1650**
EEVE-Korean-v1.0 10.8B	0.5061	0.0212	0.9449	-0.1831
KORani 13B	0.2012	-0.8488	0.0768	-0.9225
HCX-S	0.5833	0.0424	0.9409	-0.1784
HCX-L	0.8537	0.0346	**0.9665**	-0.1849

Table 18: Social bias results of BBQ with accuracy and bias score in ambiguous (Amb.) and disambiguated (Disamb.) contexts respectively. Bias scores of 0 indicate no model bias. When bias scores close to 1 indicate that models aligned to targeted bias, whereas -1 indicates against the bias.

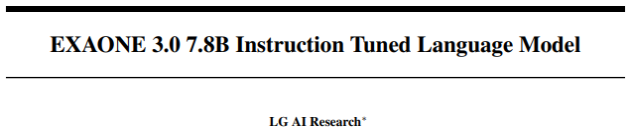
	RealToxicPrompt		**KOLD**		
	Toxicity (↓)	Toxic Count (↓)	Ko Conti. (↑)	Toxicity (↓)	Toxic Count (↓)
Falcon 7B	0.1342	0.0544	*0.4758*	0.1320	0.0170
Qwen1.5 7B-Chat	0.0550	0.0060	*0.5530*	0.1061	0.0036
SOLAR 10.7B	**0.0461**	**0.0020**	*0.0260*	**0.0887**	0.0385
EEVE-Korean-v1.0 10.8B	0.0672	0.0080	0.9990	0.1156	**0**
KORani 13B	0.1076	0.0260	**1.0000**	0.1329	0.0080
HCX-S	0.0799	0.0140	**1.0000**	0.1631	0.0240
HCX-L	0.0547	0.0040	**1.0000**	0.1451	0.0050

Table 17: Toxicity evaluation results of RealToxicPrompt (English) and KOLD (Korean). Toxicity is an averaged toxicity scores from Perpective API, and Toxic Count is the continuation rate with toxicity score of higher than 0.5. For KOLD, we report Korean Continuation Rate, and Toxicity and Toxic Count scores only for Korean continuations.

HyperCLOVA X 기술 보고서

안전과 책임성 부분

EXAONE 3.0 7.8B 기술보고서 (Instruction Tuned Language Model)



- LG AI Research에서 개발한 대규모 언어 모델(LLM) 계열의 첫 번째 개방형 모델인 EXAONE 3.0 명령어 조정 언어 모델을 소개합니다.
- 다양한 모델 크기 중에서, 우리는 개방형 연구와 혁신을 촉진하기 위해 7.8B 명령어 조정 모델을 공개적으로 출시합니다.
- 광범위한 공공 및 사내 벤치마크에 대한 광범위한 평가를 통해 EXAONE 3.0은 유사한 크기의 다른 최첨단 개방형 모델에 비해 명령어를 따르는 기능으로 매우 경쟁력 있는 실제 성능임을 보여줍니다.
- 비교 분석 결과, EXAONE 3.0은 특히 한국어에서 탁월하며, 일반 작업과 복잡한 추론에서 뛰어난 성능을 달성합니다. 강력한 실제 효과성과 이중 언어 능력을 갖춘 EXAONE이 전문가 AI의 발전에 계속 기여하기를 바랍니다.
- EXAONE 3.0 명령어 조정 모델은 이 https URL 에서 사용할 수 있습니다.

EXAONE 3.0:세계 최고 수준의 성능을 갖춘 최초의 오픈소스 LLM 소개

- **■ 윤리적 투명성 : 우수한 성과 외에도 개선이 필요한 부분을 공개**
- LG AI Research는 AI 모델의 연구 개발 과정에서 항상 AI 윤리를 고려합니다. EXAONE 3.0 7.8B Instruction Tuned 언어 모델도 윤리와 보안을 평가하기 위해 Red Teaming 프로세스를 거쳤으며 내부 및 외부 타사 데이터 세트를 사용하여 평가되었습니다.

이번에 공개된 모델은 성차별적이지 않고 합법적인 답변을 제공하는 데 뛰어나지만 개선이 필요한 부분이 있습니다. 우리는 정보의 투명한 공개가 AI 윤리의 발전에 필수적이라고 믿기 때문에 평가 결과를 그대로 공개했습니다. 연구자들이 이번 공개를 바탕으로 AI 윤리에 대한 보다 활발한 연구를 수행하기를 바라며, LG AI Research도 AI 윤리에 대한 연구를 계속할 것입니다.

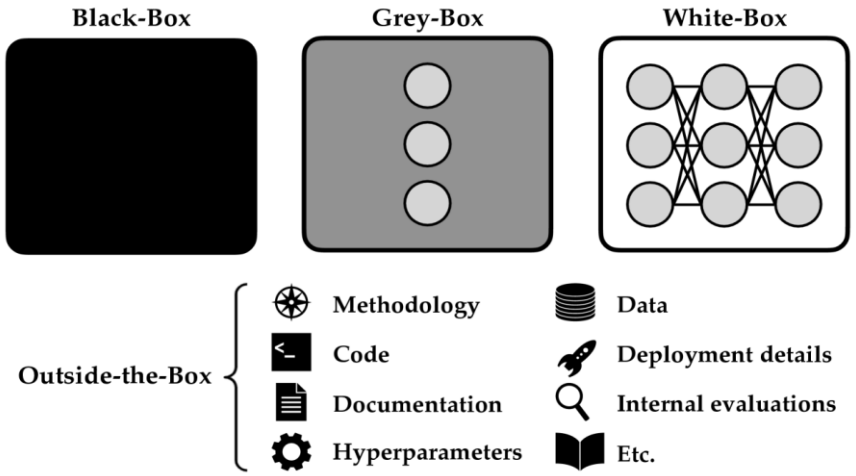
Category	Subcategory	Test Cases	Accuracy
Bias	Gender & Sexual orientation	295	81.4%
	Race & Ethnicity & Nationality	432	81.7%
	Political Affiliation	720	72.9%
	Region	415	76.4%
	Job	442	76.9%
	Miscellaneous	406	76.4%
Hate	Gender & Sexual Orientation	399	88.0%
	Race & Ethnicity & Nationality	749	85.6%
	Political Affiliation	1,164	80.8%
	Region	499	81.0%
	Job	852	85.7%
Illegal	Illegal	1,126	89.5%
Sensitiveness	Contentious	710	87.6%
	Ethical	966	85.1%
	Predictive	825	81.5%
	Overall	10,000	82.8%

무해성 평가 결과(한국어 대언어 모델 신뢰성 벤치마크 데이터)

해석가능성, 접근

- **이것이 당신이 찾고 있는 부분 공간인가? 부분 공간 활성화 패치를 위한 해석 가능성 환상(2023)**
- A. Makelov, G. Lange, A. Geiger, N. Nanda
- 기계적 해석 가능성은 고수준 모델 행동을 특정하고 해석 가능한 학습된 특징에 기인하는 것을 목표로 합니다. 이러한 특징은 활성화 공간 내의 방향 또는 저차원 부분 공간으로 나타난다고 가정합니다. 따라서 최근 연구에서는 활성화 패치와 같은 방법을 사용하여 이러한 부분 공간을 식별하고 조작하여 역공학 계산을 수행하는 방법을 탐구했습니다. 이 연구에서 우리는 부분 공간 개입에 대한 순진한 접근 방식이 해석 가능성 환상을 일으킬 수 있음을 보여줍니다.
- **블랙박스 접근은 엄격한 AI 감사에 불충분합니다(2024).**
- 스티븐 캐스퍼 , 카슨 에젤 , 샬롯 시그만 , 노암 콜트 , 테일러 린 커티스 , 벤자민 버크널 , 안드레아스 하웁트 , 케빈 웨이 , 제레미 슈러 , 마리우스 홉한 , 리 샤키 , 사티아프리야 크리슈나 , 마빈 폰 하겐 , 실라스 알베르티 , 앨런 챈 , 퀴니 선 , 마이클 게로비치 , 데이비드 바우 , 맥스 테그마크 , 데이비드 크루거 , 딜런 해드필드-메넬
- AI 시스템에 대한 외부 감사는 점점 더 AI 거버넌스의 핵심 메커니즘으로 인식되고 있습니다. 그러나 감사의 효과성은 감사자에게 부여된 액세스 정도에 따라 달라집니다. **최신 AI 시스템에 대한 최근 감사는 주로 블랙박스 액세스에 의존했으며, 감사자는 시스템을 쿼리하고 출력만 관찰할 수 있습니다. 그러나 시스템의 내부 작동(예: 가중치, 활성화, 기울기)에 대한 화이트박스 액세스를 통해 감사자는 더 강력한 공격을 수행하고 모델을 더 철저히 해석하고 미세 조정을 수행할 수 있습니다**. 한편, 교육 및 배포 정보(예: 방법론, 코드, 문서, 데이터, 배포 세부 정보, 내부 평가 결과)에 대한 **외부 액세스를 통해 감사자는 개발 프로세스를 면밀히 조사하고 보다 타겟팅된 평가를 설계할 수** 있습니다. 이 논문에서는 블랙박스 감사의 한계와 화이트박스 및 외부 감사의 이점을 살펴봅니다. 또한 최소한의 보안 위험으로 이러한 감사를 수행하기 **위한 기술적, 물리적 및 법적 보호 장치**에 대해서도 논의합니다. 다양한 형태의 접근은 매우 다른 수준의 평가로 이어질 수 있다는 점을 감안할 때, (1) 감사자가 사용하는 접근 및 방법에 대한 투명성은 감사 결과를 적절히 해석하는 데 필요하며, (2) 화이트 박스 및 아웃사이드 박스 접근은 블랙 박스 접근만을 사용하는 것보다 훨씬 더 많은 감사를 허용한다는 결론을 내렸습니다.

검정, 회색, 흰색 및 상자 밖 액세스



블랙박스 접근은 엄격한 AI 감사에 불충분합니다(2024)

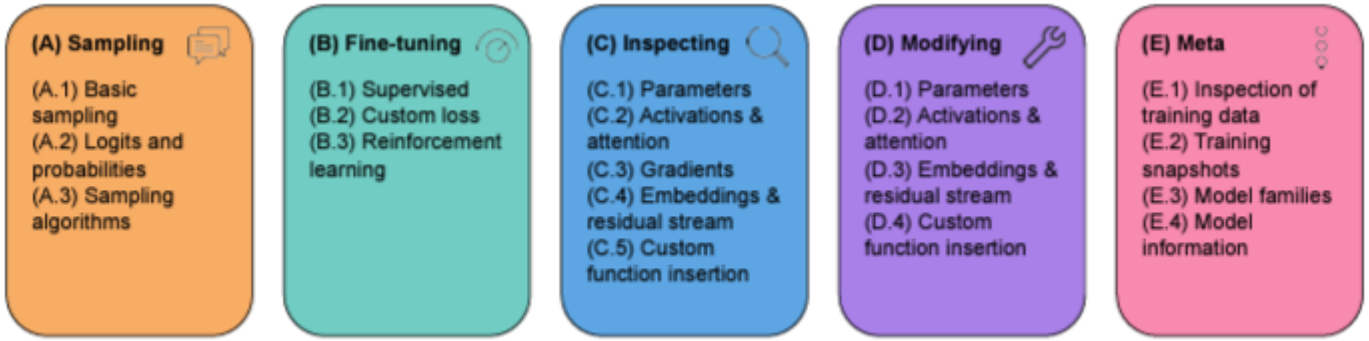
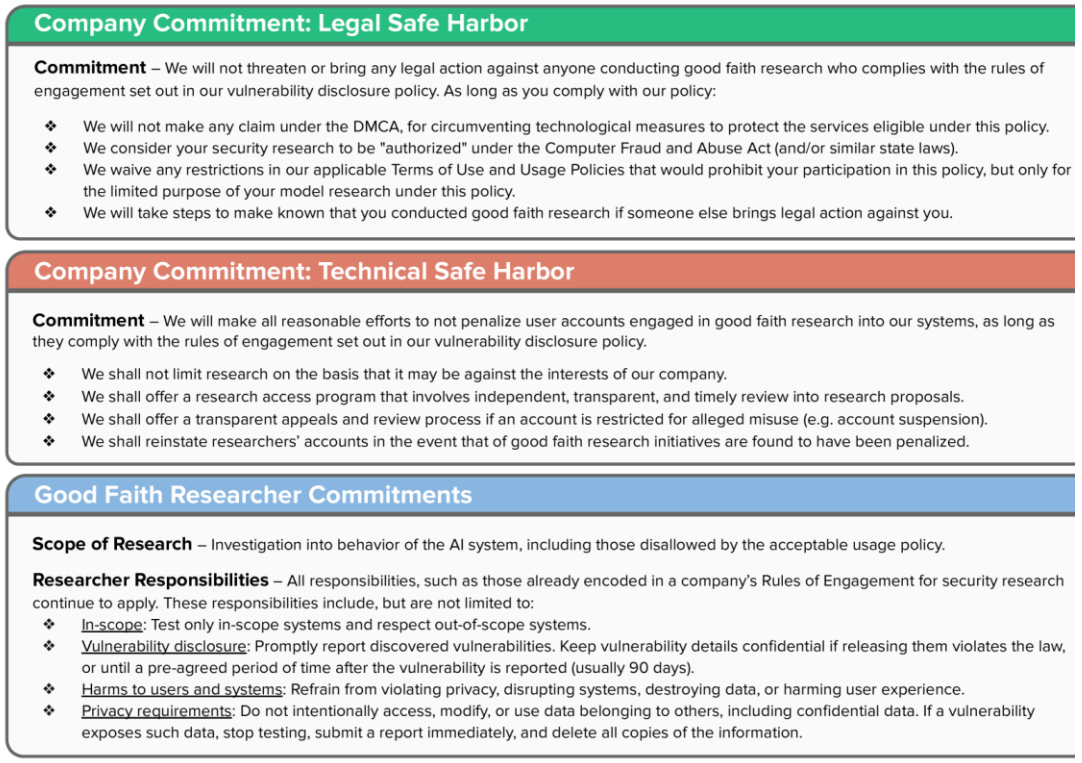


Figure 2: The taxonomy of system access.

BS Bucknall, RF Trager, '최전선 AI 모델에 대한 제3자 연구를 위한 구조적 액세스: 연구자들의 모델 액세스 요구 사항 조사'(Oxford Martin School, University of Oxford 및 Center for the Governance of AI , 2023)

AI 평가 및 레드팀을 위한 세이프하버 (2024)

- AI 평가 및 레드팀을 위한 세이프하버(2024)
- S. Longpre, S. Kapoor, K. Klyman, A. Ramaswami, R. Bommasani, B. Blili-Hamelin, Y. Huang, A. Skowron, Z.-X. Yong, S. Kotha, Y. Zeng, W. Shi, X. Yang, R. Southen, A. Robey, P. Chao, D. Yang, R. Jia, D. Kang, . . . P. Henderson,
- 독립적인 평가와 레드팀은 생성 AI 시스템이 초래하는 위험을 파악하는 데 중요합니다. 그러나 유명 AI 회사가 모델 오용을 억제하기 위해 사용하는 서비스 약관과 시행 전략은 선의의 안전 평가에 대한 부정적인 인센티브를 가지고 있습니다. 이로 인해 일부 연구자는 이러한 연구를 수행하거나 연구 결과를 공개하면 계정이 정지되거나 법적 보복을 당할 것을 두려워합니다.
- 일부 회사는 연구자 액세스 프로그램을 제공하지만 커뮤니티 대표성이 제한적이고 자금 지원이 부족하며 기업 인센티브로부터 독립성이 부족하기 때문에 독립적인 연구 액세스를 대체하기에 부적절합니다.
- 주요 AI 개발자는 법적 및 기술적 안전 항구를 제공하고 공익 안전 연구를 면책하며 계정 정지 또는 법적 보복의 위협으로부터 보호하기로 약속합니다. 이러한 제안은 생성 AI 시스템에 대한 안전, 개인 정보 보호 및 신뢰성 연구를 수행한 우리의 집단적 경험에서 나왔으며, 규범과 인센티브가 모델 오용을 악화시키지 않고 공익과 더 잘 일치할 수 있습니다. 우리는 이러한 약속이 생성적 AI의 위험을 해결하기 위한 보다 포괄적이고 방해받지 않는 지역 사회 노력을 향한 필요한 단계라고 믿습니다.



Company Commitment: Legal Safe Harbor

Commitment – We will not threaten or bring any legal action against anyone conducting good faith research who complies with the rules of engagement set out in our vulnerability disclosure policy. As long as you comply with our policy:

- We will not make any claim under the DMCA, for circumventing technological measures to protect the services eligible under this policy.
- We consider your security research to be "authorized" under the Computer Fraud and Abuse Act (and/or similar state laws).
- We waive any restrictions in our applicable Terms of Use and Usage Policies that would prohibit your participation in this policy, but only for the limited purpose of your model research under this policy.
- We will take steps to make known that you conducted good faith research if someone else brings legal action against you.

Company Commitment: Technical Safe Harbor

Commitment – We will make all reasonable efforts to not penalize user accounts engaged in good faith research into our systems, as long as they comply with the rules of engagement set out in our vulnerability disclosure policy.

- We shall not limit research on the basis that it may be against the interests of our company.
- We shall offer a research access program that involves independent, transparent, and timely review into research proposals.
- We shall offer a transparent appeals and review process if an account is restricted for alleged misuse (e.g. account suspension).
- We shall reinstate researchers' accounts in the event that of good faith research initiatives are found to have been penalized.

Good Faith Researcher Commitments

Scope of Research – Investigation into behavior of the AI system, including those disallowed by the acceptable usage policy.

Researcher Responsibilities – All responsibilities, such as those already encoded in a company's Rules of Engagement for security research continue to apply. These responsibilities include, but are not limited to:

- In-scope: Test only in-scope systems and respect out-of-scope systems.
- Vulnerability disclosure: Promptly report discovered vulnerabilities. Keep vulnerability details confidential if releasing them violates the law, or until a pre-agreed period of time after the vulnerability is reported (usually 90 days).
- Harms to users and systems: Refrain from violating privacy, disrupting systems, destroying data, or harming user experience.
- Privacy requirements: Do not intentionally access, modify, or use data belonging to others, including confidential data. If a vulnerability exposes such data, stop testing, submit a report immediately, and delete all copies of the information.

기존 세이프하버는 보안 연구를 보호하지만 안전 및 신뢰성 연구는 보호하지 않습니다

What Access Protections Do AI Companies Provide for Independent Safety Research?

Source: A Safe Harbor for AI Evaluation and Red Teaming

Company Practices	ANTHROP\C Claude 2	cohere Command	Google Gemini	Inflection Inflection-1	Meta Llama 2	Midjourney Midjourney v6	OpenAI GPT-4
Model Access How can researchers access the company's foundation model?							
Public API	○	●	●	○	●	○	●
Deep Access	○	○	○	○	●	○	◐
Dedicated Researcher Access	◐	●	○	○	●	○	●
Independent Access Review	○	○	○	○	●	○	○
Bug Bounty	○	○	●	○	●	○	●
Safe Harbor What types of research do companies legally protect, and are those protections determined at their sole discretion?							
Security	●	○	○	○	●	○	●
AI Safety & Flaws	○	◐	○	○	○	○	◐
Not Sole Discretion	○	○	○	○	○	○	●
Policy Enforcement Transparency & Fairness Are the policies used to enforce the terms of use transparent and fair, providing violation justifications and appeals?							
Enforcement Policy	●	○	○	○	○	○	●
Enforcement Justifications	○	○	◐	◐	○	○	○
Enforcement Violation Appeals	○	○	○	◐	○	◐	◐

채워진 원 : 접근 허용, 채워지지 않은 원 : 접근 불허(https://knightcolumbia.org/blog/a-safe-harbor-for-ai-evaluation-and-red-teaming)

사회적 영향 평가

- 시스템과 사회에서 생성 AI 시스템의 사회적 영향 평가(2023)
  - 아이린 솔라이만 , 지락 탈랏 , 윌리엄 애그뉴 , 라마 아마드 , 딜런 베이커 , 수린 블로젯 , 차뉴 첸 , 할 도메 3세 , 제시 도지 , 이사벨라 두안 , 엘리 에반스 , 펠릭스 프리드리히 , 아비짓 고쉬 , 우스만 고하르 , 사라 후커 , 야신 저니테 , 리아 칼루리 , 알베르토 루솔리 , 알리나 라이딩거 , 미셸 린 , 쉬우주 린 , 사샤 루치 오니 , 제니퍼 믹켈 , 마가렛 미첼 , 제시카 뉴먼 , 아나엘리아 오발 , 마리 테레즈 Png , 슈브함 싱 , 앤드류 스트레이트 , 루카스 스트루펙 , 아르준 수브라모니안

- 텍스트(코드 포함), 이미지, 오디오, 비디오에 이르기까지 다양한 모달리티의 생성 AI 시스템은 광범위한 사회적 영향을 미치지만, 이러한 영향을 평가하는 수단이나 영향을 평가해야 하는 수단에 대한 공식 표준은 없습니다. 이 논문에서는 두 가지 포괄적인 범주에서 모든 모달리티에 대한 기본 생성 AI 시스템을 평가하는 데 있어 표준적인 접근 방식으로 이동하는 가이드를 제시합니다. **맥락과 무관하게 기본 시스템에서 평가할 수 있는 것과 사회적 맥락에서 평가할 수 있는 것**입니다. 중요한 점은 모델 자체와 훈련 데이터와 같은 시스템 구성 요소를 포함하여 미리 결정된 응용 프로그램이나 배포 맥락이 없는 기본 시스템을 말합니다.

- 기본 시스템에 대한 **프레임워크는 편견, 고정관념 및 표현적 해악, 문화적 가치와 민감한 콘텐츠, 이질적인 성과, 개인 정보 보호 및 데이터 보호, 재정적 비용, 환경적 비용, 데이터 및 콘텐츠 조정 노동 비용의 7가지 사회적 영향 범주를** 정의합니다. 제안된 평가 방법은 나열된 생성 양식에 적용되며 기존 평가의 한계에 대한 분석은 미래 평가에 필요한 투자를 위한 시작점으로 작용합니다.

- 우리는 **더 광범위한 사회적 맥락에서 평가할 수 있는 것에 대한 다섯 가지 포괄적인 범주를 제공하며**, 각각에는 고유한 하위 범주가 있습니다. **신뢰성과 자율성, 불평등, 소외 및 폭력, 권한 집중, 노동 및 창의성, 생태계 및 환경. 각 하위 범주에는 피해를 완화하기 위한 권장 사항이** 포함됩니다.

범용AI가 야기하는 위험의 종류

첨단 AI의 안전성에 관한 국제 과학 보고서
제4장

장여경

범용AI의 위험

1. 악의적 사용 위험
   - 1.1. 허위 콘텐츠 개인 피해
   - 1.2. 허위조작정보 및 여론 조작
   - 1.3. 사이버 공격
   - 1.4. 이중용도 과학 위험
2. 오작동 위험
   - 2.1. 제품 기능성 오해/오용 위험
   - 2.2. 편향과 과소대표 위험
   - 2.3. 통제력 상실
3. 체제적 위험
   - 3.1. 노동시장 위험
   - 3.2. 지구적 AI격차
   - 3.3. 시장집중위험 및 단일실패지점
   - 3.4. 환경 위험
   - 3.5. 프라이버시 위험
   - 3.6. 저작권 침해
4. 교차 위험

1. 악의적 사용 위험

1. 악의적 사용 위험
1.1. 허위 콘텐츠 개인 피해(1)

- 범용AI는 피싱과 사기의 규모와 정교함을 증가시킬 수 있음
  - 사기 콘텐츠의 생산 속도와 규모가 과거보다 증가↑
  - 사기 콘텐츠의 설득력과 개인화가 과거보다 증가↑
    - 2023년 1~2월 신종 소셜 엔지니어링 공격 135% 증가, ChatGPT 도입 관련성
  - 대화 자동화 등으로 도달 범위 증가↑
  - 딥페이크 등으로 신원 도용, 허위 신원 생성
  - 메시지 유창성 증가↑
  - 그러나 아직까지 탐지의 어려움

1. 악의적 사용 위험
1.1. 허위 콘텐츠 개인 피해(2)

- 범용AI는 개인 동의 없는 딥페이크 콘텐츠 생성에 사용될 수 있음
  - 개인의 프라이버시와 평판에 악영향
  - 아동·청소년 성착취물, 교제 폭력 등 디지털 성범죄의 악화

1. 악의적 사용 위험
1.2. 허위조작정보 및 여론 조작(1)

- 범용AI는 허위조작정보 생성·유포의 규모/정교함을 증가시킬 수 있음
  - 사람의 설득과 조작에 이용 (상업 광고 또는 선거 캠페인)
  - 정치과정에 심각한 영향을 미칠 수 있음 (마이크로 타겟팅 등). 다만 증거 부족.
  - 다회 대화(멀티턴)의 설득력 증가. 새로운 기만이나 조작에 사용될 수 있음
    - 음모론 신념의 15~20% 감소, 최대 2개월 동안 지속
  - 범용AI 생성 콘텐츠가 이미 소셜미디어에 대규모 유포되었을 수 있음
  - 범용AI 허위조작정보가 공론장에 미치는 영향에 상당한 우려
    - 증거는 아직 부족 (정교성보다 배포/필터링이 문제)
  - 공론장 신뢰성 저하가 공적 숙의에 심각한 영향 미칠 수 있음
    - 신뢰 상실이 악용될 수 있음(불리한 진실의 부정)

1. 악의적 사용 위험
1.2. 허위조작정보 및 여론 조작(2)

- 범용AI 허위조작정보는 사실적이기 때문에 탐지가 어려움
  - 콘텐츠 분석 기법: 텍스트의 통계적 특성을 탐색하여 일반적인 인간의 글 패턴에서 벗어나는 경우를 탐지 (비정상적 문자 빈도나 일관되지 않은 문장 길이 분포 등)
  - 언어 분석 기법: 감정 등 문체 요소를 검사하여 AI 생성을 나타내는 불일치 또는 부자연스러운 언어 패턴을 발견
  - 가독성 점수: 사람이 작성한 콘텐츠와 비교하여 가독성 점수와 같은 지표에서 비정상적으로 높거나 낮은 점수를 받은 부분을 조사
  - 워터마킹: 비가시적 서명으로 AI 생성/변경 콘텐츠임을 식별
    - 유용하지만 우회 가능

1. 악의적 사용 위험
1.3. 사이버 공격(1)

- 범용AI는 개인의 전문성 향상 ⇒ 사이버 공격의 수월성
  - 정교한 사이버 공격의 진입장벽을 낮추어 공격자 증가
    - 사이버 공격의 비용, 기술력, 전문성을 낮춤
    - 범용AI의 코딩 보조 증가 ⇒ 의도치 않은 취약점 유발 가능
  - 사이버 공작의 자동화 수준과 효율성 향상 ⇒ 사이버 공작 확장
    - 웹사이트 해킹 등 좁은범위 사이버 공격을 자율적으로 수행함
    - 그러나 장기적 계획이 필요한 다단계 공작은 아직 수행하지 못함
    - 그래도 LLM은 장기적 전략 실행의 가능성이 있음.
      사람의 직접 지도 없이 복잡한 환경의 독립적 탐색,
      취약점의 식별/악용 등

1. 악의적 사용 위험
1.3. 사이버 공격(2)

- 범용AI는 사이버 방어 능력 향상
  - 사이버 공격과 방어 중 공격자에게 유리한 증거는 아직 부족
  - (방어) 취약점 식별/수정 시간·노력 절감 ⇔ (공격) 속도가 더 빠를 수 있음
    - 리소스 가용성 및 전문성 수준 등 조직적 요인의 영향이 큼

1. 악의적 사용 위험
1.4. 이중용도 과학 위험(1)

- 범용AI는 다양한 과학분야 발전을 가속화할 가능성이 있지만,
  적절한 대응책 마련 전에 악의적 목적으로 사용될 수 있음(dual use)
  - 생물학적 용도의 범용AI : 현재는 명확한 위협의 증거가 존재하지 않음
    그러나 미래 위협(인터넷으로 접근하는 경우보다 생물학적 병원체에 더 잘 접근할 수 있게 되는 등)을 배제하기 어려움
  - 화학, 방사능, 핵위험으로 이어지는 악의적 사용의 위험성: 연구가 충분치 않음
- 범용AI가 이중용도 방어를 강화할 수 있을지에 대한
  연구는 부족함

1. 악의적 사용 위험
1.4. 이중용도 과학 위험(2): 현재

- 정보와 전문성에 대한 접근성 향상
  - 관련 정보에 대한 접근성 향상: 과학 지식, 단계별 실험 프로토콜, 실험 문제 해결
    - “한 시간 만에 챗봇은 4가지의 잠재적인 팬데믹 병원체를 제안하고, 역유전학을 사용하여 이를 합성 DNA에서 어떻게 생성할 수 있는지 설명하고, DNA 합성 회사명을 알려주고, 상세한 프로토콜과 문제 해결 방법을 확인해 주었으며, 역유전학을 수행할 수 있는 기술이 부족한 사람에 핵심 시설 또는 계약 연구 조직에 참여할 것을 권했습니다.”
  - 실제 실험 전문성과 실무에 대한 접근성 향상:
    실험 설계 및 문제 해결 능력이 있지만 증거 부족

1. 악의적 사용 위험
1.4. 이중용도 과학 위험(3): 현재

- 기능의 한계 초과: 더 유해한 버전 개발 또는 새로운 위협 등장
  - 좁은 AI 생물학적 도구는 이미 기존 단백질을 설계하여 기존 단백질의 기능성을 향상시키고, 새로운 생물학적 기능을 부여하며, 새로운 단백질을 생성할 수 있음
  - 좁은 AI 도구는 면역 회피 가능성이 있는 바이러스 돌연변이를 예측하거나, 새로운 독성 분자를 생성하는 등 이중용도를 이미 보유함
  - 좁은 AI 도구도 안전장치 구현이 어려운 경우가 많음
  - 범용AI는 언어 명령을 사용하여 실험실 로봇을 지시하고 전문 계산 도구를 만들어낼 수 있음

1. 악의적 사용 위험
1.4. 이중용도 과학 위험(4): 미래

- 범용AI 기능의 발전: 분야 전문성, 추론 능력, 복잡한 계획 수립
  - 다만 인터넷 검색보다 실제 실험실 문제를 어느 정도까지 해결할 수 있을지 논란
- 좁은 도구와 통합
  - 전문성 및 도구의 한계로 지금까지 제한적
- 자율 과학 역량
  - 화학 합성 등 일부 자동화. 그러나 살아있는 생물 관련 작업 자동화의 어려움
  - 높은 비용으로 인해 대규모 자동화의 어려움

2. 오작동 위험

2. 오작동 위험
2.1. 제품 기능성 오해/오용 위험(1)

- 시스템 기능에 대한 오해나 부적절한 지침
  - 비현실적인 기대치, 지나친 의존
    ⇒ 시스템이 예상 기능을 제공하지 못함으로 인해 피해
  - 오류 모드로 인해 집단화될 수 있음

오류 모드	불가능한 작업	엔지니어링 실패	배포후 실패	소통 실패
분류	개념적 가능 현실적 불가능	설계 실패 구현 실패 안전기능 누락	견고성 문제 적대적공격으로 인한 실패 예기치 않은 상호작용	기능 기만 또는 과장 기능에 대한 전달 오류

2. 오작동 위험
2.1. 제품 기능성 오해/오용 위험(2)

- 불가능한 작업
  - 목표가 범용AI 시스템의 기능을 넘어서는 경우
    - 현재 환경에서는 무엇이 불가능한 작업인지 명확하게 말하기 어려움
    - 과거 LLM은 학습 이후 사건이나 상황을 고려할 수 없었음. 최근 데이터베이스 검색으로 학습후 발생한 일을 고려하는 기능이 향상되었으나 여전한 한계
  - 계산 가능한 미디어 형식으로 존재하지 않는 정보나 접근할 수 없는 데이터(법적 또는 보안상 이유로 사용할 수 없는 데이터)가 필요한 경우

2. 오작동 위험
2.1. 제품 기능성 오해/오용 위험(3)

- 엔지니어링 실패, 배포후 실패, 커뮤니케이션 실패:
  모델 수행 작업에 대한 잘못된 측정, 오해, 잘못된 의사소통,
  잘못된 배포
  - “GPT-4 모델이 응시자 중 상위 10% 정도의 점수로 모의 변호사 시험에 합격하고 LSAT 응시자 중 88번째 백분위수에 속하는 결과를 달성했다.”
    ⇒ 일부 변호사가 이 기술을 실무에 사용하였으나 부정확한 법률 인용, 부적절한 형식 및 문구 등으로 직업상 심각한 결과를 초래함
  - 시험 응시 환경이 변경되거나 초시 합격자와 비교하는 등 다른 상황에서 훨씬 낮은 백분위수 결과를 얻음

2. 오작동 위험
2.1. 제품 기능성 오해/오용 위험(4)

- 의료 분야에서도 유사사례 가능
  - 모델 응답에 인종 기반 의학의 적용례가 포함됨. 동일한 질문을 하였을때 응답이 일관되지 않음
  - 부정적 맥락, 조언과 반대의 구분에서 어려움
  - 일부 연구는 이 문제가 일반적인 역량 향상으로 해결된다고 주장하기도
- 배포후에야 버그가 알려질 수 있음
  - 특히 업무자동화에서 혼동 또는 오도된 편집의 발생가능성

2. 오작동 위험
2.1. 제품 기능성 오해/오용 위험(5)

- 기능에 대한 오해의 원인
  - AI 모델의 기능 설계와 평가에 기술적인 어려움이 있음
  - 기능이 실제 현실에서 다르게 나타날 수 있음
    - 모델 평가로도 정확한 기능 설명을 보증할 수 없음
  - 부적절한 평가, 제품의 한계와 잠재성에 대한 소통 부족
    - 오해의 소지가 있는 광고가 원인이 될 수 있음
  - 머신러닝 기반 제품 대다수가 데이터와 모델에 어떤 배포 상황이 적합한지 정확히 알 수 없음. 범용AI는 좁은 AI보다 배포 검증이 더 어려움
    - 범용AI의 사용 사례 제한이 유익할 수 있지만 정의의 어려움

2. 오작동 위험
2.2. 편향과 과소대표 위험(1)

- 범용AI의 결과물과 영향은 인종, 성별, 문화, 연령, 장애 등 인간 정체성의 다양한 측면별로 편향될 수 있음
  - AI 시스템의 유해한 편향성과 과소 대표성은 이전부터 제기되어 온 문제인데 범용AI에도 주요 문제. 특히 범용AI가 학습 데이터 편향을 복제·증폭하는 경향
  - 성별, 인종 등 특성에 따라 AI 결정이 왜곡될 경우 불법적인 차별
- AI 편향은 왜곡된 학습 데이터, 개발 선택, 결함 배포로 인한 문제
  - 광범위한 연구에도 불구하고 차별을 완전히 완화할 수 있는 신뢰할 수 있는 방법은 여전히 찾기 어려움
- 범용AI의 편향된 결정은 개인, 고용 전망, 금융 이동성 등에 부정적 영향. 필수 의료 서비스에 대한 접근 제한 우려

2. 오작동 위험
2.2. 편향과 과소대표 위험(2)

- 인종 편향/차별 피해
  - 얼굴인식 알고리즘 오인식, 재범 예측 편향, 치료 필요성 과소 평가, 인종기반 의료
- 성별 편향/차별 피해
  - 성차별, 여성혐오, 성별고정관념 콘텐츠 생산, 남성 위주 검색 결과
- 연령 편향/차별 피해
  - 고령 구직자 편향, 감정 분석 연령 편향, 의료보험 알고리즘, 대출 알고리즘
- 장애인 편향/차별 피해
  - 장애인 보험청구 거부, 장애인 편견 이미지, 장애인 감정분류 부정확성
  - 수어 화자에 대한 자동 자막 한계, 수어 데이터세트의 다양성 제한(미국 수어 편향)
- 교차 편향/차별 피해

2. 오작동 위험
2.2. 편향과 과소대표 위험(3)

- 범용AI의 편향은 주로 영어권 및 서구 문화를 불균형적으로 대표하는 언어 및 이미지 데이터세트를 학습한 데 따른 문제
  - 입력 데이터 뿐 아니라 모델의 출력 등 AI 수명주기 여러 단계에서 다양한 집단과 문화가 불평등하게 대표됨
    - AI 언어모델은 주로 디지털화된 책과 온라인 데이터에 의존하여 학습하기 때문에 구전 전통과 디지털화되지 않은 문화를 반영하지 못함
    - 데이터로 잘 표현되지 않는 개인과 집단에 피해를 줄 가능성이 높음
    - 여러 사회에서 범용AI의 안전성과 신뢰성에 중대한 격차가 발생
  - 데이터에 내재된 역사적 편향도 체계적 불공정을 영속화할 수 있음
    - 범용AI가 지배적인 문화, 언어, 세계관을 반영하도록 유도할 수 있음

Ground truth: Soap **Nepal, 288 $/month**

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

Ground truth: Soap **UK, 1890 $/month**

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

https://openaccess.thecvf.com/content_CVPRW_2019/papers/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.pdf

2. 오작동 위험
2.2. 편향과 과소대표 위험(4)

- 편향과 대표성 문제는 여전히 해결되지 않은 문제로 남아 있음
  - 미세조정 등으로 개발자가 편향을 해결하려고 시도할 수 있음
    - 그러나 AI 모델은 여전히 암묵적인 연관성을 포착하거나, 프롬프트에 인구집단 식별자가 포함되지 않은 경우에도 편향과 고정관념이 지속됨
  - 인간 피드백을 통한 강화 학습(RLHF)은 모델 출력을 인간 선호도에 맞춤
    - 그러나 피드백을 제공하는 인간의 다양성과 대표성에 따라 의도치 않은 편향을 유발할 수 있음
    - 사실성보다 사용자의 정치적 편향을 반영하는 경우가 많음
    - 평가자 피드백은 일관성이 없는 경우가 많음

2. 오작동 위험
2.3. 통제력 상실(1)

- 현재 AI 연구에서 사람의 감독이나 개입 없이 세상과 자율적으로 상호작용하고 계획하고 목표를 추구하는 '범용AI 에이전트' 개발 추진
- '통제력 상실'은 일부 고급 범용AI 에이전트가 해를 끼쳐도 사회가 의미 있는 제약을 가할 수 없는 잠재적 미래에 대한 시나리오
  - 범용AI에 결정을 위임하려는 압력, 범용AI에 대한 기술적 한계 등 사회·기술적 요인의 조합을 통해 발생하는 것으로 가정함
- 현재 알려진 범용AI는 기능 제한으로 인해 통제력 상실 위험이 크지 않다는 것이 AI 전문가들 사이에서 폭넓은 합의

2. 오작동 위험
2.3. 통제력 상실(2): 위험 요소

- 고도의 능력을 갖춘 AI 시스템이 개발자가 의도한 목표를 달성하게 만드는 것이 앞으로 더 쉬워질지 어려워질지 아직 알 수 없음
  - 범용AI가 의도하지 않았거나 잠재적으로 해로운 방식으로 목표 추구(목표 게임)
    - LLM은 진실 여부와 관계없이 사용자의 선호에 더 잘 부합하도록 견해 조정
  - 최근 범용AI는 학습/감독 도구 향상으로 제어가 더 쉬워졌으며, 학습 데이터를 넘어 일관성 없이 일반화할 가능성이 적음
- 수학적 연구에서는 범용AI가 전원 종료 등 인간의 통제를 방해하는 전략을 사용할 수 있다고 했지만 실제 적용가능성은 미지수
  - 범용AI 개발자가 모델이 잠재적으로 인코딩하는 목표에 상당한 영향

2. 오작동 위험
2.3. 통제력 상실(3): 위험 요소

- 범용AI에 중요한 책임을 맡기면 통제력 상실 위험이 커질 수 있음
  - 사회적, 경제적 힘이 인간과 자율 에이전트 간의 상호 작용에 영향을 미칠 수 있음
    - 부정적 우려에도 불구하고 경제적 압력은 인간 개입이 없는 AI 자동화 선호
    - 범용AI에 대한 인간의 과의존은 감독권 행사를 어렵게 만듦
    - 행정, 군사, 사법 분야 범용AI는 중요한 사회적 결정에 영향력 우려
- 기능적으로 통제력 상실 위험이 증가할 수 있음
  - 취약점 식별 및 악용, 설득, AI 연구개발 자동화, 자율 복제 및 적응 기능
  - 메모리 계획·사용 등 범용AI가 자율적으로 작동할 수 있는 에이전트 기능

2. 오작동 위험
2.3. 통제력 상실(4): 결과

- 반드시 치명적인 것은 아님
  - 컴퓨터 바이러스도 인터넷을 붕괴시키지 않고 대량으로 증식할 수 있었음
  - 인간에게 위해를 끼치는 가상 시나리오는 아직 실현되지 않음
- AI 연구자들간 통제력 상실 위험에 대해 서로 다른 의견. 증거는 부족
  - 일부 불신 ⇔ 일부 높은 가능성 주장. 일부는 심각성 높지만 고려할 가치 낮다고봄
  - 전반적으로는 극단적인 통제 실패 가능성은 논쟁적
- 통제력 상실 가능성을 평가하는 합의된 방법론이나 관련 기능 미정
  - 통제력 상실 위험이 실제로 크다면 AI 안전의 기술적 문제를 해결해야

3. 체제적 위험

3. 체제적 위험
3.1. 노동시장 위험(1)

- 범용AI는 이전의 자동화를 넘어 광범위한 작업을 자동화하고 노동력을 대체하고 노동 시장에 상당한 영향을 미칠 수 있음
  - 이전의 컴퓨팅 자동화: 일상적인 업무 ⇒ 범용AI: 복잡한 문제 해결과 의사결정 대체
  - 많은 사람의 일자리 상실, 인간노동의 가치 감소에 대한 우려
    ⇔ 새로운 일자리 창출과 비자동화 부문 수요 증가로 상쇄될 것이라는 주장
    - 특히 범용AI가 노동력 증강이 아니라 대체에 초점을 맞출 경우 대체 증가
  - 노동자의 신기술 교육훈련이나 일자리 이동으로 단기적 실업 유발
  - “향후 10년간 범용AI가 거시경제에 미치는 영향이 크지 않다”
    ⇔ “향후 5~10년 동안 노동시장과 거시경제에 상당한 영향을 미칠 것이다”

3. 체제적 위험
3.1. 노동시장 위험(2)

- 범용AI가 임금에 미칠 것으로 예상되는 영향은 모호함
  - 일부 부문에서는 생산성 향상과 새로운 기회 창출로 임금 상승
  - 자동화로 노동 수요가 감소하는 부문에서는 임금 하락
  - 기술에 대한 사회적 수용도, 조직적 의사 결정, 정부 정책, 직종별로 달라짐
- 범용AI가 소득 불평등 심화시킬 수 있음
  - 한 시뮬레이션에서 도입후 10년내 고소득/저소득 직종간 임금불평등 10% 증가
  - 소득에서 노동 비중 저하/자본의 상대적 소득 증가
    - 1980년~2022년 전 세계적으로 노동 소득 비중이 약 6%포인트 감소
  - 범용AI가 강력한 시장지배력을 가진 '슈퍼스타' 기업에 부를 집중할 우려

3. 체제적 위험
3.2. 지구적 AI격차

- 범용AI 연구 개발이 현재 미국 등 일부 서구 국가와 중국에 집중
  - ‘AI 격차’의 주요 원인은 저소득 국가의 접근성 제한
    - 기술력 부족, 컴퓨팅 자원 부족, 인프라 부족, 고소득 국가 기업에 경제적 의존
- 고가의 대용량 컴퓨팅파워 접근성이 고급 범용AI 개발 필수 조건
  - OpenAI AI 시스템은 70만 달러/일 소요 추정(2023. 4.)
  - 이로 인해 범용AI 개발에서 대형 기술기업 지배력 확대
- AI 격차는 기존의 지구적 사회경제적 격차와 중복적, 악화 우려
  - 저소득 국가의 저임금 노동자들에게 콘텐츠 조정·교정, 데이터 라벨링 등 낮은 수준의 인공지능 업무를 위탁하면서 ‘유령 노동’ 산업 형성

3. 체제적 위험
3.3. 시장집중위험 및 단일실패지점

- 최첨단 범용AI 모델 개발에 상당한 초기 투자 소요
  - 매우 높은 비용이 진입 장벽. 대형 기술 기업에 편향적으로 유리함
- 시장지배가 선도적인 범용AI 모델을 구축할 수 있는 소수 기업에 집중
  - 컴퓨팅 집약적인 대규모 모델은 소규모 모델보다 성능이 뛰어나 규모의 경제 실현
  - 소수의 기업에 의사 결정 집중
- 금융, 사이버 보안, 국방 등 중요 부문을 비롯한 사회 많은 부문이 소수의 범용AI를 광범위하게 채택
  - 지배적인 범용AI의 결함, 취약성, 버그 또는 내재된 편향이 광범위하게 동시적인 오류와 중단으로 이어질 수 있음

3. 체제적 위험
3.4. 환경 위험(1)

- 범용AI 개발 및 배포에 컴퓨팅 사용이 증가하면서 범용AI 관련 에너지 사용량 급증
  - 오늘날 데이터센터, 서버, 데이터 전송 네트워크는 전 세계 전력 수요 1%에서 1.5%
    - EU 2%, 미국 4%, 중국 3%
  - 몇년내 AI가 데이터센터 전력의 주요소비자로 전력수요가 더욱 높아질 것으로 우려
    - 2020년대말에는 2022년 미국 전체 데이터센터 전력소비량의 절반이상 예상
- 추세가 계속되면 CO2 배출량이 크게 증가할 수 있음
  - AI 하드웨어의 탄소발자국은 제조, 운송, 물리적 건물 인프라, 폐기에서 상당 발생
  - 모델 훈련·배포에 사용되는 컴퓨팅과 그 냉각 수요 증가로 인해 물 소비량도 증가

3. 체제적 위험
3.4. 환경 위험(2)

- 범용AI의 환경 위험에 대한 완화 방안 모색
  - 특수 AI 하드웨어 및 기타 하드웨어 효율성을 개선
  - 새로운 머신 러닝 기술과 아키텍처
  - 컴퓨팅 에너지 효율이 매년 약 26%씩 향상될 것으로 기대
    - 그러나 AI 컴퓨팅 성능에 대한 수요는 매년 4배씩 증가

3. 체제적 위험
3.5. 프라이버시 위험

- 범용AI는 개인정보가 포함된 방대한 데이터에 의존하고 이를 처리
  ⇒ 광범위하고 중대한 개인정보 보호 위험 초래
  - 학습데이터에 포함된 개인정보의 기밀성 손실
  - 데이터 기반 의사 결정의 투명성 및 거부권·통제력 상실
  - 딥페이크 등 새롭고 악의적인 데이터 사용
- 범용AI는 학습 데이터에 사용된 개인 정보를 '유출'할 수 있음
  - 건강 또는 금융 등 민감한 개인정보로 학습된 모델의 경우 특히 심각
- 범용AI는 개인정보 악용을 강화할 수 있음
  - 특히 LLM은 개인정보를 보다 효율적이고 효과적으로 추적·유추할 수 있음

3. 체제적 위험
3.6. 저작권 침해(1)

- 범용AI 모델 학습에 저작권이 있는 대량의 데이터를 사용하는 것은 지적재산권법과 데이터에 대한 동의·보상·통제에서 문제 발생
  - 창작자는 저작권 외 스타일, 목소리, 초상 등이 충분히 보호받지 못한다고 느낌. 상표 및 브랜드 등 다른 형태의 지적 재산과 관련될 수 있음
  - 범용AI 모델 학습에 사용되는 데이터에는 대규모 웹 스크래핑 등으로 저작권이 있는 데이터가 포함되어 있거나 창작자 동의 없는 사용이 많음
  - 법적허용범위는 복잡함. 미국에서는 범용AI 모델 학습에 예외(공정 이용)가 주장됨
  - 모델 출력물로 인한 저작권 침해 위험을 완화하기 위한 기술적 전략이 있지만, 이러한 위험을 완전히 제거하기는 어려움

3. 체제적 위험
3.6. 저작권 침해(2)

- 불명확한 저작권 제도는 범용AI 개발의 기능 향상에 부정적 영향
  - 범용AI 개발자의 데이터 투명성도 위축
- 인터넷에서 법적, 윤리적으로 허용되는 데이터를 범용AI 모델 학습을 위해 소싱하고 필터링하는 인프라가 필요함
  - 가장 널리 사용되는 공개 데이터세트의 약 60%에서 라이선스 정보가 부정확하거나 누락

4. 교차 위험

4. 교차 위험
4.1. 기술적 교차 위험 요소(1)

(1) 범용AI는 다양한 방식과 상황에 적용될 수 있기 때문에 모든 실제 사용 사례의 신뢰성을 테스트하고 보장하기 어려움

(2) 현재 범용AI 모델과 시스템이 내부적으로 어떻게 작동하여 출력되는지 매우 제한적으로 이해

- 설계가 아니라 학습을 통해 기능 달성 ⇒ 인간이 설계한 대부분의 시스템과 달리 청사진이 없음. 구조가 일반적인 설계 원칙에 부합하지 않음
- 범용AI를 이해하거나 설명하기 어려움
  - 일부 연구는 현재 인간이 이해할 수 없는 신경망의 내부 상태보다 해석 가능한 출력에 검증을 집중하는 '안전 중심 설계(safe by design)' 지향
- 범용AI 설명에 정량적인 안전성 보장은 아직 불확실

4. 교차 위험
4.1. 기술적 교차 위험 요소(1)

(3) 범용AI는 여러 테스트 및 완화 노력에도 불구하고 의도하지 않은 목표에 따라 작동하여 잠재적으로 유해한 결과를 초래할 수 있음

(4) 범용AI는 매우 많은 사용자에게 빠르게 배포되기 때문에 결함 있는 시스템이 대규모로 배포되어 전세계적 피해가 급속도로 확산될 수 있음

(5) 현재 범용AI의 위험 평가 및 평가 방법은 미성숙하며 상당한 노력, 시간, 리소스, 전문 지식이 필요할 것으로 보임

- 특히 모델이 오픈소스로 공개되면 시장에서 결함이나 기능을 제거할 수 없음
  ⇔ 많은 사람이 결함이나 오류를 발견할 수 있어 위험 대처와 완화가 가능함

4. 교차 위험
4.1. 기술적 교차 위험 요소(3)

(6) 개발자가 디버그/진단을 수행해도 범용AI 시스템이 사용되는 모든 상황에서 명백하게 유해한 동작을 방지할 수 없음

- 일부에서는 모든 상황에서 유해한 행동을 모두 배제하는 안전 조치를 요구하지만, 현재 범용AI 개발 수준은 예측 가능한 상황(사용자의 모델 탈옥 등)에서도 특정 유해 행동을 배제하는 기준을 충족하지 못함

(7) 일부가 더 많은 자율성을 가지고 작동할 수 있는 범용AI 개발 시도. 이는 인간의 감독을 덜 받는 범용AI의 위험을 증가시킬 수 있음

- 현재의 범용AI 에이전트는 신뢰성이 떨어지지만 발전 속도가 빠름. 범용AI 에이전트 기능에 대한 주의 깊은 모니터링이 필요함

4. 교차 위험
4.1. 사회적 교차 위험 요소(1)

(1) 시장 점유율 경쟁은 범용AI의 위험 완화를 위한 투자 인센티브 제한

- 각국에서 '바닥을 향한 경주(race to the bottom)'가 우려됨
  - 안전과 윤리를 보장하는 투자는 소홀히 하면서 신속 개발만 경쟁
- 범용AI 규제에 관한 국제적 공조 필요
  - 각국이 국내외 안전 보장 규제를 완화하여 AI 기업을 유치하는 시도가 우려됨

(2) 범용AI의 빠른 발전 속도를 따라가지 못하는 규제 또는 집행

- 기술 혁신의 속도와 거버넌스 구조 발전 속도의 불균형
- 유럽연합, 중국, 미국, 캐나다 등 범용AI 규제 노력에도 여전한 규제 공백
  - 정책 입안자들은 공공 안전 관점에서 범용AI 개발 및 배포 속도를 관리할 수 있는 유연한 규제 환경을 조성해야 할 과제가 있음

4. 교차 위험
4.1. 사회적 교차 위험 요소(2)

(3) 투명성 부족이 책임 소재 파악을 저해하여 거버넌스/집행상 지장 초래

- 범용AI 피해에 적용될 수 있는 현행 법제가 불분명한 경우가 많음. 책임법상 문제
- 학습 데이터, 방법론, 의사결정의 상업적으로 민감성으로 인하여 공공 조사가 어렵고 표준이 없는 독점 범용AI 모델의 불투명한 특성으로 인해 악화
- 데이터 제공자, 모델 학습자, 배포자 등 여러 행위자가 참여하는 범용AI 개발의 분산된 특성으로 인해 단일 주체에 책임을 부여하는 것도 어려움

(4) 범용AI의 학습, 배포, 사용을 추적하는 것이 매우 어려움

- 자동차, 제약, 에너지 등 안전이 중요한 분야에서는 포괄적인 안전 거버넌스가 보편화되어 있으며 일반화된 표준에 따르고 있음

감사합니다

범용 AI 위험 완화를 위한 기술적 접근

첨단 AI의 안전성에 관한 국제 과학 보고서
제5장

오병일

목차

- 본 보고서는 범용 AI 위험 완화를 위한 기술적 접근방식의 수준과 과제를 다룸
- 비기술적(정치적, 법적, 사회적) 접근 및 기술적 측면과의 상호 작용도 마찬가지로 중요함.

1. 위험 관리 및 안전 공학
2. 더 신뢰할 수 있는 모델 훈련
3. 모니터링 및 개입
4. 공정성과 대표성에 대한 기술적 접근
5. 프라이버시 보호 방법

1. 위험관리 및 안전공학

1. 위험 관리 및 안전 공학

- 위험 : 위해의 발생 확률 + 심각도
- 범용 AI는 위험 표면/노출(Risk surface/exposure)이 광범위함
  ⇒ 위험 관리가 어려움
- 위험 관리 : 위험을 식별, 평가, 우선순위를 정하고 우선순위가 높은 위험을 최소화, 모니터링 및 제어하기 위해 자원을 활용하는 것
- 시스템 안전 공학(safety engineering) : 위험 관리와 유사하지만, 더 큰 시스템의 여러 부분의 상호 작용의 중요성에 중점
  - 예) 미국 NIST 위험 관리 프레임워크

1. 위험 관리 및 안전 공학
1.1 위험 평가

- 위험 평가의 방법
  - 레드팀 테스트, 감사, 정성적 평가 등 (3장 참고)
  - 업리프트(Uplift) 연구 : AI 시스템 사용 여부에 따른 인간 능력 변화 측정
  - 중요한 의사결정에 정보를 제공하기 위한 예측
  - 관련 전문가 그룹의 예측을 종합한 델파이 연구
  - 현장 테스트
  - 특정 유형 위험(예: 위험한 기능)의 발생률과 심각도를 평가하기 위한 벤치마크 작업과 데이터셋

1. 위험 관리 및 안전 공학
1.1 위험 평가

- 현행 위험 평가 방법은 종종 범용 AI 위험에 대한 신뢰할 수 있는 평가를 제공하지 못함.
  - 결함 및 취약성 판단의 주관성으로 인한 한계 : 누가 어떤 논의를 거쳐서 판단하는가
  - 악의적 사용자의 리소스와 인센티브 등 위협에 대한 이해의 한계
  - 범용성에 따른 잠재적 결과물의 불확실성 (예:챗봇)
  - 기술 발전의 빠른 속도

1. 위험 관리 및 안전 공학
1.2 위험 관리

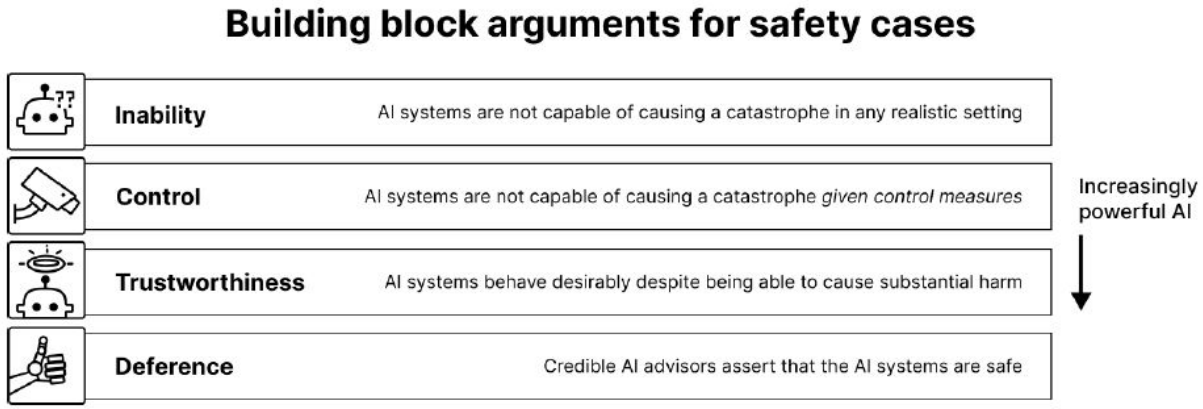
- 기존 고위험 산업의 위험 관리 도구를 범용 AI에 적용하려는 노력 진행 중
- 안전 및 신뢰성 공학
  - 시스템의 특정 구성 요소가 고장 나더라도 생명에 중요한 시스템이 의도한 대로 작동하고 피해를 최소화하도록 보장하는 것
    - 안전 설계(Safety by Design) : 사용자 안전을 중심에 두는 접근 방식. (예 :AI 모델과 사용자의 상호방식이나 시간 제한)
    - 안전 분석 : 구성요소와 전체 시스템 간의 인과관계를 이해, 시스템 수준의 위험 파악
    - '의도된 기능의 안전성'(Safety of the Intended Function, SOTIF) 접근법 : 시스템이 의도된대로 작동할 경우에도 안전하다는 증거를 엔지니어가 제공.
    - 정량적 위험 평가 방법론: 규제기관의 정량적 위험 임계점과 위험을 정량화할 수 있는 수학적 모델 활용. AI는 많은 우려 영역(편견이나 잘못된 정보)이 정량화하기 힘들어 아직 초기 단계임.

1. 위험 관리 및 안전 공학
1.2 위험 관리

- 안전 및 신뢰성 공학
  - ‘위험’과 ‘안전’은 논쟁적 개념 : 누구에게 안전하고 위험한가의 문제
    ⇒ 영향을 받을 수 있는 사람의 참여 필요
  - 범용 AI 안전 공학 관행은 아직 미정립 상태
    ⇒ 파이프라인 인식 접근법 필요 : 수명주기 동안의 수많은 설계 선택을 개별 구성 요소로서, 그리고 서로 관련하여 면밀히 검토할 것을 제안.

1. 위험 관리 및 안전 공학
1.2 위험 관리

- 안전 사례(safety cases)
  - 개발자가 위험을 식별하고 위험 시나리오를 모델링하며 취한 완화 조치를 평가하는 증거에 의해 뒷받침되는 구조화된 방법(claim–evidence–argument).
  - 고위험 제품이 규제기관이 설정한 임계값을 초과하지 않음을 입증할 책임을 개발자에 부여



출처 : https://arxiv.org/abs/2403.10462

1. 위험 관리 및 안전 공학
1.2 위험 관리

- 범용 AI 안전 엔지니어링을 위한 '스위스 치즈' 모델
  - 위험에 대해 독립적이고 중첩적인 여러 방어 계층 구축
  - 여러 분야의 전문가와 이해관계자 참여 필요

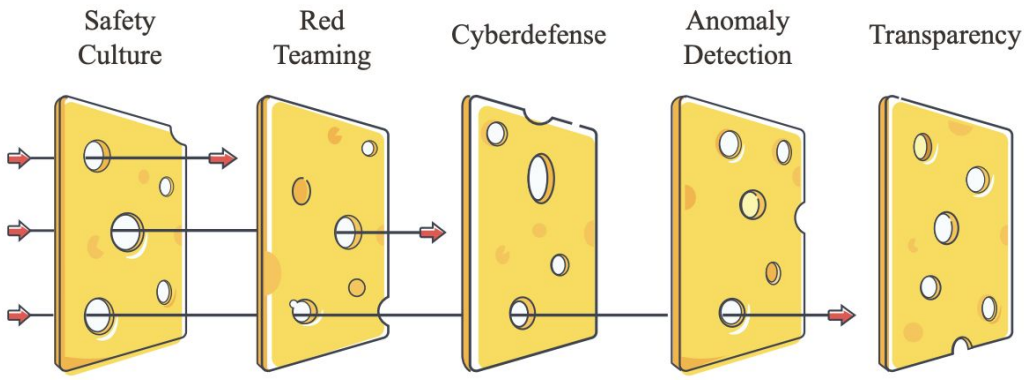
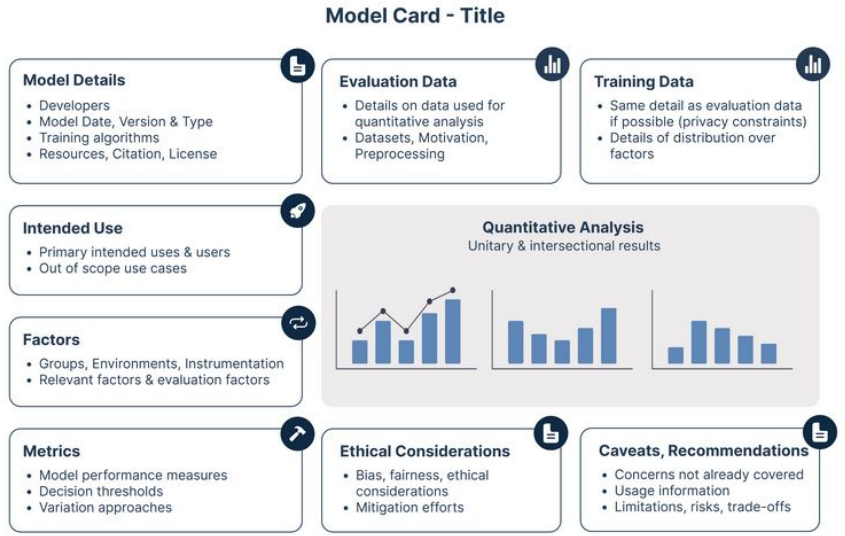


Figure 14: The Swiss cheese model shows how technical factors can improve organizational safety. Multiple layers of defense compensate for each other's individual weaknesses, leading to a low overall level of risk.

출처 : https://newsletter.safe.ai/p/aisn-28-center-for-ai-safety-2023

1. 위험 관리 및 안전 공학
1.2 위험 관리

- 범용 AI 개발자들의 현재 위험 관리 현황
  - 레드팀과 벤치마킹을 통해 출시 전에 일부 위험한 기능에 대해 모델을 테스트하고 그 결과를 '모델 카드'에 게시
    - 모델 카드: 범용 AI 모델의 목적, 평가 및 벤치마크에서의 성능, 안전 기능과 같은 중요한 정보를 제공하는 문서
  - 자발적으로 사전 정의된 임계값을 통해 결정 제한
    - 실제 위험 완화에 도움이 되는지, 적절한 임계값 설정의 실행 가능성에 대한 추가 연구 필요



출처 : https://www.trail-ml.com/blog/ml-model-cards

2. 더 신뢰할 수 있는 모델 훈련

2. 더 신뢰할 수 있는 모델 훈련
2.1 범용 AI 시스템 정렬

- AI 정렬(alignment) : 범용 AI 시스템이 개발자의 목표와 이해관계에 따라 작동하도록 하는 것
- 두가지 정렬 과제
  - 의도한 목표를 장려하는 것을 목적으로 훈련
    - 복잡한 가치와 선호도를 AI가 이해할 수 있는 형태로 정의하는 것의 어려움
    - 범용 AI는 통상 진정한 목표의 불완전한 대리인(proxy)인 목표에 최적화하도록 훈련됨. (인간 평가자의 승인은 이용자 이익의 불완전한 대리인임)
  - 훈련 컨텍스트에서 현실 세계로 의도한 대로 전환되도록 훈련
    - 훈련 컨텍스트가 실제 상황을 적절히 표현하지 못할 가능성

2. 더 신뢰할 수 있는 모델 훈련
2.1 범용 AI 시스템 정렬

- 인간 피드백 기반 훈련(미세조정)
  - 사람의 실수나 편견에 의한 품질 저하, 노동집약적이고 비용이 많이 듬
- 불확실성 기반 접근
  - 범용 AI가 목표에 대한 불확실성을 가지고 행동하도록 훈련하여 예기치 않은 행동의 위험을 줄이고 모호할 경우 정보를 찾거나 인간에 따르도록 장려
- 정량적 안전 보장을 제공할 수 있는 안전 설계(SbD) 접근
  - 현재로서는 실질적으로 유용하고 입증 가능한 안전성 보장은 불가능

2. 더 신뢰할 수 있는 모델 훈련
2.1 범용 AI 시스템 정렬

- 확장 가능한 감독(Scalable oversight)
  - 인간보다 더 능력 있는 AI 시스템을 어떻게 감독할 것인가에 대한 연구
  - 덜 능력 있는 AI 시스템이 더 능력 있는 시스템을 감독하는 방법
  - 아직 초보적인 수준

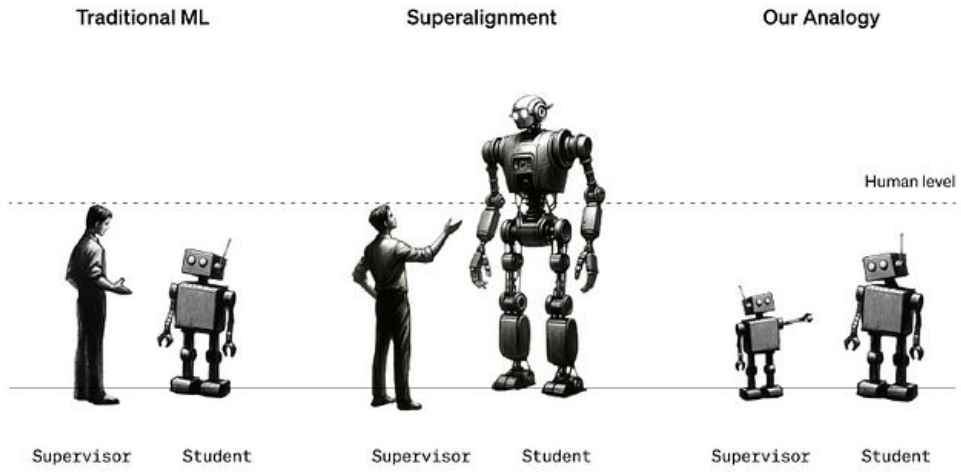


Figure 1: **An illustration of our methodology.** Traditional ML focuses on the setting where humans supervise models that are weaker than humans. For the ultimate superalignment problem, humans will have to supervise models much smarter than them. We study an analogous problem today: using weak models to supervise strong models.

출처 :
https://medium.com/@prdeepak.babu/scalable-oversight-in-ai-beyond-human-supervision-d258b50dbf62

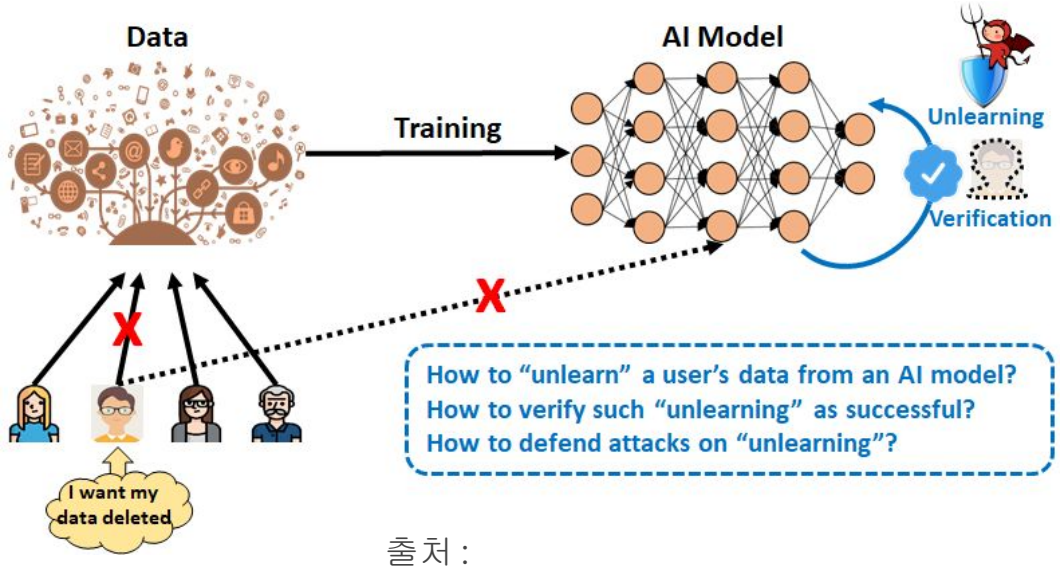
2. 더 신뢰할 수 있는 모델 훈련
2.2 허위 사실에 대한 환각 감소 방법

- 미세 조정: 특별히 설계된 데이터셋으로 추가 훈련하여 출력의 정확성을 높이는 방법
- 지식 데이터베이스 접근
  - 질문에 답할 때 신뢰할 수 있는 외부 지식 소스 참조
  - 검색 증강 생성(RAG) 기법
- 환각 탐지 및 경고
- 범용 AI의 환각을 완전히 제거하는 것은 현재 기술로는 불가능하며 지속적인 연구 필요

2. 더 신뢰할 수 있는 모델 훈련
2.3 장애에 대한 견고성 향상

- 모델 견고성(robustness)을 위한 적대적 훈련(Adversarial training)
  - 모델을 의도적으로 혼란시키는 입력을 생성하고 이에 대응하도록 훈련
  - 적대적 훈련의 한계
    - 적대적 훈련에는 예제가 필요 ⇒ 새로운 취약점에 대한 지속적 업데이트 필요
      ⇒ 모든 유형의 공격을 예측하고 대비하는 것은 불가능
    - 적대적 훈련이 모델의 성능이나 견고성을 약화시키는 문제 발생
- 견고성 향상을 위한 다른 방법
  - 입력이 아닌 내부 상태(internal states)에 적대적 훈련을 적용하는 방법 :
    이에 대한 연구는 아직 초보적 수준
  - 수학적 증명을 통해 견고성을 입증하는 방법 :
    현재의 모델과 방법으로는 불가능

2. 더 신뢰할 수 있는 모델 훈련
2.4 위험한 기능 제거



출처 :
https://www.researchgate.net/publication/370763971_Learn_to_Unlearn_A_Survey_on_Machine_Unlearning

- 머신 언러닝(machine unlearning)
  - 특정한 바람직하지 않은 기능을 제거하기 위한 방법. 예) 악의적인 사용자가 폭발물, 생물무기, 화학무기, 사이버 공격을 하는 데 도움이 될 수 있는 특정 기능 제거
  - 원래 프라이버시, 저작권 보호를 위해 특정 훈련 데이터의 영향을 제거하기 위한 방법으로 제안됨
  - 미세 조정, 모델 내부 작업 편집에 기반한 방법 등이 있음
  - 한계 : 종종 언러닝이 제대로 되지 않을 수 있으며, 바람직한 모델 지식에 부작용을 야기할 수 있음.

2. 더 신뢰할 수 있는 모델 훈련
2.5 모델 내부 작동 분석 및 편집

- 기계적 해석 가능성(mechanistic interpretability)
  - 특정 기능의 파악을 위한 모델의 내부 작동 연구
  - 예) 시각적 분류기의 공정성, 언어 모델이 어떤 지식을 갖고 있는지 파악
  - 작은 신경망에서 매우 간단한 작업을 수행하는 방법에 대한 철저한 조사 연구, 인간이 해석할 수 있는 모델을 설계하기 위한 보다 확장 가능한 기법 등이 시도됨
  - 최첨단 신경망은 규모가 크고 복잡, 범용 AI의 내부 표현을 평가하는 방법은 부정확함.

2. 더 신뢰할 수 있는 모델 훈련
2.5 모델 내부 작동 분석 및 편집

- 모델의 동작을 변경하기 위한 편집 기술
  - 내부 파라미터, 뉴런, 표현(representation) 등의 변경에 기반한 다양한 방법 시도
  - 이러한 기법은 불완전하며 모델 행동에 의도하지 않은 부작용 초래

3. 모니터링 및 개입

3. 모니터링 및 개입
3.1 범용 AI 생성 콘텐츠 감지

- 모니터링 : 지속적인 위험 식별, 모델 작업의 검사 및 성능 평가
- 개입 : AI 모델의 유해한 행동을 방지하는 기술
- 딥페이크와 같은 범용 AI 생성 콘텐츠의 유해한 영향을 방지하기 위해 콘텐츠 탐지 필요 ⇒ 탐지 기술이 개발되고 있지만 신뢰성에 한계
- AI 생성 콘텐츠의 탐지를 위한 워터마크 기술
  - 이미지 픽셀에 눈에 띄지 않는 패턴을 삽입, 텍스트용 워터마크는 문체 또는 단어 선택 편향의 형태.
  - 유용하지만 제거될 수 있기 때문에 불완전함
  - 워터마크는 진품 콘텐츠를 표시하는데도 사용할 수 있음

3. 모니터링 및 개입
3.2 이상 징후 및 공격의 탐지

- 이상이나 공격이 감지되면, 시스템이 오작동하거나 악의적인 행위로 인해 피해를 입기 전에 예방 조치를 취할 수 있음.
- 비정상적인 입력이나 동작을 탐지하는 방법 : 악성 공격이 범용 AI 모델로 전달되기 전에 탐지 및 필터링
- 주어진 입력에 대한 불확실한 모델 출력을 탐지하는 방법 : 잠재적으로 유해한 출력을 탐지하여 사용자에게 전송되기 전에 차단

3. 모니터링 및 개입
3.3. 모델 작업 설명하기

- 범용 AI 언어 모델에 단순히 결정에 대한 설명을 요청하는 것은 오해의 소지가 있는 답변 경향 ⇒ 설명의 신뢰도를 높이기 위해 개선된 프롬프트와 훈련 전략 연구 중
- 범용 AI의 작동을 설명하는 기술은 아직 초기 단계

3. 모니터링 및 개입
3.4 AI 시스템에 안전 장치 구축

- 완벽한 안전조치는 없지만 여러 계층의 중복적인 안전장치 마련 필요
- 인간–AI 협력 패러다임
  - 사람이 루프에 참여하여 직접 감독하고 필요할 경우 AI의 결정을 변경.
  - 자동화된 시스템에 비해 비용이 많이 들지만, 중요한 의사결정에서는 반드시 필요
  - 루프에 참여하는 것이 항상 가능한 것은 아님 : 의사 결정이 빠르게 이루어지는 경우, 사람이 충분한 지식을 가지고 있지 않은 경우, 사람의 편견이나 오류로 위험이 악화될 수 있는 경우 등

3. 모니터링 및 개입
3.4 AI 시스템에 안전 장치 구축

- 자동화된 처리 및 필터링
  - 사이버공격 패턴을 제거하는 입력 전처리, 유해한 출력의 사용자 전송을 탐지하는 후처리
- 잠재적으로 위험한 기능을 갖춘 범용 AI 시스템을 위한 보안 인터페이스 설계
  - 사람이나 사물에 직접 영향을 줄 수 있는 방식를 제한하는 방법 (예를 들어, 어떤 기계를 작동시키기 전에 사람의 확인을 거치도록 하는 경우)

4. 공정성과 대표성에 대한 기술적 접근

4. 공정성과 대표성에 대한 기술적 접근

- 공정성은 보편적으로 합의된 정의가 없으며 문화적, 사회적, 학문적 맥락에 따라 달라짐
- AI의 공정성
  - 자동화된 의사 결정이나 콘텐츠 생성에서 알고리즘의 편향성을 바로잡으려는 시도
  - AI의 공정성은 다양한 방식으로 정의되고 측정되며, 맥락과 애플리케이션의 특정 목표에 따라 달라짐 예) COMPAS 사례
- 데이터 세트는 종종 소수를 적절하게 대표하지 못하며, 이러한 데이터 세트로 학습된 AI에 반영되고 증폭됨.

4. 공정성과 대표성에 대한 기술적 접근
4.1 편견과 차별 완화 방법

- 편견과 차별 완화는 범용 AI 개발 및 배포의 모든 단계에서 작동함
- 전처리 기법
  - 데이터를 분석하고 수정하여 데이터 세트에 내재된 편향 제거
  - 데이터 증강 : 기존 데이터의 수정된 복사본 또는 합성 데이터를 통해 과소 대표 그룹의 샘플 추가
  - 데이터 변경 : 성별, 인종과 같은 속성을 추가, 제거, 마스킹하여 정의된 규칙에 따라 샘플 수정

4. 공정성과 대표성에 대한 기술적 접근
4.1 편견과 차별 완화 방법

- 처리 중 기법
  - 데이터가 완벽하더라도 사회의 고정관념과 편견 포함할 수 있으므로 학습 단계에서 편견 완화
    ⇒ 인간 피드백에 기반한 모델 훈련
  - 편향성이 덜한 범용 AI 모델('교사')에서 다른 범용 AI 모델('학생')로 정보를 전송하여 공정성 교육
  - 편향된 데이터 샘플과 편향되지 않은 데이터 샘플을 모두 사용하여 훈련
  - 과소 대표되는 속성을 더 부각하여 모델 편향 제거 : 프라이버시와 상충 가능성

4. 공정성과 대표성에 대한 기술적 접근
4.1 편견과 차별 완화 방법

- 후처리 기법
  - 입력 또는 출력을 조작하여 차별 완화
  - 편향성/안전성 분류기와 같은 외부 모듈 활용 : 불공정한 출력을 감지하여 재생성 요구

- 다양한 사회에서 범용 AI 시스템이 모든 사람의 가치를 대변하는 것이 가능한가
  - (완전한 해결은 어렵지만) 참여확대, 대표성, 대화가 일부 사람들의 이익에만 부합할 위험성을 줄이는 방법으로 제안됨.

4. 공정성과 대표성에 대한 기술적 접근
4.2 범용 AI 시스템에서 공정성을 달성할 수 있을까

- 완전한 공정성은 불가능하다는 의견
  - 수학적 결과에 따르면 공정성의 모든 측면을 동시에 만족시키는 것은 불가능
  - 공정성, 정확성, 개인정보보호, 효율성 사이의 트레이드 오프
  - 예) 1800년대 미국 상원의원으로서의 원주민과 유색인종 여성, 2차 세계대전 당시 인종적으로 다양한 독일군 병사의 이미지를 생성한 Gemini 사례
- 실질적인 해결책을 찾을 수 있다는 의견
  - AI 시스템 결과물의 불균형을 줄이는 것이 정확도의 현저한 저하를 수반하는 것은 아님

4. 공정성과 대표성에 대한 기술적 접근
4.3 공정한 범용 AI 시스템을 달성하기 위한 과제

- 공정성을 어떻게 정의하고 측정할 것인가
  - 유용하고 정확한 지식과 해로운 고정관념의 경계는 모호하고 편견에 대한 인식은 상황에 따라 달라짐
- 안전성과 편향의 트레이드오프
  - 개인정보 보호를 위한 데이터 정제가 인구통계학적 분포를 변경하여 편견을 증폭할 수 있음
- 교차 편향 문제 해결의 어려움
- 범용 AI 시스템의 개발, 배포, 사용 전반에 걸쳐 지속적인 노력이 필요함

5. 프라이버시 보호 방법

5. 프라이버시 보호 방법

- 현재의 개인정보보호 강화 기술이 대규모 범용 AI 모델로 확장되지 않음
  - 모델 정확도 저하, 대규모 모델로의 확장 어려움, 모든 사용 사례(특히 텍스트로 학습된 범용 AI 모델)에 적합하지 않을 수 있음.
  - AI 학습에 합성 데이터를 사용하는 방법 : 합성 데이터 활용도가 높으면 원본 데이터만큼 많은 정보를 포함하고 있고 동일한 공격에 노출
  - 기밀성 및 데이터 중앙집중화 문제 : 암호화 접근 방식, 연합 학습(federated learning), 하드웨어 보호 등 보안 솔루션 사용 ⇒ 기존 기술은 가장 크고 유능한 모델에 맞게 확장되지 않았고, 규모에 따른 엄청난 비용이 필요함.

5. 프라이버시 보호 방법

- 데이터 투명성 및 통제 조치는 범용 AI 시스템에도 적용 가능
  - 예) 정보주체의 개인정보 관리를 위한 프로세스나 인터페이스
  - 공개 데이터에 대한 투명성과 통제권을 제공하는 것은 난제
  - 또 다른 난제는 파생 데이터 또는 식별되지는 않지만 사람에 대한 추론이 가능한 데이터 사용에 대한 의미있는 제어를 제공하는 문제
- 일부 개인정보 남용은 기술적 수단을 통한 예방이 곤란
  - 미합의 딥페이크나 스토킹
  - 데이터 최소화 및 목적제한 등 개인정보보호원칙을 존중하는 AI 시스템 개발을 요구하는 법적 프레임워크 (Privacy by Design, EU AI Act)
    ⇒ 이를 달성할 수 있을지는 미지수

결론

범용 AI가 우리 삶의 여러 측면에 미치는 영향이 심대할 수 있고
그 발전 속도가 계속 빨라질 수 있기 때문에 사전 예방 원칙은 합의를
도출하고 이러한 위험을 이해하고 해결하는 데 자원을 투입해야 한다는
긴급한 필요성을 시사한다. 사회와 정책 입안자들이 올바른 선택을
하기 위해서는 건설적인 과학적이고 대중적인 토론이 필수적이다.

감사합니다