AI 윤리란 무엇인가?
원칙에서 행동으로

이상욱
한양대학교 철학과 & 인공지능학과
유네스코 세계과학기술윤리위원회 의장단
HY 과학기술윤리·법·정책센터장

UNESCO on AI Ethics

- 유네스코의 AI 윤리 논의는 다른 기관에서 이루어지는 윤리 논의에 비해 '포괄적(comprehensive)'이고 '근본적(fundamental)'이라는 특징을 갖는다.
- UNESCO의 권한/책무: Education, Science, Culture, information/Communication, Africa, Gender
- COMEST(과학기술윤리위원회)가 유네스코 내에서 AI 윤리 논의를 주도할 이상적인 기관이라고 결정

AI 글로벌 거버넌스의 현황 (1)

- **AI 기술경쟁 가속화**: 기술 선진국 사이의 경쟁만이 아니라 개발도상국 및 저개발국의 관심 증대

- **원칙에서 행동으로**: 여러 국제 기구(OECD, UNESCO, IEEE)와 국가(유럽연합, G20, 미국, 한국 등)의 AI 윤리 원칙 선언에서 출발하여, 2023년 이후부터는 다양한 규제의 흐름이 나타났고 현재 시점에서는 유럽연합(<AI 법>)과 미국(<AI 행정명령>)이 제정되면서 각국의 입법 활동도 활발해질 것으로 예측됨

AI 글로벌 거버넌스의 현황 (2)

- **AI 기술 개발에서 '윤리 비용' 문제 제기**: 기술선진국의 AI 관련 법률 및 제도에 대응하기 위해서는 상당한 비용이 들기에 이것이 '무역장벽'이나 '사다리 걷어차기' 기능을 하는 것이 아닌가라는 비판이 제기되고 있음

- **자율규제와 인증/표준 제정**: 현재까지는 AI 글로벌 거버넌스는 자율규제에 기반하되 문제의 소지가 있을 때 정부가 추가조사를 수행하거나 행정 조치를 취하는 방식을 채택. 그와 동시에 AI 기술발전을 반영하는 '신뢰성/투명성/안전성' 인증이나 관련 기술표준 제정 노력이 함께 이루어지고 있음.

신뢰할 수 있는 AI와 안전한 AI

- 글로벌 AI 거버넌스는 우선 '신뢰할 수 있는(trustworthy)' AI를 만드는 데 주목 (혁신과 윤리의 균형을 강조한 여러 국제 문건에서 확인됨.)

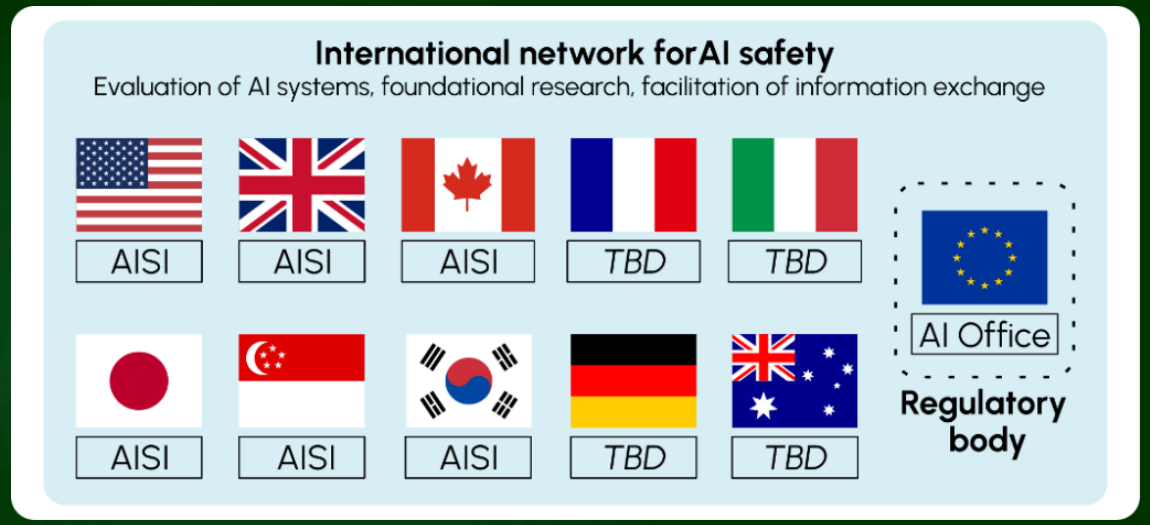
- 최근 생성형 인공지능이 제기하는 여러 위험에 주목하면서 '안전한(safe)' AI를 위한 국제 거버넌스에 대한 관심이 커지고 있음.

Cf.) 2024 AI Seoul Summit, 2025 Paris AI Action Summit

유럽은 규제, 미국은 혁신?

- 유럽의 <AI 법>이 위험 기반 접근에 기반하여 규제만 담았다?
- 미국은 기술 혁신을 도모하기 위해 (특히 트럼프의 재집권 이후) 모든 규제를 철폐했다?
- 모든 국가는 기술 혁신으로부터 얻을 수 있는 '잠재적' 혜택과 그 과정에서 발생할 수 있는 '잠재적' 위험에 대응하기 위한 정책을 '자국의 이익을 중심으로' 제시하고 실천한다!
- 이 기조는 현재까지도 유지되는 AI 국제 거버넌스의 핵심!

AI 안전연구소 네트워크와 글로벌 AI 거버넌스 경쟁

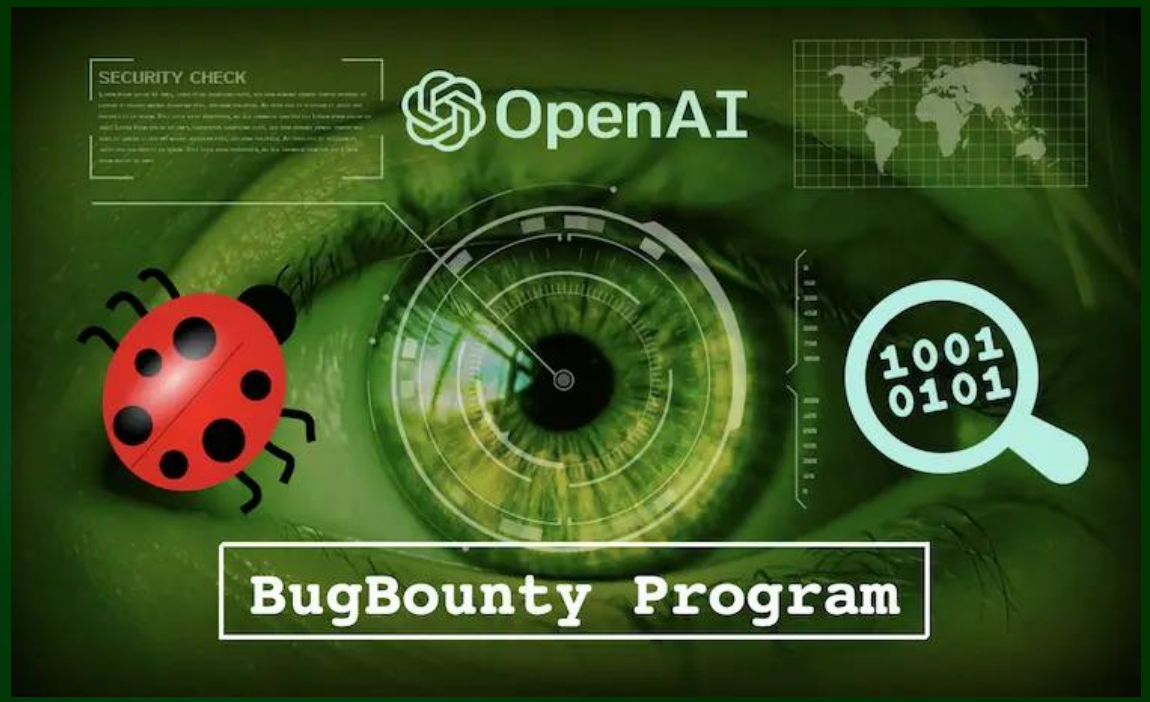


- 현재 우리나라까지 포함된 11개국의 AI 안전 연구소 네트워크가 구축되어 있음
- 이 네트워크는 각국의 AI 윤리, 안전 관련 연구 및 제도적 성과를 공유하고 국제 거버넌스를 구축하려는 시도임
- 중요한 점은 IEEE, ISO 등의 국제 기술표준기구에서는 AI 관련 기술표준만이 아니라 신뢰성이나 안전성 같은 '윤리적' 판단에 대한 기술 표준도 마련 중이라는 사실
- 이러한 기술 표준은 미래의 무역 장벽으로 작용할 가능성이 높음

환각과 거짓/기만 정보의 문제

- 환각, 거짓/기만 정보(mis/dis-information)
- 허위/기만 정보를 만들기 너무 쉽다는 현실과 제도적 대응은 상대적으로 느리다는 문제
- 주식 시장이 출렁거리고 선거 결과가 뒤바뀔 수 있다.
- 보다 심각한 점은 생성형 인공지능을 활용한 거짓/기만 정보가 개인에게 심각한 피해를 줄 수 있다는 점이다. (예: 성착취물, 학교 폭력 동영상)
- 2023년에 열린 미상원 청문회

환각은 버그가 아니라 특성이다!



- ChatGPT를 공개한 OpenAI의 Big Bounty Program
- 자사의 제품의 버그(프로그램 오류)를 찾아주는 사용자에게 상금을 주는 경진대회
- 여기에서 '환각'과 '탈옥'은 상금을 주지 않는다!

- 중요한 점은 '환각'은 프로그램의 버그가 아니라 기술적 특징(feature)라는 사실이다.

생성형 AI의 훈련과 '탈옥'

- 인공 신경망에 엄청난 양의 사전 훈련 데이터를 활용하여 비지도 학습 시행
- 그 다음 사람이 보기에 더 좋은 답을 가르쳐 주는 인간 피드백 강화학습(HFRL) 시행 (Cf. GPT-3.0과 GPT-3.5의 차이)
- HFRL의 결과는 원칙적으로 가역적이어서, 악의 의도를 갖고 인공지능을 '퇴화(degeneration)' 혹은 '탈옥'시키는 일이 가능하다.

저작권과 데이터 주권의 문제

- 생성형 AI의 학습 과정에서 활용된 데이터에 대한 '적절한 보상'에 대해 현재 법적 분쟁이 진행 중이다.
- 생성형 AI의 결과물에 대해 현재까지는 인공지능 자체가 저작권을 가질 수 없다는 점에는 합의가 있다.
- 하지만 결과물 작성 과정에 참여한 사람 중 누가 저작권을 갖는지에 대해서는 합의가 없다.
- 인공지능 개발과 활용에서 데이터의 중요성이 강조되면서 자국에서 생산된 데이터에 대한 통제권과 이익공유권으로 이해된 데이터 주권 의식이 널리 퍼지고 있다.

이중 사용(dual use) 문제

- 신약 개발을 위한 후보물질 탐색처럼 '착한' 목적으로 개발된 인공지능을 아주 약간만 변형해도 독성 물질 발견처럼 사악한 목적에 활용될 수 있다!
- '착한'/'나쁜' 인공지능 구별의 모호성
- 그러므로 생성형 AI 활용이 '안전하게' 이루어질 수 있게 하기 위한 국제 거버넌스 구축이 현재 진행 중이다.

(https://www.nature.com/articles/s42256-022-00465-9)

인공지능의 활용과 탈숙련의 문제

- 현재 교육과 학술 연구 영역에서 생성형 인공지능의 활용이 활발하게 진행되고 있다.
- 특히 이미 전문성을 획득한 전문가가 인공지능을 활용하면 작업 능력이 상승하는 강화(augmentation)가 발생한다.
- 하지만 전문성을 채 획득하기 전에 생성형 인공지능을 과다하게 사용하면 탈숙련(deskill)의 문제가 발생할 수 있다.
- 결국 어떤 능력을 인류가 핵심 역량(core skill)으로 보존할 것인지에 대한 구체적인 고민이 필요하다.

강의 내용 정리

- 현재 세계 각국은 인공지능 기술의 잠재적 혜택과 위험성을 모두 고려하는 국제 거버넌스(관리 체계)를 만들어 나가고 있다.
- 인공지능 국제 거버넌스 구축에서 2021년 유네스코의 <AI 윤리 권고>와 유럽 의회의 <AI 기본법> 통과는 중요한 의미를 지닌다.
- 인공지능 국제 거버넌스의 핵심은 신뢰가능성과 안전성으로 요약될 수 있다.
- AI 윤리의 다양한 쟁점은 AI의 기술적 특징인 환각과 기능저하 가능성과 관련이 있다.
- 저작권과 이중 사용 문제는 현재 첨예한 이해관계가 충돌하고 있으며 탈숙련의 문제는 인간이 쌓아올린 '전문성'을 AI 시대에 재규정할 필요성을 제기한다.