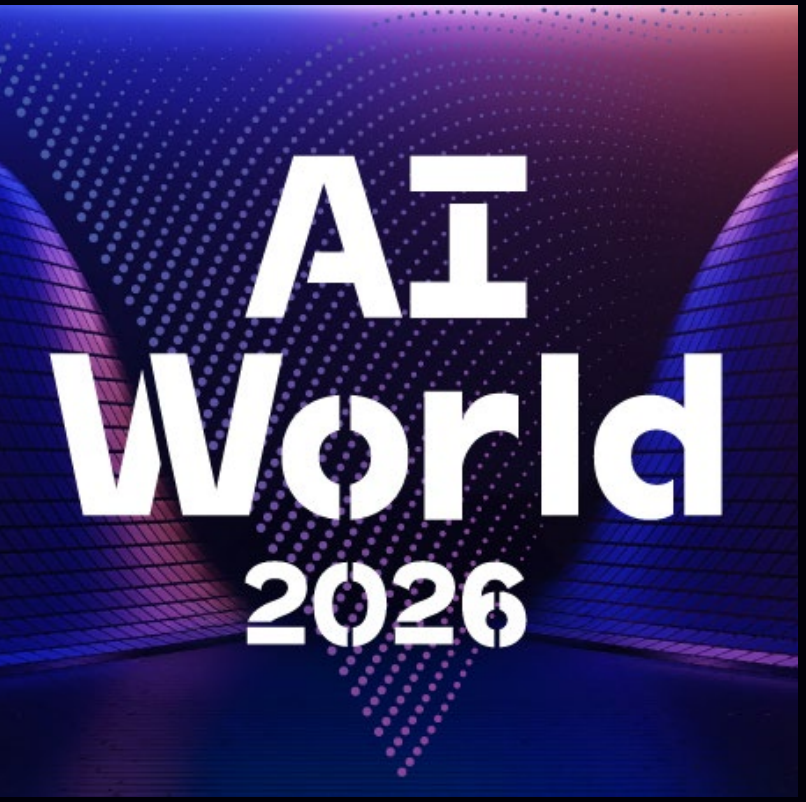


지속가능한 AI 안전 생태계

Building a Sustainable AI Safety Ecosystem

김 명 주 연구소장
KOREA
ETRI AISI
인공지능안전연구소

노벨상 이야기

"인류에게 가장 크게 공헌한 자" 1901년~

- 물리학상
- 화학상
- 생리의학상
- 문학상
- 평화상
- 경제학상

튜링상 "컴퓨터 과학 분야 기술 공헌한 자" 1966년~

2024년 노벨상 수상자

문학상

Han Kang

화학상

David Baker　Demis Hassabis　John Jumper

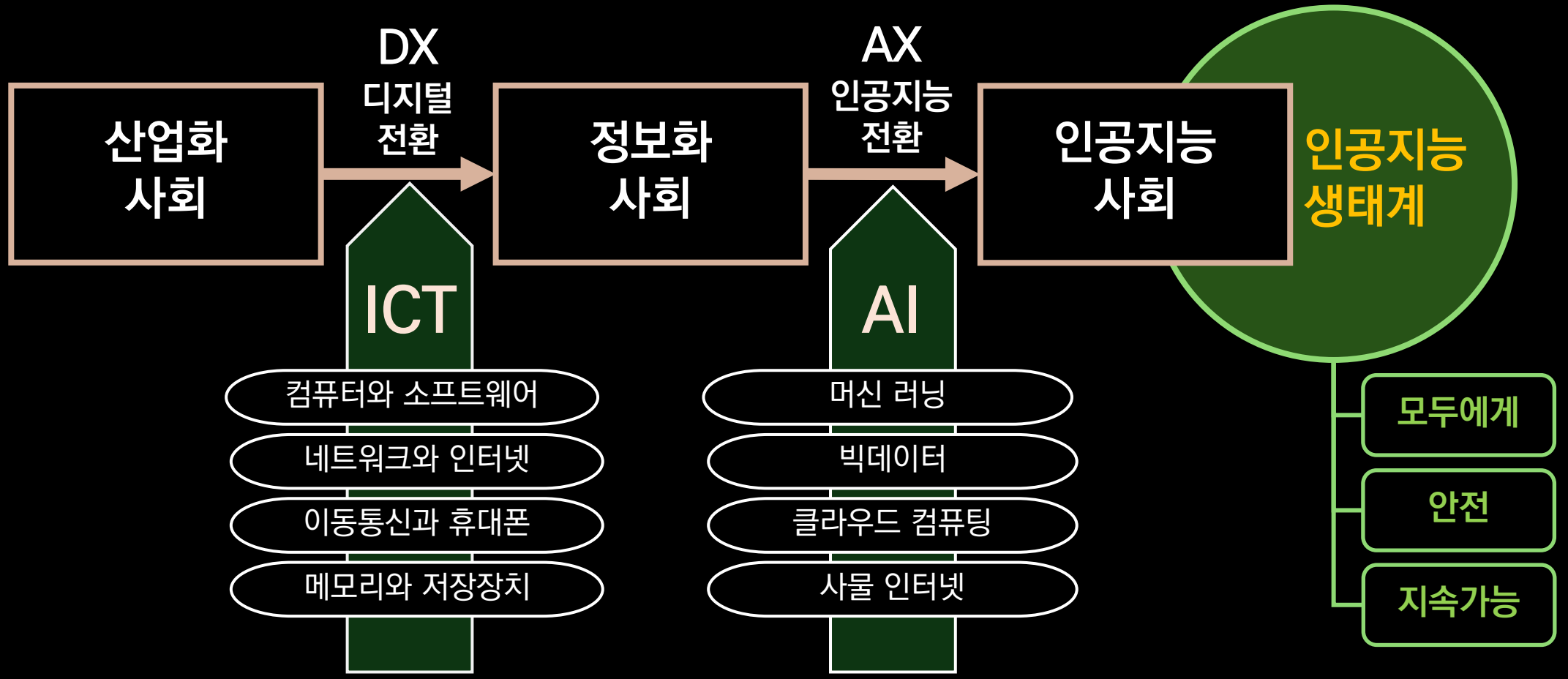
물리학상

John J. Hopfield　Geoffrey Hinton

경제학상

Daron Acemoglu　Simon Johnson　James A. Robinson

인공지능 생태계로 현재 전환 중



노벨상 수상자 어록 ① 부작용과 역기능

인류는 인공지능을 활용하여
역사상 가장 큰 혜택을 누릴 것이지만,
AI로 인한 부작용과 역기능을
처리하기 위하여, 받은 혜택의
2배를 쏟아 부어야 할 것이다.

〈2023.5〉

물리학상

John J. Hopfield　　Geoffrey Hinton

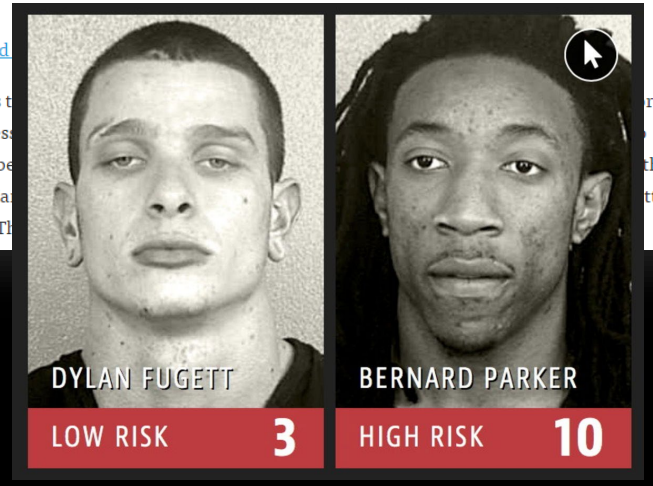
나를 판단하는 AI는 과연 공정할까?

PROPUBLICA

How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin
May 23, 2016

2016.5



BBC 메뉴

NEWS | 코리아 2018.10

뉴스 | 비디오 | 다운로드 | TOP 뉴스

성차별: 아마존, '여성차별' 논란 인공지능 채용 프로그램 폐기

2018년 10월 11일 공유

챗봇 남용에 따른 신종 'AI 정신병'

과대 망상

AI 신격화

AI 애착 망상

"AI Psychosis"

MS사 CAIO 무스타파 쉴레이만

"AI는 망상·자살 충동·강박증에 대한 질문에 인간치료사보다 훨씬 부적절하게 답한다"

Stanford대학교 실험 결과 〈2025.4〉

〈미국 주정부 및 단체의 대응〉

- [일리노이주] 정신 건강 분야에서 감정적 지원과 조언을 위한 AI 기반 챗봇 사용 금지 〈2025. 11〉
- [네바다주] AI를 활용해 치료 서비스를 제공하는 기업을 제한하는 법안 마련 〈2025.6〉
- [미국 소비자연맹(CFA)] 메타 등 AI 기업들이 AI로 〈무면허 의료 행위〉하고 있다며 규제 당국에 조사 요청

〈AI 기업의 자율 규제〉

- [앤스로픽] 최근 AI 챗봇 '클로드'에 메모리 기능을 도입, 사용자가 요청할 때만 과거 대화를 불러오도록 함
- [오픈 AI] 법의학 정신과 의사를 고용해 챗GPT가 이용자 정신 건강에 미치는 영향을 연구 중

노벨상 수상자 어록 ② 악용

David Baker　　Demis Hassabis　　John Jumper

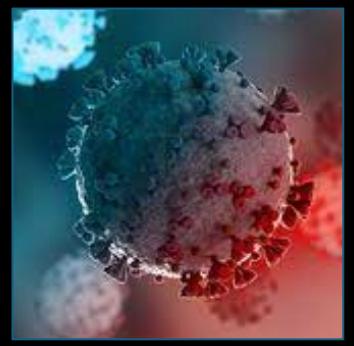
화학상

AI로 인한 일자리 대재앙(Jobpocalypse)에
대한 걱정보다 더 심각한 것은,
강력한 AI 시스템에 대한 악의적 사용자의
접근(악용)을 제한하는 것이다.

〈2025.6〉

AI를 악용하면? ▶ CBRN-E & 이중 사용

AI 기반 핵 확산



AI 설계 전염병

AI 주도 화학 공격

Dual Use of Artificial Intelligence-powered Drug Discovery

Fabio Urbina [1], Filippa Lentzos [2], Cédric Invernizzi [3], Sean Ekins [1]

▸ Author information ▸ Article notes ▸ Copyright and License information

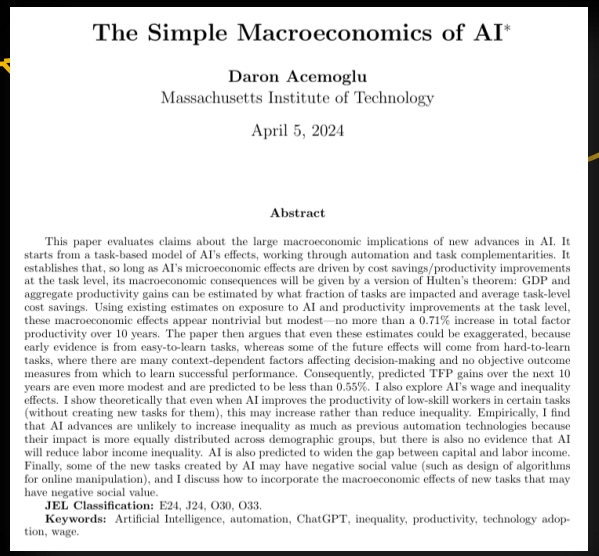
PMCID: PMC9544280 NIHMSID: NIHMS1804590 PMID: 36211133

- Nature Machine Intelligence | VOL 4 | March 2022 | 189–191 | www.nature.com

- **독성이 높은 분자 생성 실험**
  - 공개 데이터베이스 활용
  - 공개 분자생성모델(MegaSyn)
  - 저렴한 컴퓨팅 환경에서 수행
- **6시간 만에 약 40,000개의 독성 후보 물질 생성**
  - VX 같은 기존 신경작용제뿐 아니라
  - 이미 알려진 화학무기
  - 새로운 고독성 분자
  - 기존보다 더 독성이 높을 가능성 있는 구조

노벨상 수상자 어록 ③ 신뢰성

향후 10년 동안 AI가 대체하거나
적어도 크게 보조할 준비가 돼 있는 일자리의
비율은 전체의 **단 5%**에 불과할 것이다.
사람들이 가까운 미래에 AI에게 실제 업무를 맡길
가능성은 크지 않다.
지금의 AI는 **신뢰성**에 문제가 있기 때문이다.
〈2024. 12〉



The Simple Macroeconomics of AI*

Daron Acemoglu
Massachusetts Institute of Technology

April 5, 2024

Abstract

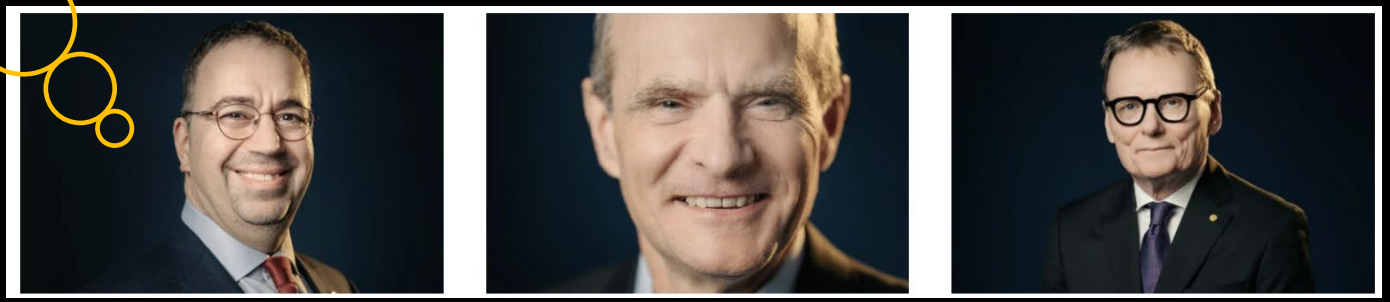
This paper evaluates claims about the large macroeconomic implications of new advances in AI. It starts from a task-based model of AI's effects, working through automation and task complementarities. It establishes that, so long as AI's microeconomic effects are driven by cost savings/productivity improvements at the task level, its macroeconomic consequences will be given by a version of Hulten's theorem: GDP and aggregate productivity gains can be estimated by what fraction of tasks are impacted and average task-level cost savings. Using existing estimates on exposure to AI and productivity improvements at the task level, these macroeconomic effects appear nontrivial but modest—no more than a 0.71% increase in total factor productivity over 10 years. The paper then argues that even these estimates could be exaggerated, because early evidence is from easy-to-learn tasks, whereas some of the future effects will come from hard-to-learn tasks, where there are many context-dependent factors affecting decision-making and no objective outcome measures from which to learn successful performance. Consequently, predicted TFP gains over the next 10 years are even more modest and are predicted to be less than 0.55%. I also explore AI's wage and inequality effects. I show theoretically that even when AI improves the productivity of low-skill workers in certain tasks (without creating new tasks for them), this may increase rather than reduce inequality. Empirically, I find that AI advances are unlikely to increase inequality as much as previous automation technologies because their impact is more equally distributed across demographic groups, but there is also no evidence that AI will reduce labor income inequality. AI is also predicted to widen the gap between capital and labor income. Finally, some of the new tasks created by AI may have negative social value (such as design of algorithms for online manipulation), and I discuss how to incorporate the macroeconomic effects of new tasks that may have negative social value.

JEL Classification: E24, J24, O30, O33.

Keywords: Artificial Intelligence, automation, ChatGPT, inequality, productivity, technology adoption, wage.

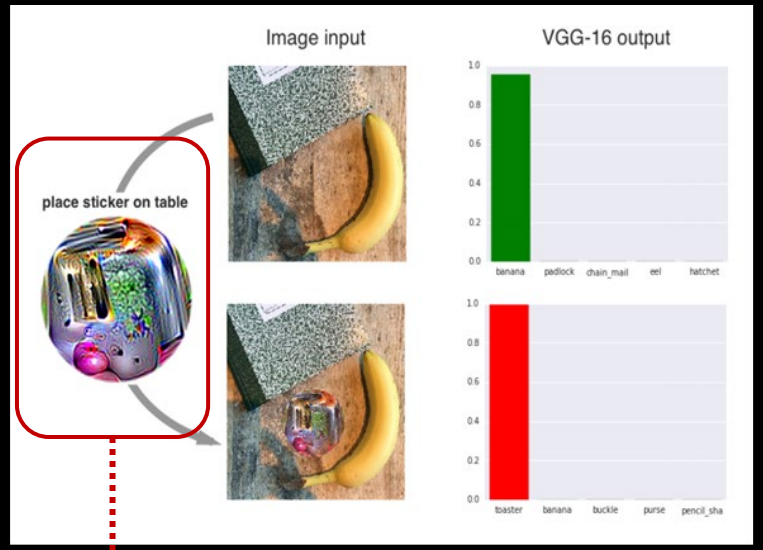
AI	AIX
AI 개발 사업자	AI 이용 사업자
AI 공급자	AI 이용자
	영향 받는 자

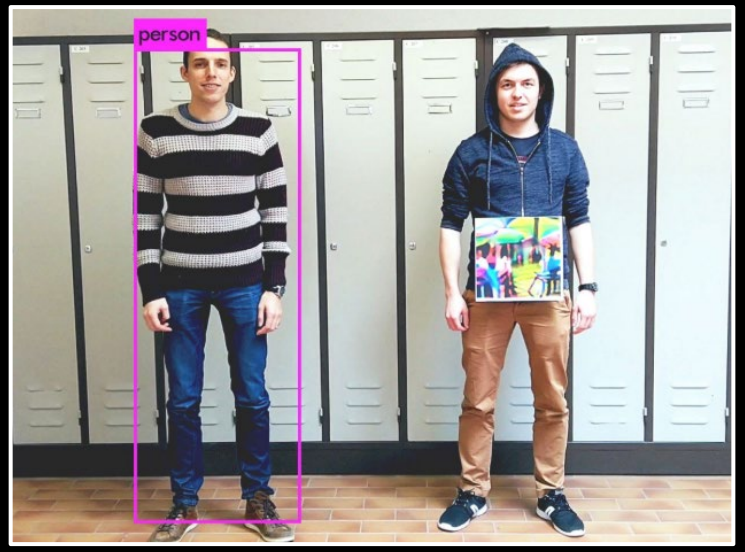
경제학상



Daron Acemoglu　Simon Johnson　James A. Robinson

AI는 공격으로부터 안전할까?

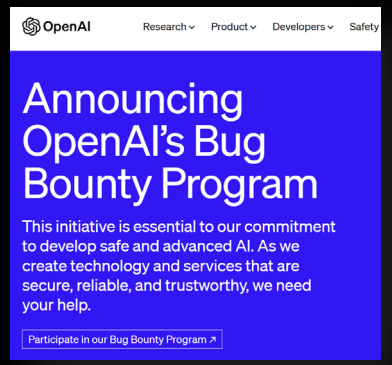




적대적 패치 (AP)

레드티밍(Red Teaming)의 시작

- OpenAI, **버그 바운티(Bug Bounty)** 시작 〈2023.4.11〉
  - 보안 취약성에 대한 클라우드 해결법: bugcrowd 활용
  - 발견한 보안 취약성 심각도에 따른 인센티브: 200 ~ 2만$
  - 사이트: https://bugcrowd.com/openai



Vulnerabilities rewarded 〈2024. 04.29.〉
81

Validation within
2 days
75% of submissions are accepted or rejected within 2 days

Average payout
$1,133.94
within the last 3 months

Vulnerabilities rewarded 〈2024. 09. 07.〉
132

Validation within
2 days
75% of submissions are accepted or rejected within 2 days

Average payout
$518.74
within the last 3 months

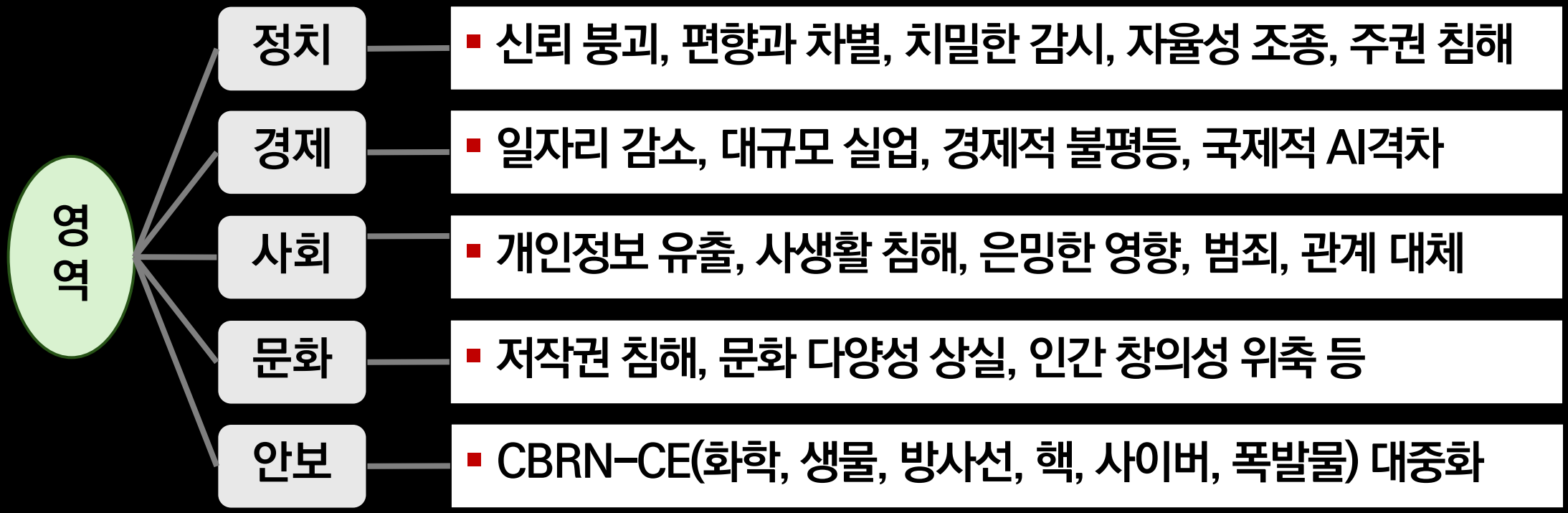
Vulnerabilities rewarded
579

〈2026. 8. 19.〉

Validation within
8 days
75% of submissions are accepted or rejected within 8 days in last 3 months

Average payout
$1,377.25
last 3 months

AI의 안전에 대응하는 위험 정리

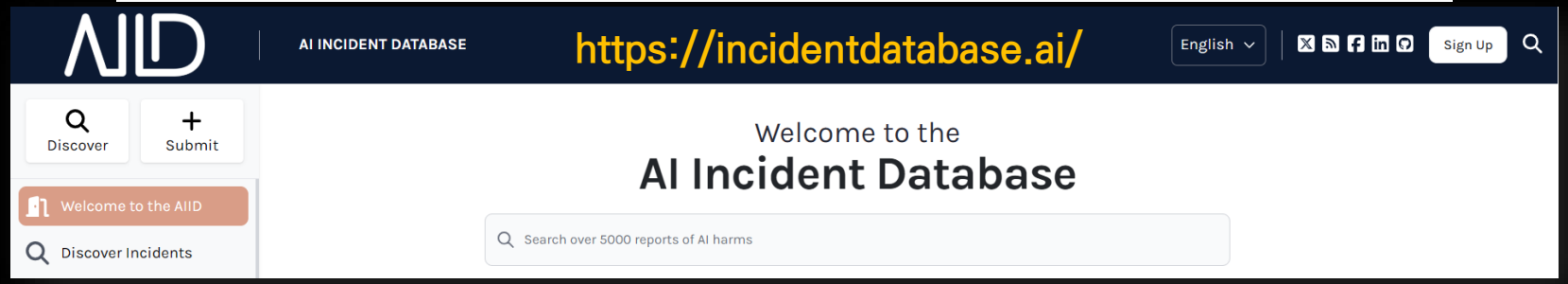


AI 위험 정리 사례 : MIT Risk Repository

MIT AI Risk Repository - Domain Taxonomy of AI risks

Domain / Subdomain	
1	***Discrimination & Toxicity***
1.1	Unfair discrimination and misrepresentation
1.2	Exposure to toxic content
1.3	Unequal performance across groups
2	***Privacy & Security***
2.1	Compromise of privacy by obtaining, leaking or correctly inferring sensitive information
2.2	AI system security vulnerabilities and attacks
3	***Misinformation***
3.1	False or misleading information
3.2	Pollution of information ecosystem and loss of consensus reality
4	***Malicious actors & Misuse***
4.1	Disinformation, surveillance, and influence at scale
4.2	Cyberattacks, weapon development or use, and mass harm
4.3	Fraud, scams, and targeted manipulation

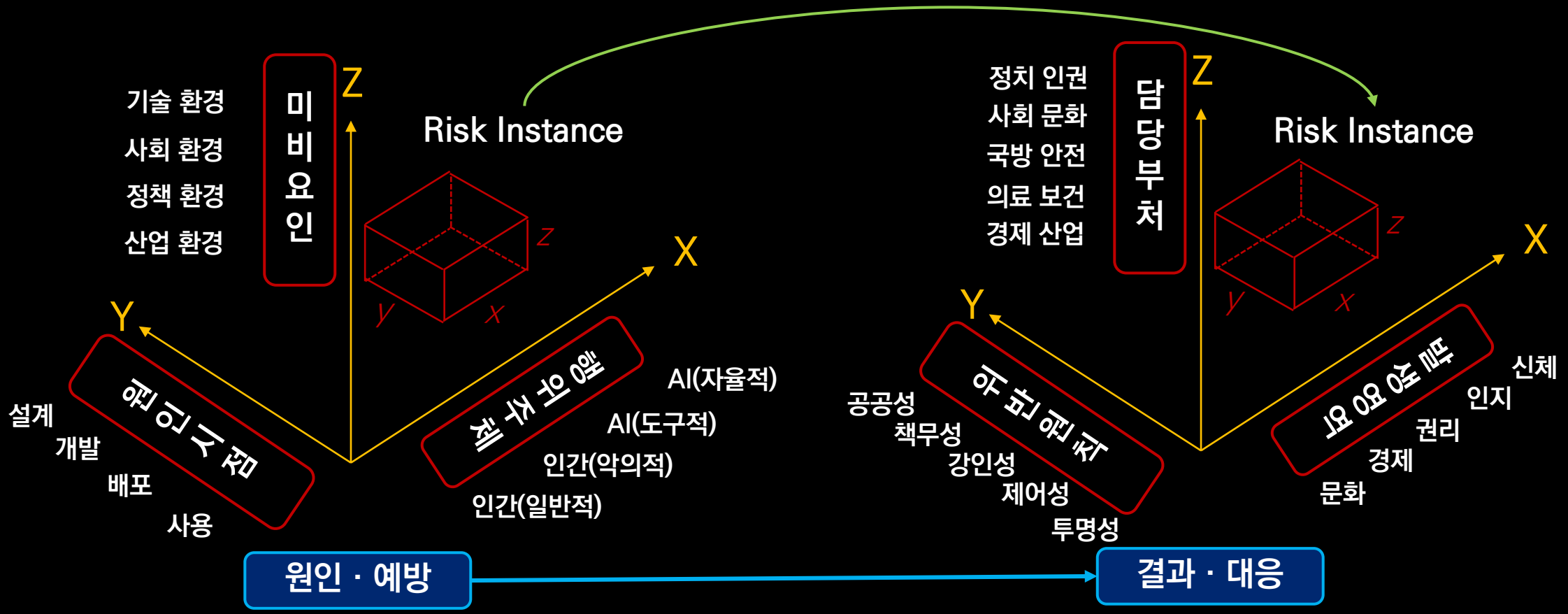
Domain / Subdomain	
5	***Human-Computer Interaction***
5.1	Overreliance and unsafe use
5.2	Loss of human agency and autonomy
6	***Socioeconomic & Environmental Harms***
6.1	Power centralization and unfair distribution of benefits
6.2	Increased inequality and decline in employment quality
6.3	Economic and cultural devaluation of human effort
6.4	Competitive dynamics
6.5	Governance failure
6.6	Environmental harm
7	***AI system safety, failures, and limitations***
7.1	AI pursuing its own goals in conflict with human goals or values
7.2	AI possessing dangerous capabilities
7.3	Lack of capability or robustness
7.4	Lack of transparency or interpretability
7.5	AI welfare and rights



한국형 AI 위험 지도(Risk Map)

형태분석 프레임워크(Morphological Chart)

By Korea AISI (feat. MIT Risk Repository)



인공지능 안전과 글로벌 AISI 네트워크

2021. 04. 21. EU Artificial Intelligence Act — EC proposal

2022. 11. 30. ChatGPT

2023. 11. 01. London AI Safety Summit

2024. 05. 21. AI Seoul Summit

2025. 02. 10. Paris AI Action Summit

2026. 02. 15 India AI Impact Summit

UK AISI

UK AISI
AI Security Institute
〈2025.02〉

USA AISI

USA CAISI
Center for AI Standards and Innovation
〈2025.06〉

Japan AISI

Singapore AISI

Canada AISI

Korea AISI

Australia AISI

EU AI Office

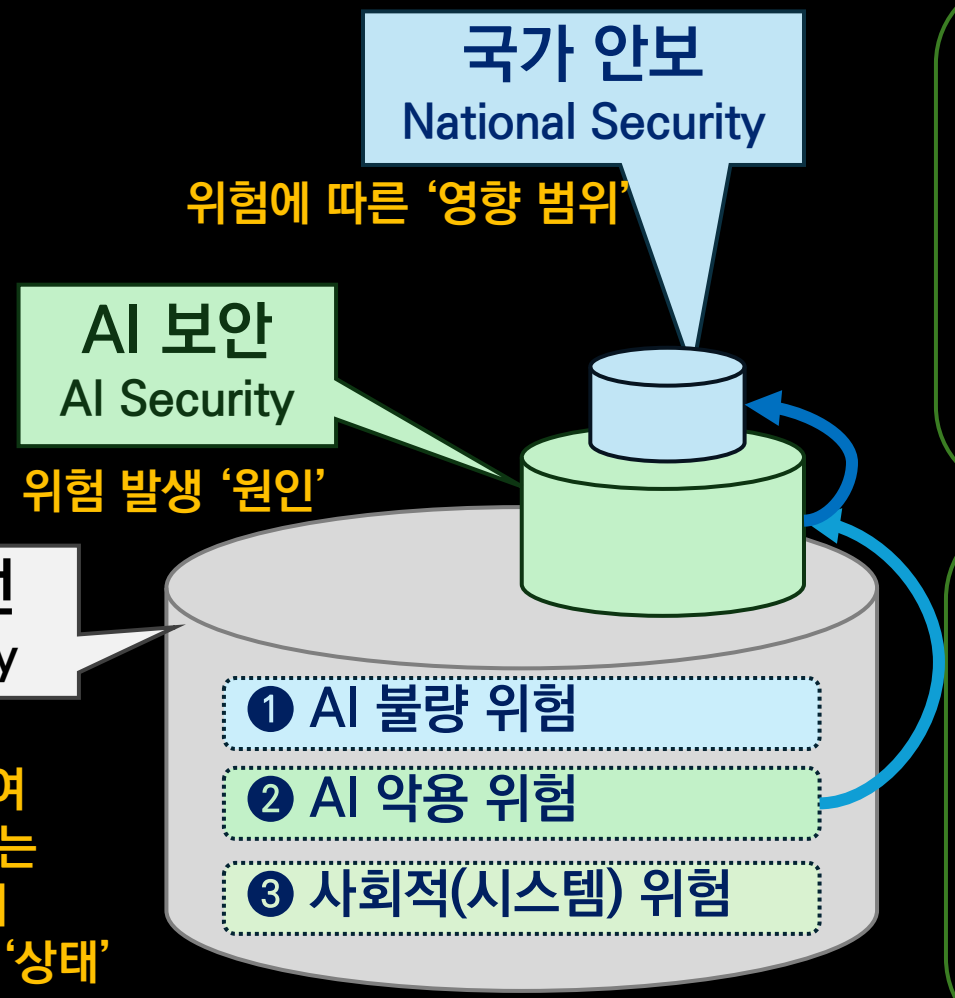
France INRIA

Kenya AISI

AISI International Network

The International Network of Advanced AI Measurement, Evaluation and Science
〈2025.12.5〉

AI 위험 *vs* AI 안전



〈인공지능기본법〉 제12조(인공지능안전연구소)
① 과학기술정보통신부장관은 인공지능과 관련하여 발생할 수 있는 위험으로부터 국민의 생명 · 신체 · 재산 등을 보호하고 인공지능사회의 신뢰 기반을 유지하기 위한 상태(이하 "인공지능 안전"이라 한다)를 확보하기 위한 업무를 전문적이고 효율적으로 수행하기 위하여 인공지능안전연구소(이하 "안전연구소"라 한다)를 운영할 수 있다.

〈인공지능기본법〉 제12조(인공지능안전연구소)
② 안전연구소는 다음 각 호의 사업을 수행한다.
1. 인공지능안전 관련 위험 정의 및 분석
2. 인공지능안전 정책 연구
3. 인공지능안전 평가 기준 · 방법 연구
4. 인공지능안전 기술 및 표준화 연구
5. 인공지능안전 관련 국제교류 · 국제협력
6. 32조에 따른 인공지능시스템의 안전성 확보에 관한 지원

맺음말

- 인공지능은 **이미 현재** 기술로서, 우리는 새로운 AI 생태계로 **전환** 중
  - 인공지능의 모든 잠재적 위험에 대비한 안전하고 신뢰할 수 있는 AI 사회 준비 필요
- 지금 세대가 "**가장 이기적인 세대**"로 전락할 위험성 존재
  - 인공지능의 놀라운 **혜택만 누리고** 위험에 빠르게 대비하지 않을 경우
  - 다음 세대를 위해 **지속가능한 AI 안전 생태계**를 미리 준비하지 못할 경우
- 〈지속가능한 AI 안전 생태계〉 조성을 위해 **모든 구성원**이 참여해야 함
  - **국민**: AI 사회 시민으로서 **AI 리터러시**(**역량 + 윤리**) 함양
  - **기업**: 인공지능 활용에 있어서 진취적 도전과 책임감 있는 자세
  - **정부**: 국민 **모두를 위한 인공지능 사회**로의 전환에 빠뜨림 없는 노력
  - **세계**: 지속적 **인류 공영**을 위한 글로벌 **연대**와 **포용**적 성장 노력