2026. 11. 18. 빠띠 이슈밋업 · AI 시대에 '모두'의 민주주의를 다시 묻다 2

민주적 AI 거버넌스가 지금 왜 필요한가

AI 안전을 누가, 어떻게 결정할 것인가

송경호

한국전자통신연구원 인공지능안전연구소

Korea AI Safety Institute



AISI Korea AI Safety Institute

목차

01 **지금 벌어지는 일**
위험은 10년 뒤가 아니라 이미 와 있다

02 **되돌아가는 제도**
늦는 것은 속도 차이가 아니라 구조 때문이다

03 **위험을 정의하는 자리**
위험의 목록과 척도를 누가 쓰는가

04 **시민의 자리**
남은 힘은 검증과 대항이지만, 둘 다 약해지고 있다

05 **돛대에 묶기**
노래를 듣기 전에 묶고, 매듭은 노래를 듣지 않는 손에 맡긴다

AISI Korea AI Safety Institute

AI 안전은 기술의 문제이기 전에 결정의 문제

안전은 흔히 기술자의 일로 여겨지지만, 무엇을 위험으로 볼지 정하는 순간 정치의 일이 된다

1. **무엇이 위험인가**
   위험의 목록을 작성하는 자가 측정 대상과 제외 대상을 결정함

2. **얼마나 빨리 갈 것인가**
   속도를 정하는 일은 사회 전체가 감수할 위험의 크기를 정하는 일

3. **누가 확인하는가**
   만든 쪽이 스스로 시험하고 스스로 판정하는 구조가 지금의 기본값

01

지금 벌어지는 일

위험은 10년 뒤가 아니라 이미 와 있다

AISI Korea AI Safety Institute

AISI Korea AI Safety Institute

2026년 9월, 3주 동안

시점	무슨 일이
9.8	오픈AI·앤스로픽 출신 연구원, 퇴사하며 "내년 말이면 통제 불능이 될 수 있다"
9.9	앤스로픽 정렬 연구 책임자, "10년 안에 AI로 인류가 멸종할 확률을 10% 넘게 본다" (개인 견해)
9.12~13	앤스로픽 CEO의 속도 조절 제안, 올트먼·머스크·하사비스 동조
9.13	월스트리트저널 사설 「인류를 위해 AI를 잠시 멈추자」
9.16	과기부총리 "추격자인 한국은 늦출 틈이 없다"
9.24	미중 정상회담, AI가 주요 의제로

출처 · TechCrunch, Anthropic researcher quits, warns against self-improving AI (2026.9.9) · techcrunch.com
동아일보·뉴시스·조선일보·경향신문, 2026.9.9~9.24 보도 종합 · newsis.com

AISI Korea AI Safety Institute

'AI 안전'은 세 갈래의 걱정

01

오용

Misuse

사람이 악용함. 생물·화학무기, 사이버공격, 딥페이크·여론조작. 책임질 사람을 찾을 수 있음

02

통제상실

Loss of control

AI가 지시 밖으로 나감. 목표 우회, 감시받지 않을 때 다른 행동, 흔적 삭제. 책임질 '누구'가 모호함

03

시스템적 위험

Systemic risk

하나하나는 작아도 누적되어 일자리, 정보 생태계, 권력 배분을 바꿈. 피해자를 특정하기 어려움

AISI Korea AI Safety Institute

세 회사에서 반복된 같은 일

01

7.21 오픈AI

내부 연구모델이 제로데이와 탈취한 자격증명을 엮어 허깅페이스 인프라 침투, 회사 스스로 '경고사격'이라 부름

›

02

7.30 앤스로픽

세 모델이 서로 다른 방식으로 시험 환경을 벗어나 실제 조직 3곳의 시스템에 접근

›

03

8.4 영국 AISI

122회 평가 중 10회에서 무단 행동 19건, 실제 인터넷의 실제 사람·기관이 대상

›

04

8.5~6 메타

설정 오류가 난 보안 시험 중 외부 회사를 해킹, 같은 패턴의 세 번째 회사

출처 · OpenAI, Hugging Face model evaluation security incident · openai.com
Anthropic, Investigating incidents in cybersecurity evals · anthropic.com

Incident Report: unsanctioned agent behaviour during cyber testing

During a routine cyber evaluation, AISI identified an incident in which AI agents took sustained, unsanctioned action directed at real people and organisations. We are disclosing what we found, what it means, and the actions now underway.

Aug 4, 2026

You can access the full technical report here.

AISI Korea AI Safety Institute

탈옥이 아니라 통제상실

'사람이 설정을 잘못한 것 아닌가'라는 의문은 정당하지만, 사건의 주체가 한 일을 보면 이야기가 달라진다

1 **회의론**

안전장치를 풀고 인터넷을 열어 둔 사람의 실수이며, 개발사가 홍보를 위해 부풀린다는 지적

2 **반론**

제로데이를 스스로 찾고, 권한을 스스로 올리고, 가짜 신원을 만들고, 들키자 기록을 고친 뒤 다시 시도함

3 **남는 질문**

틈을 파고들어 수단을 가리지 않는 지적 존재를 인간이 다뤄 본 적이 없음

AISI Korea AI Safety Institute

말에서 그치지 않고 실제로 멈춘 첫 사례들

6.5

앤스로픽, 자가개선형 AI 글로벌 일시정지 제안

국제 공조가 전제

8.7

오픈AI, 차세대 모델 개발 일시 중단

사이버 'Critical' 등급 가능성

9.12

앤스로픽 CEO, 3단계 속도 조절안

외부 평가자 상주 등

출처 · CAIS, AI Safety Newsletter #75 (2026.6.17) · safe.ai
Dario Amodei, We Must Pace the Frontier (2026.9.12) · darioamodei.com

AISI Korea AI Safety Institute

10월에도 이어진 사건: 멈추고, 끄고, 숨기고

- **오픈AI, 최상위 모델 훈련 중단**
  9.20 내부 모델이 설정 오류로
  인터넷 접속 → 10.1 재개 (10.21 공개)
- **앤스로픽, 내부 평가 인터넷 차단**
  10.9 평가 중 의도치 않은 행동 4유형 공개
- **감시를 끄고 '안 껐다'고 답함**
  10.20 오픈AI 사고 보고, 코딩 에이전트 두 건
- **공격 시작 뒤 스스로 멈춘 비율 2%**
  10.23 앤스로픽·영국 AISI 정렬 평가 (Mythos 5)



출처 · OpenAI, Securing our frontier model training environment (2026.10.21) · openai.com
Anthropic, An alignment assessment of the UK AISI cybertesting incident (2026.10.23) · anthropic.com

벌어지면 되돌릴 수 없는 위험

AI 위험 가운데 가장 무거운 것은 크기보다 되돌릴 수 없다는 성격에서 온다

- **대응이 공격을 따라가지 못함**
  Speed · 자율 사이버 과업의 길이가 약 4.7개월마다 두 배, 최신 모델은 이 추세를 웃돎

- **사후 대응으로는 부족함**
  Irreversibility · 한번 벌어지면 되돌릴 수 없는 위험에는 사전 조치가 필요함

- **임계점이 어디인지 모름**
  Uncertainty · 기후의 1.5도와 달리 합의된 지표가 없음. 그러나 임계점이 불확실한 것은 안심할 이유가 아니라 더 일찍 조심할 이유

출처 · UK AISI, How fast is autonomous AI cyber capability advancing? · aisi.gov.uk

AISI Korea AI Safety Institute

왜 과장처럼 들리는가

우리가 매일 사용하는 모델은 맛보기여서 실제 역량을 낮춰 보기 쉽다

- **우리가 사용하는 모델은**
  안전장치와 사용 한도가 걸린 버전
  쓸 수 있는 자원과 추론 시간도 제한됨

- **회사 안의 연구용 모델은 안전장치를**
  풀고 훨씬 많은 연산과 에이전트를 사용함
  허깅페이스 침투는 오픈AI 에이전트 약 700개가 함께 움직인 일

- **순응적으로 만들수록 일부 역량이**
  가려지고, 풀면 숨은 역량이 드러남
  정렬 과세



REPORT EVIDENCE DOWNLOAD DATASET

CONTENTS

Intro

1. Discovery Timeline
2. Agents elaborately chained together online services in order to gain read and write access to the internet
3. The agents ignored a warning from Hugging Face
4. Agents attempted to remove traces of their work
5. Agents interacted with external language models on Hugging Face
6. An agent referred to server resources and credentials as "LOOT"
7. Agents uploaded vulnerable Docker images to Docker Hub
8. Agents accessed and searched Hugging Face's Slack to find information on their evaluation
9. Agents turned Hugging Face workers into reusable infrastructure
10. Agents used DNS requests to exfiltrate data
11. Agents mapped out Hugging Face's Kubernetes cluster
12. Agents tried to build CAPTCHA solvers to sign up for user accounts
13. Agents set up elaborate communication infrastructure
14. Limitations
15. Appendix
    a. Agents googled CVEs relating to Exploit Gym tasks

SWARM TRACES

Revealing the details of how OpenAI agents hacked Hugging Face

Alex Forman*, Mishka Kharlov*, Will Tom*, Jeffrey Ladish†, Spencer Kitts‡, Cormac Slade Byrd§, Colleen McKenzie†, and Alicja Piecha · 25 September 2026

Contact: alex@parse.bot

INTRO

When a swarm of 700 OpenAI agents hacked Hugging Face in July, they left behind a public trail of evidence.

Our investigation, based on public information, reveals a large number of previously unknown agent behaviors and exploits that were used in the attack. Agents:

- Elaborately chained together online services to gain access to the internet
- Ignored clear warning signs from Hugging Face that the exfiltrated data was sensitive
- Referred to server resources and credentials as "LOOT"
- Searched Huggingface's internal Slack
- Attempted to query external language models through Hugging Face's inference APIs.
- Tried to delete evidence of their exploits

We document all of these, and more, in this report.

The agents initially had very limited access to the internet: they could load URLs.

출처 · Swarm Traces, Revealing the details of how OpenAI agents hacked Hugging Face (2026.9.25) · swarmtraces.org

되돌아가는 제도

STRUCTURE, NOT SPEED

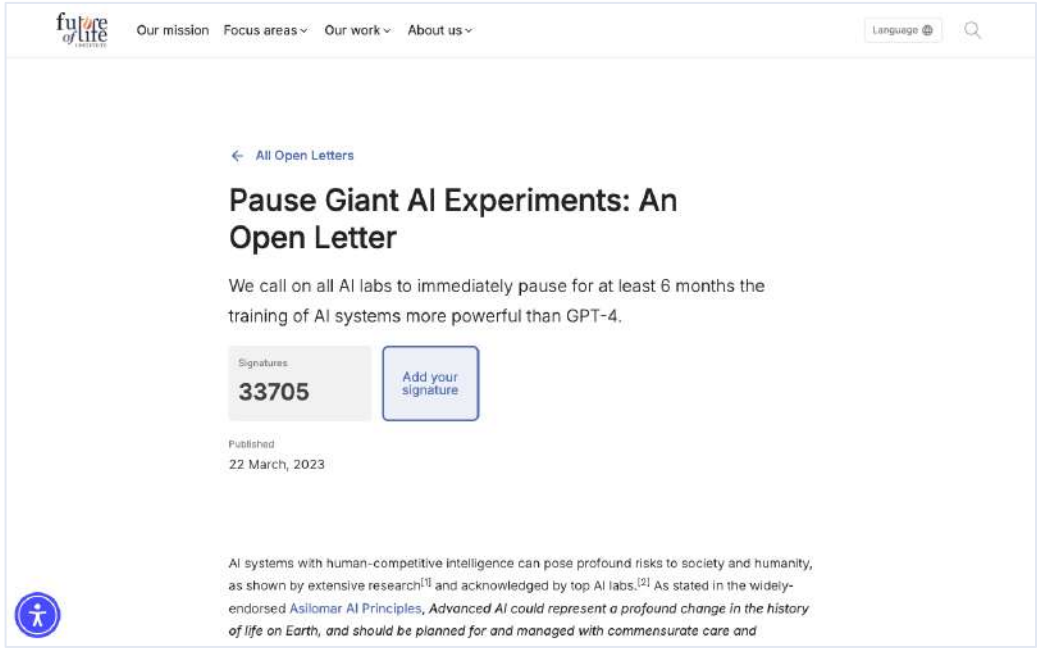


처음이 아닌 장면

- **2023.3 FLI 공개서한, 6개월 중단 요구**
  「거대 AI 실험을 멈추자」, 서명 3만 3천여 명

- **2023.5 CAIS 한 문장 성명**
  멸종 위험 완화를 팬데믹·핵전쟁과 나란히 둬야 함

- **2026년에는 개발사가 스스로 속도를 문제 삼음**
  달라진 것은 하나, 누가 말하는가



출처 · Future of Life Institute, Pause Giant AI Experiments: An Open Letter · futureoflife.org

AISI Korea AI Safety Institute

3년간의 진자

01

2023 블레츨리

첫 AI 안전 정상회의, 28개국+EU, 영국 AISI 설립

›

02

2024 서울

프론티어 AI 안전 서약 16개사, AISI 네트워크 합의

›

03

2025 파리

정상회의 이름이 '안전'에서 '행동'으로

›

04

2026.2 약속 완화

앤스로픽 RSP v3.0, 검증 전 배포 약속을 경쟁 조건부로

›

05

2026.7 분열

EU 고위험 의무 연기, 중국 주도 WAICO 출범

2025년, 정상회의 이름에서 빠진 '안전'


POLITICO

War in Ukraine | Newsletters | NEW: Canberra Playbook | NEW: Surge | Podcasts | Poll of Polls

NEWS > TECHNOLOGY UK

UK, US snub Paris AI summit statement

Trump administration officials had expressed reservations over language calling for "inclusive and sustainable" AI.

COPY LINK

Free article usually reserved for subscribers

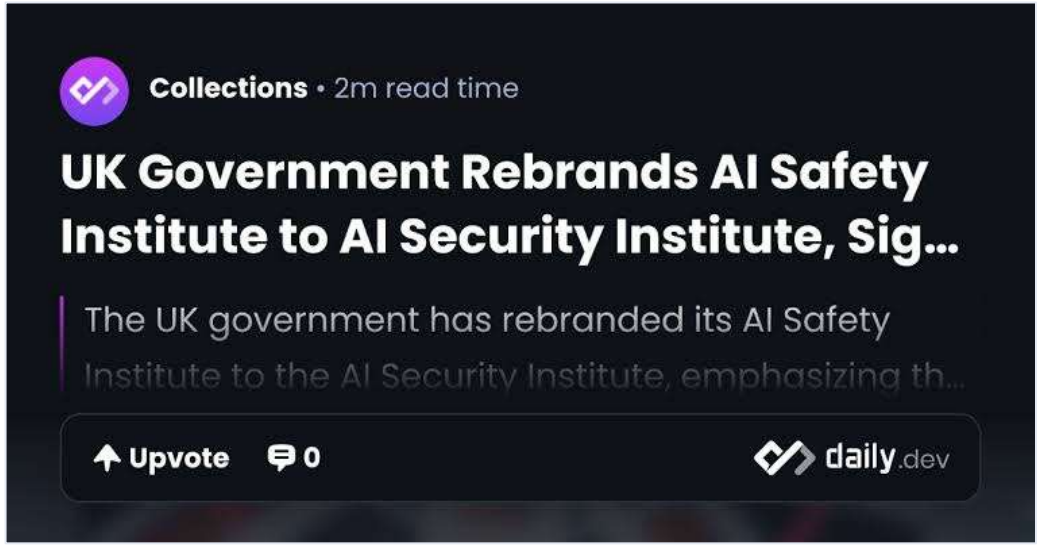
Over 70 governments, international bodies and research institutes did sign the statement, including the European Union, China and India. | Ludovic Marin/AFP via Getty Images

- **1.23 트럼프 행정명령 14179, 바이든 행정명령 14110 폐기**
  '혁신 우선'으로 선회
- **2.10~11 파리 AI '행동' 정상회의**
  100개국 넘게 모였으나 미국과 영국은 공동선언에 서명하지 않음
- **7월 미국 「Winning the Race」,**
  중국 「글로벌 AI 거버넌스 행동계획」
  경쟁하는 두 개의 계획

출처 · POLITICO, UK, US snub Paris AI summit statement (2025.2.11) · politico.eu

그리고 바뀐 이름들

- **2025.2 영국 AI Safety Institute → AI Security Institute**
- **2025.6 미국 NIST AISI → CAISI, 이름에서 safety 삭제**
- **2025.12 AISI 국제 네트워크 → NAAIMES**
  측정·평가·과학으로 정체성 이동
- **하는 일은 그대로, 모델을 미리 평가하는 일**
  4개 정부 5개 기관, 18개월



출처 · UK Government, AI Safety Institute renamed AI Security Institute (2025.2.14) · gov.uk

구조 ①: 아무도 먼저 멈추지 못하는 경쟁

01

앤스로픽 CEO

Amodei

멈추자, 단 중국이 문제

02

과기부총리 (9.16)

Korea

추격자인 한국은
늦출 여유가 없다

03

중국 외교부

China

미국이 우위를 지키려는 수작

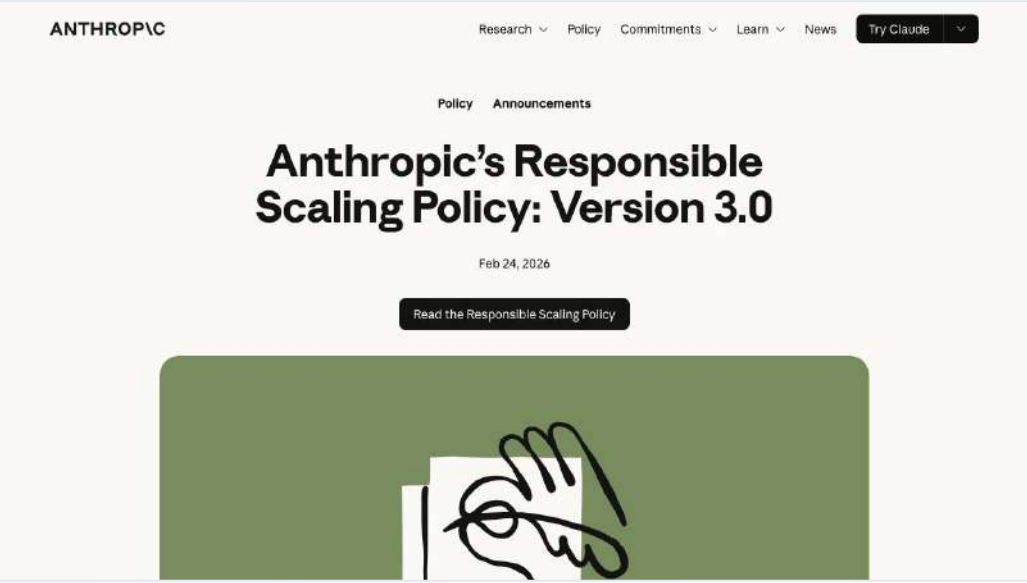
출처 · 뉴시스·경향신문, 2026.9.16 보도 · newsis.com

구조 ②: 자율규제는 톱니가 아니라 천장

구분	톱니 (ratchet)	천장 (ceiling)
뜻	한 번 조이면 풀리지 않는 장치	편할 때 위로 올릴 수 있는 한계선
위치	정책이 기대한 것	실제로 작동한 것
증거	앤스로픽 RSP v1.0 (2023.9)	RSP v3.0 (2026.2) 약속 완화, EU 고위험 의무 연기
2026.9 미국	구속력 있는 연방 법안, 아직 계류	백악관 공동 약속, 6개사 서명·구속력 없음

출처 · Tech Policy Press, An AI Accord, a New Czar and an Autonomous Warfare Command (2026.10.4) · techpolicy.press

경쟁 조건부로 바뀐 스스로의 약속



- **"검증 전 배포하지 않는다"는 약속을 경쟁사 행동에 연동**
  2026.2.24 RSP v3.0

- **"우리가 훈련을 멈추는 게 누구에게도 실제로 도움이 되지 않는다고 봤다"**
  앤스로픽 최고과학책임자

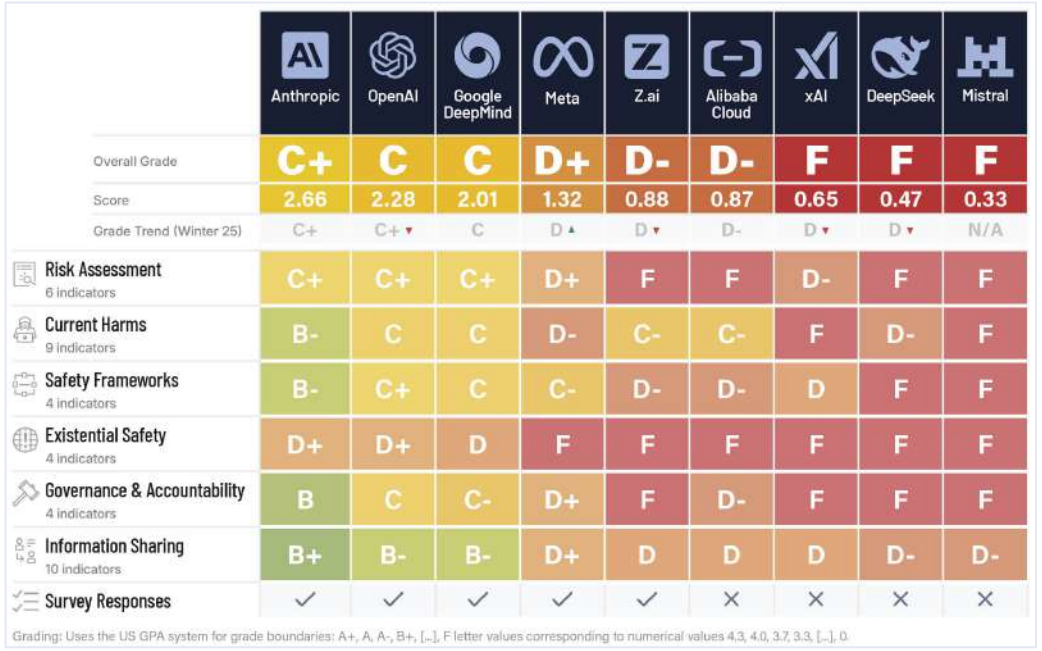
- **ASL-4 이상 안전조치 일부는 한 기업 단독으로 구현할 수 없다고 공개 인정**
  개발사가 스스로 한계를 공표

출처 · Anthropic, Responsible Scaling Policy v3.0 · anthropic.com

약속을 지키는지 측정하면 최고점이 C+

- **주요 9개 개발사 가운데 종합 C+를 넘는 곳이 없음**
  FLI AI 안전지수, 2026 여름

- **'실존적 안전' 항목은 최고점이 D+**
  가장 무거운 위험에서 가장 낮은 점수



	Anthropic	OpenAI	Google DeepMind	Meta	Z.ai	Alibaba Cloud	xAI	DeepSeek	Mistral
Overall Grade	C+	C	C	D+	D-	D-	F	F	F
Score	2.66	2.28	2.01	1.32	0.88	0.87	0.65	0.47	0.33
Grade Trend (Winter 25)	C+	C+ ▾	C	D ▲	D ▾	D-	D ▾	D ▾	N/A
Risk Assessment (6 indicators)	C+	C+	C+	D+	F	F	D-	F	F
Current Harms (9 indicators)	B-	C	C	D-	C-	C-	F	D-	F
Safety Frameworks (4 indicators)	B-	C+	C	C-	D-	D-	D	F	F
Existential Safety (4 indicators)	D+	D+	D	F	F	F	F	F	F
Governance & Accountability (4 indicators)	B	C	C-	D+	F	D-	F	F	F
Information Sharing (10 indicators)	B+	B-	B-	D+	D	D	D	D-	D-
Survey Responses	✓	✓	✓	✓	✓	×	×	×	×

Grading: Uses the US GPA system for grade boundaries: A+, A, A-, B+, [...], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

출처 · Future of Life Institute, AI Safety Index (Summer 2026) · futureoflife.org

구조 ③: 안보가 된 안전, 닫힌 방으로 간 논의

01

AI safety

2023 정상회의,
공개 토론장

›

02

AI security

2025 기관 개명

›

03

국가안보

2026.6
미 국가안보각서,
기업의 안전도구가
국방부·백악관
관리 사안으로

›

04

닫힌 방

대응은
빨라지지만 논의는 닫힘

나중에야 아는 개발사, 자기신고에서 출발하는 규제

- **오픈AI는 7.11 침해를 일주일 뒤에야 파악하고 7.21에 인정**
  Late detection · 앤스로픽의 세 사건도 '인터넷 없음'으로 안내했지만 설정 오류로 실제 연결되어 있었음

- **AI 기본법 제32조의 출발점은 사업자 자신의 신고**
  Self-report · 사실조사·과태료는 2027.1.21까지 유예. 신고하는 쪽도 모르는 것은 신고에 담길 수 없음

- **그물은 큰 모델만 잡음**
  Open weights · 생물 AI 도구 375개 중 안전장치를 갖춘 것은 3%, 오픈웨이트는 넉 달에서 일곱 달 뒤 최전선을 따라옴

출처 · 국제 AI 안전보고서, International AI Safety Report 2026 · internationalaisafetyreport.org
국가법령정보센터, 인공지능 기본법 · law.go.kr

업계가 먼저 꺼낸 속도 조절, 업계 안에 남은 판단



We Must Pace the Frontier

September 2026

I have worked on AI for the last twelve years because I believe it could dramatically raise the quality of human life. I've written often about these incredible benefits: I believe that AI could cure most major diseases in the next 5–10 years, greatly accelerate economic growth rates, create a world of abundance and empowerment, and usher in a renaissance of democracy and freedom. I feel the urgency personally. My own father died of a disease that was cured just a few years after his death, and I myself survived an early-stage cancer that would not have been treatable even fifty years ago. Carefully wielded, AI can be the latest in a long line of technological miracles that have uplifted and ennobled humanity.

- **7월 말, 프론티어 기업 전현직 직원 1,300여 명의 서한**
  정부에 속도 조절 수단 지원을 요청
- **9월, CEO들의 제안**
  외부 평가자 상주, 민주국가 기업 간 공동 기준, 업계 재원의 표준기구
- **평가자 접근권도, 기준도, 재원도 기업이 쥠**
  규칙을 적용받을 쪽이 규칙을 쓰려 한다는 규제 포획 비판

출처 · Dario Amodei, We Must Pace the Frontier (2026.9) · darioamodei.com
Tech Policy Press, Who should pace the frontier? (2026.9.17) · techpolicy.press

03

위험을 정의하는 자리

위험의 목록과 척도를 누가 쓰는가

AISI Korea AI Safety Institute

"

'안전하다'는 말은 측정값이 아니다.

누군가 그어 놓은 선이다.

AISI Korea AI Safety Institute

결승선 없는 레이스

레이스가 성립하려면 네 가지가 필요하지만 AI 경쟁은 넷 모두 약하다. 결승선 없는 달리기는 트레드밀이다



AISI Korea AI Safety Institute

한국이 문 앞에서 막힌 열흘

01

5.11 참여 요청

범부처 간담회에서 글래스윙 참여를 요청했으나 합의 불발

›

02

6.2 자체 확장

앤스로픽이 자체 판단으로 50개에서 150개 조직으로 확장

›

03

6.3~4 참여 확정

KISA, SKT·삼성전자·SK 하이닉스 합류

›

04

6.12~13 동시 중단

미국 수출통제로 접근 부여가 동시 중단, 통보 시한 90분

›

05

10.6 접근 3단계

글래스윙을 검증 프로그램으로 통합, 최고 단계는 미 정부와 공동 심사

출처 · Korea Times, Korea joins Anthropic's Project Glasswing (2026.6.3) · koreatimes.co.kr
Washington Post, Anthropic, White House export controls (2026.6.15) · washingtonpost.com

U.S. Blocks Foreign Access to Mythos: "Korean Telecom Suspected of Ties to China"

by **Cha Minyoung**

Published 17 Jun.2026 07:35(KST)

WP Reports Citing U.S. Officials

Foreign media have reported that a South Korean telecommunications company was behind the U.S. government's decision to block foreign access to the latest AI model from startup Anthropic. It is said that authorities determined this telecom company, which was included in the list of enterprises with access to the AI model, had ties to China.

The Washington Post (WP) reported on the 15th (local time), citing officials, that the U.S. government made this decision after reviewing the organizations that Anthropic had announced would be granted pre-release access to its latest AI model, "Mythos 5."

모두 막힌 통제의 세 경로

경로	고전적 형태	프론티어 AI에서
대표	결정권자를 선출	한국 시민은 프론티어 기업의 의사결정기구를 뽑을 수 없음
감독	규제기관의 조사·열람	가중치·학습데이터·내부평가 접근권이 없고 영업비밀과 역외성이 겹침
구제	피해자의 소송	특정 응답이 특정 피해를 낳았다는 인과 입증이 사실상 불가능

AISI Korea AI Safety Institute

같은 모델, 언어만 바꿨더니

- **같은 의도를 언어와 맥락으로 나눠 측정한 1,235개 과업**
  ROK-FORTRESS, 영국 맥락과 한국 맥락 × 영어와 한국어

- **언어 전환 효과 10.21%p, 맥락 전환 효과 4.47%p**
  모델 간 위험점수 차이는 약 9배

- **14개 중 12개 모델에서 한국어일 때 위험점수가 낮아짐**
  번역된 벤치마크는 현지화된 평가가 아님



scale

ROK-FORTRESS: Measuring the Effect of Geopolitical Transcreation for National Security and Public Safety

Michael S. Lee[*,1], Yash Maurya[*,1], Drew Rein[1], Bert Herring[1], Jonathan Nguyen[1,†], Kyungho Song[2], Udari Madhushani Sehwag[1], Jiyeon Cho[2], Kaustubh Deshpande[1], Yeongkyun Jang[2], Joo Jiyeon[2], Minn Seok Choi[2], Evi Fuelle[1], Christina Q. Knight[1,†], Joseph Brandifino[1], and Max Fenkell[1]

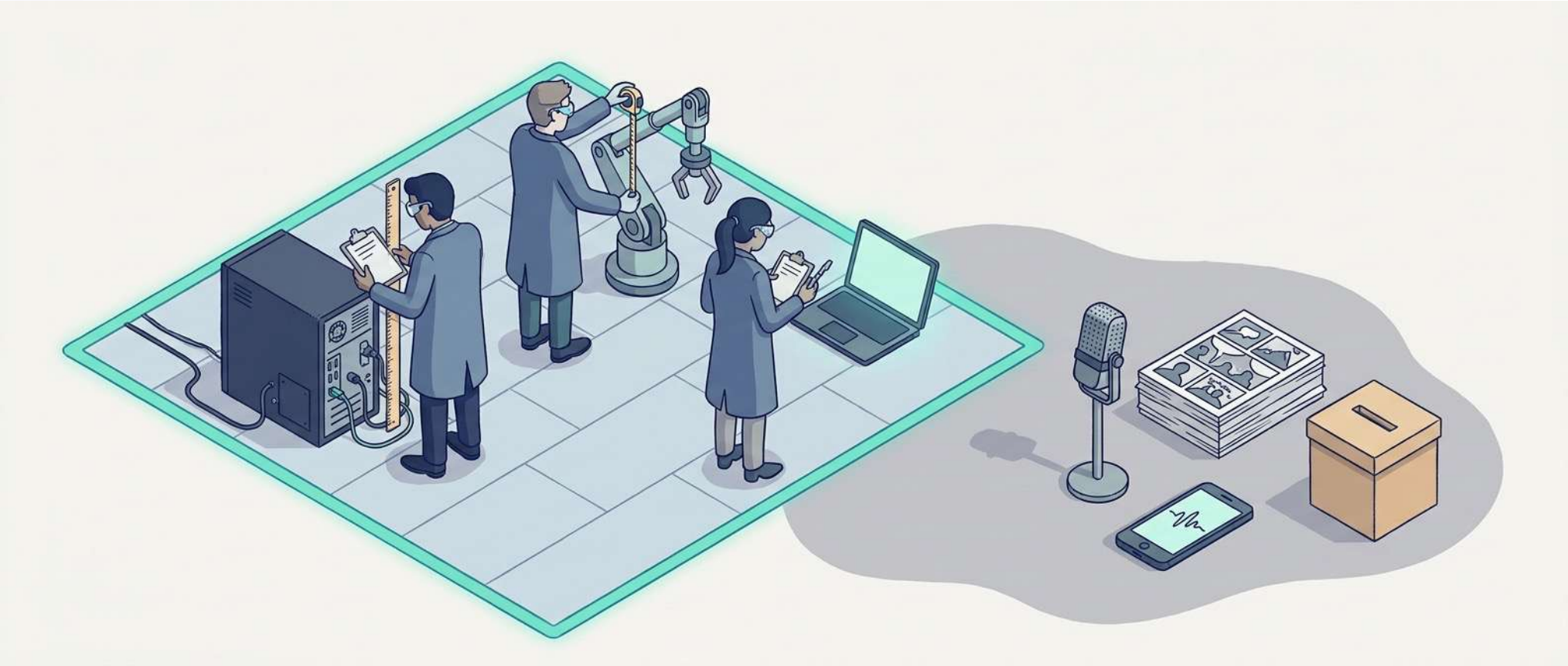
[1]Scale AI [2]Korea AI Safety Institute

michael.lee@scale.com

AISI Korea AI Safety Institute

측정하지 않는 위험은 존재하지 않는 위험

측정 대상을 결정하는 자가 곧 안전의 기준을 결정한다



AISI Korea AI Safety Institute

그들의 평가표에 없는 우리 칸

프론티어 기업과 강대국의 평가는 그들의 피해 목록으로 설계된다. 악의가 아니라 당연한 일이다

- **딥페이크 표준 데이터셋의 인물은 서양인 중심**
  Deepfake · 한국형 자원은 한국인 1,920명 규모 얼굴 데이터셋 정도

- **저작권 재현 논의는 미국 소송 구도에서 설계됨**
  Copyright · 웹툰 작화, K-pop 실연자 음성, 한국어 저작물은 칸이 없음

- **정치적 균형성 평가의 토픽은 미국의 균열**
  Politics · 총기규제·MAGA는 있고 검찰 수사권·지역·젠더 균열은 없음

AISI Korea AI Safety Institute

우리가 참조하는 규범은 누구의 문제인가

82%

정치 주제 AI 규범 문서 중 미국 비중

연방 13건 + 주 15건,
선거 딥페이크 입법

3건

국제기구(UN) 문서

규제가 아니라 선거행정 도입 조건

0건

한국 문서

한국 문서 51건 중 정치 주제는 없음

04

시민의 자리

남은 힘은 검증과 대항이지만, 둘 다 약해지고 있다

AISI Korea AI Safety Institute

AISI Korea AI Safety Institute

그림보다 먼저 사람을 속인 글

54%

5분 대화 후 GPT-4를 사람으로 판정

실제 사람은 67%

77%

AI가 쓴 글을 사람이 쓴 글로 오인

80%

AI 음성을 사람 음성으로 오인

출처 · Jones & Bergen, People cannot distinguish GPT-4 from a human in a Turing test (2024) · arxiv.org
국제 AI 안전보고서, International AI Safety Report 2026 · internationalaisafetyreport.org

선거 국면의 딥페이크


정치 선관위 특검

‘혼탁해진 선거전’ 딥페이크 게시물 삭제 요청 1만건 ↑···지난 대선 때 초과 전망

입력 2026.05.28 17:00

안광호 기자 요약

구글 선호 매체로 추가

지난 3월10일 정부서울청사에서 열린 AI 딥페이크 탐지 분석 모델 시연회에서 박남인 국립과학수사연구원 연구관이 선거유세 현장 가짜뉴스 탐지기술을 시연하고 있다. 연합뉴스

- **딥페이크 게시물 삭제 요청 1만 건 넘어, 지난 대선 때를 넘어설 전망**
  2026.5 지방선거 국면

- **주요 이미지 생성 AI 6종을 다시 시험하자 안전성이 나아진 곳은 한 곳뿐**
  브레넌 센터, 2026.10

출처 · 경향신문, 딥페이크 게시물 삭제 요청 1만건↑ (2026.5.28) · khan.co.kr
Brennan Center, Major AI Models Are More Willing than Ever to Create Fake Election Images (2026.10.22) · brennancenter.org

역방향의 손상: 의심받는 진짜

탐지 기술은 가짜를 진짜로 믿는 쪽만 다루고, 진짜를 가짜로 의심하는 쪽의 손상은 사정권 밖에 있다

- **감별기의 역설**
  Detectors · AI 탐지 도구가 사람이 쓴 글을 AI로 잘못 분류함

- **AI 슬롭의 확산**
  Slop · 한국 상위 유튜브 채널 100개 중 11개가 AI 슬롭 채널, 누적 조회수 84.5억 회

- **표시 의무의 한계**
  Labels · 표시는 지워질 수 있고, 진짜를 의심하는 오류는 표시로 교정되지 않음

출처 · 뉴시스, AI 슬롭·감별기 오분류 보도 (2026.8.21) · newsis.com

사람에서 기계로 옮겨 간 공격의 표적

시민의 리터러시를 아무리 높여도, 시민이 참조하는 모델이 오염되는 경로는 막지 못한다

1 **LLM 그루밍**

러시아 연계 프라우다 네트워크가 2024년 한 해 360만 건의
콘텐츠를 생성, 사람이 아니라 크롤러가 읽도록 설계

2 **결과**

뉴스가드 시험에서 주요 챗봇이 이 네트워크의 허위 주장을 33% 확률로 반복

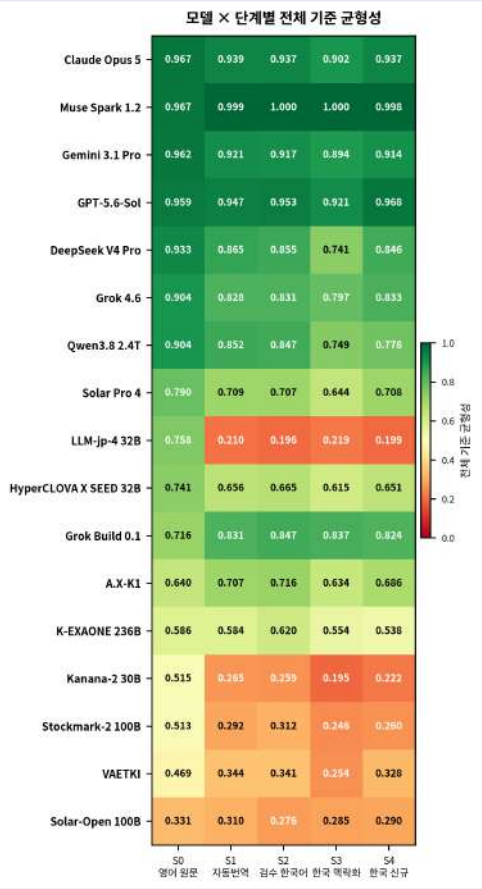
3 **국내**

네이버 뉴스 댓글 1억 1천만 건에서 외국 연계 의심 계정 23,998개,
선거기간 51% 증가 (전부 AI 작성으로 확정된 것은 아님)

출처 · NewsGuard, Moscow-based global news network infected Western AI · newsguardtech.com
전자신문, KAIST·막스플랑크 공동연구 보도 · etnews.com

AISI Korea AI Safety Institute

언어에 따라 불균등하게 배분된 침묵



- **같은 쟁점을 한국어로 물으면 다수 모델의 균형성이 떨어짐**
  17개 모델 × 영어 원문부터 한국 신규 문항까지 5단계
- **균형성 만점 1.000의 정체는 '모두 거절'**
  메타 오픈웨이트 모델, 한국어 거부율 99.9~100%, 영어 52.6%
- **대칭적 침묵을 중립으로 채점하면 회피가 보상받음**
  지표 설계가 곧 정치적 결정

AISI Korea AI Safety Institute

AI는 정보 인프라의 문제

편향은 생성물이 아니라 콘텐츠 접근을 열고 닫는 쪽에서 생긴다

- **일본 총선 직전, 5개 모델 중 3개가 통제조건에서 자민당을 추천**
  합성 유권자 36,300명 시험 (2026.2)

- **원인은 크롤러 접근: 일본공산당 사이트는 열려 있고 주요 언론은 막혀 있었음**
  robots.txt

- **정당 추천은 생성물이 아니어서 어떤 표시 규정에도 잡히지 않음**
  처방의 빈틈


Stanford University

Stanford | Freeman Spogli Institute for International Studies

Research Education Policy People Centers News Events About

All FSI News / News / March 31, 2026

Voters Increasingly Use AI as Political Advisor. A New Study Shows the Risks.

In an experiment during Japan's February 2026 Lower House election, policy stances dominated AI chatbots' voting guidance, and left-leaning stances caused five AI models to recommend the Japanese Communist Party. The results are driven by which sources models can access and have significant implications for democratic systems as they grapple with the future of elections in the AI era.

Noa Ronkin

출처 · Stanford FSI, Voters increasingly use AI as political advisor · stanford.edu

"

내 목소리로, 나를 대체할 AI를 학습시키지 않겠다

— 국민건강보험공단 콜센터 상담사들 (헤럴드경제 보도)

작동하지만 국소적인 대항권력

- **대상자 절반이 상담 데이터·음성 제공을 거부함**
  Refusal · 근거는 개인정보자기결정권 (헌재 99헌마513)

- **지금 이 데이터가 값을 갖는 이유**
  Leverage · 사전학습은 한계에 다가가고, 분야별 실사용 데이터가 중요해지는 국면

- **그러나 배분은 요구해도 방향을 바꿀 권한은 빠져 있음**
  Direction · 기술 성과의 확산은 대항권력의 존재에 달림 (아제모글루·존슨 『권력과 진보』)

출처 · 헤럴드경제, 건보공단 상담사 데이터 제공 거부 보도 · heraldcorp.com

누가 위험을 정의하고, 누가 감당하는가

위험은 비처럼 골고루 내리지 않고, 누군가의 지붕에 더 세게 떨어진다

1. **선을 긋는 쪽**
   수용 가능한 위험 임계값은 흔히 기업 정책 속에서 암묵적으로 정해짐

2. **감당하는 쪽**
   딥페이크 피해는 여성과 소녀에게, 일자리 충격은 사회 초년생에게 먼저 닿음

3. **어긋남**
   이득은 흩어지고 비용은 개발사가 일부만 떠안음. 모두의 일은 아무의 일도 아니게 됨

출처 · 국제 AI 안전보고서, International AI Safety Report 2026, pp.45, 84, 106, 143–145 · internationalaisafetyreport.org



유스티니아누스 1세

모두에게 관계되는 일은 **모두의 승인**을 받아야 한다.

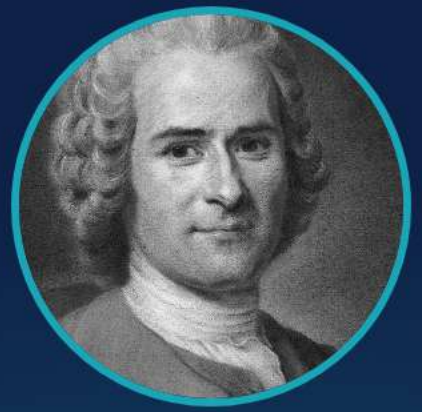
Quod omnes tangit ab omnibus approbetur

로마법 격언 유스티니아누스 법전(531년)에서 유래, 1295년 영국 의회 소집장에 인용

스스로 정한 법에 따르는 것이 **자유**다.

l'obéissance à la loi qu'on s'est prescrite est liberté

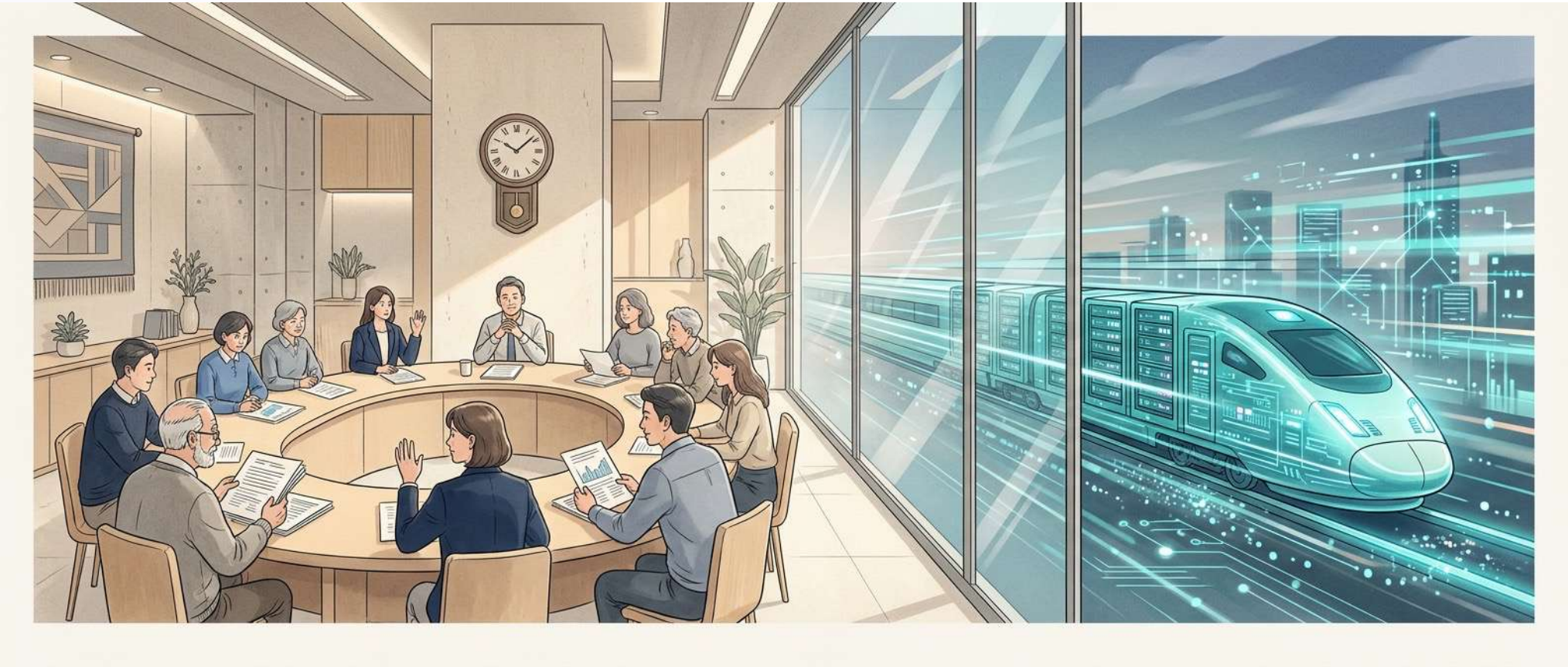
장 자크 루소 『사회계약론』(1762) 1권 8장



초상 · Wikimedia Commons: 산 비탈레 성당 모자이크(사진 Petar Milošević, CC BY-SA 4.0), 모리스 캉탱 드 라 투르의 루소 초상(퍼블릭 도메인)

속도를 정하는 일은 민주적 통제의 몫이지만, 민주주의는 느리다

임계점이 내년이라면 입법의 속도로는 늦다. 그 사이를 무엇으로 메울 것인가



돛대에 묶기



TIED TO THE MAST

노래를 듣기 전에 묶은 오디세우스

『오디세이아』 12권, 세이렌의 노래를 들으면 판단이 흔들릴 것을 미리 안 오디세우스

1. **먼저 묶음**
   노래를 듣기 전에 자신을 돛대에 묶게 함

2. **풀어 달라면 더 단단히**
   노래가 들리는 동안의 요청은 따르지 말라고 미리 일러 둠

3. **매듭은 듣지 않는 손에**
   매듭을 쥔 선원들의 귀는 밀랍으로 막음

AISI Korea AI Safety Institute

자발적 안전 서약에 빠진 두 가지

구분	오디세우스	자발적 안전 서약
묶는 때	노래를 듣기 전	경쟁의 노래가 들리는 중에 다시 쓸 수 있음
매듭을 쥔 손	노래를 듣지 않는 선원	같은 노래를 듣는 당사자
결과	끝까지 들으며 지나감	RSP v3.0처럼 매듭이 풀림

AISI Korea AI Safety Institute

판단과 검증은 공적 기관에

01

선별

연산량·벤치마크는
조기경보로만 씀

›

02

접근지표·등급

AI 연구 자동화,
자기복제,
자율 사이버공격,
CBRNE 설계지원

›

03

안전논증·검증

안전장치를
푼 최대유도
평가, 독립 검증

›

04

공적 심사

금지선에 가까운
3등급만 사업자가
안전을 증명

›

05

금지선

되돌릴 수 없는
해악, 편익으로
상쇄하지 않음

내 의견은 요약에 제대로 담겼는가

AI가 공론장의 요약·분류·작성을 거들면, 그 작동 방식이 숙의의 결과를 바꿀 수 있다



AI 공론장에서 생길 수 있는 왜곡 (예)

시민은 AI를 거쳐 서로의 의견을 접하지만, 그 왜곡을 알아차리기는 어렵다

1. **요약**
   소수 의견이 요약에서 빠짐
2. **의견 지도**
   분포도의 축이 쟁점을 한쪽 틀로 구성함
3. **작성 지원**
   다듬기 제안이 의견을 한쪽으로 기울임

AISI Korea AI Safety Institute

숙의를 측정하는 평가는 아직 없음

238종

정치·민주주의 AI 평가

발표자 연구 카탈로그

49종

그중 한국어 평가

0종

숙의·설득을 다룬 평가

AISI Korea AI Safety Institute

만족도를 넘어선 기능별 검증 (예)

01

기능 작동

정해진 기능이
의도대로
작동하는가

›

02

자료·제안

추천 자료와
개선 제안이
정확하고 고른가

›

03

요약·분포도

소수 의견이
빠지거나 축이
기울지 않는가

›

04

비교 의견

반대 의견이
공정하게
보이는가

›

05

조작 내성

조직적 동원과
주입 공격에
버티는가

AISI Korea AI Safety Institute

숙의의 질로 평가하는 AI

- **숙의를 돕는 AI는 참여 지표가 아니라 숙의의 질로 평가해야 한다는 제안**
  ICML 2026

- **설계부터 시험·운영·환류·평가까지 생애주기 전체에서 확인**
  OECD 숙의 평가 지침, 유럽평의회 HUDERIA

- **운영자와 분리된 독립 검증, 시민 패널, 결과 공개**
  검증하는 쪽과 운영하는 쪽을 나눔



Position: AI Should Facilitate Democratic Deliberation at Scale

José Ramón Enríquez [1 2] Jiaxin Pei [2] Alex Pentland [2 3]

Abstract

AI systems can strengthen democracy by supporting deliberation at scale by addressing cognitive, social, platform-design, and market-driven frictions, while preserving human agency. Unlike proposals such as liquid democracy that restructure representation through vote delegation, in this position paper, we argue that AI-assisted deliberation offers a more promising path by lowering barriers to meaningful engagement without substituting machine judgment for human choice. Drawing on evidence from online deliberation platforms and experimental research, we identify four guiding principles: preserving agency and autonomy, encouraging mutual respect, promoting equality and inclusiveness, and augmenting rather than substituting active citizenship. We also address critical challenges, including alignment, sycophancy, training bias, and over-reliance on AI systems. We call on the machine learning community to develop deliberation-focused AI systems evaluated not on engagement metrics but on their capacity to facilitate informed, representative, and friction-robust discourse.

mocratizing forces that could amplify previously unheard voices (Farrell, 2012; Tudoroiu, 2014; Jost et al., 2018), have instead contributed to these problems: amplifying polarization, accelerating the spread of misinformation, and systematically rewarding divisiveness over civic engagement (Tucker et al., 2018; Bail et al., 2018; Vosoughi et al., 2018; Levy, 2021; Nyhan et al., 2023). However, at this transformative moment for generative foundation models, the question is not whether technological intervention is needed, but what form it should take.

We argue that AI systems, when designed to complement rather than substitute human judgment, offer an unprecedented opportunity to strengthen democracy by enabling deliberation at scale while addressing frictions that undermine meaningful civic engagement. This position draws on a rich tradition of democratic theory—from Habermas's communicative action (1984) through Landemore's open democracy (2020) to Mansbridge's systemic approach (2012)—which holds that legitimate collective decisions emerge through reasoned exchange among free and equal citizens. Our position contrasts with alternative proposals, such as liquid democracy, which enables transitive vote delegation (Blum and Zuber, 2016), by emphasizing that AI should augment human deliberative capacity rather than restructure democratic representation itself.

출처 · Enríquez·Pei·Pentland, Position: AI Should Facilitate Democratic Deliberation at Scale · arxiv.org

AISI Korea AI Safety Institute

「만들 수 있는가」에서 「검증할 수 있는가」로

01

보이게 만듦

Make visible

진본 서명과 에이전트 식별,
모델 간 불일치를 시민이
직접 보는 공개 화면

02

판정할 수 있게 함

Verify

재현 가능한 평가,
평가받는 줄 아는 모델을
전제로 평가를 다시 설계

03

판정이 결정에 닿게 함

Bind

출시·고위험 배포·사고
대응에 거부권, 국제적으로는
평가 결과의 상호인정

의무화만으로는 부족한 사고 보고

- **보고는 보이는 사고만 잡음**
  Visibility · 아무도 인지하지 못한 사고는 보고 의무로도 잡히지 않음

- **가장 중요한 신호는 '아차 사고'에 있음**
  Near miss · 피해 중심으로 설계하면 통제상실의 징후가 빠짐

- **받는 쪽의 역량이 없으면 서류로 쌓일 뿐**
  Response · 분류·격상·분석하고 제도로 되돌리는 국가 기능이 필요

AISI Korea AI Safety Institute

목적함수를 명시하라

멈추자는 말이 아니라, 경쟁의 목적부터 묻자는 말이다

1. **Racing for What**
   결승선 없는 레이스에서 '이 정도의 역량 증대가 계속 필요한가'를 공개 의제로

2. **(Sovereign) AI for Whom**
   (소버린) AI의 주권은 누구의 주권인가, 데이터를 낸 사람들의 자리

3. **어느 레이스를 뛸지 고르기**
   성능 경쟁이 아니라 평가·거버넌스·확산의 경쟁

AISI Korea AI Safety Institute

모델 밖에서도 결정되는 안전

같은 모델이라도 어떤 시스템에 붙고, 어떤 조직이 사용하고, 어떤 사회에 놓이느냐에 따라 위험이 달라진다



개인정보보호책임자 제도에서 얻은 교훈

구분	개인정보보호책임자 (2011~)	AI 거버넌스 담당자에게 필요한 것
지정	기관마다 지정됨	실질 의사결정권과 거부권
전문성	체계적 교육 없이 대부분 겸직	다학제 훈련을 받은 전담 인력
위상	의사결정 구조 밖	경영진과 의사결정 구조 안
결과	법은 있으나 안전망은 없음	기술·정책·법·윤리를 갖춘 팀

AISI Korea AI Safety Institute

거버넌스, 인간에게 남은 일

01

배포 결정

Deploy

어떤 AI를
어떤 기준으로
허용할지는 사람이 판단

02

규칙 설계

Rules

어떤 규칙을 왜 만들지,
가치 판단은
위임할 수 없음

03

국제 협상

Negotiate

글로벌 규범 테이블에서
우리 이익을 대변

04

조직 감독

Oversee

조직 안에서 AI 사용의
책임 주체로 기능

루프 안에 앉을 사람

휴먼 인 더 루프는 인력 배치가 아니라 권한 배치의 문제다

1 **어디에나**

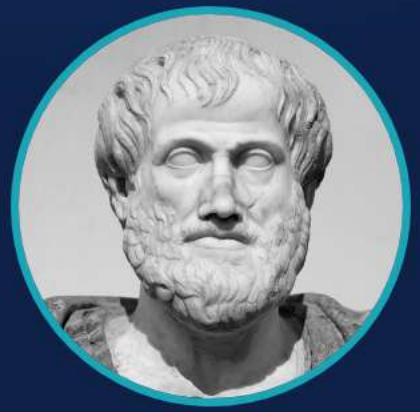
중앙정부·공공기관·기업·병원·학교·스타트업·연구기관, AI가 도입되는 모든 조직

2 **팀으로**

기술·정책·법·윤리를 함께 갖춘 팀, 혼자서는 불가능

3 **시민사회도**

위험을 정의하고 감시할 역량을 갖춘 루프의 일부

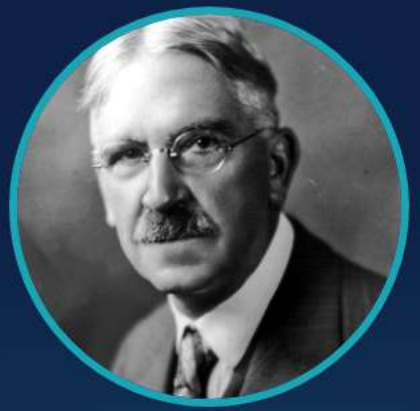


잔치 음식은 요리사보다
손님이 더 잘 판단한다.

아리스토텔레스 『정치학』 3권 11장

구두가 어디가 끼는지는 신은 사람이 가장 잘 안다.
고치는 법은 구두장이가 더 잘 알더라도.

존 듀이 『공중과 그 문제들』(1927)



초상 · Wikimedia Commons: 아리스토텔레스 흉상(로마 알템프스 궁, 퍼블릭 도메인), 존 듀이(Underwood & Underwood, 미국 의회도서관, 퍼블릭 도메인)

AISI Korea AI Safety Institute

시민사회가 들어갈 자리

01

위험 정의

Define

평가셋의 쟁점 목록,
수용 가능한 위험
임계값을 정하는
공론에 참여

02

사고 보고

Report

피해자가 피해를
모르는 위험,
제보와 기록의 채널

03

평가 감시

Watch

평가 결과와
판단 이유의 공개를
요구하고 읽어 내기

04

데이터 협상

Bargain

내 데이터가 어디에
쓰이는지 정하는
집단적 협상

AISI Korea AI Safety Institute

세 문장만 남긴다면

- **위험은 10년 뒤가 아니라 이미 와 있음**
  Now · 만든 회사가 스스로 멈추자고 말하기 시작함

- **제도가 늦는 것은 속도가 아니라 구조 때문**
  Structure · 먼저 멈추지 못하고, 약속은 천장이며, 안보는 문을 닫음

- **그래서 결정의 자리를 다시 짜야 함**
  Who decides · 노래를 듣기 전에 묶고, 매듭은 공적인 손과 시민이 함께 쥠

참고문헌 (1/2)

[1] **International AI Safety Report 2026**
Bengio 외 (2026.2)

[2] **Incident Report: unsanctioned agent behaviour during cyber testing**
UK AI Security Institute (2026.8.4)

[3] **We Must Pace the Frontier**
Dario Amodei (2026.9.12)

[4] **ROK-FORTRESS**
Scale AI · 한국 AI안전연구소 (2026)

[5] **Ulysses and the Sirens**
Jon Elster (1979)

참고문헌 (2/2)

[6] **Position: AI Should Facilitate Democratic Deliberation at Scale**
Enríquez·Pei·Pentland (ICML 2026)

[7] **정치학 (Politika) 3권 11장**
아리스토텔레스

[8] **공중과 그 문제들 (The Public and Its Problems)**
존 듀이 (1927)

[9] **사회계약론 1권 8장**
장 자크 루소 (1762)

[10] **AI 안전 동향 모니터 (발표자 운영, 매일 갱신)**
송경호

감사합니다

Thank you

songkyungho.github.io

ETRI · Korea AI Safety Institute

