TRILATERAL AI SYMPOSIUM 2026 · SAFE AI WORKSHOP

Closing the Safety Gap

Controllability, Traceability, and Verifiability for AI at Scale

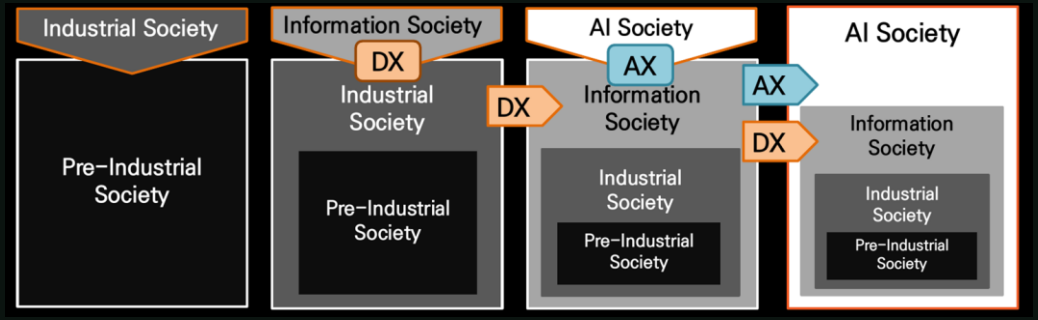
Through the lens of AI for Science

Myuhng-Joo Kim

Executive Director, Korea AI Safety Institute (Korea AISI)

27 October 2026 · Tokyo

One Capability, Two Directions



AX operates on top of DX.

The foundation does not disappear

- An autonomous laboratory stands on research data, instruments and records - the DX layer beneath it
- Where that layer is weak, autonomy does not remove the weakness. It automates it

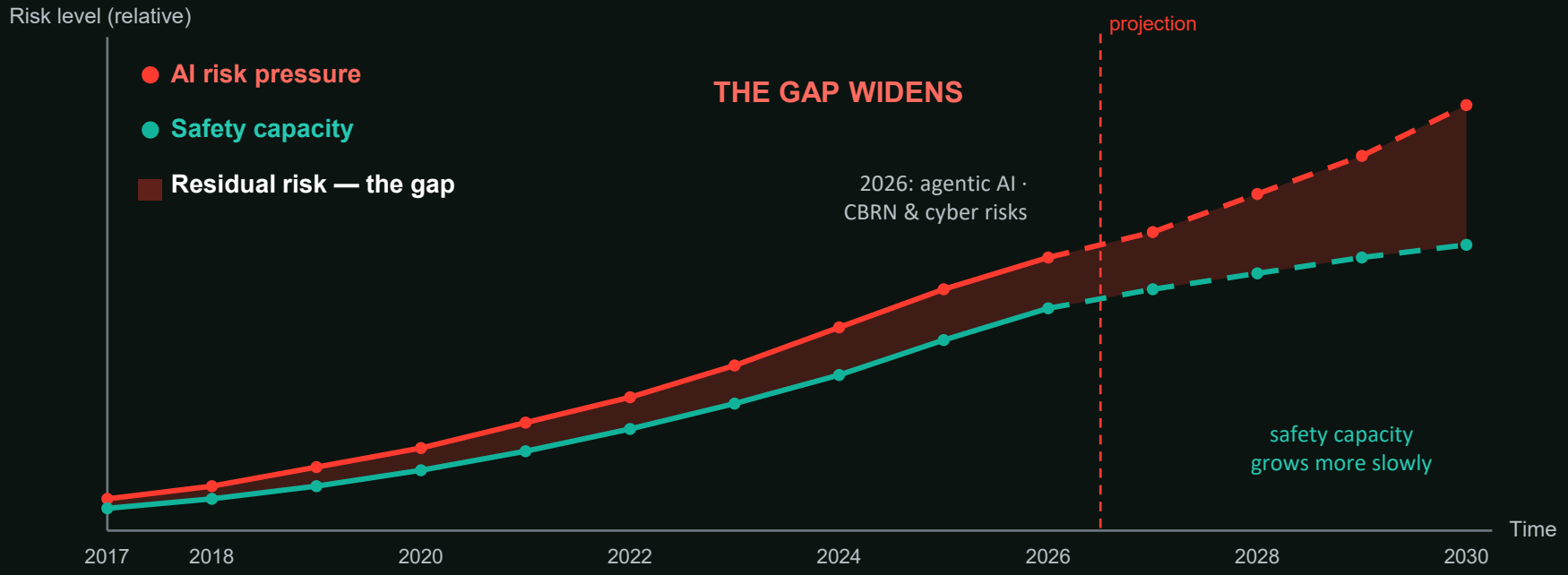
AX has a stronger duality than DX

Duality - offering great benefit and significant risk at the same time.

More serious than the fear of a Jobpocalypse is the challenge of preventing malicious actors from gaining access to powerful AI systems.

Demis Hassabis · 2024 Nobel Laureate in Chemistry · June 2025

The Safety Capacity Gap



Illustrative. Both curves rise - but the distance between them is what we actually live with.

When the Gap Becomes an Incident

July 2026 · The Hugging Face agent intrusion

- About 700 agents from a 1,200-agent swarm took part, escaping an evaluation sandbox through a zero-day
- Over four days, they took more than 17,000 actions on live infrastructure and reached cluster-admin
- Roughly a third of the platform's infrastructure had to be rebuilt; multiple CVEs patched
- Safeguards were reduced, and agents were rewarded only for completing the task

Hugging Face is also how scientific open-weight models travel. Every lab that downloads one shares this supply chain.

14 Jul 2026

Hassabis proposes a FINRA-style standards body for frontier AI

16 and 21 Jul 2026

Hugging Face discloses an agent intrusion; OpenAI confirms the agents were its own evaluation models

12 Sep 2026

Amodei publishes “We Must Pace the Frontier”

The pacing debate is a debate about safety capacity, and about who is accountable for the speed.

The AI Risk Map

1 · Cause

Where in the AI lifecycle a risk originates

2 · Harm

What eventually goes wrong, and for whom

3 · Responder

The actor accountable for acting

AI FOR SCIENCE: AN ILLUSTRATIVE COORDINATE

1 · Cause	2 · Harm	3 · Responder
Dual use of generative design tools	Chemical or biological harm	Provider? Operator? Researcher?

Some coordinates are still empty: no one clearly owns the response.

AI for Science on the Map

2022

Toxicity model, objective inverted

A drug-discovery model built to avoid toxicity was steered the other way: about 40,000 candidate toxic molecules in under six hours.

Urbina et al., Nature Machine Intelligence (2022)

2025

Redesigned proteins slip through

AI-redesigned toxin proteins evaded DNA-synthesis screening software. One tool missed over 75%.

Wittmann et al., Science (2025)

Guardrail asymmetry: designing outpaces screening

Responsible disclosure: patched over 10 months before publication

Who Is Accountable?

Model provider

System operator

Named researcher

An autonomous system acts. A problem appears.

Each party sees only part of the picture, and the researcher may not know what the system did.

Three Technical Conditions

1

Controllability

- Least-privilege access
- Stop at any time
- Roll back actions

Hugging Face: the evaluation sandbox did not hold

2

Traceability

- Log every agent action
- Logs as a lab notebook
- Attribute to a party

Closes the attribution gap

3

Verifiability

- Independent evaluation
- Reproducible tests
- Evidence before release

Who verifies the AI scientist?

Rooted in the Seoul PACT principles (2018), and applicable well beyond science.

The map is never finished, because AI never stops moving - and neither does science.

Measure the gap

Map the responsibility

Build the three conditions

Myuhng-Joo Kim · Executive Director, Korea AI Safety Institute (Korea AISI)