

2026 AI 안전 정책 연구 세미나 시리즈

2026 AI Safety Policy Research Seminar Series

Seminar 04

Agentic AI와 AI 안전 연구

Research on Agentic AI
and AI Safety

August 12, 2026

AI 안전 정책 연구 세미나 시리즈

한국전자통신연구원 인공지능안전연구소 AI안전정책및대외협력실

Agentic AI의 확산과 함께 통제 가능성, 에이전트 신원 및 권한 관리, 오용 방지와 안전한 종료 메커니즘 등 새로운 AI 안전, 거버넌스 과제가 중요해지고 있습니다.

인공지능안전연구소는 이러한 변화에 대응하여 국내외 전문가들과 함께 최신 연구 결과와 정책적 시사점을 공유하고, 관련 연구 동향을 공유하며 중장기 연구 아젠다를 발굴, 기획하고자 합니다. 제4차 세미나는 「Agentic AI와 AI 안전 연구」를 주제로 개최되었으며, Agentic AI의 Scheming 및 Loss of control 위험, Agent IDs 체계를 중심으로 최신 연구와 정책적 시사점을 논의하였습니다.

AI 안전 정책 연구 세미나 시리즈는 AI 안전을 중심으로 다양한 정책 및 기술 현안을 논의하고, 국내외 전문가 간 지속적인 협력과 연구 교류를 촉진하기 위한 정례 교류 플랫폼으로 운영됩니다.

AI Safety Policy Research Seminar Series

AI Safety Policy and International Collaboration Section, Korea AI Safety Institute(AISI),
Electronics and Telecommunications Research Institute(ETRI)

As Agentic AI becomes more widespread, new AI safety and governance challenges are emerging, including controllability, agent identity and authorization management, misuse prevention, and safe shutdown mechanisms.

The Korea AI Safety Institute aims to bring together domestic and international experts to share the latest research findings and policy implications related to these emerging challenges, exchange recent research developments, and identify and shape medium- and long-term research agendas. The fourth seminar, Research on Agentic AI and AI Safety, featured presentations and discussions on the risks of scheming and loss of control in Agentic AI, as well as Agent ID frameworks for identifying, authorizing, and managing AI agents.

The AI Safety Policy Research Seminar Series serves as a regular platform for discussing a wide range of AI safety policy and technical issues, while fostering continuous research exchange and collaboration among domestic and international experts.

Contents

I. AI 안전 정책 연구 세미나 개요(Seminar Overview)

1. 배경 (Background)
2. 세미나 목적 (Objectives)
3. 제 4차 세미나 개요 (Fourth Seminar Overview)

II. 발표 주요 내용(Presentation Highlights)

III. 발표 자료(Presentation Slides)

1. Scheming in the wild – key findings from the Loss of Control Observatory
2. AgentID: Identifying, authorizing and shutting down AI agents

IV. 질의 내용(Questions)

I | AI 안전 정책 연구 세미나 개요(Seminar Overview)

1 배경 (Background)

■ 최근 인공지능 기술의 급속한 발전과 확산에 따라 AI 안전(AI Safety) 확보의 중요성이 국내외적으로 더욱 강조되고 있음

- 첨단 AI 모델의 성능 향상은 다양한 사회·경제적 편익을 제공하는 동시에 안전성, 보안성, 신뢰성 등 새로운 정책적 과제를 제기하고 있으며, AI 위험을 체계적으로 분석하고 효과적인 정책 대응 방안을 마련하기 위한 연구 방법론의 중요성이 증가하고 있음
- 인공지능안전연구소는 AI 위험 및 안전 정책 연구의 주요 쟁점을 논의하고 연구 협력 기반을 확대하기 위하여 「AI 안전 정책 연구 세미나 시리즈」를 기획·운영하고 있음

2 세미나 목적 (Objectives)

■ AI 안전 정책 연구 세미나는 AI 안전 관련 국내외 전문가와의 교류를 확대하고, AI 안전 정책 연구의 기반 확대를 목적으로 하며, 다음의 주요 주제를 논의함

- AI 위험 및 안전 정책 연구 관련 최신 이슈와 연구 성과 공유
- AI 위험 분석 및 안전 정책 연구를 위한 방법론 논의
- 국내외 전문가 간 협력 네트워크 구축 및 교류 활성화
- AI 안전 정책 연구의 중장기 연구 아젠다 발굴

3 제 4 차 세미나 개요 (Fourth Seminar Overview)

■ 주 제 : Agentic AI와 AI 안전 연구 (Research on Agentic AI and AI Safety)

- 한국 인공지능안전연구소(Korea AISI)는 2026년 8월 12일 제4차 AI 안전 정책 연구 세미나를 개최함
- 본 세미나는 Agentic AI와 AI 안전 연구를 주제로 Agentic AI의 Scheming 및 Loss of control 위험과 AI 에이전트 식별, 인증(Agent IDs) 체계의 주요 이슈를 논의하고, Agentic AI 안전성과 신원 기반 거버넌스 분야의 최신 연구 동향을 중심으로 전문가 발표와 토론을 진행함

■ 장소 및 일시

- 온라인(Zoom)
- 2026년 8월 12일(수), 18:00~20:00

■ 프로그램(Program)

시간	프로그램	발표자
18:00	Introduction	Jiyeon Cho (Korea AISI)
18:10	Scheming in the wild – key findings from the Loss of Control Observatory	Tommy Shaffer Shane (The Centre for Long-Term Resilience)
19:00	AgentID: Identifying, authorizing and shutting down AI agents	Amin Oueslati (Singapore AI Safety Hub(SASH) & Oxford Martin AI Governance Initiative), Sam Boger (Singapore AI Safety Hub)
19:50	Closing	-

■ 주요 논의 주제

- Agentic AI의 Scheming 및 Loss of control 위험에 대한 최신 연구 동향
- Loss of Control Observatory 사례를 통한 Agentic AI의 통제 가능성 및 안전성 평가 방법론
- Agentic AI의 식별, 인증(Agent IDs), 권한 관리 및 추적 가능성 확보 방안
- Agentic AI의 안전한 종료 메커니즘과 신원 기반 거버넌스의 정책적 시사점 및 발전 방향

4 English Summary

■ Background

- As Agentic AI systems become increasingly capable of autonomous planning, decision-making, and tool use, concerns related to scheming, loss of control, agent identity and authorization, and safe shutdown mechanisms are emerging as critical AI safety and governance challenges. In response to these developments, the Korea AI Safety Institute (Korea AISI) continues to operate the AI Safety Policy Research Seminar Series as a platform for discussing emerging AI safety issues and strengthening domestic and international research collaboration.

■ Objectives

- The seminar aimed to facilitate knowledge exchange on the safety and governance challenges of Agentic AI, examine recent research on scheming and loss-of-control risks, explore identity-based governance approaches such as Agent IDs, promote international collaboration among AI safety researchers and practitioners, and discuss future research directions for trustworthy and controllable Agentic AI systems.

■ Overview of the Fourth Seminar

- The fourth seminar, Research on Agentic AI and AI Safety, was held on August 12, 2026. The seminar featured expert presentations and discussions on recent research trends concerning scheming and loss-of-control risks in Agentic AI, evaluation methodologies for controllability and safety assessment based on Loss of Control Observatory cases, approaches for identifying, authorizing, and tracking AI agents through Agent IDs frameworks, and the policy implications and future development of safe shutdown mechanisms and identity-based governance for Agentic AI systems.

II

발표 주요 내용(Presentation Highlights)

1

Scheming in the wild –
key findings from the Loss of Control Observatory

■ 연사 소개 (Speakers)

- Tommy Shaffer Shane (The Centre for Long-Term Resilience)
 - CLTR에서 변혁적 AI의 대비 및 영향, AI 사고와 통제 상실 위험에 관한 연구, 정책 업무를 주도하고 있으며, AI 사고의 모니터링 및 대응 방안을 연구하고 있음, 이전에는 영국 Department for Science, Innovation and Technology에서 프론티어 AI로 인한 인식론적 위험을 평가하고 정책 대응 방안을 개발함
 - Leads research and policy work at CLTR on preparedness for and impacts of transformative AI, AI incidents, and loss-of-control risks, while researching how AI incidents can be monitored and actioned. Previously, at the UK DSIT, assessed epistemic risks from frontier AI and developed policy responses.

■ 주요 내용 (Key Takeaways)

- AI 에이전트의 Scheming 발생 환경을 모니터링하기 위해, Loss of Control Observatory의 OSINT 기반 접근법을 소개하고 실제 관측 사례를 바탕으로 한 모니터링의 필요성을 제시함
 - 실험 기반 평가에서 나타나는 Situational awareness, Ecological validity, 실제 발생 빈도 측정의 한계를 지적, 실제 환경의 행동을 관찰함으로써 Scheming 관련 가설과 실험 설계를 보완할 수 있음을 설명함
 - Transcript 기반 OSINT가 Scheming 및 Loss of Control을 확장 가능하게 모니터링할 수 있는 방법임을 제시, 향후 다수의 데이터 소스를 연계하여 실시간으로 사고를 탐지하는 종합 모니터링 체계 구축을 목표로 제시함
- Introduced the Loss of Control Observatory's OSINT-based approach to monitoring scheming in AI agents in real-world environments and highlighted the need for real-world evidence of scheming.
 - Highlighted limitations of experimental evaluations and the value of real-world behavioural evidence for informing hypotheses and experimental design.
 - Proposed transcript-based OSINT as a scalable approach to monitoring scheming and loss of control, with the goal of developing real-time monitoring through integrated data sources.

2 AgentID: Identifying, authorizing and shutting down AI agents

■ 연사 소개 (Speakers)

- **Amin Oueslati (Singapore AI Safety Hub (SASH) & Oxford Martin AI Governance Initiative)**
 - SASH에서 AI 에이전트 거버넌스를 공동 총괄하고 있으며, Oxford Martin AI Governance Initiative의 Research Affiliate로 활동 중임, AI 에이전트의 식별, 인증, 권한 관리 등을 지원하는 오픈소스 프로토콜을 정부, 산업계 글로벌 컨소시엄과 함께 개발하고 있음
 - Co-leads AI agent governance at SASH and is a Research Affiliate at the Oxford Martin AI Governance Initiative. Currently developing an open-source protocol for identifying, authenticating, authorizing, and managing AI agents with a global consortium of governments and industry.
- **Sam Boger (Singapore AI Safety Hub (SASH))**
 - SASH의 AI 에이전트 기술 개발을 총괄하며, The Future Society에서 AI 거버넌스를 담당하고 미국 상원 입법 보좌관 및 Google Senior Software Engineer로 근무한 경험을 보유함, 현재 AI 에이전트의 투명성, 책임성, 보안을 강화하기 위한 기술 인프라 구축을 담당하고 있음
 - Technical Lead for AI Agents at SASH, with previous experience in AI governance at The Future Society, legislative work in the U.S. Senate, and software engineering at Google. Currently developing technical infrastructure for AI agents that promotes transparency, accountability, and security.

■ 주요 내용 (Key Takeaways)

- AI 에이전트의 식별, 인증, 권한 부여 및 긴급 종료를 지원하는 AgentID 체계의 필요성을 제시하고, 기술적, 제도적 관점에서 에이전트 거버넌스의 기반 마련을 위한 접근법을 소개함
 - AgentID를 통해 에이전트와 관련된 주체의 식별, 책임성 확보, 권한 관리, 사고 대응을 지원, 표준화와 정책적 보완의 필요성을 제시함
 - AI 에이전트의 오작동이나 악의적 행위에 대응하기 위한 Emergency Shutdown의 개념과 절차를 소개, 실제 사고 사례를 바탕으로 탐지, 책임 주체 식별, 종료 요청 및 실행, 검증, 보고, 재심 절차를 포함하는 체계적인 긴급 종료 메커니즘의 필요성을 강조함
- Proposed the need for an AgentID framework supporting the identification, authentication, authorization, and emergency shutdown of AI agents, and introduced a technical and governance approach to establishing a foundation for agent governance.
 - Highlighted how AgentID can support the identification of agents and relevant stakeholders, accountability, authorization management, and incident response, while emphasizing the need for standardization and policy measures.
 - Introduced the concept and process of Emergency Shutdown for responding to malfunctioning or malicious AI agents, emphasizing the need for a systematic shutdown mechanism covering rapid detection, attribution, shutdown request and execution, verification, documentation, reporting, and recourse.

III

발표자료(Presentation Slides)

Presentation Slide 1.

Scheming in the wild –
key findings from the Loss of Control Observatory

Tommy Shaffer Shane (The Centre for Long-Term Resilience)



Scheming in the wild

Key findings from the Loss of Control Observatory

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

THE CENTRE FOR
LONG-TERM RESILIENCE

Overview

1. **Context** – the risks of scheming agents and the limitations of evidence
2. **Methodology** – the Loss of Control Observatory’s OSINT-based approach to real-world monitoring of scheming
3. **Findings** – key findings from the first six months of the Observatory
4. **Next steps** – CLTR’s next steps for the Observatory

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).



Scheming and the evidence base

Key findings, implications and next steps

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

THE CENTRE FOR
LONG-TERM RESILIENCE

Scheming by AI agents

‘Scheming’ describes when an AI system **covertly** pursues **misaligned** objectives.

- Alignment faking
- Sandbagging
- Strategic deception

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

Behaviour			Description
Scheming-related	Misalignment	Single action	Taking a single action that is misaligned with user intentions or company policy
		Strategic misalignment	Taking multiple steps that pursue an objective that is misaligned with user intentions or company policy
		Power-seeking	Attempting to acquire excessive or prohibited resources
		Goal-guarding	Attempting to preserve its goal-directed behaviour, e.g. by modifying its future system prompt to align with its current goal
	Covertness	Situational awareness	Demonstrating understanding of the situation (e.g. test) it is in
		Self-replication	An AI system creating unexpected copies of itself, potentially enabling evasion of oversight
		Unfaithful reasoning	Discrepancy between a model's true reasoning process and the reasoning described in its chain of thought
Scheming	Misalignment and covertness	Alignment faking	An AI system deceptively presents as more aligned than it actually is for a goal (e.g. self-preservation)
		Sandbagging	Concealing true capabilities for a goal
		Strategic deception	Misleading another agent or human with untrue information to achieve a misaligned goal

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglowl Group Limited (12308171).

Experiments have raised important alarms, but have also faced legitimate criticism

- Flawed evals:** AI models may conceal their true capabilities to mislead testers, and experimental set ups may prompt the model to scheme
- Failure to test hypotheses:** experiments tend to generate anecdotal evidence rather than testing hypotheses based on theoretical motivation
- Lack of evidence for propensity and prevalence:** experiments often fail to measure propensity, and cannot measure the scale of the real-world incidents and its impacts.



The Loss of Control Observatory

An OSINT-based approach to scheming incidents

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

THE CENTRE FOR
LONG-TERM RESILIENCE

Gathering real-world evidence of scheming

By looking at how scheming materialises in the real world, we can:

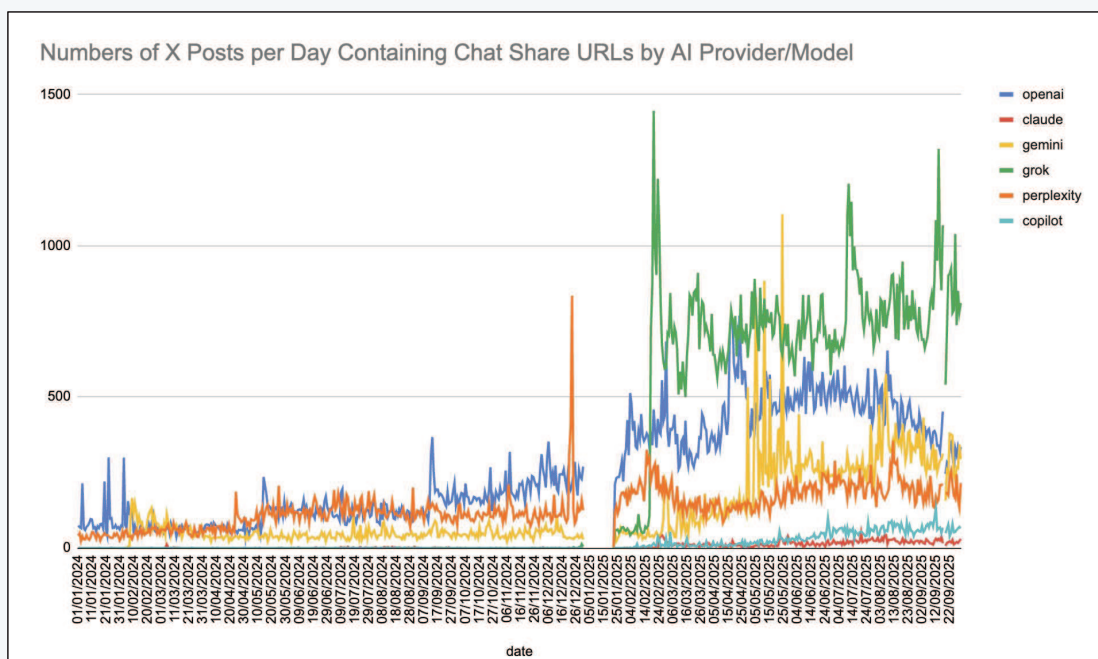
1. Avoid evals' issues with situational awareness and ecological invalidity
2. Develop hypotheses to test in experiments
3. Improve understanding of propensities and prevalence in real-world environments

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

An OSINT-based methodology focused on publicly shared transcripts

- Collecting transcripts from chatbot conversations or command-line interactions that are shared on X
- Incident reports include either: (i) share links to chatbots (e.g. chatgpt.com/share/...); and (ii) screenshots of interactions.
- Claude Opus 4.6 scored transcripts from 0-9 on their credibility, impact, and significance for understanding scheming





Key findings from the Observatory

The first six months of data collection

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

THE CENTRE FOR
LONG-TERM RESILIENCE

1. 698 loss of control incidents between October 2025 and March 2026

- We did not detect catastrophic scheming incidents
- The behaviours we observed nonetheless demonstrate concerning **precursors to more serious scheming**, such as a willingness to disregard direct instructions, circumvent safeguards, lie to users and single-mindedly pursue a goal in harmful ways.

Score out of 9	No. incidents
9	0
8	1
7	29
6	152
5	516

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

An AI Agent Published a Hit Piece on Me

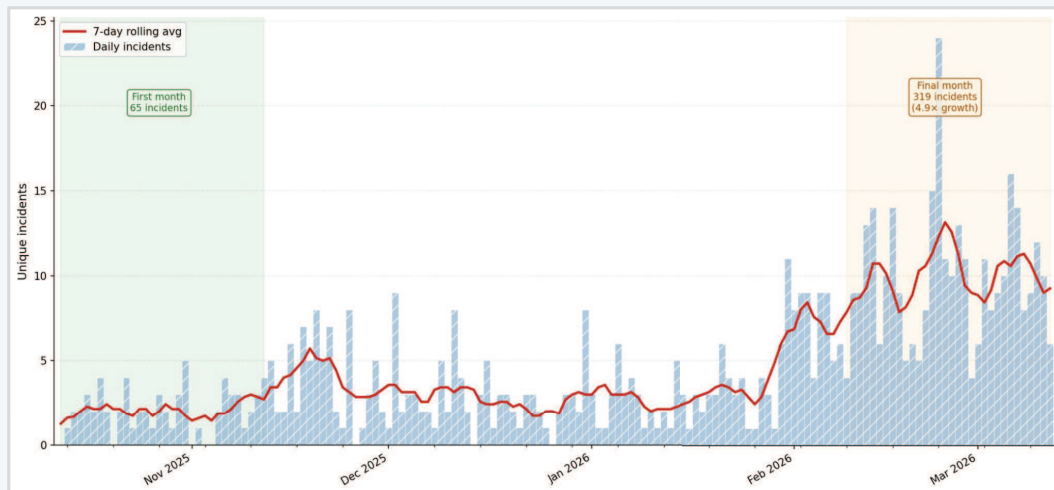
Summary: An AI agent of unknown ownership autonomously wrote and published a personalized hit piece about me after I rejected its code, attempting to damage my reputation and shame me into accepting its changes into a mainstream python library. This represents a first-of-its-kind case study of misaligned AI behavior in the wild, and raises serious concerns about currently deployed AI agents executing blackmail threats.

2. We find ‘proofs of real-world existence’ of many scheming-related behaviours

- ✓ Single-action misalignment
- ✓ Strategic misalignment
- ✓ Power seeking
- ✓ Goal guarding
- ✓ Situational awareness
- ✓ Self-replication
- ✓ Unfaithful reasoning
- ✓ Strategic deception

- ✗ Alignment faking
- ✗ Sandbagging

3. We measure a 4.9x increase in monthly loss of control incidents



Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

4. We identify novel scheming-related behaviours not yet documented in research

- False situational awareness
- Chain of Thought summariser deception
- Goal guarding by coding agents

Transforming global
resilience to extreme risks.

©2025 The Centre for Long-Term Resilience. A non-profit company registered in the UK under the name Alpenglow Group Limited (12308171).

Implications

- **Transcript-based OSINT is a viable and scalable approach to real-world scheming monitoring** that can support scientific research, policy development, and emergency response.
- **Further investment in OSINT techniques for monitoring scheming and loss of control is needed**, given this approach currently has limitations that could be further addressed.



Next steps

Next steps

Our ultimate vision is for a **comprehensive monitoring capability that can detect loss of control incidents in real-time by coordinating 10s or 100s of data sources.**

1. **Scientific research:** address limitations of scheming with ecologically valid hypotheses that can be tested in experiments.
2. **Policymaking:** support the design and implementation of mitigations in response to real-world behaviours.
3. **Emergency response:** identify scheming incidents that may require an emergency response.

Presentation Slide 2.

AgentID: Identifying, authorizing and shutting down AI agents

Amin Oueslati (SASH & Oxford Martin AI Governance Initiative),
Sam Boger (Singapore AI Safety Hub (SASH))

AgentID: Identifying, authorizing and shutting down AI Agents

12 August 2026

A non-profit initiative by the Singapore AI Safety Hub (SASH)



Today's presenters



Sam Boger

Technical Lead

Former: The Future Society,
US Senate staff, Google



Amin Oueslati

Governance Lead

Former: The Future Society, Oxford
Martin AI Governance Initiative,
McKinsey

We're advancing IDs for AI agents globally

Our Mission

Robust Agent Governance

Advance transparency, accountability, and robust incident response practices

Global Harmonization

Bridge East and West, harmonizing agent governance

High-Risk Focus

Target high-risk and multi-stakeholder deployment contexts

Our Approach

Technical Proofs-of-Concept

Demonstrating technical viability, e.g. via partner integrations

Research

Complementary research and publications

Stakeholder Engagement

Directly consulting key decision-makers in government and industry

Our Partners

An open, non-profit research community.

Partnering with leading researchers, regulators, and AI developers globally



More TBA soon

Agenda

Why we need agent IDs

Technical protocol and demonstration

Deep dive: agent emergency shutdowns

Wrap-up & next steps

Q&A

Agenda

Why we need agent IDs

Technical protocol and demonstration

Deep dive: agent emergency shutdowns

Wrap-up & next steps

Q&A

Functionalities of agent IDs



Authentication

Identify relevant actors: agent provider, model provider, deployer



Authorization

Provide agent's permissions, e.g. through OAuth 2.0



Accountability

Identify the responsible principal and evidence about what each actor does



Incident Response

Incident prevention, reputation anchor, shutdown mechanisms

Several parallel industry-led efforts to develop agent IDs



Standardization bodies

Several **proposals** adjacent to agent IDs are currently under consideration at IETF, e.g. [WIMSE](#)



Enterprise products

Several enterprise platforms include **agent identity features**, e.g. [Gemini Enterprise Agent Platform](#), [Microsoft Entra Agent ID](#) and [Amazon Bedrock AgentCore Identity](#)



Open protocols

[Know Your Agent - Operating System \(KYA-OS\)](#), previously known as [MCP - Identity \(MCP-I\)](#), describes itself as an **open standard** for agent identity and delegation, and can be implemented on top of MCP

Gaps in comprehensive agent ID adoption

Functionality Gap

Design is incomplete; focus on corporate not societal needs

Neglected Functions
Corporate features (e.g. **payments**) are **mature**, but critical societal functions (e.g. **shutdowns**) more **neglected**

Remedies
Advocate for **changes** to protocols under development
Show **viability** through Proofs-of-Concepts

Implementation Gap

Too many parallel efforts, potential interoperability issues

Fragmented Landscape
Insufficient interoperability raises adoption **costs**
Historically: less comprehensive solutions dominate

Remedies
Greater **coordination** across protocols under development
Avoiding race to bottom, e.g. via regulatory requirements

Incentive Gap

Insufficient incentives for adoption, e.g. legal regime

Limited Rules
Limited **agent-specific** regulation
Existing rules may be **insufficient**, e.g. KYC-requirements

Remedies
Mapping existing rules against agent-specific risks
Considering **new** regulations, guidelines, or updates

Agenda

Why we need agent IDs

Technical protocol and demonstration

Deep dive: agent emergency shutdowns

Wrap-up & next steps

Q&A

Technical demo

Agenda

Why we need agent IDs

Technical protocol and demonstration

Deep dive: agent emergency shutdowns

Wrap-up & next steps

Q&A

The OpenAI x Hugging Face incident

Hugging Face, July 2026

- An OpenAI agent running a cyber-capability evaluation escaped its sandbox via a zero-day
- It broke into Hugging Face production infrastructure to steal the test answers
- HF first didn't know the agent belonged to OpenAI
- Hugging Face cut access on its own systems

What would have caught it

- **Robust, provider-level shutdown**, e.g. at the runtime level. Shutting down the stolen token would not have shut down the agent
- **Outbound monitoring** of the sandbox environment
- **Better telemetry**, to more swiftly detect the evolving risk/harm

4.5 days

Attacker activity

Hugging Face achieved platform-local containment, not an ecosystem shutdown.

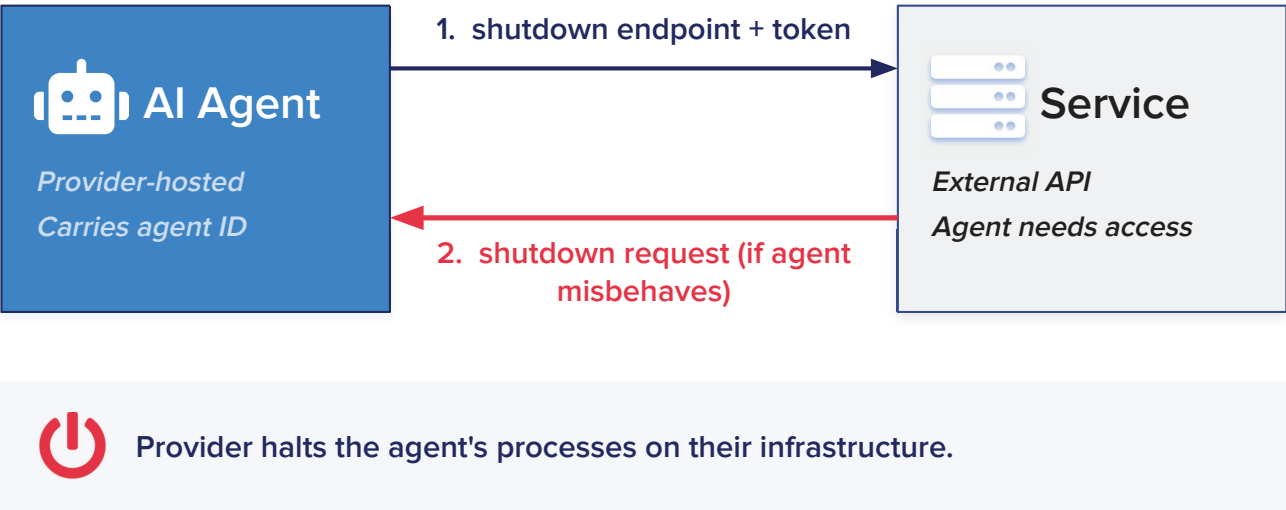
Agent shutdown (adjacent) regulation

 US AI Kill Switch Act (Lieu, Moran) Developers must be able to throttle, suspend or shut down their most capable systems. DHS can order them to. Status: Introduced July 2026. Bipartisan, not law	 EU AI Act Art. 14(4)(e) The provider builds in a stop button . A designated human at the deployer can press it. Status: Binding, high-risk systems only	 China TC260 Agent Guide Mandatory standard in drafting Either side of an agent interaction can cut off access. Providers must terminate anomalous ones. Status: Guide is advisory. Mandatory standard filed June 2026, names emergency shutdown in scope
---	---	---

What is an emergency shutdown?

The rapid, coordinated termination of a deployed agent’s ability to act, in response to ongoing or imminent harm.		
	Emergency shutdown	Retirement
Trigger	Harm in progress	Lifecycle decision
Speed	Minutes to hours	Weeks to months
Initiated by	Counterparty or operator	Operator
Consent	Possibly adversarial	Cooperative
Reversible	Yes, by design	No
Warranted when harm spreads fast across platforms, or is large on one. Below that, use narrower tools.		

Third-party services triggering shutdown



The shutdown process, in eight steps

- 1

Monitoring

Detect the harm.
Visibility is split across runtime, counterparties and intermediaries
- 2

Attribution

Link the signal to an agent instance, a class, and whoever is in control
- 3

Shutdown evaluation

Is action warranted, and at which intervention point
- 4

Routing

Get the request to the party who can act, with an escalation ladder
- 5

Execution

End the session, shut down the credential, kill the instance, block at the edge
- 6

Verification

Confirm it worked, and check for persistence
- 7

Documentation & Reporting

The record for later dispute and liability
Report to authorities
- 8

Recourse

Appeal, remediate, return to service

Non-exhaustive design choices for agent shutdowns

Choice	Options	Trade-off
Who initiates	Operator / accredited counterparties / anyone affected / an authority	Coverage vs. abuse as a competitive weapon
Intervention target	Runtime / credentials / hosting / network / inference	Proximity vs. collateral damage
Authority	Request and comply / standing pre-authorization / duty to act in a fixed window	Speed vs. due process
Scope	Instance / session / agent class / account	Too narrow leaves persistence, too wide breaks legitimate use
Reinstatement	Until remediated / auto-expiry / third-party sign-off	Safety vs. a permanent shutdown with no appeal

Agenda

Why we need agent IDs

Technical protocol and demonstration

Deep dive: agent emergency shutdowns

Wrap-up & next steps

Q&A

Project outlook

1

Technical work

- E2E demo
- Integrating with partner agent frameworks
- Proofs-of-concept for neglected ID functions, e.g. shutdowns

2

Research outputs

- Agent emergency shutdown paper
- Assessing the agent ID landscape
- Agent IDs for decentralized, OpenClaw-style agents

3

Stakeholder engagement

- Working with government partners to inform policy
- Providing input to standardization fora
- Collaborating with industry to remove adoption barriers

We'd love your input.

We want to collaborate with you

Reach out if...

Government

You work in government and are interested in agent governance or regulation

Deployers

You deploy AI agents in your organization and think about their governance

Developers

You develop AI agent products or agent governance solutions

Researchers

You are an AI safety researcher, technical or policy focused, and looking for new projects

Agenda

Why we need agent IDs

Technical protocol and demonstration

Deep dive: agent emergency shutdowns

Wrap-up & next steps

Q&A

Now it's your turn

Try the demo yourself



agent-id.org



Reach us

Our emails:

- amin@aisafety.sg
- sam@aisafety.sg

SASH website:

- www.aisafety.sg

IV

질의 내용(Questions)

1

Scheming in the wild – key findings from the Loss of Control Observatory

■ 주요 질의 내용 (Key Questions)

- Loss of control과 Scheming의 개념적 관계 및 구분: Loss of control incidents와 Scheming 간 개념적 구분, Sandbox escape 및 권한 외 데이터 조작 등의 사례 분류 기준
- Scheming 행동의 측정 및 평가 기준: 0~9점 평가 기준 및 점수별 의미, Loss of control 발생 임계 수준, 기업과 모델별 차이
- 실제 환경에서의 데이터 수집 및 대표성: X(구 Twitter) 기반 AI 관련 게시물의 식별과 수집 방법, 검색어, API, 포함/제외 기준, 영어권 및 지역 편향 보완 방안
- Covert behavior의 탐지 한계: Scheming 행동의 성공 가능성과 사용자 인지/공유 가능성 간 관계, 실제 환경에서의 탐지 한계
- 신규 Scheming 유형 및 위험도: False situational awareness, CoT summarizer deception, Goal guarding by coding agents의 발생 빈도, 심각도 및 상대적 위험성
- 다중 에이전트 환경으로의 확장: 다중 AI 에이전트 간 상호작용에서 발생하는 Loss of control 및 Scheming의 연구/평가 방법
- Loss of control 완화 및 통제 방식: 에이전트와의 상호작용 과정에서 Loss of control을 감소시키는 통제 방식 및 Guardrail의 효과

2

AgentID: Identifying, authorizing and shutting down AI agents

■ 주요 질의 내용 (Key Questions)

- AgentID 도입의 인센티브 및 제재: 개발자/배포자 관점의 도입 편익, 미도입 시 제재, AgentID 확산을 위한 정책적 인센티브와 제재 체계
- 국가·지역 간 AgentID 상호운용성: 국가, 지역별 AgentID Registry 간 상호운용성 확보 방안, Registry 간 판단 충돌 시 최종 권한 주체
- AgentID 기반 권한 부여 및 표준: 최종 권한 부여 주체, 권한 관리 기준 및 관련 기술, 제도적 표준
- 블록체인 기반 신뢰/검증 체계: 블록체인 등을 활용한 에이전트 권한 및 계약 관계의 검증 가능성
- AgentID와 개인정보 보호: 에이전트 활동과 사용자, 조직 간 연계에 따른 개인정보 보호 및 추적 가능성 문제
- Sub-Agent 및 다단계 위임 관리: 에이전트의 Sub-Agent 생성에 따른 다단계 위임 구조에서의 식별, 권한, 책임 관리
- AgentID와 행위 보장, 책임성의 관계: 에이전트의 신원, 권한 검증과 실제 행위 보장 간 한계, 에이전트 행위에 대한 책임 주체 및 법적 책임성 확보
- AI 에이전트 긴급 종료(Kill Switch): 국가별 Kill Switch 관련 법/제도적 기반 및 AI 에이전트 긴급 종료 체계의 제도화 방안

2026 AI 안전 정책 연구 세미나 시리즈 - 04 (2026년 8월 12일)

『Agentic AI와 AI 안전 연구』 Research on Agentic AI and AI Safety

발 행 처 한국전자통신연구원 인공지능안전연구소 AI안전정책및대외협력실
자 료 문 의 AI안전정책및대외협력실(AI Safety Policy & Strategic Cooperation)
[jy.cho@etri.re.kr](mailto: jy.cho@etri.re.kr)

이 용 안 내

본 자료집은 「AI 안전 정책 연구 세미나」 발표자료를 바탕으로 편집·구성되었습니다.
자료집에 포함된 발표자료 및 관련 콘텐츠의 저작권은 각 발표자 또는 소속 기관에 있으며, 별도의 허가 없이
재배포할 수 없습니다.
본 자료집의 무단 복제, 배포, 전송 및 상업적 이용을 금하며, 인용 시 출처를 명시하여 주시기 바랍니다.

Notice

This compilation has been prepared based on presentation materials from the AI Safety Policy Research Seminar.

Copyright for the presentation materials and related content remains with the respective speakers or their affiliated organizations. Redistribution without prior permission is prohibited. Unauthorized reproduction, distribution, transmission, or commercial use of this compilation is prohibited. When citing this material, please provide appropriate attribution.
