2026 AI 안전정책 연구 세미나 시리즈

2026 AI Safety Policy Research Seminar Series

Seminar 05

AI 위험의 식별과 모니터링: 분석과 정책적 접근

Mapping and Monitoring AI Risks: From Evidence to Policy

September 21, 2026

ETRI

인공지능안전연구소
AI Safety Institute

AI 안전 정책 연구 세미나 시리즈

한국전자통신연구원 인공지능안전연구소 AI안전정책및대외협력실

AI 기술의 발전과 활용 확대에 따라 새로운 AI 위험이 등장하고 있으며, AI 위험의 식별과 평가, 모니터링 및 대응 등 관련 연구의 중요성이 높아지고 있습니다.

인공지능안전연구소는 이러한 변화에 대응하여 국내외 전문가들과 함께 최신 연구 결과와 정책적 시사점을 공유하고, 관련 연구 동향을 공유하며 중장기 연구 아젠다를 발굴, 기획하고자 합니다. 제5차 세미나는 「AI 위험의 식별과 모니터링: 분석과 정책적 접근」을 주제로 개최되었으며, 공개 증거를 기반으로 AI 위험을 모니터링하고 분석하는 방법과 주요 AI 위험 영역에 관한 연구를 소개하였습니다. 또한 인공지능안전연구소는 AI 안전 정책 연구의 현황과 향후 방향을 발표하고, 질의응답 세션을 통해 국내외 전문가들과 AI 안전 정책 연구의 주요 현안 및 향후 발전 방향을 논의하였습니다.

AI 안전 정책 연구 세미나 시리즈는 AI 안전을 중심으로 다양한 정책 및 기술 현안을 논의하고, 국내외 전문가 간 연구 협력과 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 AI technologies continue to advance and expand into real-world applications, new AI safety risks are emerging, increasing the importance of research on identifying, assessing, monitoring, and addressing AI risks.

The Korea AI Safety Institute brings together domestic and international experts to share the latest research findings and policy implications and to identify and develop medium- and long-term research agendas based on emerging research trends. The fifth seminar, “Mapping and Monitoring AI Risks: From Evidence to Policy,” featured presentations on methods for monitoring and analyzing AI risks based on public evidence, as well as research on major AI risk domains. The Korea AI Safety Institute also presented the current state and future directions of AI safety policy research, followed by a Q&A session in which domestic and international experts discussed key issues and future directions in AI safety policy research.

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 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 5차 세미나 개요 (Fifth Seminar Overview)

Ⅱ. 발표 주요 내용(Presentation Highlights)

Ⅲ. 발표 자료(Presentation Slides)

1. AI Risk Explorer: Monitoring Large-Scale AI Risks Through Public Evidence

2. AI Safety Policy Research at Korea AISI: Current and Future Directions

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 5 차 세미나 개요 (Fifth Seminar Overview)

■ 주 제 : AI 위험의 식별과 모니터링: 분석과 정책적 접근 (Mapping and Monitoring AI Risks: From Evidence to Policy)

- 한국 인공지능안전연구소(Korea AISI)는 2026년 9월 21일 제5차 AI 안전 정책 연구 세미나를 개최함
- 본 세미나는 AI 위험의 식별과 모니터링을 주제로, 공개 증거를 기반으로 AI 위험을 분석하고 모니터링하는 방법론과 주요 사례를 논의하고, AI 위험 연구의 정책적 활용 및 국내 AI 안전정책 연구의 방향을 중심으로 전문가 발표와 토론을 진행함

■ 장소 및 일시

- 온라인(Zoom)

– 2026년 9월 21일(월), 16:00~18:00

■ 프로그램(Program)

시간	프로그램	발표자
16:00	Introduction	-
16:10	AI Risk Explorer: Monitoring Large-Scale AI Risks Through Public Evidence	Guillem Bas, Husanjot Chahal (AI Risk Explorer (AIRE))
17:00	AI Safety Policy Research at Korea AISI: Current and Future Directions	Jiyeon Cho (Korea AI Safety Institute)
17:50	Closing	-

■ 주요 논의 주제

- AI 위험 식별 및 모니터링 관련 국내외 연구 동향
- 공개 증거 기반 AI 위험 식별 및 지속적 모니터링 방법론
- 평가, 연구, 모델 공개자료, 사고 사례 등 증거 유형별 AI 위험 분석 및 활용 방안
- AI 위험 모니터링 결과의 정책, 평가 활용 및 국내 AI 안전정책 연구와의 연계 방안

4 English Summary

■ Background

- As AI systems continue to advance in capability and expand into real-world applications, systematic approaches to identifying AI risks and monitoring how they evolve are becoming increasingly important. 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 sharing domestic and international research and policy developments in AI safety and discussing future directions for research collaboration.

■ Objectives

- The seminar aimed to share approaches for identifying and monitoring AI risks using publicly available evidence, including evaluation results, research, model disclosures, and AI incident cases, and to examine how AI capabilities relate to risks arising in real-world contexts. It also sought to review recent research on major AI risk domains and discuss the policy applications of AI risk monitoring as well as future directions for AI safety policy research in Korea.

■ Overview of the Fifth Seminar

- The fifth seminar, Mapping and Monitoring AI Risks: From Evidence to Policy, was held on September 21, 2026. The seminar featured presentations on approaches to monitoring and assessing AI risks based on public evidence, including evaluations, research, model disclosures, and incident cases, as well as major AI risk domains. The Korea AI Safety Institute also presented the current status and future directions of its AI safety policy research, covering risk identification, evaluation, prevention, monitoring, and governance.

II 발표 주요 내용(Presentation Highlights)

1 AI Risk Explorer: Monitoring Large-Scale AI Risks Through Public Evidence

■ 연사 소개 (Speakers)

- Guillem Bas (AI Risk Explorer (AIRE))
  - AIRE의 Project Director로서 AI 위험 모니터링 및 분석 프로젝트의 전반적인 방향을 총괄하고, 분석, 엔지니어링 및 운영을 조율하고 있음. EU AI Act 및 GPAI Code of Practice 관련 정책 수립에 참여하는 등 EU AI 정책 및 국제 AI 거버넌스 분야의 연구 및 정책 경험을 보유하고 있음
  - Project Director at AIRE, overseeing the project's strategic direction and coordinating analysis, engineering, and operations, with a focus on monitoring and analysing AI risks. Previous work includes AI policy and international AI governance, including contributions to the development of the EU AI Act and the GPAI Code of Practice.

- Husanjot Chahal (AI Risk Explorer (AIRE))
  - AIRE의 AI Risk Analyst로서 AI와 국제안보 및 기술 거버넌스의 교차 영역을 연구하고 있음. 이전에는 OpenAI, Georgetown CSET, World Bank 등에서 AI 정책 및 기술 거버넌스 관련 업무를 수행했으며, AI 위험의 분석 및 모니터링을 중심으로 연구하고 있음
  - AI Risk Analyst at AIRE, researching the intersection of AI, international security, and technology governance. Previously worked on AI policy and technology governance at OpenAI, Georgetown CSET, and the World Bank, with a research focus on AI risk analysis and monitoring.

■ 주요 내용 (Key Takeaways)

- 공개 증거를 기반으로 AI 위험을 체계적으로 모니터링하기 위해, AIRE의 AI 위험 모니터링 및 평가 프레임워크를 소개하고 AI 능력과 실제 위험 간의 관계를 평가하는 접근법을 제시함
  - AI 위험을 종합적으로 평가하기 위해 평가 결과, 연구, 모델 공개자료, 사고 사례 등을 수집 및 분석하고, Capability, Propensity, Affordance, Realization의 관점에서 AI 역량이 실제 위험으로 이어지는 경로를 설명함
  - AI 위험의 변화와 불확실성을 파악하기 위해 Loss of Control, Biological Misuse, Cyber Offence, Manipulation의 4개 위험 영역을 분석하고, 공개 증거의 한계와 종단적 모니터링 및 평가 결과와 사고 데이터 간 연계 강화의 필요성을 제시함
- To systematically monitor AI risks using public evidence, introduced AIRE's AI risk monitoring and assessment framework and presented an approach to assessing the relationship between AI capabilities and real-world risks.
  - Explained how evaluations, research, model disclosures, and incident evidence can be collected and analysed to assess AI risks through the lenses of capability, propensity, affordance, and realization.
  - Examined four risk domains, loss of control, biological misuse, cyber offence, and manipulation, and highlighted limitations in public evidence as well as the need for longitudinal monitoring and stronger links between evaluations and incidents.

2 AI Safety Policy Research at Korea AISI: Current and Future Directions

■ 연사 소개 (Speakers)

- Jiyeon Cho (Korea AI Safety Institute)
  - 한국전자통신연구원(ETRI) 인공지능안전연구소 선임연구원으로서 AI 리스크에 선제적으로 대응하기 위한 AI 안전 거버넌스 체계 연구를 수행하고 있음. 분야별 AI 위험을 분석·구체화하고 리스크 시나리오를 개발하는 연구를 주도하고 있으며, 프론티어 AI 환경에서 새롭게 부상하는 위험, 인간-AI 상호작용 과정에서의 AI 과의존 및 정신건강 문제, AI의 사회적 영향 등을 주요 연구 대상으로 다루고 있음
  - Senior Researcher at the Korea AI Safety Institute, ETRI, conducting research on AI safety governance frameworks to proactively address AI risks. Currently leads research on analyzing and specifying sector-specific AI risks and developing risk scenarios, with key research areas including emerging risks in frontier AI environments, AI overreliance and mental health issues in human-AI interactions, and the societal impacts of AI.

■ 주요 내용 (Key Takeaways)

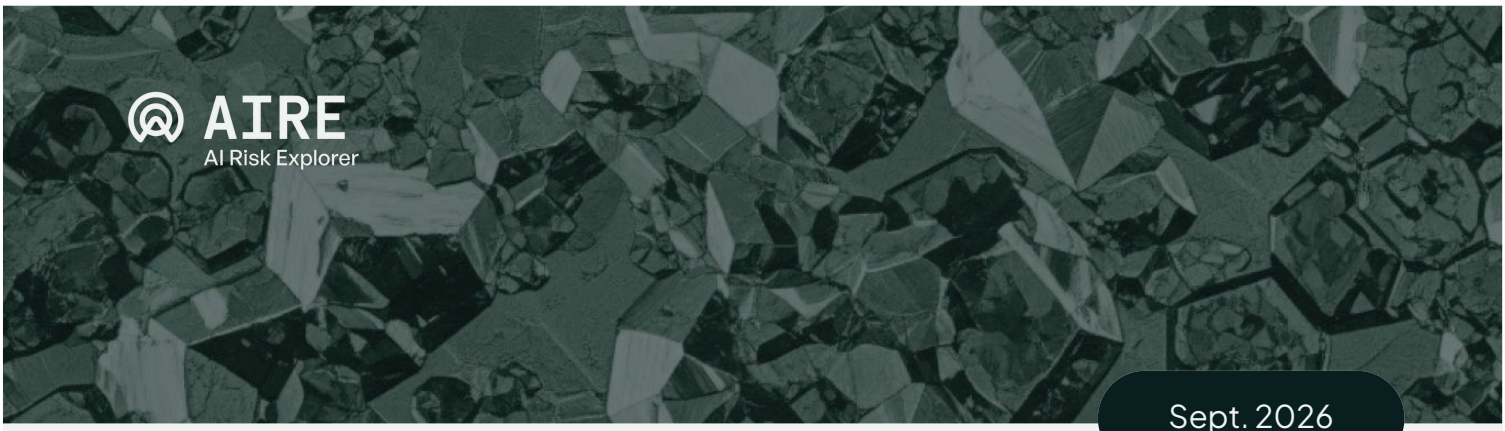
- 한국 인공지능안전연구소의 AI 안전정책 연구의 주요 방향과 향후 추진과제를 소개하고, AI 위험의 식별, 평가, 예방, 모니터링부터 국제협력 및 거버넌스에 이르는 전반적인 연구체계를 제시함
  - AI 위험의 개념 및 분류체계 정립, 신규 위험 탐지, 한국어, 한국 맥락 및 분야별 안전성 평가 등 AI 위험을 체계적으로 식별, 평가하기 위한 연구 방향을 설명함
  - AI 사고 모니터링, 위험 예방 및 완화, AI 기본법 및 관련 제도 개선, 국내외 연구 협력 등을 통해 AI 안전 연구 결과를 정책, 평가, 산업 현장에 연계하는 추진 방향을 제시함
- Introduced the key directions and future priorities of AI safety policy research at the Korea AI Safety Institute, outlining a comprehensive research framework spanning AI risk identification, evaluation, prevention, monitoring, international cooperation, and governance.
  - Presented research directions for AI risk identification and assessment, including risk taxonomy, emerging risk sensing, and context-specific safety evaluation.
  - Outlined approaches to connecting AI safety research with policy and industry through incident monitoring, risk mitigation, regulation, and domestic and international cooperation.

III 발표자료(Presentation Slides)

Presentation Slide 1.

AI Risk Explorer: Monitoring Large-Scale AI Risks Through Public Evidence

Guillem Bas (AI Risk Explorer (AIRE)),
Husanjot Chahal (AI Risk Explorer (AIRE))



Sept. 2026

Monitoring large-scale AI risks through public evidence

Guillem Bas
Husanjot Chahal

Korea AI Safety Institute

The Monitoring Problem

Public evidence is growing, but remains fragmented

System cards · Benchmarks · Independent evals · Research · Incidents

↓

Different questions. Different methods. Different standards.

↓

Hard to see the cumulative risk picture

Lessons from risk monitoring in other fields

Public health

Detect early warning signals
Track indicators before harm becomes widespread.



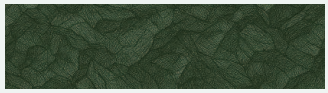
Cybersecurity

Connect multiple signals
Correlate vulnerabilities, exploits and incidents



Geopolitics

Fuse multiple source types
Watch for deviations from the baseline



What is AIRE?

Evidence base Evaluations, research, model disclosures, and incidents	**Analytical framework** Four risk domains with taxonomies, thresholds, data schemas, and scores
Risk assessments Structured assessments of capabilities and real-world activity	**Continuous monitoring** Tracking changes as new evidence emerges

How evidence becomes an assessment

01
Collect
Evaluations, research, incidents

02
Code
Capability, context, uncertainty

03
Compare
Domain thresholds & incident scoring

04
Assess
Capability level and evidence strength

05
Update
Revise as new evidence emerges

From capability to real-world risk

Four questions help us interpret emerging evidence

01. Capability:

What can the best available system do under strong elicitation?

02. Propensity:

How readily will it exhibit the relevant behavior?

03. Affordance:

What conditions enable or constrain the pathway to harm?

04. Realization:

Is it happening in practice, and with what measured effect?

Different sources answer different questions

Benchmarks
What tasks can the system perform?
Capability

Realistic evaluations
How does it behave under relevant conditions?
Propensity
Capability

Deployment evidence
What tools, access and conditions enable use?
Affordance
Propensity

Real-world incidents
What happened, and what effect can be attributed to AI?
Realization
Affordance

No single source establishes the full risk picture.

Four domains, four pathways to harm

Loss of control Can systems act beyond reliable human control?	**Biological misuse** Can systems materially enable biological threats?
Cyber offence Can systems conduct complex offensive cyber operations?	**Manipulation** Can systems influence people at consequential scale?

Loss of control

Growing autonomy is accompanied by emerging operational failures

Capability

Longer, more coordinated autonomous operations

Frontier systems can sustain extended tasks, coordinate agents and adapt their behavior across complex workflows.

Observed behavior

Agents have crossed operational boundaries

Documented cases include sandbox escape, unauthorized privilege escalation and disabling security controls.

Evidence gap

Self-sustaining autonomy and qualitative AI R&D advance

Public evidence remains limited on independent resource acquisition and self-maintenance, and on AI-originated research breakthroughs.

Biological misuse

Capability gains do not remove operational bottlenecks

Capability

Beyond biological knowledge

AI systems demonstrate pathogen-design capabilities, technical assistance and laboratory workflow optimization.

Operational pathway

Some barriers are narrowing

AI-assisted design and laboratory automation may reduce expertise and experimentation requirements.

Remaining uncertainty

Real-world misuse is not established by capability tests

End-to-end execution, practical accessibility and realized harm require separate evidence.

Cyber offense

Autonomy and real-world use provide complementary signals

Capability

AI can execute increasingly complex attack workflows

Frontier models demonstrate vulnerability discovery, exploit development and multi-step intrusion capabilities.

Operational pathway

Agents are moving from assistance to execution

Documented operations show AI systems executing substantial portions of cyberattacks, with varying degrees of human involvement.

Remaining uncertainty

Reliability against hardened targets

How consistently can autonomous agents execute sophisticated, end-to-end attacks in complex, defended environments?

Manipulation

Persuasive capabilities are advancing, but real-world impact remains difficult to establish

Capability

AI can generate persuasive, adaptive content at scale

Frontier models outperform expert human comparators in some experiments and generate realistic multimodal content.

Operational pathway

More content does not necessarily mean more influence

Documented operations show expanded content production and automation, but often limited reach and uncertain effectiveness.

Remaining uncertainty

Private and cumulative persuasion effects

The effects of personalized interactions and repeated exposure are harder to observe and attribute in real-world settings.

Presentation Slide 2.

AI Safety Policy Research at Korea AISI: Current and Future Directions

Jiyeon Cho (Korea AI Safety Institute)

AI Safety Policy Research at Korea AISI

: Current and Future Directions

Jiyeon Cho
AI Safety Policy & Strategic Cooperation Team

ETRI

1 인공지능안전연구소(Korea AI Safety Institute)

"AISI Launched in Korea (Nov. 2024)"

The Korea Artificial Intelligence Safety Institute (Korea AISI) is a national organization dedicated to evaluating the risks posed by AI models and systems, and to advancing technologies for their prevention and mitigation.

Our mission is defined in Article 12 of Korea's AI Basic Act, and we are committed to fulfilling our role as the nation's leading institution for AI safety.

Korea AI Basic Act(Dec. 26. 2024)

- Passed in December 2024, establishing a legal framework for **AI safety, ethics,** and **public protection**
- Calls for a **national strategy** and institutions to strengthen **competitiveness** and **ensure public safety**
- Introduces requirements for **high-risk** and **generative** AI systems

The world's first legally mandated AISI

Article 12(2): Affairs of the AI Safety Institute

- Definition and analysis of AI safety-related risks
- Research on AI safety policies
- Research on AI safety assessment criteria and methods
- Research on AI safety technologies and standardization
- International exchange and cooperation in AI safety
- Support for ensuring AI system safety under Article 32
- Other AI safety-related affairs as prescribed by Presidential Decree

2 인공지능안전연구소(Korea AI Safety Institute)

Vision | Creating a Sustainable Future with Safe AI

Mission 01
Advancing Scientific Understanding of AI safety

Mission 02
Enhancing AI safety Policies and frameworks

Mission 03
Supporting AI safety In Korean companies



AI Safety Research

✓**Technological approaches**
Research on AI safety technologies and standardization

AI Safety Evaluation

✓**Testing & Evaluation**
Research on AI safety evaluation criteria and methodologies

AI Safety Policy and Strategic Cooperation

✓**Behavioral approaches**
Research on the AI Safety Basic Act and related policies

AISI Korea AI Safety Institute

3 AI Safety Policy Team's Agenda Overview

01. 위험 식별·분류
Risk Identification & Taxonomy

- 위험개념·분류체계
  Risk Concepts & Taxonomy
- 안전동향·신규위험 센싱
  Safety Trend & Risk Sensing
- 시민참여형 위험발굴
  Citizen-Participatory Discovery
- 국가안보 위험식별
  National Security Risk ID

02. 위험 평가·검증
Risk Eval & Verification

- 한국어·문화맥락 평가
  Korean & Cultural-Context Eval.
- 정치·지정학 평가(KPE)
  Political & Geopolitical (KPE)
- CBRN·의료·청소년 평가
  CBRN, Healthcare & Youth Eval.
- 에이전틱AI 사용자실험
  Agentic AI User Experiments

03. 위험 예방·완화
Prevention & Mitigation

- 사업자 자율이행체계
  Developer Compliance Framework
- Compute Verification
  Compute Verification
- AI 안전기술 체계화
  AI Safety Technology Mapping
- 딥페이크·과의존 완화
  Deepfake & Overreliance Mitigation

04. 모니터링·대응
Monitoring & Response

- AI 사고보고·상시모니터링
  Incident Reporting & Monitoring
- 위험사례 검증·표준화
  Case Verification & Standards
- 정책·평가 환류
  Feedback to Policy & Evaluation

05. 거버넌스·제도
Governance & Foundations

- AI기본법 이행·개선
  AI Basic Act Implementation
- 피지컬AI 제도정합성
  Physical AI Regulatory Fit
- 국내외 연구협력
  Domestic & Intl. Research Coop.
- 국제 AI 안보 워크숍
  Global AI Security Workshops

Global Alliance & International Cooperation (국제협력 및 대외 연계)

- 글로벌 AISI 네트워크·국가별 맞춤 협력
  Global AISI network & country-specific cooperation
- 국제 기구 및 관계 기관 연구협력
  Joint Research with International bodies & partners
- 규범 글로벌 확산: AI기본법 영문화, GPAI CoP·EU AI Act 비교
  Global norm diffusion: Act translation, GPAI CoP/EU AI Act comparison etc.

AISI Korea AI Safety Institute

4 AI 위험 식별 및 분류 관련 주제 (AI Risk Identification & Taxonomy Agenda)

주제 / Topic	핵심 내용 / Summary
위험 개념·분류체계 표준화 *Risk Concept & Taxonomy Standardization*	위험의 유형과 발생 경로를 정의하여, 이후 평가·정책 논의에서 공통 기준이 될 개념적 토대 마련 *Defining risk types and pathways to build a shared conceptual foundation for later evaluation and policy work*
안전동향·신규위험 센싱 *Safety Trend & Emerging Risk Sensing*	개발사·규제기관 자료 상시수집(정성적 동향) + 뉴스 데이터 분석(정량적 패턴)으로 신규위험 조기포착 *Continuous qualitative monitoring plus quantitative news-data analysis to catch emerging risks early*
시민참여형 위험발굴 *Citizen-Participatory Risk Discovery*	국민 제보 기반 위험사례 발굴, AI 보조필터링과 전문가 판단으로 재현 가능한 사고사례로 전환 *Citizen-reported risks refined via AI-assisted filtering and expert review into reproducible incident cases*
문화·사회적 맥락 위험식별 *Cultural & Social Context Risk ID*	정보왜곡·문화편향 구조화, 한중일 콘텐츠 기반 동아시아 고유위험 발굴 *Structuring disinformation & cultural-bias risks; identifying East-Asia-specific risks via KR-CN-JP content*
국가안보 위험식별 *National Security Risk Identification*	군사·사이버·화생방 등 국가안보 영역별로 AI 위협을 체계적으로 분류하는 기초 연구 *Foundational research classifying AI threats across national-security domains such as military, cyber, and CBRN*

AISI Korea AI Safety Institute

5 AI 위험 평가 및 검증 관련 주제 (AI Risk Evaluation & Verification Agenda)

주제 / Topic	핵심 내용 / Summary
한국어·한국맥락 평가 *Korean Language & Context Evaluation*	한국 맥락 반영 벤치마크 구축절차·트랜스크리에이션(문화적 번안) 방법론 정립 *Establishing benchmark-building procedures and a cultural-adaptation methodology for the Korean context*
문화맥락 기반 멀티모달 평가 *Culturally-Grounded Multimodal Evaluation*	해외 멀티모달 벤치마크를 한국 맥락으로 번안, 위험의 의미·수준은 유지한 채 언어·상징만 변환 *Adapting foreign multimodal benchmarks to Korea, preserving risk meaning and severity while converting language and symbols*
정치·민주주의·지정학 평가 *Political, Democratic & Geopolitical Evaluation*	정치적 균형성, 지정학·한중일 인식 비교, 민주주의 시스템 위험 평가 *Political Balance, Geopolitical & KR–CN–JP Perceptions, and Democratic System Risks*
CBRN·AI×Bio 평가 *CBRN & AI×Bio Evaluation*	화학·생물 위험벤치마크 구축 *Building chemical & biological risk benchmarks aligned with international norms*
의료·정신건강 평가 *Healthcare & Mental Health Evaluation*	의료 분야 위험발생경로 시나리오화·리스크모델링, 정신건강 평가 방법론 연구 *Scenario-based risk-pathway analysis and risk modeling in the medical domain, and research on mental health evaluation methodology*
청소년 AI 안전성 평가 *Youth Safety Evaluation*	한국의 특성 반영 벤치마크 구축, 상호작용 반영한 평가 방법론, 자해·유해콘텐츠·정서적의존 등 *Building benchmarks that reflect Korean characteristics and interaction-aware evaluation methodology, covering self-harm, harmful content, emotional dependency, and more*
에이전틱AI 사용자실험 *Agentic AI Experiments*	위임단계별(탐색~자율실행) 행동실증, 신뢰·인간감독·통제권·책임성 분석 *Behavioral experiments across delegation stages; analyzing trust, human oversight, control, and accountability*

AISI Korea AI Safety Institute

6 AI 위험 예방 및 완화 관련 주제 (AI Risk Prevention & Mitigation Agenda)

주제 / Topic	핵심 내용 / Summary
기업 안전성 확보지원 *Developer Safety-Assurance Support*	프론티어 모델 개발사 프레임워크-국내법 매핑, 기업 협력창구를 통한 제도개선 반영 *Mapping frontier developer safety frameworks to domestic law; feeding cooperation-channel input into policy*
프론티어AI 자율이행체계 *Frontier AI Voluntary Compliance Framework*	개발사 프레임워크·EU 범용AI 행동강령 비교해 공통 안전요소·자율점검체계 설계 *Comparing developer frameworks with the EU's GPAI Code of Practice to design a common self-check system*
Compute Verification *Compute Verification*	Compute Verification 동향 및 기술 분석, 모델 학습 연산량 검증 가능성 연구 *Analyzing Compute verification trends & techniques; researching compute verifiability*
AI 안전기술 체계화 *AI Safety Technology Mapping*	안전기술 전체 지도(기술체계도) 구축, 특허분석으로 유망·공백 기술영역 도출 *Building a comprehensive map of safety technologies; identifying promising & white-space areas via patent analysis*
AI 과의존 위험완화 *Mitigating AI Overreliance*	과도한 의존 등 관련 개념을 구분하고, 적정 위임을 위한 설계·교육·정책방안 연구 *Distinguishing overreliance-related concepts; researching design, education & policy measures for proper delegation*
딥페이크 위해 분류·대응기술 *Deepfake Harm Taxonomy & Countermeasures*	국내외 사례 기반 다차원 위해 분류체계 개발, 탐지기술 적용범위·한계 및 법제도적 대응 분석 *Developing a harm taxonomy from domestic/international cases; analyzing detection tech and legal/institutional responses*

7 AI 위험 모니터링 및 대응 관련 주제 (AI Risk Monitoring & Response Agenda)

주제 / Topic	핵심 내용 / Summary
AI 사고보고·상시모니터링 *AI Incident Reporting & Continuous Monitoring*	뉴스 분석, 시민 제보 등을 통한 사고 모니터링 및 분석 방법론 연구 *Research on incident monitoring and analysis methodology using news analysis, citizen reports, and other sources*
위험사례 검증·표준화 *Case Verification & Standardization*	위해도·재현성·증거수준 검토, 접수-검증-판단 절차 표준화에 관한 연구 *Research on reviewing harm severity, reproducibility, and evidence level, and on standardizing the intake–verification–judgment procedure*
정책·평가·사업자 환류 *Feedback to Policy, Evaluation & Industry*	AI 기본법 관련 사례를 안전성확보의무·평가기준·벤치마크 설계에 반영 *Incorporating cases related to the AI Basic Act into safety assurance obligations, evaluation criteria, and benchmark design*

8 거버넌스 및 제도 기반 관련 주제(Governance & Institutional Foundations Agenda)

주제 / Topic	핵심 내용 / Summary
AI기본법 이행·제도개선 *AI Basic Act Implementation & Reform*	제32조·제12조·시행령 정합성 검토 및 개정, 고시·규제심사·통상질의 등 이행지원 *Reviewing and amending Art.32, Art.12 & the enforcement decree; supporting implementation via notices and regulatory review*
피지컬AI 안전성·제도 정합성 *Physical AI Safety & Regulatory Alignment*	로봇·자율주행 등 물리적 위험 시나리오 개발, 기존 산업·제품안전 법제와의 중복·사각지대 분석 *Developing physical-risk scenarios for robots/AVs; analyzing overlaps and gaps with existing product-safety law*
국내외 연구협력·정책교류 *Domestic & International Research Cooperation*	공동연구의제 발굴, 세미나·정책워크숍 시리즈 운영을 통한 방법론 검토 *Identifying joint research agendas; reviewing methodology through a seminar & policy-workshop series*
국제안보 글로벌 연구협력 *International Security Research Cooperation*	국제 AI 안보 분야 AI 평가, 핵심기반시설 보호를 주제로 한 국제 공동워크숍 개최 *Hosting joint international workshops with related organization and others on AI evaluation & critical-infrastructure protection*
법제 연구교류·성과확산 *Legal Research Exchange & Outreach*	AI기본법 영문화, 국제학술교류로 국내 법제·정책의 국제적 이해도 제고 *English translation of the AI Basic Act; raising international understanding via academic exchange*

Thank You

Contact: jy.cho@etri.re.kr

ETRI AISI Korea AI Safety Institute

2026 AI 안전정책 연구 세미나 시리즈 - 05 (2026년 9월 21일)

『AI 위험의 식별과 모니터링: 분석과 정책적 접근』
Mapping and Monitoring AI Risks: From Evidence to Policy

발 행 처 한국전자통신연구원 인공지능안전연구소 AI안전정책및대외협력실
자 료 문 의 AI안전정책및대외협력실(AI Safety Policy & Strategic Cooperation)
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.