[Workshop Update] AI Security Workshop Co-Hosted by the AI Safety Institute, U.S. Department of State, RAND, and MITRE

The Artificial Intelligence Safety Institute (AISI) in Korea, in collaboration with the U.S. Department of State, RAND, and MITRE, announced that it successfully convened the "AI Security Workshop" from Wednesday, February 25 to Friday, February 27. Approximately 60 participants from 17 companies and 4 institutions attended the workshop.

The event was designed to share and discuss practical approaches to AI risk management, encompassing not only AI safety but also broader security and national security perspectives. As AI technologies continue to expand rapidly across both industry and the public sector, ensuring safety and trustworthiness alongside innovation has emerged as a critical requirement for sustainable development. In this context, the workshop provided a timely and meaningful platform for dialogue and collaboration.

▲ Day 1: AI Threat Landscape and Presentation on the "AI Security Guide" and "MITRE ATLAS"

The first day (February 25) consisted of three sessions. The opening session provided a comprehensive overview of the evolving AI threat landscape. Beginning with a presentation on AI risks from a nonproliferation perspective, the discussion highlighted the dual-use nature of advanced AI systems. Participants agreed on the importance of embedding safety across the entire AI lifecycle—from development and deployment to operation—and of strengthening organizational security governance through proactive risk management approaches.

In the second session, RAND introduced its AI Security Guide, outlining methodologies for assessing and improving organizational AI security posture. A lifecycle-based threat analysis framework and risk prioritization tools were presented to help identify critical risks and derive practical mitigation measures. The final session featured MITRE's presentation of MITRE ATLAS and the AI Assurance Process. The assurance framework structures key activities such as system characterization, risk assessment, and continuous monitoring into an iterative lifecycle model. The discussion emphasized how structured assurance practices can enhance the safety, security, and reliability of AI systems.

▲ Day 2: Tailored 1:1 Consultations for Participating Companies

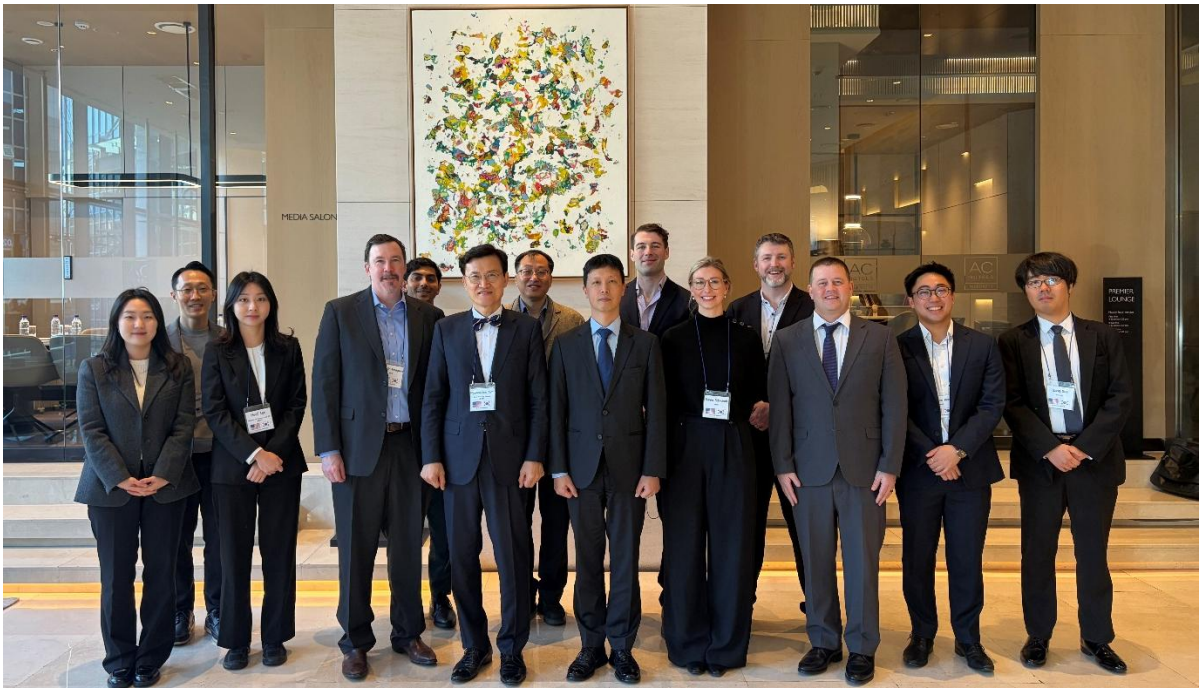
On the second day (February 26), RAND and MITRE experts conducted one-on-one tailored consultations with pre-registered companies. The sessions examined each company's AI system architecture, data flows, deployment environments, and existing control mechanisms. Together, participants identified priority actions and outlined directions for developing assurance and risk management plans.

These consultations provided companies with a practical foundation for building integrated governance frameworks that place safety at the core while incorporating security and accountability considerations.

▲ Day 3: Visit to a Leading Korean AI Startup

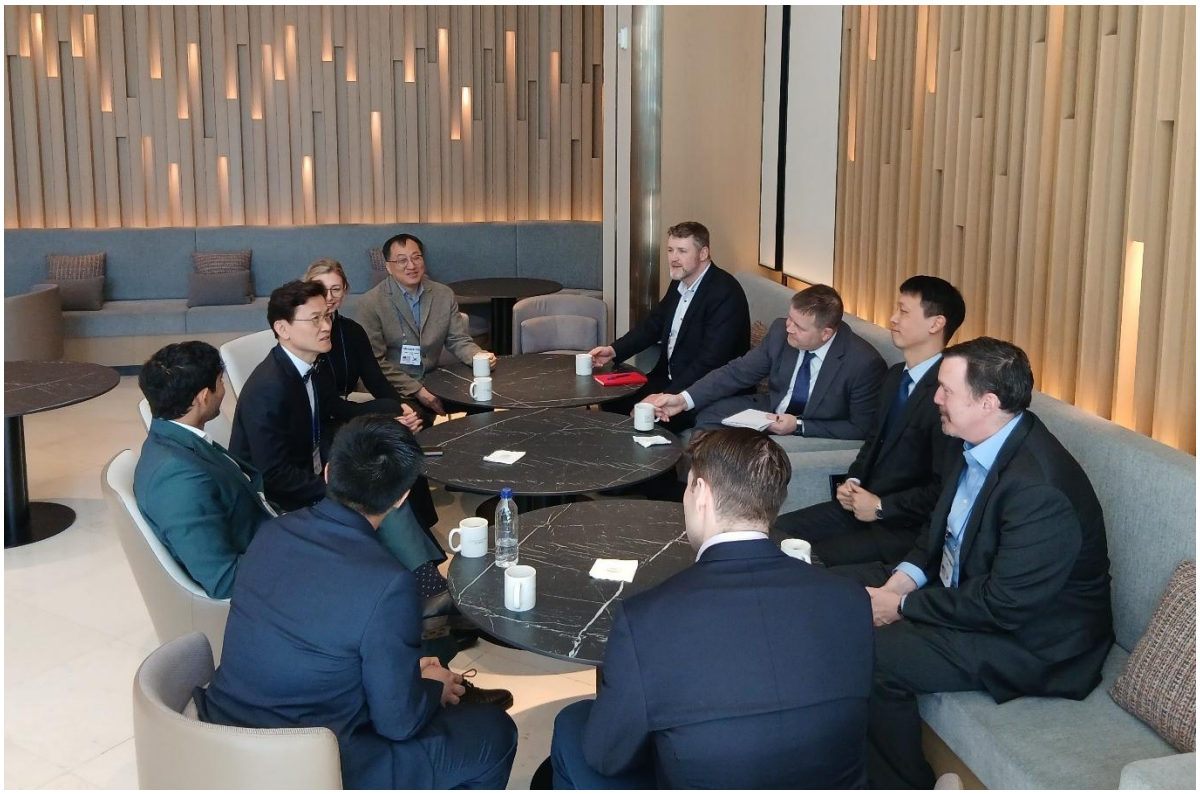
On the final day (February 27), participants visited AIM Intelligence, a leading Korean startup in AI safety and security. The visit included discussions on current technology development and the application of internationally aligned safety and security management frameworks.

The AISI will continue to strengthen AI safety capabilities in industry through international collaboration and to support the development of a trustworthy AI ecosystem.


MEDIA SALON
PREMIER LOUNGE


보안 통제의 적용
가능성을 검토하고 이를
감리하는 차입금
노동집약적인 과정이다


AI Security Workshop



[워크숍 업데이트] 인공지능안전연구소 – 미국 국무부 – RAND – MITRE 공동 주관, 'AI Security Workshop' 개최

인공지능안전연구소는 미국 국무부, RAND 및 MITRE와의 공동 주관으로 17개 기업과 4개 기관에서 약 60여 명이 참석하는 「AI Security Workshop」을 2월 25일(수)부터 2월 27일(금)까지 개최했다고 밝혔다. 이번 워크숍은 인공지능(AI) 시스템의 안전(Safety) 확보를 중심으로 보안(Security) 및 국가안보(National Security) 관점까지 아우르는 위험 관리 체계를 공유하고 적용 방안을 논의하기 위해 기획되었다. AI 기술이 산업 및 공공 영역 전반으로 빠르게 확산되는 가운데 기술 혁신과 함께 안전성과 신뢰성을 동시에 확보하는 것이 지속 가능한 발전의 핵심 과제로 부상하고 있다는 점에서 이번 행사의 의미가 크다.

▲ 1일차: AI 위협 환경 조망과「AI Security Guide」및「MITRE ATLAS」등 논의

워크숍 첫째 날(2월 25일)은 총 3개 세션으로 구성되었다. 첫 번째 세션에서는 AI 기술 확산에 따른 위험 환경을 종합적으로 조망했다. 비확산(nonproliferation) 관점에서 본 AI 위협 환경 발표를 시작으로 고성능 AI의 이중용도 특성이 논의되었다. 참석자들은 AI 시스템의 개발·배포·운영 전 단계에서 안전을 내재화하고, 조직 차원의 보안 관리 체계를 정비하는 선제적 접근이 필요하다는 데 공감했다. 두 번째 세션에서는 미국 RAND가「AI Security Guide」를 중심으로 조직의 AI 보안 태세를 진단하는 방법론을 소개했다. 생애주기 중심 위협 분석과 위험 우선순위화 도구를 통해 핵심 위험을 식별하고 실질적 통제 방안을 도출하는 체계적 접근이 공유되었다. 마지막 세션에서는 미국 MITRE가「MITRE ATLAS」와「AI Assurance Process」를 발표했다. 해당 프로세스는 시스템 특성 문서화, 위험 평가, 모니터링 등 보증 활동을 구조화한 반복적 생애주기 모델로, AI의 안전성과 신뢰성을 확보하는 방안을 제시했다.

▲ 2일차: 기업 맞춤형 1:1 컨설팅

둘째 날(2월 26일)에는 사전 신청 기업을 대상으로 RAND와 MITRE 전문가가 참여하는 1:1 맞춤형 컨설팅이 진행되었다. 각 기업의 AI 시스템 구조, 데이터 흐름, 배포 환경 및 기존 통제 체계를 점검하고, 우선 조치 사항과 보증 계획 수립 방향을 구체화했다. 이를 통해 기업이 안전을 핵심 원칙으로 하되 보안과 책임성까지 아우르는 통합 관리 체계를 구축할 수 있는 기반이 마련되었다.

▲ 3일차: AI 분야 대표 스타트업 방문 및 교류

마지막 날(2월 27일)에는 AI 안전 및 보안 분야를 대표하는 한국 스타트업 AIM Intelligence을 방문하여 기술 개발 현황을 공유하고, 국제 기준에 부합하는 안전 및 보안 관리 체계 적용 방안을 논의했다. 인공지능안전연구소는 앞으로도 국제 협력을 바탕으로 산업 현장의 AI 안전 역량을 강화하고, 신뢰 가능한 AI 생태계 조성을 지속적으로 지원할 계획이다.