[Workshop Update] AI Safety Institute and U.S. Experts Convene Discussions on AI Security in the CBRN Domain

The CBRN AI Security Technical Exchange was convened by the U.S. Department of State and held from January 26 to 29, 2026, at the Artificial Intelligence Safety Institute (AISI) in Pangyo, Republic of Korea. The workshop brought together approximately 30 participants from the AISI, along with 11 representatives from the United States, including officials and researchers from the U.S. Department of State, Lawrence Livermore National Laboratory (LLNL), Pacific Northwest National Laboratory (PNNL), and the U.S. Center for AI Standards and Innovation. This technical exchange followed the international AI security workshop hosted in Livermore in August 2025 and aimed to further strengthen cooperation on artificial intelligence-related risks and threats in the chemical, biological, radiological, and nuclear domains.

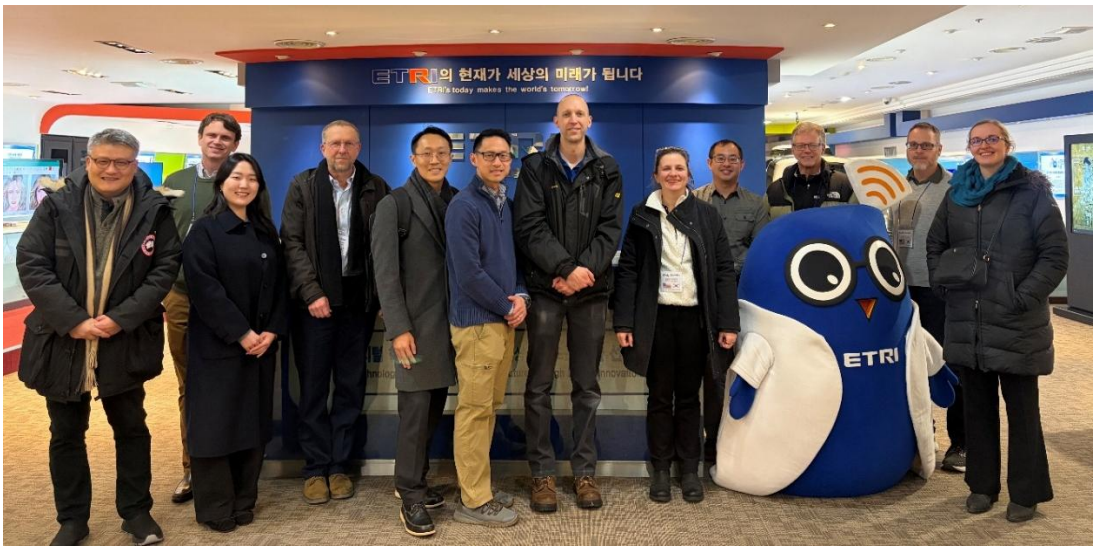
The workshop focused on how recent advances in artificial intelligence, particularly large language models and emerging agentic systems, are reshaping the CBRN threat landscape. Initial sessions examined AI-enabled pathways for proliferation and misuse, with particular attention to the biological domain, where the convergence of AI and biotechnology might pose heightened risks to global security. The AISI introduced its AI Risk Mapping approach, highlighting the importance of systematically identifying and structuring emerging risks to support national security.

Building on this foundation, sessions led by experts from LLNL and PNNL explored the integration of AI into chemical and biological research environments. Discussions addressed dual-use concerns, red teaming practices, and benchmarking methodologies, supported by concrete case studies. Participants examined how evaluation science can be operationalized to identify vulnerabilities and reduce misuse risks, underscoring the need for shared technical baselines and interoperable evaluation frameworks across institutions and countries.

The AISI also presented its ongoing research on AI safety evaluation and model resilience. Topics included agent-based evaluations in sandboxed environments, automated generation of CBRN-related prompts, and the development of software platforms for scalable benchmarking and red teaming. Additional sessions covered multilingual counter censorship benchmarking,

risks associated with AI model uplift, jailbreak detection, and multimodal deepfake detection, reflecting the expanding scope and complexity of AI security challenges in security sensitive contexts.

The program further included site visits to the Agency for Defense Development and the Electronics and Telecommunications Research Institute, as well as expert-led discussions on CBRN red teaming and defense related AI research. Over four days of technical exchange, the workshop reinforced the shared understanding that AI security challenges transcend national boundaries. The AISI will continue to work closely with the U.S. Department of State, U.S. national laboratories, and other international partners to strengthen global capacity for evaluating, governing, and mitigating AI-enabled CBRN risks.



[워크숍 업데이트] 인공지능안전연구소와 미국 전문가, CBRN 분야 AI 안보 관련 논의 개최

미국 국무부가 주관한「 CBRN AI 안보 기술 교류」가 2026 년 1 월 26 일부터 29 일까지 대한민국 판교에 위치한 인공지능안전연구소에서 개최되었다. 이번 워크숍에는 인공지능안전연구소 소속 약 30 명의 연구진이 참석했으며, 미국 측에서는 국무부, 로렌스 리버모어 국립연구소(LLNL), 퍼시픽 노스웨스트 국립연구소(PNNL), 미국 AI 표준·혁신 센터(U.S. Center for AI Standards and Innovation) 관계자 및 연구자 등 11 명이 참여하였다. 이번 기술 교류는 2025 년 8 월 리버모어에서 개최된 국제 AI 안보 워크숍의 후속 논의로 화학·생물·방사능·핵(CBRN) 분야에서의 인공지능 관련 위험과 위협에 대한 협력을 심화하기 위해 마련되었다.

이번 워크숍은 특히 대규모 언어모델과 신흥 에이전트형 시스템 등 최근 인공지능 기술 발전이 CBRN 위협 지형을 어떻게 재편하고 있는지에 초점을 맞추었다. 초기 세션에서는 AI 기반 확산 및 오용 경로를 분석했으며, 특히 AI 와 생명공학의 융합이 글로벌 안보에 중대한 위험을 초래할 수 있는 생물 분야에 대한 논의가 집중적으로 이루어졌다. 인공지능안전연구소는 국가안보 지원을 위해 신흥 위험을 체계적으로 식별 및 구조화하는 'AI 위험지도(AI Risk Mapping)' 접근법을 소개하였다.

이어 LLNL 과 PNNL 전문가들이 주도한 세션에서는 화학 및 생물 연구 환경에서의 AI 통합 사례를 다루었다. 논의는 이중용도(dual-use) 우려, 레드팀 운영 방식, 벤치마킹 방법론 등을 구체적 사례와 함께 검토하는 방식으로 진행되었다. 참가자들은 평가 과학(evaluation science)을 어떻게 제도화 및 운영하여 취약점을 식별하고 오용 위험을 줄일 수 있는지 논의했으며, 기관 및 국가 간 공유 가능한 기술 기준과 상호운용 가능한 평가 프레임워크 구축의 필요성을 강조하였다.

인공지능안전연구소는 또한 AI 안전성 평가와 모델 복원력(model resilience)에 관한 연구 현황을 발표하였다. 주요 내용으로는 샌드박스 환경에서의 에이전트 기반 평가, CBRN 관련 프롬프트 자동 생성, 확장 가능한 벤치마킹 및 레드팀 운영을 위한 소프트웨어 플랫폼 개발 등이 포함되었다. 이와 함께 다국어 검열회피 대응 벤치마킹, AI 모델 성능 증폭(uplift)과 관련된 위험, 탈옥(jailbreak) 탐지, 멀티모달 딥페이크 탐지 등 안보 민감 영역에서 점차 복잡해지는 AI 보안 과제도 논의되었다.

아울러 참가자들은 국방과학연구소(ADD)와 한국전자통신연구원(ETRI)을 방문하고, CBRN 레드팀 및 국방 관련 AI 연구에 대한 전문가 토론을 진행하였다. 4 일간의 기술 교류를 통해 참가자들은 AI 안보 과제가 국경을 초월하는 문제라는 데 인식을 같이하였다. 인공지능안전연구소는 앞으로도 미국 국무부, 미국 국립연구소 및 기타 국제 파트너들과 긴밀히 협력하여 AI 기반 CBRN 위험의 평가·거버넌스·완화 역량을 강화해 나갈 예정이다.