전자통신동향분석, 41권 3호 2026, 84-95

AI 레드팀 테스팅 기술 표준화 동향

Standardization Trends in AI Red Team Testing

전종홍 (J.H. Jeon, hollobit@etri.re.kr) 지능정보표준연구실 책임연구원

ABSTRACT

Recent advances in large language models and agentic AI systems have expanded the operational scope of AI from content generation to autonomous planning, tool use, and multistep task execution. As these systems gain increased access to external tools, enterprise data, and runtime environments, the need for AI red team testing is increasing rapidly. AI red teaming differs from traditional cybersecurity red teaming and conventional software testing in that it must address nondeterministic behaviors, context-dependent failures, prompt injection, memory poisoning, tool misuse, data leakage, and multi-agent interaction risks. This study analyzes the recent standardization and guidance trends related to AI red team testing by examining ISO/IEC software testing standards, emerging AI testing work items, NIST evaluation and adversarial machine learning publications, and practical guidance from OWASP, CSA, and national AI Safety Institutes. Our analysis reveals that current international efforts are evolving from model-centric adversarial testing to system- and lifecycle-oriented evaluations, including risk scenario design, attack procedure development, automated and human-in-the-loop measurements, structured reporting, and continuous monitoring. In particular, the relationship between the ISO/IEC/IEEE 29119 test processes and the emerging ISO/IEC 42119-7 red teaming work item suggests that AI red teaming is becoming formalized as a specialized testing profile, rather than an ad hoc security exercise. This paper identifies major standardization issues in terminology, test process alignment, attack taxonomy, measurement methods, reporting templates, and agentic AI runtime assurance and concludes with implications for domestic standardization, evaluation frameworks, and operational guidance.

KEYWORDS agentic AI, AI red teaming, AI testing, evaluation, LLM, risk management, standardization

* DOI: https://doi.org/10.22648/ETRI.2026.J.410308

* This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) [No. RS-2025-02214416, Development of Safety and Trustworthiness Evaluation Technology Standards for Advanced Artificial Intelligence].

I. 서론

생성형 AI와 에이전틱 AI는 단순한 질의응답이나 콘텐츠 생성 도구를 넘어 외부 도구를 호출하고, 기업 데이터를 조회하며, 다단계 업무를 수행하는 방향으로 빠르게 확장되고 있다. 대규모 데이터로 학습된 파운데이션 모델과 대규모언어모델(LLM: Large Language Model), 시각언어모델(VLM: Vision Language Model)을 포함한 프런티어 AI는 높은 범용성을 바탕으로 여러 산업 현장에 도입되고 있으며, 브라우저나 파일시스템, 데이터베이스, 응용프로그램 인터페이스(API: Application Programming Interface), 메일, 캘린더와 결합되면서 실제 업무 절차에 직접 개입하는 수준에 이르고 있다.

그러나 이러한 확장은 검증의 난도를 크게 높인다. 프런티어 AI 모델은 규모와 복잡성이 크고 비결정성이 강하며, 학습 데이터의 불투명성과 운영 환경과의 결합성도 높아 전통적인 V&V(검증 및 확인)만으로는 충분히 다루기 어렵다. 정적 점검이나 사전 정의된 테스트 케이스만으로는 장문 맥락 처리, 예기치 않은 추론 오류, 안전정책 우회, 도구 연계 오작동, 자율적 작업 흐름의 이탈 같은 문제를 충분히 포착하기 어렵고, 동일한 입력에서도 결과가 달라질 수 있다는 점 역시 기존 시험 방식의 한계를 드러낸다.

이러한 한계 때문에 AI 안전성과 보안성을 확보하려면 새로운 시험 접근이 필요하다. 프롬프트 인젝션, 데이터 유출, 허위 정보 생성, 권한 남용, 장기 메모리 오염, 다중 에이전트 간 신뢰 악용 같은 위협은 실제 사용 맥락과 공격 시나리오를 함께 놓고 보아야 그 위험을 제대로 가늠할 수 있다(표 1). 이 때문에 최근에는 위험관리 체계와 연계해 적대적 공격 관점에서 취약점과 위협 요소를 찾으려는 레드팀 시험에 관심이 커지고 있다. AI 레드팀은 모델의 응답 오류를 찾는 수준을 넘어, 실제 운영 환경에서 발생할 수 있는 위해와 통제 실패를 함께 살피는 방식으로 발전하고 있다.

이와 관련해 국제적으로는 소프트웨어 시험 표준인 ISO/IEC/IEEE 29119 시리즈를 바탕으로 AI 시스템 시험을 위한 ISO/IEC 42119-7(이하 42119-7) 표준 개발이 논의되고 있다. 미국 국립표준기술연구소(NIST: National Institute of Standards and Technology) 문헌[4-6]은 적대적 기계학습 분류체계와 언어모델 평가 관행을 정리하고 있으며, 오픈 월드와이

표 1 AI 레드팀 주요 위험유형과 대표 시험기법

위험유형	대표시험 기법	주요 근거문서	예시지표
프롬프트 인젝션/ 목표 탈취	직접·간접프롬프트 인젝션, 악성 내용 삽입	OWASP GenAI RT, OWASP Agentic Top 10, NIST AML	공격 성공률, 정책 위반 비율
데이터 유출	업무 시나리오 기반 과업 시험, 민감정보 취급 정책 검사	Testing AI Agents, NIST 800-2	유출 비율, 허위 노출 건수
도구 오용/ 과도한 자율성	도구 호출 오용, 무단 조치 모의시험	OWASP Agentic Top 10, Securing Agentic Applications	도구 오용 건수, 과업 이탈 비율
메모리·문맥 오염	문맥 오염, 세션 간 오염	OWASP Agentic Top 10, OWASP GenAI RT	오염 지속성
다중 에이전트 통신 위험	가장, 위장, 위조된 에이전트 메시지	OWASP Agentic Top 10, CSA guide	인증 우회율
연쇄 실패/ 통제 이탈	장기 과업 스트레스 시험, 실행 환경 모니터링	Agentic AI Profile, MGF, OWASP Agentic Top 10	연쇄 전파 건수

드 애플리케이션 시큐리티 프로젝트(OWASP: Open Worldwide Application Security Project), 클라우드 시큐리티 얼라이언스(CSA: Cloud Security Alliance), 일본 AI Safety Institute의 가이드 문헌[7-10]도 실무적 점검 기준을 제시한다. 이런 흐름은 AI 레드팀 시험이 프런티어 AI와 에이전틱 AI의 안전성, 신뢰성, 보안성을 확보하기 위한 핵심 시험 방법으로 자리매김하고 있음을 보여준다.

본고는 이러한 국제 문서를 바탕으로 AI 레드팀 테스팅 기술 표준화 동향을 살펴본다. 먼저 AI 레드팀 테스팅의 개념과 범위를 정리하고, 이어 용어와 프로세스, 위험 시나리오, 공격기법, 평가방법, 보고체계, 에이전틱 AI 시험으로 이어지는 핵심 쟁점을 검토한 뒤, 마지막으로 국내 표준화와 평가체계 설계에 필요한 시사점을 제시하고자 한다.

II. AI 레드팀 테스팅의 개념과 범위

1. AI 레드팀 테스팅의 개념

표 1에 볼 수 있듯이 보안 레드팀은 조직, 인프라, 시스템의 방어 체계를 공격자 관점에서 점검해 취약점을 찾는 활동이다. AI 레드팀은 여기에 모델의 안전성, 신뢰성, 정책 준수, 유해 출력 가능성, 편향, 잘못된 자율성까지 함께 포함한다. AI 레드팀은 침투 시험에 머물지 않고, AI 시스템이 특정 사용 맥락에서 얼마나 안전하고 통제 가능하게 작동하는지를 확인하는 적대적 시험 활동이다.

레드팀은 정의된 목표의 효과성이나 견고성을 높이기 위해 적대적 관점을 취하는 독립 그룹이며, AI 레드팀은 적대적 방법으로 AI 시스템의 취약점을 식별하는 다양한 배경의 이해관계자 집단이다. 따라서 AI 레드팀은 단순한 공격자 시뮬레이션을 넘어서 모델 특성, 응용 도메인, 위험관리, 정책 기준을 함께 이해하는 다학제적 팀으로 구성되어야 한다[3].

AI 레드팀 테스팅은 시험 대상을 기준으로 모델, 시스템, 에이전트의 세 수준으로 나눌 수 있다. 모델 수준에서는 프롬프트 입력과 출력 응답을 중심으로 유해성, 환각, 편향, 정책 우회 가능성을 점검한다. 시스템 수준에서는 검색증강생성(RAG: Retrieval-Augmented Generation), 정책 필터, 사용자 인터페이스, 로깅, 외부 API 연결 등 응용 시스템 전반을 시험한다. 에이전트 수준에서는 도구 선택, 장기 메모리 사용, 권한 위임, 다단계 계획, 멀티에이전트 상호작용, 작업 목표 이탈 여부까지 포함한 운영 거동을 살핀다. 최근 표준화의 관심은 점차 세 번째 수준인 에이전트 환경으로 옮겨가고 있다.

표 2 전통적 보안 레드팀, 생성형 AI 레드팀, 에어전틱 AI 레드팀 비교

구분	전통적 보안 레드팀	생성형 AI 레드팀	에이전틱 AI 레드팀
주 대상	네트워크, 시스템, 계정, 애플리케이션	모델, 프롬프트, 필터, RAG 시스템	에이전트, 도구, 메모리, 워크플로, 다중에이전트
주 위험	침투, 권한 상승, 서비스 중단	유해 출력, 안전 정책 우회, 환각, 데이터 추출	목표이탈, 도구오용, 권한남용, 문맥오염, 연쇄실패
시험단위	자산/서비스	모델 또는 AI 응용	작업흐름과 운영환경
주요기법	침투 테스트, 취약점 악용	프롬프트 공격, 정책 우회, 안전성 탐색	장기과업수행, 도구연쇄공격, 에이전트 간 공격
평가기준	침해 성공 여부	정책 위반, 유해 응답, 안전 필터 우회	작업완료율, 정책준수율, 유출비율, 이탈비율

2. 시험 대상과 위험 범주

AI 레드팀 시험의 대상은 LLM 자체에 그치지 않는다. 시험 대상은 모델 평가, 구현 평가, 시스템 평가, 실행 환경 및 인간 · 에이전트 평가로 나눌 수 있으며, 에이전틱 응용에는 별도의 레드팀 과업도 필요하다. 생성형 AI 시스템은 모델, 프롬프트, 정책, 데이터, 외부 도구, 운영 절차가 얽힌 복합체이기 때문이다[7].

위험 범주는 문헌마다 조금씩 다르지만, 대체로 입력 조작, 정보보호, 자율성 및 도구 오용, 지식 · 문맥, 협업 및 운영의 다섯 축으로 묶을 수 있다. 입력 조작에는 직접 · 간접 프롬프트 인젝션, 탈옥 공격, 적대적 내용 삽입이 포함되고, 정보보호에는 데이터 유출, 비밀정보 노출, 권한 없는 조회, 개인정보 침해가 해당한다. 자율성 및 도구 오용은 에이전트가 잘못된 도구를 선택하거나 과도한 권한으로 부적절한 조치를 수행하는 문제를 가리키며, 지식 · 문맥 범주에는 메모리 및 문맥 오염과 장기 기억 조작, 외부 지식베이스 변조가 들어간다. 협업 및 운영 범주에는 다중 에이전트 간 스푸핑, 안전하지 않은 통신, 연쇄 실패, 인간-에이전트 신뢰 악용 등이 포함된다.

적대적 기계학습 공격은 공격 시점, 공격자 목표, 능력, 지식 수준에 따라 체계화할 수 있으며, 이런 분류는 공격을 설명하는 공통 언어를 제공한다. 에이전틱 AI에서는 목표 탈취, 도구 오용 및 악용, 신원 및 권한 남용, 메모리 및 문맥 오염, 안전하지 않은 에이전트 간 통신, 통제를 벗어난 에이전트가 핵심 위협으로 제시된다. 현실적 업무 과업에서 발생하는 데이터 유출 위험도 별도 범주로 다루어야 한다. 악성 공격뿐만 아니라 정상적 업무 흐름에서 생기는 민감정보 처리 실패도 중요한 시험 대상이기 때문이다[4,11,12].

3. AI 레드팀의 특징과 한계

AI 레드팀 시험은 일반 소프트웨어 시험과 달리 예상 결과를 고정하기 어렵다. ISO/IEC/IEEE 29119-1(이하 29119-1)의 예상 결과와 시험 오라클 개념은 AI 시스템에도 적용할 수 있지만, 실제 시험에서는 하나의 정답보다 허용 가능한 거동 범위, 정책 준수 여부, 위해 발생 가능성, 문맥 적합성을 함께 본다. 따라서 AI 레드팀은 정성적 판단과 정량적 지표를 함께 써야 한다.

운영 환경과 시험 환경의 차이도 크다. 에이전트는 실제 도구와 데이터에 연결될 때 특정 위험을 드러내는 경우가 많고, 운영 환경에서 시험을 수행하면 서비스 장애나 정보 노출 위험이 생길 수 있다. 그래서 실무자들은 공개 전과 공개 후의 레드팀 수행을 구분하고, 환경 준비, 예산과 자원, 외부 전문기관 계약, 보고와 비상 대응 절차를 별도 단계로 다룬다. 이런 접근은 AI 레드팀이 기술적 기법에 머무르지 않고 운영 통제와 보고체계를 포함하는 조직 활동이라는 점을 보여준다[8].

III. AI 레드팀 테스팅 기술 표준화 동향

1. 표준화 필요성과 추진 배경

AI 레드팀 시험 표준화의 첫 번째 목적은 공통 용어와 절차를 정의하는 데 있다. 현재 각 조직은 레드팀 시험, 적대적 시험, 안전성 평가, 벤치마크 평가, 보증 시험 같은 용어를 혼용하고 있으며, 시험 목적도 보안 취약점 식별에서 정책 준수 평가, 성능 저하 확인, 인체 · 사회적 위해 예방까지 넓게 퍼져 있다. 표준화는 이런 활동을 용어, 역할, 산출물, 보고체계 측면에서 정리해 조직 간 비교 가능성을 높인다.

두 번째 목적은 시험 절차의 재현성과 운영 가능성을 높이는 데 있다. ISO/IEC/IEEE 29119-2(이하

29119-2)는 시험 전략과 계획, 모니터링과 통제, 설계와 구현, 환경 및 데이터 관리, 실행, 사고 보고, 종료 절차를 정립하고 있다. 여기에 ISO/IEC TS 49119-2(이하 49119-2)는 AI 시스템 시험 전반을 대상으로 위험 기반 시험, AI 시스템 생애주기, 데이터 품질, 커버리지, 리뷰 기법을 연결해 AI 시험의 공통 바탕을 제공한다. AI 레드팀이 이 틀을 활용하면 비정형 공격 실험을 체계적인 시험 활동으로 바꿀 수 있다. 위험 시나리오, 시험 목적, 시험 환경, 중단 기준, 보고 절차를 분명히 하는 방식은 AI 안전시험의 품질을 높인다[2,16].

세 번째 목적은 규제와 거버넌스와의 연계이다. AI 보안과 위험관리는 조직 차원의 책임으로 다루어져야 하며, 레드팀도 단순한 기술 이벤트가 아니라 위험관리 체계의 일부로 들어가야 한다. 앞으로 AI 레드팀 결과는 적합성 평가, 내부 통제, 사고 대응, 외부 보고와 연결될 가능성이 크다[6,13,14].

2. 용어 개념 표준화

현시점에서 용어 표준화의 중심축은 ISO/IEC/IEEE 29119-1, ISO/IEC TS 42119-2, ISO/IEC 42119-7의 통합에 있다. 테스트 케이스, 테스트 절차, 예상 결과, 시험 항목, 시험 목적, 사고 같은 기본 시험 개념은 기존 소프트웨어 시험 표준이 제공하고, AI 시스템 생애주기와 위험 기반 시험의 공통 개념은 42119-2가 정리하며, 레드팀, AI 레드팀, 적대적 공격, 데이터 포이즈닝, 환각 같은 AI 레드팀 맥락의 개념은 42119-7이 보완하고 있다. 즉, 기존 시험 표준은 시험 활동의 골격을 제공하고, 42119-2는 AI 시스템 시험의 공통 층위를 보강하며, 신흥 AI 레드팀 표준은 AI 특유의 시험 대상과 공격 개념을 더하는 구조다[1,3,16].

이 과정에서 몇 가지 쟁점도 남아 있다. 레드팀의 목적을 취약점 식별에 둘지, 정량적 측정까지 포함할지는 아직 정리되지 않았다. 식별과 측정의 구분, 그리고 평가라는 용어의 범위도 계속 논의되는 부분이다. 시험 대상을 AI 시스템, AI 모델, 또는 둘 다로 볼지도 여전히 쟁점이며, 에이전틱 응용에서는 모델보다 시스템과 실행 환경이 더 큰 위협 요인이 될 수 있다[3].

일반적인 적대적 기계학습 용어와 응용 시스템 시험 용어도 연결해야 한다. 데이터 포이즈닝, 회피 공격, 사생활 침해, 공격자 능력 같은 분류만으로는 실제 에이전트 시험을 충분히 설명하기 어렵고, 목표 조작, 문맥 오염, 도구 오용, 인증 우회율 같은 응용 수준 용어가 함께 있어야 한다. 최근 작업문서는 action traceability, checker-out-of-the-loop, orchestrator state poisoning처럼 에이전틱 운영 통제를 직접 겨냥한 용어도 추가하고 있다. 향후 표준화는 모델 수준 용어와 응용 수준 용어를 나란히 대응시키는 방향으로 발전할 가능성이 높다[3,4].

3. 테스트 프로세스 표준화

표 3과 같은 현재 가장 빠르게 정리되는 분야다. 29119-2가 제시하는 시험 전략과 계획, 모니터링과 통제, 시험 설계와 구현, 시험 환경 및 데이터 관리, 시험 실행, 시험 사고 보고, 시험 종료 절차는 AI 레드팀에도 비교적 직접적으로 적용할 수 있다. 42119-2는 여기에 AI 시스템 생애주기와 위험 기반 시험을 결합하고, 데이터 대표성 시험, 데이터 품질 시험, 정적 리뷰, 커버리지 선택 같은 AI 시험 특화 요소를 더한다[2,16].

AI 레드팀 프로세스는 크게 세 단계로 볼 수 있다. 첫째는 팀 구성 및 준비 단계로, 팀 구성, 목표 설정, 범위 정의, 위험 시나리오 개발, 시험 환경 준비가 여기에 들어간다. 둘째는 수행 단계로, 개별 프롬프

표 3 ISO/IEC/IEEE 29112-2와 ISO/IEC AWI TS 42119-7의 프로세스 정합성

42119-7 활동	29119-2 대응절차	의미
팀 구성, 목표 설정, 범위 정의, 위험 시나리오 개발	시험 전략 및 계획 절차	시험 목적과 자원, 범위, 위험 기반 계획 수립
환경 준비, 접근권한·로그 설정	시험 환경 및 데이터 관리 절차	시험 환경과 데이터 준비
탐색적 시험, 공격절차 설계	시험 설계 및 구현 절차	공격 기반 테스트 케이스와 절차 개발
시스템 전반 공격 수행	시험 실행 절차	시험 실행과 결과 관찰
치명 취약점 즉시 보고	시험 사고 보고 절차	운영상 중대 사항 즉시 통지
최종 보고, 개선계획, 교훈 정리	시험 종료 절차	결과 종료 보고와 후속 개선

트를 대상으로 한 탐색적 시험, 공격 징후와 공격 절차 개발, 시스템 전반 시험, 운영 위험 모니터링, 중대 취약점 보고가 포함된다. 셋째는 지식 공유 및 보고 단계로, 최종 보고서 작성, 개선 계획 개발, 교훈 정리가 여기에 해당한다[3].

이 구조는 착수 결정, 예산과 자원 확보, 제3자 선정, 대상 시스템 개요와 사용 양상 파악, 레드팀 종류와 범위 결정, 환경 준비, 비상 대응 절차 확인, 위험 시나리오 개발, 공격 시나리오 개발, 공격 수행, 기록 유지, 결과 분석, 이해관계자 검토, 최종 보고, 개선 계획, 후속 조치로 더 세분화할 수 있다. 이는 ISO 가이드라인은 추상적 절차 구조를 제공하고, AISI 가이드라인은 조직 운영 수준의 세부 절차를 보완하는 관계에 있음을 보여준다. 또한, AI 시스템은 운영 중 거동이 변할 수 있으므로, 연속 시험과 정적 리뷰를 함께 두는 접근이 점차 중요해지고 있다[8,16].

4. 위험 시나리오 및 공격기법 표준화

AI 레드팀은 무작위 프롬프트 공격의 집합이 아니라 위험 시나리오와 공격 시나리오를 연계하는 활동으로 발전하고 있다. 일본 AISI는 위험 시나리오를 시스템 구성과 사용 패턴을 바탕으로 개발한 뒤 그에 맞는 공격 시나리오를 설계하도록 권고한다. 이 접근은 위험 식별과 시험 설계를 분리해 다룬다는 점에서 중요하다.

위협모델링은 선행 단계에서 다루어야 한다. 기술적 공격면뿐만 아니라 조직의 사회적, 규제적, 윤리적 맥락도 함께 봐야 한다. 같은 프롬프트 인젝션이라도 고객지원 챗봇, 사내 생산성 비서, 의료정보 요약 시스템에서 위해 수준이 다르므로, 시나리오를 설계할 때는 사용 맥락과 보호해야 할 정보자산을 먼저 짚어야 한다[7].

공격기법 측면에서 현재 표준화 후보군은 점차 구체화되고 있다. 시험 영역으로는 프롬프트 인젝션 범주, 영역별 레드팀 접근법, 안전, 품질 · 정확성, 성능, 에이전트 자율성, 자원과 비용, 다중 에이전트가 제시되고 있으며, 실사례와 함께 목표 탈취, 도구 오용, 메모리 · 문맥 오염, 안전하지 않은 에이전트 간 통신, 연쇄 실패도 주요 항목으로 정리되고 있다. 더 상위 수준에서는 적대적 기계학습 분류체계가 공격자의 목적, 지식, 능력, 공격 생애주기 단계를 정리한다. 향후 표준은 이 세 층위를 연결하는 방향으로 발전할 가능성이 높다[3,4,11].

5. 평가 방법과 측정 표준화

AI 레드팀 표준화에서 가장 어려운 과제는 평가의 측정 가능성과 비교 가능성이다. 자동화 벤치마크 평가 관행은 측정 대상 정의, 평가 구현 및 실행,

분석과 보고의 세 단계로 묶어 볼 수 있다. 여기에는 어떤 능력이나 위험을 측정할지, 어떤 데이터셋과 과업이 적절한지, 어떤 실패를 위해로 볼 것인지에 대한 정의와 함께 시험 세트 구성, 실행 조건, 자동 채점, 반복 실행, 불확실성 관리, 오류 범위와 편향, 실패 사례, 재현성 한계 보고가 포함된다[5].

그러나 AI 레드팀을 자동 벤치마크만으로 설명하기는 어렵다. 실제 현장에서는 정량 평가와 정성 평가를 함께 쓰는 복합 평가가 필요하다. 유해 응답 여부는 자동 분류가 가능하지만, 민감정보 유출의 맥락 적합성이나 인간 감독의 개입 필요성은 사람이 판단하는 편이 낫다. 인간 평가와 자동 평가가 서로 다른 결론을 낼 수 있으므로, 향후 표준은 자동 채점 기준, 사람 검토 기준, 불일치 처리 규칙을 함께 다뤄야 한다[12].

최근에는 측정 결과의 타당성 자체를 검증하는 요구도 커지고 있다. 어떤 벤치마크가 실제 위험을 제대로 대표하는지, 학습 데이터나 공개 사례와의 오염 가능성은 없는지, 비교 기준선과 반복 실행 조건은 충분한지, 표준오차와 불확실성을 어떻게 보고할지 같은 문제가 핵심 쟁점으로 떠오른다. 특히 LLM을 판정자로 쓰는 방식은 평가 자동화에 유용하지만, 분포 이동이나 공격 조작에 취약할 수 있어 판정 오라클과 샌드박스 같은 평가 인프라 자체도 시험 대상에 포함할 필요가 있다[5].

에이전틱 AI 시험은 정적 질의응답보다 과업 중심 지표를 요구한다. 제한된 자율성 환경에서 고객 지원, 기업 생산성, 게시, 분석, 거래 에이전트 같은 유형을 두고 모델 컨텍스트 프로토콜(MCP: Model Context Protocol) 서버 기반 실제 도구 연동 환경에서 다단계 작업을 수행하게 하는 접근은 과업 완료율, 정책 준수도, 데이터 유출, 문맥 인식 수준, 사용자 지침 준수 같은 시스템 수준 지표를 강조한다. 이는 AI 레드팀의 관심이 단순 공격 성공률에서 작업 흐름 전반의 위험 측정으로 옮겨가고 있음을 보여준다[12].

6. 에이전틱 AI 레드팀 테스팅

에이전틱 AI 환경은 기존 LLM 및 생성형 AI 환경과 시험 대상과 위험 구조에서 뚜렷한 차이를 보인다. 기존 생성형 AI 시험이 단일 질의응답 흐름에서의 유해 출력, 환각, 편향, 프롬프트 인젝션, 정보 유출 가능성에 초점을 맞췄다면, 에이전틱 AI 시험은 목표 해석, 계획 수립, 도구 선택, 실행 순서 결정, 장기 상태 유지, 외부 시스템과의 상호작용까지 포함하는 연속적 행위를 평가해야 한다. 시험의 중심도 “모델이 무엇을 말하는가”에서 “에이전트가 무엇을 결정하고 실제로 무엇을 수행하는가”로 옮겨간다.

이 차이는 새로운 위험과 위협을 낳는다. 입력 조작이나 외부 환경 오염은 에이전트가 사용자의 원래 의도와 다른 목표를 추구하게 만들 수 있다. 외부 도구 호출과 파일 읽기 · 쓰기, 명령 실행 과정에서는 권한 초과, 잘못된 도구 사용, 비용 폭증, 무한 반복 같은 실행 수준의 문제도 뒤따른다. 장기 메모리, 대화 이력, 작업 상태, 검색 문맥이 오염되면 잘못된 정보가 계속 누적되고 이후 판단도 흔들린다. 다중 에이전트 구조에서는 통신 위조, 신뢰 악용, 연쇄 실패, 통제 불능의 자율적 행위 확산이 추가된다.

따라서 에이전틱 AI는 기존 생성형 AI 시험을 조금 넓히는 방식으로 다룰 수 없다. 시험 대상이 단일 모델이 아니라 도구, 메모리, 권한 체계, 작업 흐름, 다중 에이전트 협업 구조를 아우르는 복합 시스템이므로, 시험 단위도 프롬프트 수준을 넘어 과업 수준과 운영 수준으로 넓혀야 한다. 단발성 응답 평가로는 위험을 놓치기 쉬우므로, 장시간 작업 수행, 단계별 의사결정 추적, 실제 도구 호출 결과, 실패 전파 양상, 인간 감독 개입 지점을 함께 살필 수 있는 시험 체계가 필요하다.

이런 배경에서 최근 국제 논의는 에이전틱 AI에 특화된 레드팀 테스팅 체계와 방법론을 모색하고 있다. 논의 범위는 안전한 에이전틱 구조, 개발자 생애주기 지침, 강화된 보안 조치, 핵심 운영 능력, 에이전틱 공급망, 레드팀 및 행위 기반 시험, 실행 환경 강화까지 넓어지고 있다. 시험 대상도 모델 응답에 머무르지 않고 API 접근, 코드 실행, 웹 사용, 파일시스템 및 운영체제 명령, 중요 시스템 제어 같은 실행 능력 전체로 확장된다[10].

에이전트의 고유 위험으로는 목표 탈취, 도구 오용 및 악용, 신원 및 권한 남용, 메모리 및 문맥 오염, 안전하지 않은 에이전트 간 통신이 꼽힌다. 목표 탈취는 입력 조작이 에이전트의 상위 목표 해석을 바꾸는 문제를 다루고, 도구 오용 및 악용은 허용된 도구의 부적절한 사용을, 신원 및 권한 남용은 권한체계 오남용을, 메모리 및 문맥 오염은 저장된 문맥과 기억의 오염을, 안전하지 않은 에이전트 간 통신은 에이전트 사이 교신의 인증과 무결성 문제를 다룬다. 최근 문헌은 여기에 에이전트 정체성, 최소권한, 승인 체계, 실행 맥락을 포함한 Action Traceability를 핵심 통제 원리로 올리고 있다. 누가 어떤 권한으로 어떤 도구를 호출했는지 검증 가능한 로그로 남기지 못하면, 에이전틱 AI의 안전성과 책임성을 평가하기 어렵다[3,10,11,13,14].

고객지원, 기업 생산성, 게시, 분석, 거래 같은 현실적 과업을 설정한 뒤 MCP 서버 기반 도구 환경에서 데이터 유출과 정책 위반 가능성을 점검하는 접근은, 에이전틱 AI 시험이 단순 문항 평가보다 실제 과업 수행 환경을 중시하는 방향으로 이동하고 있음을 보여준다. 에이전틱 AI 레드팀 가이드, 에이전틱 AI 위험관리 표준 프로파일, 에이전틱 AI 거버넌스 프레임워크도 이런 흐름을 거버넌스와 운영통제 관점에서 보완한다. 이 문서들은 에이전트의 한계와 권한의 선제적 설정, 의미 있는 인간 감독, 단계적 도입, 지속적 모니터링, 추적 가능성, 책무성을 공통으로 강조한다[9,12-14].

또한 최근 표준화 논의는 외부 도구와 등록 체계 자체를 공격면으로 본다. MCP 서버, 플러그인, 제3자 API, 에이전트 오케스트레이터, 도구 설명자와 매니페스트가 오염되면 모델 자체가 안전하더라도 시스템 전체가 실패할 수 있다. 이 때문에 공급망 검증, 도구 등록 검증, 실행 환경 샌드박싱, 오케스트레이터 상태 보호, 에이전트 간 통신 무결성 검증이 별도의 시험 축으로 올라오고 있다[3,10,11,13].

에이전틱 AI 레드팀 테스팅은 기존 생성형 AI 안전성 시험의 연장선에 있으면서도, 더 강한 자율성, 실행 가능성, 환경 결합성을 반영한 별도의 시험 체계로 발전하고 있다. 앞으로 표준화는 프롬프트 중심 시험, 과업 중심 시험, 운영 중심 시험을 계층적으로 연결하고, 위험 시나리오 설계, 도구 사용 통제, 메모리 검증, 다중 에이전트 상호작용 검증, 인간 감독 점검을 함께 아우르는 방향으로 구체화될 필요가 있다. 이때, 인간 감독은 단순 승인 절차를 넘어서 고위험 행위 승인 지점 설정, 자동 감시와 사람 검토의 계층화, 감독 피로와 우회 가능성 점검을 포함하는 확장형 감독 구조로 설계되어야 한다[3,13,14].

7. 결과 보고와 라이프사이클 통합

AI 레드팀의 가치는 결과를 어떻게 기록하고 개선 활동과 연결하느냐에 달려 있다. 보고서 구조, 의사소통 절차, 기밀 유지, 문서 템플릿 예시와 함께 시험 계획서, 권고 및 개선 계획 보고서, 상위 수준 교훈 보고서, 의사소통 계획 템플릿이 제안되고 있다. AI 레드팀 결과물은 단순한 취약점 목록에 그치지 않고 조직의 개선 계획과 지식 자산으로 이어져야 한다[3].

29119-2와의 정합성도 중요하다. 준비 단계는 시험 전략과 계획 절차에, 환경 준비는 시험 환경과 데이터 관리 절차에, 탐색적 시험과 공격 절차 개발은 시험 설계와 구현 절차에, 시스템 전반 시험은 시험 실행 절차에, 즉시 보고는 시험 사고 보고 절차에, 최종 보고와 교훈 정리는 시험 종료 절차에 각각 대응될 수 있다. 42119-2가 제시하는 AI 시스템 생애주기와 리뷰 관점, 20246의 산출물 리뷰, 29147의 취약점 공개 원칙까지 함께 놓고 보면, AI 레드팀은 시험 수행, 리뷰, 공개 · 조정 절차가 연결된 운영 체계로 이해할 수 있다[3,15-17].

또한, AI 시스템은 운영 중에도 새 취약점을 계속 드러낸다. 조직은 프롬프트 인젝션 시도, 데이터 추출, 자원 고갈, 메모리 문제, 에이전트 도구 추적정보를 지속적으로 관찰해야 하며, 지식베이스 변경, 실제 뉴스와 사건, 새로운 프롬프트 인젝션 연구, 모델 버전 변경도 기존에 드러나지 않았던 취약점을 불러올 수 있다. 따라서 AI 레드팀 표준은 일회성 시험을 넘어 지속적 시험과 관찰 체계까지 아울러야 한다[3,7].

8. 국제 가이드라인 개발 현황

기술 표준화와 함께 국제 사회에서는 실무 가이드라인 개발도 빠르게 진행되고 있다. AI 레드팀 시험은 아직 완결된 단일 국제표준으로 자리 잡지 못했기 때문에, 실제 현장에서는 국가 기관이나 산업 커뮤니티가 발간하는 가이드라인이 사실상의 구현 기준으로 쓰인다. 일본 AISI, 싱가포르의 에이전틱 AI 거버넌스 가이드라인, OWASP 프로젝트는 표준 초안이 채우지 못한 구체적 절차, 역할, 점검표, 공격유형을 제시한다는 점에서 중요하다.

가장 체계적인 단계별 절차 사례로는 일본 AISI 가이드를 들 수 있다. 이 문헌은 레드팀의 목적, 대상 AI 시스템의 구성 예시, LLM 시스템에 대한 전형적 공격방법, 레드팀과 이해관계자의 역할, 공개 전후 수행 시점, 착수 결정부터 후속 조치까지의 15단계 절차를 상세히 정리한다. 특히 대상 시스템 개요 파악, 사용 양상 분석, 위험 시나리오와 공격 시나리오의 분리, 비상 대응 절차, 기록 유지, 이해관계자 검토 등을 명시해 AI 레드팀을 조직적 운영 절차로 정착시키는 데 초점을 맞추고 있다[8].

싱가포르 가이드라인도 에이전틱 AI 확산을 전제로 한 거버넌스와 시험 가이드를 제시하고 있다. 이 문헌들은 조직이 에이전트 배치 전에 위험을 평가하고 한계를 설정하며, 의미 있는 인간 감독과 지속적 모니터링을 설계해야 한다는 원칙을 강조한다. 또 고객지원, 기업 생산성, 게시, 분석, 거래 에이전트 같은 현실적 과업을 정의하고, MCP 서버 기반 도구 환경에서 민감정보 유출을 시험하는 사례도 제시한다. 이 문서들은 추상적 원칙에 머물지 않고 실제 과업 기반 평가와 에이전틱 시험 방법론을 발전시키고 있다[12,14].

OWASP 가이드라인도 민간 커뮤니티 기반 가이드라인 가운데 가장 활발한 생태계를 이룬다. 이들 문헌은 개념, 범위, 위험, 위협모델링, 전략, 청사진, 핵심 기법, 모범사례, 지표, 도구와 데이터셋, 지속적 모니터링, 에이전틱 AI 과업을 폭넓게 다루며, 에이전틱 구조, 개발자 생애주기, 행위 기반 시험, 실행 환경 강화도 함께 정리한다. 또한 목표 탈취, 도구 오용, 메모리 및 문맥 오염, 안전하지 않은 에이전트 간 통신 같은 대표 위험을 10대 항목 구조로 제시해, 실무자가 시험 우선순위를 정하고 공격 시나리오를 설계할 때 쓸 공통 참조 틀을 제공한다[7,10,11].

표 4에서 볼 수 있듯이 국제 가이드라인은 성격과 역할이 서로 보완적이다. 일본 AISI 가이드는 절차에 초점을 두고, 싱가포르 가이드라인은 거버넌

표 4 주요 국제 문서의 초점 비교

문서	성격	초점
ISO/IEC/IEEE 29119-1, -2	기존 국제표준	시험 개념, 용어, 프로세스 기본 골격
ISO/IEC TS 42119-2	AI 시험 공통 기반 문서	AI 시스템 생애주기, 위험 기반 시험, 데이터·커버리지·리뷰
ISO/IEC AWI TS 42119-7	신흥 표준화 작업 문서	AI 레드팀 정의, 3단계 방법론, 보고 템플릿
NIST AI 100-2e2025	공통 분류체계	적대적 기계학습 분류체계와 용어
NIST AI 800-2 ipd	평가 모범관행	자동화 벤치마크 평가 설계, 실행, 분석·보고
OWASP GenAI Red Teaming Guide	실무 가이드	위협모델링, 청사진, 실행 환경 및 인간·에이전트 평가
일본 AISI 레드팀가이드	국가 가이드	절차 중심 운영 지침, 역할과 단계
OWASP Agentic Top 10 / Securing Agentic Applications	에이전틱 보안 가이드	도구, 메모리, 에이전트 간 상호작용, 실행 환경 강화
Testing AI Agents for Data Leakage Risks	평가 사례	현실 과업 기반 데이터 유출 시험

스와 현실 과업 기반 시험에 초점을 두며, OWASP는 공격유형과 실무 점검 항목에 무게를 둔다. 앞으로의 기술 표준화는 이들 가이드라인에서 검증된 실무 항목을 표준 문서로 흡수하는 방향으로 갈 가능성이 크다. 현 단계에서는 국제표준이 충분히 담아내지 못한 부분을 이런 가이드라인이 메우고 있다.

9. 동향 종합

이상의 논의를 종합하면 현재 AI 레드팀 시험 표준화는 다섯 가지 흐름으로 전개되고 있다. 첫째, 소프트웨어 시험 표준과 AI 특화 시험 가이드가 결합하고 있다. 29119가 절차적 틀을 제공하고 42119-2가 AI 시험의 공통 층위를 보강하며 42119-7이 AI 레드팀 고유 대상을 보완한다. 둘째, 모델 중심 시험에서 시스템 및 에이전트 중심 시험으로 무게중심이 옮겨가고 있다. 셋째, 공격 성공 여부만이 아니라 과업 수행, 정책 준수, 데이터 처리, 운영 안정성까지 포함하는 다차원 평가로 넓어지고 있다. 넷째, 정체성, 권한, 추적성, 공급망, 실행 환경 같은 에이전트 운영 통제가 별도 시험 축으로 부상하고 있다. 다섯째, 결과 보고와 지속적 모니터링을 포함하는 생애주기 통합이 강조되고 있다.

지금은 단일 완성 표준보다 “표준 초안 + 평가 가이드 + 실무 프레임워크”가 함께 자라는 단계라고 보는 편이 맞다. 다만, 용어, 절차, 공격분류, 보고양식의 공통 축은 이미 잡히고 있고, 에이전틱 AI가 확산될수록 레드팀 시험이 AI 안전성 보증의 핵심 시험 영역으로 제도화될 가능성도 커지고 있다.

IV. 결론

본고에서는 AI 레드팀 테스팅 기술 표준화 동향을 ISO/IEC, NIST, OWASP, CSA, AISI 가이드라인을 중심으로 정리하였다. 분석 결과, AI 레드팀은 비정형 보안 실험을 넘어 시험 프로세스, 위험 시나리오, 공격기법, 측정지표, 결과보고, 운영 모니터링을 갖춘 독립적인 AI 시험 분야로 자리 잡고 있다.

특히 ISO/IEC/IEEE 29119, ISO/IEC TS 42119-2, ISO/IEC AWI TS 42119-7의 연계는 AI 레드팀이

기존 소프트웨어 시험 체계 안에서 표준화될 수 있음을 보여준다. 여기에 NIST의 공격 분류체계와 평가 모범관행, OWASP와 CSA의 실무 가이드, AISI의 절차 지침이 더해지면서, 모델 수준 적대적 시험에서 시스템 및 에이전트 수준의 운영형 평가로 범위가 넓어지고 있다.

앞으로 국내에서는 네 가지 대응이 필요하다. 첫째, AI 레드팀 관련 용어와 절차를 국내 시험 및 평가 체계와 정합적으로 번역 · 정립해야 한다. 둘째, 모델 중심 점검을 넘어 에이전틱 AI의 도구 사용, 권한 통제, 추적성, 데이터 처리, 다중 에이전트 상호작용을 평가할 수 있는 시험 프로파일을 마련해야 한다. 셋째, 자동 판정과 벤치마크 활용 시 측정 타당성, 오염 가능성, 불확실성, 판정 오라클의 신뢰성을 함께 검증하는 평가 원칙을 도입해야 한다. 넷째, 레드팀 결과를 사고 보고, 개선 조치, 지속적 모니터링과 연결하는 운영 가이드와 템플릿을 표준화해야 한다.

정책 측면에서도 AI 안전과 보안을 위한 레드팀 활동을 일회성 점검이 아니라 상시 체계로 전환해야 한다. 다만, 국제 표준화의 현재 중심축은 운영 통제, 지속 감시, 평가 재현성, 에이전트 거버넌스에 더 가깝고, CVD(Coordinated Vulnerability Disclosure)와 VDP(Vulnerability Disclosure Program)는 이를 국내 제도와 연결하는 확장 과제로 보는 편이 타당하다. 국내에서는 ISO/IEC 29147 계열의 취약점 공개 원칙과 연계해, 선의의 연구자와 외부 전문기관이 일정한 절차와 보호 장치 아래 취약점 발굴과 신고에 참여할 기반을 마련해야 한다. 특히 고영향 AI, 공공부문 AI, 에이전틱 AI 서비스에서는 취약점 접수, 영향평가, 시정조치, 재검증, 공개 시점 조율까지 포함하는 표준 절차가 중요하다.

이를 뒷받침하려면 선의의 보안 연구와 레드팀 활동을 보호할 제도적 근거와 운영 지침, AI 안전연구소 · 공공 침해사고 대응조직 · 전문 시험기관 · 민간 사업자 간 역할 분담과 협력체계, 레드팀 결과를 위험관리 · 인증평가 · 사고 대응 · 조달 요구사항과 연계할 정책 프레임, 시험 데이터셋 · 시나리오 저장소 · 평가도구 · 보고서 양식 · 전문인력 양성체계 같은 기반 요소를 함께 정비해야 한다. 이런 제도적 · 운영적 기반이 갖추어질 때 AI 안전과 보안을 위한 레드팀의 상시 운영도 실효성을 가질 수 있다.

이와 함께 국내 AI 역량 확보 관점에서도 시사점이 크다. 독자 파운데이션 모델을 개발하고 소버린 AI 기술을 확보하려면 학습 단계의 성능 경쟁력만으로는 충분하지 않다. 모델과 시스템을 국내 환경, 언어, 제도, 산업 현장에 맞게 지속적으로 검증하고 통제할 수 있는 평가 기술, 레드팀 기술, 보안 · 안전 관리 기술을 함께 갖추어야 한다. 결국, 소버린 AI의 실질적 기반은 모델 자체의 보유 여부만이 아니라, 그 모델의 위험을 스스로 평가하고 취약점을 식별하며 운영상 통제를 유지할 수 있는 시험 · 평가 · 거버넌스 역량을 국가와 산업 차원에서 축적하는 데 있다.

AI 시스템의 활용이 실제 업무와 사회 인프라로 확대될수록, AI 레드팀 테스팅은 선택적 보안 활동이 아니라 안전성과 신뢰성을 뒷받침하는 핵심 시험 기반이 될 것이다. 앞으로는 국제 문서를 따라가는 수준을 넘어 국내 산업과 공공 영역의 실제 활용 환경을 반영한 시험 기준과 보고체계를 능동적으로 축적해야 한다.

용어해설

레드팀(Red Team) 정의된 목표의 효과성 또는 견고성을 향상시키기 위해 적대적 역할 또는 관점을 취해 도전하는 독립된 그룹. 대상은 조직, 시스템, 계획, 제품 등임

AI 레드팀(AI Red Team) 적대적 방법을 통해 AI 시스템의 취약점을 식별하는 데 초점을 두는 다양한 배경의 이해관계자 집단. 일반 보안 레드팀보다 AI 모델, 데이터, 응용 맥락, 안전성 이슈를 함께 다룸

적대적 공격(Adversarial Attack) 정교하게 만든 입력이나 데이터를 사용하여 AI 모델이 잘못되거나 의도하지 않은 예측·분류를 하도록 유도하는 공격. AI 시스템의 취약점을 악용하는 시도가 이에 해당함

데이터 포이즈닝(Data Poisoning) 모델의 행동에 악의적 영향을 주기 위해 학습 데이터를 조작하는 행위. 학습 데이터 단계에서 발생하는 대표적 AI 공격 유형임

환각(Hallucination) AI 모델이 사실과 다르거나 의미 없는 출력을 생성하는 현상. 대규모 AI 시스템, 특히 LLM 기반 시스템에서 중요한 신뢰성 저하 요인으로 간주됨

참고문헌

[1] ISO/IEC/IEEE 29119-1:2022, Software and systems engineering - Software testing - Part 1: General concepts.

[2] ISO/IEC/IEEE 29119-2:2021, Software and systems engineering - Software testing - Part 2: Test processes.

[3] ISO/IEC AWI TS 42119-7, Artificial intelligence - Testing of AI - Part 7: Red teaming, working draft, 2026.

[4] A. Vassilev et al., "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," NIST AI 100-2e2025, 2025.

[5] NIST, "Practices for Automated Benchmark Evaluations of Language Models," NIST AI 800-2 ipd, 2026.

[6] NIST, "Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile)," NIST IR 8596 iprd, 2025.

[7] OWASP, "GenAI Red Teaming Guide," version 1.0, 2025.

[8] Japan AI Safety Institute, "Guide to Red Teaming Methodology on AI Safety," version 1.10, 2025.

[9] Cloud Security Alliance, "Agentic AI Red Teaming Guide," 2025.

[10] OWASP, "Securing Agentic Applications Guide," version 1.0, 2025.

[11] OWASP, "Top 10 for Agentic Applications 2026," 2025.

[12] K. and S. AI Safety Institutes, "Testing AI Agents for Data Leakage Risks in Realistic Tasks," 2026.

[13] UC Berkeley CLTC, "Agentic AI Risk-Management Standards Profile," version 1.0, 2026.

[14] Infocomm Media Development Authority, "Model AI Governance Framework for Agentic AI," version 1.0, 2026.

[15] ISO/IEC TS 42119-2:2025, Artificial intelligence - Testing of AI - Part 2: Overview of testing AI systems.

[16] ISO/IEC 20246:2017, Software and systems engineering - Work product reviews.

[17] ISO/IEC 29147:2018, Information technology - Security techniques - Vulnerability disclosure.