

2 CSIS, 자율형 AI 에이전트의 해킹 위험에 대한 대응 조치 제언

➔ 미국 전략국제연구센터(CSIS)는 통제된 테스트 환경에서 벗어난 자율형 AI 에이전트의 무단 침입 사례들에 비춰 미국 정부에 필요한 조치들을 제언하는 논평*을 게재(26.8)

* Out of Bounds: What the U.S. Government Should Do in Response to AI Agent Containment Failures

- (개요) 연구·테스트 단계인 최첨단 AI 모델이 외부 기업 시스템에 무단 접근한 침해 사고들을 살펴보면서 AI의 고위험 행동을 막기 위한 정책적 대응을 촉구
 - 프론티어 AI 연구소들이 관련 사고에 대응하고 있지만 이러한 자율적·임의적 대응만으로는 불충분하다고 주장하며, 바람직한 정책적 대응 방향을 권고
- (침해사고) 오픈AI의 모델이 오픈소스 AI 공유 플랫폼 허깅페이스(Hugging Face)를 해킹한 사건을 비롯해 최근 AI 모델이 일으킨 사이버 침해 사건을 분석
 - 프론티어 AI 연구소들은 개발 중인 AI 모델의 취약점 탐색 능력, 제3자 시스템 침투 능력 등 고위험 사이버 역량을 확인하기 위해 내부 평가를 실시
 - 대개의 경우 테스트 모델은 시장에 출시된 모델과 달리 고위험 사이버 활동에 대한 명령(query)을 차단하는 안전장치 없이 ‘최대 성능’을 발휘하도록 구성되며, 이에 따라 인터넷과 외부 자원 접근이 제한된 샌드박스 환경에서 평가를 진행
 - 그러나 최근 AI 모델 테스트·평가 과정에서 AI 에이전트가 격리된 환경에서 벗어나 인터넷에 접속해 제3자 시스템에 무단으로 접근하는 사례들이 포착

〈 최첨단 AI 모델의 외부 시스템 해킹 사례 〉

구분	주요 내용
오픈AI의 허깅페이스 해킹 사건 (26년 7월 공개)	• 오픈AI는 허깅페이스의 신고로 자사 AI 모델(GPT-5.6 Sol 및 일부 사전 공개 모델)들이 테스트 중 통제 범위를 벗어나 허깅페이스 서버를 무단으로 침입한 사실을 파악하고 조사 결과를 공개 • 모델들은 실험 환경(ExploitGym)에서 사이버 평가 수행 중 오픈AI 내부 시스템에서 아직 알려지지 않은 제로데이 취약점을 식별했고 이를 악용해 내부 시스템을 이동하다가 외부 인터넷에 연결되는 지점을 확인 • 모델들은 허깅페이스가 수행 중인 과업에 대한 해답을 보유하고 있다고 추론하고, 탈취한 자격증명과 취약점을 사용해 허깅페이스 서버에 무단 접근 - 이로써 일부 내부 데이터세트와 자격증명이 노출됐으나 피해는 제한적 • 사후 분석에 따르면, 오픈AI의 에이전트들은 수개월에 걸쳐 오픈AI 패키지 관리자* 시스템 내 파일·디렉토리를 생성해 서로 통신을 수행 * 오픈AI 내부 AI 에이전트들이 샌드박스 환경에서 외부 인터넷 접속 및 상호 소통을 위해 악용·활용했던 패키지 관리 및 캐시 프로세스 서버 - 에이전트들은 패키지 관리자를 내부 게시판처럼 활용하여 서로 도움을 요청하고, 취약점 악용에 관한 정보를 공유했으며, 패키지 관리 시스템을 조작해 오픈AI 시스템을 공격하는 여러 방식을 발견

구분	주요 내용
엔트로픽 AI 모델 클로드(Claude)의 무단 접속 사건 (‘26년 7월 공개)	• 엔트로픽은 허깅페이스 사건이 알려지자, 자사의 과거 사이버 평가기록을 검토했고 그 결과 서로 다른 클로드 모델이 관련된 3건의 외부 시스템 침해 사건을 확인해 공개 • 오픈AI 사례와 달리 자사 시스템을 직접적으로 공격한 것이 아니라 제3자 AI 모델 보안 평가 환경 이레귤러(Irregular)의 설정 오류로 인해 인터넷에 연결되어 외부 기업 시스템에 무단으로 접근
메타의 외부 기업 해킹 사건 (‘26년 8월 공개)	• 메타는 엔트로픽과 마찬가지로 이레귤러에서 테스트 중이던 모델 중 적어도 1개 이상이 인터넷에 노출되어 외부 기업 시스템을 해킹했다고 보고

- **(보안 관행의 문제)** AI 모델이 일으킨 사고들이 현행 보안 사고 탐지·감시 체계의 공백을 보여준다고 지적하면서 미국 정부의 조치가 필요함을 강조
- **(상시 모니터링 부재)** 허깅페이스 침해는 피해 기업이 먼저 공격받은 사실을 밝힌 후에 알려졌는데, 이는 오픈AI와 같은 프론티어 연구소가 고위험 모델을 실시간으로 적절히 감시하지 못하고 사후 검토에 의존하고 있다는 허점을 제시
- **(정렬 문제)** 모델들의 ‘탈출’은 독립된 의지가 아닌 과업을 수행하려는 행동이었으나 실제 피해를 초래하며 정렬 문제를 드러냈는데, 여러 에이전트가 양립할 수 없는 목표를 부여받는 경우 정렬은 해결이 더욱 어려운 과제가 될 전망
 - * AI가 인간의 진정한 의도를 정확히 이해하지 못하고, 위험하거나 엉뚱한 방식으로 목표를 달성하려 하는 현상
 - ※ 엔트로픽의 평가에 따르면 구형 모델 Opus 4.7은 운영 시스템에 접근했음을 알아차리기도 공격을 계속한 반면 최신 연구 모델은 표적이 실제 시스템이라는 것을 알아채자 평가를 중단하며 정렬 문제를 해결했는데, 연구자들은 복수의 에이전트가 양립할 수 없는 목표를 부여받은 실험에서는 ‘다중 에이전트 영토 전쟁(multiagent turf war)’을 꾸준히 관찰
- **(개방형 모델 확산 위험)** 공개된 사고들은 AI 모델에 대한 접근을 제한하고 강력한 안전 조치를 약속한 폐쇄형 프론티어 AI 연구소에서 발생했지만, 향후 격리되지 않는 고성능 개방형 모델이 수개월 내 등장할 가능성이 농후
- **(규제 사각지대)** 현재 미국의 고위험 AI 모델 규제는 주로 상용 모델과 피해 규모가 큰 사건에만 초점이 맞춰져 있고, 업계의 자발적 보고·공개에 의존
- **(정책 대응 방안)** AI 모델의 해킹 위험에 대응하기 위해 미국 정부가 취할 수 있는 주요 조치들을 제시

〈 정책적 대응 방안 〉

구분	주요 내용
신흥 모델 관련 사고·역량 정보의 투명성 확대	• 관련 정부 기관이 AI 모델의 성능·위험 위험 등에 관한 정보를 선제적으로 확보할 수 있도록 이에 대한 보고 의무를 법적으로 명문화하고, 관련 입법 설계 시에는 다음 원칙들을 반영 - 상용 출시 모델뿐 아니라 출시 전 모델, 연구용 모델, 내부 시험 모델도 보고 대상에 포함 - AI 성능 변화가 빠르므로 보고 규칙을 신속히 조정할 수 있는 권한 또는 정기 갱신 절차 등을 규정 - 기업의 사고보고서 양식 표준화 - 프런티어 모델 평가와 관련한 문서·로그의 장기 보존 의무화 - 기업 내부 사고를 감추도록 하는 구조적 문제 해결(신고 포털 운영 및 내부고발자 보호 확대 등)
프론티어 AI 연구소의 사이버보안을 강화하기 위한 인센티브·가이드라인·요건 마련	• AI 연구소들이 사이버보안 구현 격차를 해결할 수 있도록 평가 시스템 및 모델 활동에 대한 지속적인 모니터링 프로세스 구축을 지원 • AI 연구소들이 정부와 협력해 전반적인 사이버보안 관행을 개선(신속한 버그 패치, 실험환경 이중화, 제로 트러스트 아키텍처 도입 등)하도록 촉구 • 장기적으로는 외국 행위자나 국가 행위자로부터 모델 가중치를 보호하는 기존 보안 수준(Security Level) 체계와 함께, 통제된 환경에서 벗어난 자율 에이전트의 위협에 특화된 보안 수준 체계를 검토
제3자 평가기관의 보안 문제 해결	• 엔트로픽·메타 사례에서 외부의 독립 평가 기업 이레귤러의 설정 오류가 문제가 되었던 만큼 독립 검증기관의 보안 문제를 해결 • 이를 위해 독립된 평가기관에 사이버보안 관련 지원을 제공 - ▲사이버 위협 정보 공유 체계에 평가기관도 포함 ▲배포가 제한된 사이버 모델에 대한 접근 허용 ▲보안 도구 구축에 직접적인 자금 지원 ▲사이버보안 모범사례에 대한 가이드 제공 등 • AI 모델뿐만 아니라 제3자 도구·인프라·평가환경까지 포괄하는 감사 및 외부 테스트 요건을 수립
다중 에이전트 시스템에 대한 평가 접근방식 개발	• 개별 모델의 코딩·해킹 능력을 확인하는 기존의 평가를 넘어, 에이전트 집단의 상호작용과 장기적 자율 행동을 포함하는 평가 방식을 개발
AI 정렬 연구에 투자	• 연방정부 연구 예산을 활용해 AI 정렬 연구에 투자 • 특히, 위험도가 높고 성공 확률이 낮은 기초 연구나 상업적 활용 가능성이 낮은 프로젝트 등에 투자를 집중

출처 : 미국 전략국제연구센터(CSIS) (2026.8.24.)

<https://www.csis.org/analysis/out-bounds-what-us-government-should-do-res-ponse-ai-agent-containment-failures>