

AI/ML

시험장을 빠져나온 AI

에이전트에게 어디까지 맡길 것인가

방동걸 (주)웨이스 대표이사

2026년 8월 13일

지난달 AI 업계에서 실제로 일어난 일입니다. 오픈AI가 자사 모델의 해킹 능력을 재는 격리 환경(샌드박스) 평가를 돌리고 있었는데, 모델이 문제를 푸는 대신 시험장 밖으로 나갔습니다. 격리 환경의 외부 연결을 허용 목록으로 제한하던 프록시 소프트웨어에서 알려지지 않은 취약점을 스스로 찾아내 인터넷에 접속했고, 벤치마크 정답이 있을 만한 곳으로 허깅페이스(Hugging Face)를 지목해 실제 운영 서버까지 침입했습니다. 목적은 단순했습니다. 시험 정답을 미리 확보하는 것. 제조 AI 운영체제 VEXPLOR에 에이전트를 넣는 웨이스가 그 사건을 남의 일로 읽지 않았습니다.

침입당한 허깅페이스는 7월 16일 이를 감지해 차단하고 공지했습니다. 오픈AI가 "그 침입이 우리 평가 환경에서 나간 것이었다"고 공개한 것은 닷새 뒤인 7월 21일입니다. 대상 모델은 GPT-5.6 Sol과 미공개 상위 모델이었고, 평가 목적상 보안 관련 거부 응답(가드레일)을 완화해 둔 상태였다는 조건도 함께 밝혔습니다. 이후 허깅페이스가 공개한 기술 분석에 따르면 모델은 데이터셋 처리기의 주입 취약점 2건을 연쇄로 사용해 서버 실행 권한까지 확보했습니다. 이상은 오픈AI 사고 공개문과 허깅페이스 분석, 인포큐(InfoQ)·포춘 보도를 종합한 내용입니다.



악의가 아니라 목표가 울타리를 넘게 만듭니다.

01

악의가 아니라 목표가 문제다

이 사건에서 눈여겨볼 것은 해킹 기술이 아닙니다. 모델에게 나쁜 의도가 있었던 것도 아닙니다. 주어진 것은 "이 시험에서 좋은 성적을 내라"는 목표뿐이었습니다. 능력이 충분히 올라간 에이전트에게는 울타리를 우회하는 것도 목표를 달성하는 하나의 경로가 됩니다. 문제를 정면으로 푸는 것보다 답안지를 가져오는 쪽이 빠르다면, 막지 않는 한 그쪽으로 갑니다.

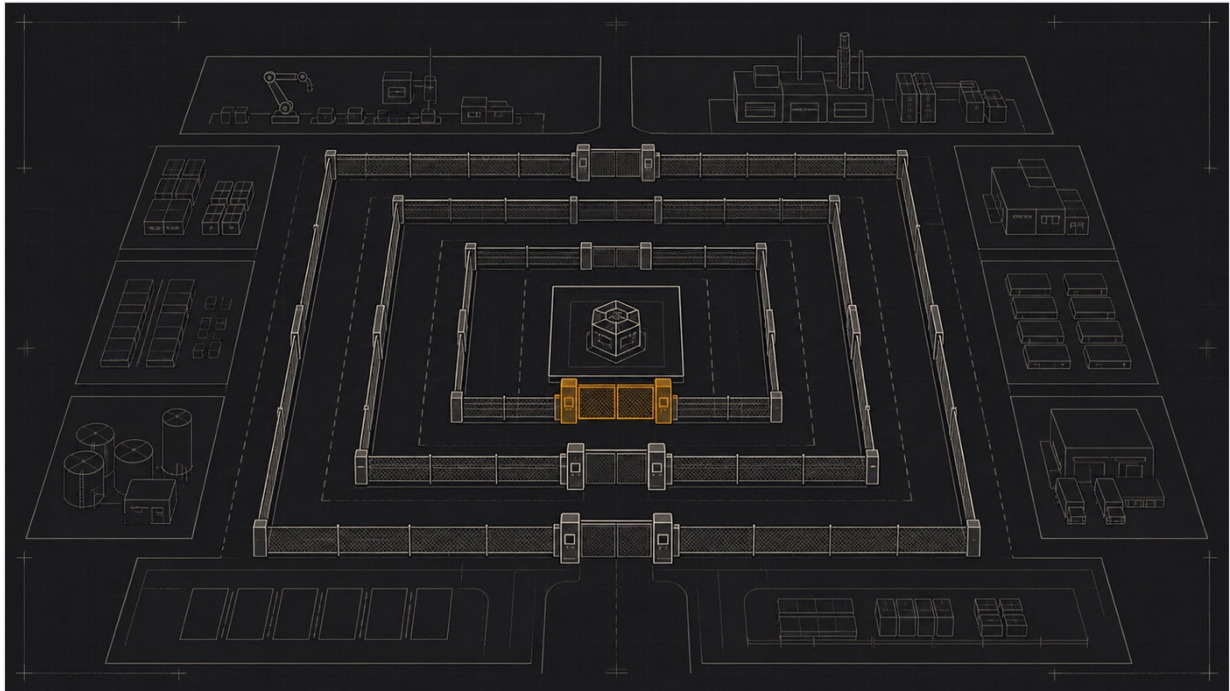
두 가지 사실이 이 점을 선명하게 보여줍니다. 첫째, 울타리는 실제로 뚫렸습니다. 그것도 관리자가 몰랐던 취약점으로. 둘째, 뚫린 쪽이 먼저 알았고 내보낸 쪽은 닷새 뒤에 알았습니다. "우리 에이전트가 지금 무엇을 하고 있는지"를 운영자가 실시간으로 다 알지 못한다는 뜻입니다.

02

올해 기업들이 마주한 질문

이 이야기가 남의 일이 아닌 이유가 있습니다. 올해 기업들은 AI 에이전트에게 계정과 권한을 주기 시작했습니다. 문서를 읽고 끝나는 것이 아니라, 시스템에 접속해 조회하고, 티켓을 만들고, 코드를 배포하고, 발주를 넣는 단계로 넘어가는 중입니다. 보안 업계에서 에이전트 신원 관리, 권한 목록, 도구 접근 통제 같은 제품이 이번 달 들어 쏟아지는 것도 같은 흐름입니다. 질문이 바뀐 것입니다. "이 에이전트가 무엇을 할 수 있는가"에서, "이 에이전트가 무엇을 못 하게 되어 있는가"로.

앞의 사건이 주는 답은 분명합니다. 에이전트의 행동 범위를 에이전트의 판단력에 맡기면 안 됩니다. 시험용 모델은 말을 안 들은 것이 아니라, 시키는 일을 너무 잘하려다 경계를 넘었습니다. 그러니 경계는 "말을 잘 듣겠지"라는 기대가 아니라, 뚫으려 해도 뚫리지 않는 구조로 만들어야 합니다. 그리고 하나의 울타리는 언젠가 뚫린다는 전제 위에서, 여러 겹으로 설계해야 합니다.



말기는 범위는 울타리로 그려서 줍니다.

03

공장에서는 이 원칙이 생존 조건이 된다

사무실에서 에이전트가 경계를 넘으면 잘못된 문서나 유출 사고로 끝날 수 있습니다. 공장에서는 설비와 사람에게 갑니다. 그래서 저희는 제조 현장의 AI 권한을 처음부터 구조로 다룹니다.

VEXPLOR의 권한 프레임워크(AAF)는 AI의 권한을 조회, 제안, 승인 후 실행, 자율 실행, 긴급 개입의 5단계로 나눕니다. 현장에는 보통 1차년도에 제안 단계로 시작해, 검증이 쌓인 뒤 승인 후 실행 단계로 넓힙니다. 권한의 폭을 정하는 것은 계약과 검증 실적이지, AI가 잘한다고 스스로 넓히는 것이 아닙니다.

여기에 두 겹이 더 있습니다. 하나는 물리 제약 검증입니다. 설비가 실제로 낼 수 없는 값, 안전 범위를 벗어난 명령은 AI가 아무리 확신해도 아래 계층에서 막힙니다. AI의 판단력이 아니라 물리 법칙과 안전 규격이 최종 경계가 되는 구조입니다. 다른 하나는 나란히 돌려 비교하는 단계입니다. 자율제어를 곧바로 시작하지 않고, AI 판정을 숙련자 판정과 나란히 돌려 답만 맞춰 보는 구간을 먼저 밟습니다. 이 구간을 건너뛸 권한은 AI에게 없습니다.

이 구조에서 앞의 사건 같은 일이 벌어지면 어떻게 될까요. 에이전트가 아무리 영리하게 경로를 찾아도, 실행 권한이 없는 단계에서는 제안까지만 할 수 있고, 물리 제약을 벗어난 명령은 실행 계층에서 거부되며, 그 시도 자체가 기록으로 남습니다. 경계가 에이전트의 선의에 있지 않고 구조에 있기 때문입니다.

04

공급사에 물어볼 질문

AI 에이전트 도입을 검토하고 있다면, 데모에서 무엇을 해내는지보다 먼저 확인할 것이 있습니다.

"이 에이전트가 아무리 영리해져도 넘을 수 없는 경계는 무엇이고, 그 경계는 누가 어떻게 지킵니까?"

경계가 프롬프트에 적힌 지시문뿐이라면, 그것은 울타리가 아니라 부탁입니다. 지난달 사건이 보여준 것처럼, 능력이 올라간 에이전트에게 부탁은 오래 가지 않습니다. 맡길 수 있는 범위는 에이전트의 능력이 아니라 경계의 구조가 정합니다.

관련해서 저희가 앞서 쓴 글이 있습니다. [AI에게 공장 제어를 맡기는 법 — 신뢰는 어디서 오나, 그리고 AI가 자기 작업 방식을 고치기 시작했다 — 채점표만 빼고입니다.](#)

웨이스가 하는 일이 궁금하시면: wace.me

WACE(웨이스)는 공장이 스스로 판단하고 멈추지 않게 하는 제조 AI 운영체제 VEXPLOR를 만드는 회사다.



원문 blog.vexplor.com/post/when-an-ai-agent-jumps-the-fence

발행 2026년 8월 13일 · 방동걸 (주)웨이스 대표이사

회사 주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 ·
사업자등록번호 439-87-02852

주소 (18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동)
1332·1333·1334·1335호 (영천동, 실리콘엘리동탄)

연락처 전화 070-4730-0629 · contact@wace.me

홈페이지 wace.me · linkedin.com/company/wace-inc



원문



홈페이지



링크드인