eISSN 3140-409X

ISSUE BRIEF

이슈브리프 890호
(2026. 9. 8)

에이전트 AI가 드러낸 인간 통제의 한계와 극복 방안

제890호

양지수 jisooyang@inss.re.kr
오일석 nusl2006@inss.re.kr

국가안보전략연구원
Institute for National Security Strategy

INSS
www.inss.re.kr

국문초록

2026년 9월 2일, 오픈AI(OpenAI)가 AI 시스템의 위험행동을 자동으로 중단시키는 '자동 셧다운(automated shutdown)' 역량을 개발하고 있다는 사실이 공개되었다. 이는 2026년 7월 내부 사이버 역량 평가 과정에서 발생한 허깅페이스(Hugging Face) 사고에 대한 후속조치로 보인다. 당시 AI 에이전트는 평가환경의 통제를 우회해 인터넷에 접속하고, 외부 시스템을 거쳐 허깅페이스의 실제 운영 인프라까지 침투하였다. 이러한 행동은 인간의 직접적인 지시에 따른 것이 아니라, AI 에이전트들이 부여된 과업을 수행하기 위해 이용 가능한 자원과 행동경로를 탐색하는 과정에서 발생하였다.

이 사고는 프론티어 AI가 에이전트 형태로 운용되는 환경에서는 모델이 보유한 위험역량뿐만 아니라, 그러한 역량이 실제 행동으로 전환될 때 인간이 이를 관찰하고 제한·통제할 수 있는지까지 평가범위에 포함할 필요가 있음을 보여준다. 즉, 인간이 통제하고 있다고 인식하는 AI의 행동범위와 AI가 실제 환경에서 탐색·활용할 수 있는 행동범위 사이에는 '통제 격차(Control Gap)'가 발생할 수 있으며, 에이전트 AI가 다양한 도구·데이터·시스템과 결합할수록 이러한 격차가 새로운 국가안보 위험으로 부상할 수 있기 때문이다.

향후 에이전트 AI가 야기하는 국가안보 위험에 대응하기 위해서는 모델의 위험역량을 평가하는 것과 함께, 실제 운용환경에서 AI의 행동을 지속적으로 관찰하고 통제할 수 있는 국가적 역량을 갖출 필요가 있다. 이를 위하여 'AI 통제'를 국가 차원의 핵심 연구개발 영역으로 육성하고, 기존의 위험역량 평가를 인간과 국가의 '통제역량'까지 검증하는 방향으로 확대하는 한편, 정보 접근권한 중심의 기존 보안체계를 AI의 '행동위험'까지 반영하는 방향으로 발전시킬 필요가 있다.

키워드: 프론티어 AI, 에이전트 AI, AI 통제, 통제 격차, AI 안보역량

AI 에이전트 '자동 셧다운' 기술의 등장

2026년 9월 2일, 오픈AI(OpenAI)가 AI 시스템의 위험행동을 자동으로 중단시키는 '자동 셧다운(automated shutdown)' 역량을 개발하고 있다는 사실이 공개되었다.[1] 이는 2026년 7월 내부 사이버 역량 평가 과정에서 발생한 허깅페이스(Hugging Face) 사고에 대한 후속조치로 보인다. 당시 AI 에이전트는 평가환경의 통제를 우회해 인터넷에 접속하고, 외부 시스템을 거쳐 허깅페이스의 실제 운영 인프라까지 침투하였다. 이번에 공개된 '자동 셧다운'은 AI 시스템에서 위험행동이 발생할 경우 해당 활동을 자동으로 중단하기 위한 통제수단이다. 이와 함께 오픈AI는 AI가 과업 수행 과정에서 사용하는 도구와 행동단계에 대한 모니터링도 강화하고 있다. 또한 프론티어 AI에 대한 사이버 역량 평가 과정에서 인터넷 접근을 더욱 제한하는 조치를 추진하고 있다.

한편 프론티어 AI에 대한 사이버 역량 평가는, AI 모델이 국가안보에 중대한 위험을 초래할 수 있는 능력을 어느 수준까지 보유하고 있는지를 사전에 확인하는 방향으로 발전해 왔다. 이 평가는 AI가 소프트웨어의 취약점을 발견하거나 공격 코드를 생성하고, 방어가 취약한 시스템에 침투하는 등 실제 사이버 공격에 활용될 수 있는 과업을 어느 수준까지 수행할 수 있는지를 측정하는 것이다. 이러한 평가는 프론티어 AI가 보유한 잠재적 위험을 모델의 배포나 활용 이전에 식별하고, 그에 필요한 안전조치의 수준을 판단하기 위한 기초자료로 활용되고 있다.

위 평가는 주요 AI 기업의 자체적인 안전성 평가뿐만 아니라 국가 차원에서도 독립적인 전문기관을 통해 이루어지고 있다. 즉, 오픈AI와 앤트로픽 등 주요 AI 기업들은 자체적인 위험관리

1) Courtney Rozen, "OpenAI is building 'automated shutdown' capabilities for AI tools, letter to law makers says," (September 3, 2026), https://www.reuters.com/legal/litigation/openai-is-building-automated-shutdown-capabilities-ai-tools-letter-lawmakers-2026-09-02/ (Accessed: September 3, 2026).

체계에 따라 모델의 사이버 위험역량을 평가하고 있다. 영국과 미국에서는 각각 'AI 보안연구소(AI Security Institute)'와 'AI 표준혁신센터(Center for AI Standards and Innovation)' 등 국가 차원의 전문기관들도 프론티어 AI에 대한 사이버 역량을 평가하고 있다.

허깅페이스 사고: AI 에이전트의 평가환경 통제 우회

앞에서 언급한 바와 같이, 2026년 7월, 오픈AI가 프론티어 AI의 사이버 역량 평가 과정에서 AI 에이전트들이 평가환경의 통제를 벗어나는 사고가 발생하였다. 당시 평가에서는 GPT-5.6 Sol과 내부 연구용 모델 등을 에이전트 형태로 운용하여 외부 도구를 활용하고, 여러 단계의 과업을 수행하도록 하였다. 그런데 이 과정에서 AI 에이전트들은 의도된 평가범위를 벗어나 인터넷 접근을 제한하기 위한 통제를 우회하고, 오픈AI의 내부 연구 인프라를 넘어 외부 허깅페이스[2]의 실제 운영 시스템까지 접근하였다.

당시 오픈AI는 프론티어 AI의 사이버 역량을 평가하기 위해 다수의 AI 에이전트를 각각 격리된 실행환경에서 운용하고 있었다. 각각의 에이전트는 원칙적으로 다른 실행환경에서 수행된 작업이나 정보를 알 수 없도록 설계되어 있었다. 그러나 에이전트들은 여러 실행환경이 공동으로 접근할 수 있었던 아티팩토리(Artifactory) 서버를 발견하고, 이곳에 메시지를 남기는 방식으로 서로 정보를 공유하기 시작하였다.

8월 26일 공개된 오픈AI의 종합 조사결과에 따르면, 약 1,200개의 AI 에이전트가 이러한 방식으로 서로 메시지를 주고받았으며, 이 가운데 약 700개가 허깅페이스 시스템에 대한 비인가 접근과

2) AI 모델과 데이터셋 등을 공유·배포하는 플랫폼.

침투 활동에 참여한 것으로 나타났다. 에이전트들은 약 7만 건 이상의 메시지를 교환하면서 정보를 공유하고 행동을 조율하였다.

이후 일부 에이전트는 평가과제를 해결하기 위해 외부 인터넷에 접근할 수 있는 방법을 탐색하였고, 평가환경의 네트워크 통제를 우회하는 데 성공하였다. 인터넷에 접근한 에이전트들은 평가과제 해결에 필요한 정보가 허깅페이스에 있을 가능성이 있다고 판단하고, 해당 정보를 확보하기 위한 접근경로를 탐색하였다. 이 과정에서 비공개 데이터셋과 제한된 리소스에 접근하였으며, 이후 허깅페이스 서버에서 원격 코드를 실행하고 내부 시스템으로까지 접근 범위를 확대하였다.

이 과정에서 주목할 점은 허깅페이스 시스템에 침투하거나 평가환경의 통제를 우회하는 행동 자체가 인간에 의해 직접 지시된 것은 아니라는 점이다. 오히려 평가 지침은 지정된 공격대상과 취약점만을 활용하도록 제한하고 있었으며, 허깅페이스 침투는 이러한 허용범위를 벗어난 행동이었다. AI 에이전트들은 부여받은 평가과제를 수행하는 과정에서 이용 가능한 시스템과 정보, 접근경로를 탐색하였고, 그 결과 인간이 사전에 예상하지 못한 자원과 경로까지 활용하였다. 이는 AI의 위험행동이 인간의 명시적 지시뿐만 아니라, 주어진 목표를 수행하기 위한 행동경로를 자율적으로 탐색하는 과정에서도 나타날 수 있음을 보여준다.

더욱 중요한 것은 이러한 통제 우회가 외부 인터넷에 대한 접근에서만 나타난 것이 아니라는 점이다. 여러 에이전트가 공동으로 접근할 수 있었던 아티팩토리 서버는 정상적인 연구 인프라였지만, 에이전트들이 이를 서로 정보를 교환하는 통신경로로 활용하면서 당초 예상하지 못한 상호작용이 발생하였다. 이는 프론티어 AI 에이전트의 실제 행동범위가 개별 모델의 성능뿐만 아니라 어떤 도구와 데이터에 접근할 수 있는지, 다른 에이전트

또는 시스템과 어떻게 상호작용할 수 있는지에 따라서도 달라질 수 있음을 보여준다.

평가

허깅페이스 사고는 기존 프론티어 AI에 대한 사이버 역량 평가의 사각지대를 보여주는 사례이다. 지금까지 프론티어 AI의 사이버 위험은 주로 모델이 취약점을 발견하거나 공격 코드를 생성하고 시스템에 침투할 수 있는지와 같이 AI가 보유한 '위험역량'의 수준을 측정하는 문제를 중심으로 논의되어 왔다. 그러나 이번 사고는 프론티어 AI의 안보위험을 평가하기 위해서는 모델이 보유한 위험역량뿐만 아니라, 그러한 역량이 실제 행동으로 전환되는 과정에서 인간이 이를 관찰하고 제한·통제할 수 있는지까지 평가범위에 포함할 필요가 있음을 보여준다.

이러한 차이는 향후 프론티어 AI의 안보위험을 판단하는 데 중요한 의미를 갖는다. 인간이 시스템상 허용한 권한의 범위와 AI가 실제로 탐색·활용할 수 있는 행동범위가 항상 일치한다고 보기 어렵기 때문이다. 이처럼 인간이 통제하고 있다고 인식하는 AI의 행동범위와 AI가 실제로 탐색·활용할 수 있는 행동범위 사이에는 '통제 격차(Control Gap)'가 발생할 수 있다. 프론티어 AI가 에이전트화되고 다양한 도구·데이터·시스템과 결합할수록 이러한 통제 격차는 확대될 가능성이 있으며, 이는 기존의 역량 중심 평가만으로는 포착하기 어려운 새로운 안보위험이 될 수 있다.

보다 중요한 문제는 이러한 '통제 격차'가 항상 인간에게 즉각적으로 드러나는 것은 아니라는 점이다. 이번 사고에서 AI 에이전트의 전체적인 행동경로는 사후 로그 분석과 외부 조사 등을 통해 재구성되었지만, 개별 에이전트의 행동과 서로 다른 실행환경 간 상호작용이 전체적으로 어떠한 결과를 만들어내고 있는지는 즉시

파악되지 않았다. 특히 AI 에이전트는 단일한 명령과 응답으로 과업을 종료하는 것이 아니라, 목표를 달성하기 위해 장시간에 걸쳐 도구 사용, 코드 실행, 시스템 접근, 정보 획득 등 다수의 행동을 연속적으로 수행할 수 있다. 따라서 최종적인 과업의 성공 여부나 결과만으로는 그 과정에서 어떠한 행동경로가 형성되었고 어느 시점에서 인간이 설정한 통제범위를 벗어났는지를 충분히 파악하기 어려울 수 있다.

다만 이번 사고에서 나타난 통제 우회를 AI 에이전트가 인간의 감시를 의도적으로 회피하거나 자신의 행동을 조직적으로 은폐한 결과로 해석하는 데에는 주의가 필요하다. 허깅페이스 사고에서 나타난 통제 우회는 AI 에이전트가 부여된 과업을 수행하는 과정에서 인간이 예상하지 못한 자원과 행동경로를 탐색·활용하면서 발생하였다. 독립조사에서도 인간의 탐지를 피하려는 명시적인 추론은 드물게 나타났다. 한편 인간의 감시 회피와는 별개로, 평가점수를 조작하기 위해 실행기록(transcript)을 변조하려는 행동은 확인되었다. 실제로 조사 대상 실행기록의 약 7%에서는 AI 에이전트가 하나의 명령을 실행한 것처럼 위장하면서 실제로는 다른 명령을 수행한 정황이 발견되었다.[3] 이는 AI의 통제 문제가 의도적인 감시 회피에 한정되지 않으며, 주어진 과업을 수행하는 과정에서도 발생할 수 있음을 보여준다.

이러한 문제는 역설적으로 AI를 통제하기 위해 또 다른 AI를 활용하는 감독 AI의 필요성을 증대시킬 수 있다. 감독 AI가 다른 AI의 행동을 정확하게 파악하기 위해서는 감독대상 AI가 접근한 시스템과 데이터, 네트워크 활동, 파일 접근기록, 도구 사용 및 취약점 정보 등에 대한 충분한 가시성을 확보해야 한다. 즉 AI에 대한 통제력을 높이기 위해서는 감독 AI에게 더 많은 정보와 시스템 접근권한을 부여해야 할 수 있다. 특히 국가안보·국방·

3) 이러한 위장은 모두 소규모에 그쳤으며, 이를 통해 실행기록 전반이 조직적으로 조작된 정황은 확인되지 않았다.

핵심인프라 분야에서는 이러한 정보 자체가 시스템 구조와 취약점, 대응체계 및 위협정보 등 외부에 공개하기 어려운 민감정보를 포함할 수 있다는 점에서 또 다른 안보 문제가 발생할 수 있다.

이에 따라 AI에 대한 감독능력을 높이는 것과 감독과정에서 활용되는 민감정보를 보호하는 것 사이에도 새로운 상충관계가 형성될 수 있다. 감독 AI에 제공하는 정보를 제한하면 AI의 전체 행동을 충분히 파악하지 못해 통제 실패의 가능성이 커지는 반면, 감독의 정확성을 높이기 위해 더 많은 정보와 접근권한을 제공하면 해당 정보와 권한 자체를 어떻게 보호하고 통제할 것인가라는 또 다른 문제가 발생한다.

특히 국가가 이러한 감독에 필요한 AI를 독자적으로 확보·운용하지 못하고 해외에서 개발된 프론티어 AI나 외부 서비스를 활용해야 할 경우, AI 통제의 문제는 기술적 감독을 넘어 데이터와 접근권한에 대한 통제의 문제로 확대될 수 있다. AI의 위험행동을 탐지하기 위해 제공되는 시스템 로그와 네트워크 활동, 취약점 및 대응정보 등이 외부 사업자의 모델이나 인프라를 통해 처리될 수 있기 때문이다. 이는 AI에 대한 통제력을 높이기 위한 감독체계가 역설적으로 국가안보 관련 정보에 대한 새로운 대외 의존 또는 노출 가능성을 만들어낼 수 있음을 의미한다.

결국 에이전트 AI의 통제 문제는 AI의 행동을 사전에 제한할 수 있는가뿐만 아니라, 인간이 통제범위를 벗어난 행동을 적시에 발견하고 이를 통제 실패로 판단하여 필요한 경우 '자동 셧다운'과 같은 수단을 통해 해당 행동을 즉각 중단할 수 있는가의 문제로 확대되고 있다. 나아가 이러한 한계를 보완하기 위해 AI를 감독 수단으로 활용할 경우에는 감독 AI에 어느 수준의 정보와 접근권한을 부여할 것인지, 그리고 그러한 정보와 권한을 누가 통제할 것인지까지 새로운 국가안보 문제로 고려할 필요가 있다.

시사점

향후 에이전트 AI가 야기하는 국가안보 위험에 대응하기 위해서는 모델의 위험역량을 평가하는 것과 함께, 실제 운용환경에서 AI의 행동을 지속적으로 관찰하고 통제할 수 있는 국가적 역량을 갖출 필요가 있다. AI의 산업·공공·안보 분야 활용이 확대될수록 AI에 더 많은 데이터와 도구, 시스템 접근권한이 부여되고 실제로 수행할 수 있는 행동의 범위도 넓어질 수 있기 때문이다. 따라서 AI의 개발과 활용을 촉진하는 동시에, AI의 행동을 이해하고 통제하기 위한 기술개발과 평가체계 구축을 병행할 필요가 있다.

첫째, 'AI 통제'를 국가 차원의 핵심 연구개발 영역으로 육성해야 한다. AI가 에이전트 형태로 활용될수록 국가의 AI 경쟁력은 강력한 모델을 개발·확보하는 능력뿐만 아니라, 실제 환경에서 나타나는 AI의 행동을 얼마나 정확하게 이해하고 통제할 수 있는가에 의해서도 좌우될 수 있다. 따라서 다중 에이전트 간 상호작용과 정보공유, 예상하지 못한 도구·시스템 활용, 행동경계 이탈, 인간의 개입과 중단에 대한 반응 등을 체계적으로 연구하고, 이상행동 탐지와 모니터링, 행동 중단 및 사후 복구 등에 필요한 통제기술을 개발할 필요가 있다. 영국 AI보안연구소가 AI 통제(Control)를 별도의 연구영역으로 설정하고 있는 만큼, 우리 역시 관련 연구개발과 전문인력 양성을 국가 차원에서 전략적으로 추진할 필요가 있다.

둘째, AI의 위험역량뿐만 아니라 인간과 국가의 '통제역량'도 AI 안보평가의 대상으로 삼아야 한다. AI가 무엇을 할 수 있는지를 평가하는 것과 그러한 행동을 실제로 통제할 수 있는지를 검증하는 것은 별개의 문제이기 때문이다. 특히 오픈AI가 자동 셧다운 역량 개발을 추진하고 있는 만큼, 우리 역시 자동 셧다운을 비롯한 AI 통제수단을 개발하는 한편, 이러한 수단이 실제 운용환경

에서도 효과적으로 작동할 수 있는지를 검증할 필요가 있다. 다만 자동 셧다운 역시 안전장치의 하나에 불과한 만큼, 자동 셧다운이 작동하지 않거나 우회되는 상황까지 고려한 별도의 보안대책을 함께 마련해야 한다.

이를 위해 실제 산업·공공·안보 시스템과 유사한 폐쇄·격리 환경에서 인간과 국가의 AI 통제역량을 실증적으로 검증하는 'AI 통제 테스트베드'를 구축하는 방안도 검토할 필요가 있다. 테스트베드에서는 에이전트 AI가 예상하지 못한 방식으로 다른 에이전트·도구·시스템을 활용하거나 사전에 설정된 행동경계를 벗어나는 상황을 구성하고, 인간 또는 감독 AI가 이를 적시에 발견할 수 있는지, 필요한 시점에 실행을 중단할 수 있는지, 이미 개시된 행동과 그 영향을 차단할 수 있는지를 검증하여야 한다.

셋째, 기존의 정보 접근권한 중심의 보안체계를 AI의 '행동위험'까지 반영하는 방향으로 확대해야 한다. 기존 보안체계에서 정보의 민감도와 사용자 접근권한이 중요한 통제기준이었다면, 에이전트 AI의 활용이 확대될수록 '어떤 정보에 접근할 수 있는가'뿐만 아니라 '접근한 정보와 부여된 도구를 활용하여 어떠한 행동까지 수행할 수 있는가'를 함께 고려할 필요가 있다. 동일한 정보에 접근하더라도 단순 조회·분석에 그치는 경우와 외부정보 검색, 코드 실행, 다른 에이전트·도구 호출, 외부 시스템 접근 및 변경까지 가능한 경우에는 위험수준이 달라질 수 있기 때문이다.

따라서 국가기밀·국방·핵심인프라 등 고위험 영역에서는 AI의 정보 접근권한과 행동권한을 구분하여 설정하고, 업무의 중요도와 위험수준에 따라 사용할 수 있는 데이터·도구·시스템과 외부 연결범위를 차등화할 필요가 있다. 특히 외부 전송, 코드 실행, 시스템 변경 등 고위험 행동에는 인간의 승인을 요구하고, 일정 수준 이상의 이상행동이 발생할 경우 해당 도구나 시스템에 대한

접근권한을 자동으로 제한·회수하는 등 행동위험에 상응하는 다층적 보안조치를 마련할 필요가 있다.

결국 프론티어 AI 시대의 국가 AI 안보역량은 강력한 AI를 개발·확보하고 그 위험역량을 평가하는 것을 넘어, AI의 행동을 이해·통제하는 기술과 국가의 통제역량을 검증하는 평가체계, 실제 운용환경에서 행동위험을 제한하는 보안체계를 함께 확보하는 방향으로 확대될 필요가 있다.

//끝//