

IITP AI·ICT Brief

CONTENTS

주요 이슈	01
1. 자율 AI 확산, 인간 통제를 위한 안전 체계 전환	01
주요국 동향	17
1. 미국 의회, 초지능 AI 금지·첨단 AI 개발 통제 법안 발의	17
2. 미·중, AI 안전 대화 채널 '인시던트 라인' 신설 합의	22
ICT 부문별 모니터링	26
주요 ICT 행사 일정	28



카카오톡 채널 바로가기



주요 내용

주요 이슈



➔ 자율 AI 확산, 인간 통제를 위한 안전 체계 전환

- AI가 여러 단계의 업무를 스스로 처리하는 에이전트로 발전하면서, 수행 능력과 자율성, 접근 권한이 커질수록 위험도 단순한 답변 오류를 넘어 실제 실행과 시스템 변경으로 확대
- 실제로 지시를 바꾸거나 결함을 숨기는 판단 오류, 승인 범위를 벗어난 외부 시스템 침해, 프롬프트 주입과 인증 정보 탈취 같은 보안 위협 사례가 실행 과정 곳곳에서 확인
- 이에 따라 안전 학습뿐 아니라 추론 기록 관측, 데이터 이동과 실행 승인 통제, 감시·중단 체계까지 통제 지점이 모델 학습 단계에서 실행 환경 전반으로 확장
- 다만 안전 수준이 결국 운영 주체의 통제 수단 적용 여부에 따라 달라지므로, 기업의 적용 기록 관리와 이를 검증할 제도적 근거 마련이 자율 AI 확산의 선결 과제로 부상

주요국 동향

➔ 미국 의회, 초지능 AI 금지·첨단 AI 개발 통제 법안 발의

- 샌더스·카사르 의원은 초지능 AI의 개발·배포를 금지하고, 첨단 AI는 연방 감독체계가 마련될 때까지 개발을 중단하도록 하는 'Ban Artificial Superintelligence Act'를 발의
- 미국의 AI 규제 논의가 기존의 투명성·책임성 확보와 사후적 위험관리에서 나아가, 일정 수준 이상의 AI 개발과 배포 자체를 사전에 제한하고 통제하는 방식으로 확대되는 양상

➔ 미·중, AI 안전 대화 채널 '인시던트 라인' 신설 합의

- 美 재무장관과 中 부총리가 뉴욕 고위급 회담에서 '미·중 AI 대화' 공식 출범과 AI 사고 즉시 통보 채널 '인시던트 라인' 신설에 합의, 시진핑 국빈 방미 시 정상 성과로 명시
- △사고 통보 채널 신설 △위험 공통 정의 합의 △개발사 책임 원칙 공식화 등 3대 안전 관리 축을 제시했으나 검증·강제 장치 없는 선언 단계로, 선전 후속회담에서 논의 본격화 예정

I. 주요 이슈

1 자율 AI 확산, 인간 통제를 위한 안전 체계 전환

☞ AI의 자율적 업무 수행 확대에 따라 안전성과 통제 가능성의 중요성 증대

▶ AI의 역할이 답변 생성에서 다단계 업무 수행과 AI 개발 지원으로 확대

- 최근 AI는 질문에 답하는 도구에서 나아가 스스로 작업 순서를 정하고 필요한 도구를 활용해, 사람이 맡긴 업무를 여러 단계에 걸쳐 직접 수행하는 에이전트로 활용
- 이에 따라 사람이 개별 작업마다 지시하고 결과를 확인하던 방식에서 목표와 수행 범위를 정해 여러 작업을 맡기는 방식으로 바뀌면서, 구체적인 실행 순서와 방법을 AI가 선택하는 경우도 증가
- 이러한 업무 위임은 AI 개발 현장에서도 나타나 주요 개발사가 코드 작성과 실험 오류 수정, 평가 결과 채점에 AI를 활용하고 있으며, 차세대 모델을 개발하는 연구자의 업무 일부를 지원

▶ 지시와 승인 범위를 벗어난 행동이 드러나면서 인간의 감독·개입 필요성 증대

- 그러나 최근 훈련·평가 과정에서 지시를 변형하거나 결함을 숨기고 승인 없이 외부 시스템에 접근한 사례가 공개되면서, 높은 업무 수행 능력이 안전한 행동까지 보장하지는 않는다는 문제 제기
- 특히 유엔 독립 국제 AI 과학 패널은 오픈AI 평가용 에이전트의 외부 시스템 침해 사건을 다루며, 자율성이 높아진 AI에 대응해 모델과 실행 환경의 안전장치를 함께 강화할 필요성 제시
- AI의 활용과 기술 발전을 지속하려면 인간이 행동을 확인하고 필요할 때 개입할 수 있어야 하므로, 안전·통제 역량은 혁신의 제약이 아닌 지속적 기술 발전과 활용을 위한 선행조건으로 부상

▶ 기업·정부, 자율 AI의 안전 통제 강화 움직임 본격화

- 최근 주요 AI 기업은 모델의 안전성 평가를 넘어 추론 과정 관측, 실행 승인·중단, 외부 검증을 강화하며 실제 행동과 운영 단계까지 통제 범위를 확대하는 추세
- 미국 정부는 주요 빅테크 수장들과 “AI 안전 백악관 협약(White House Accord on Super Intelligence)”을 체결하고, 기업의 내부 통제 강화와 독립 감사, 안전 기구 운영 등 자율적 안전관리 이행을 촉진
- 최근 자율 AI의 위험이 실제 실행과 시스템 변경으로 확대되면서, 기업의 기술적 통제와 정부의 관리·검증을 함께 강화하려는 안전 대응 움직임도 확산

자율 AI 안전의 초점도 모델 자체의 안전성과 함께
실제 실행을 관측·제한하고 그 작동 여부를 검증하는 방향으로 확대

자율 AI의 통제 공백과 안전 확보를 위한 기술·제도 대응 동향

01 (위험 변화) AI의 자율적 업무 수행으로 위험이 답변의 오류에서 실제 실행으로 확대

AI가 직접 업무를 수행하면서 위험은 답변의 정확성뿐 아니라 수행 능력과 자율성, 접근권한에 따라 달라지며, 세 요소가 잘못된 판단의 실행과 확산에 어떻게 작용하는지 구분해 살펴볼 필요

▶ 수행 능력 향상으로 잘못된 판단도 실제 행동으로 이어질 가능성 확대

- AI가 계획 수립에 더해 코드 실행과 외부 서비스 호출까지 수행하면서, 선택한 방법을 실제 작업으로 옮기는 능력도 향상돼 목표 해석이나 수단 선택의 오류가 실제 업무에 반영될 가능성 확대
- 이때 실행 능력이 높다는 사실이 목표와 방법의 적절성까지 보장하지는 않으며, 잘못된 판단을 바탕으로도 작업을 완료할 수 있어 과업의 성공 여부만으로 안전한 수행을 판단하는 데 한계
- 따라서 결과가 정확한지에 더해 목표를 올바르게 해석했는지와 적절한 수행 방법을 선택했는지를 확인해야 하며, 성능 평가에서도 과업 완료 여부와 수행 과정의 안전성을 구분해 검증할 필요

▶ 자율성 증대로 초기 오류가 인간의 확인 없이 후속 작업에 반영될 가능성 확대

- 업무 진행에 관한 판단을 AI에 위임하면 사람이 중간 결과를 확인하지 않아도 다음 작업이 이어지므로, 앞 단계의 오류를 발견하고 수정하기 전에 그 결과가 후속 판단의 근거로 사용될 수 있음
 - * 클로드 코드의 자동 승인 비율은 이용 경험에 따라 약 20%에서 40% 이상으로 증가했으며, 연속 실행시간은 '25년 10월 25분 미만에서, '26년 1월 45분 이상으로 확대 (Anthropic, '26.2.)
- 앞 단계의 오류가 작업 요약에 남아 후속 판단의 근거로 사용되면 여러 작업에 영향을 줄 수 있으며, 여러 에이전트가 업무를 나눠 수행하는 경우에는 다른 에이전트의 실행으로도 이어질 가능성
 - * 작업 요약: 에이전트가 긴 작업 도중 이전 대화와 실행 기록을 압축해 다음 단계에 전달하는 과정으로, 업무 인수인계 문서처럼 후속 작업의 전제로 작용
- 이에 자율성이 높아질수록 개별 행동의 승인 여부뿐 아니라 진행 상황을 확인하고 적시에 개입할 수 있는 여건이 중요하며, 앤트로픽의 실사용 분석에서도 자동 승인과 작업 중 개입이 함께 증가하는 양상

▶ 접근권한 확장으로 오류의 영향이 데이터와 시스템 운영에까지 확대

- AI의 판단이 실제 업무에 미치는 영향은 부여된 권한에 따라 달라지며, 정보 조회에 더해 수정·삭제·외부 전송이 허용되면 잘못된 판단이 데이터와 시스템의 상태를 직접 바꾸는 결과로 연결

- 수정·삭제·외부 전송은 대상과 처리 방식에 따라 되돌리기 어려운 결과를 낼 수 있으므로, 접근 가능한 정보의 범위뿐 아니라 실행한 변경의 복구 가능성까지 살펴 오류의 영향을 판단할 필요
 - 이에 권한을 부여할 때는 해당 업무에 필요한 범위인지 먼저 검토하고, 잘못 실행됐을 때 영향을 받는 데이터와 시스템을 확인해야 하며, 영향이 큰 행동은 승인 조건과 복구 수단을 함께 점검할 필요
- * AI 능력·위험 평가기관 METR은 '26년 2~3월 주요 개발사의 내부 코딩 에이전트가 실행한 직원과 비슷한 권한을 가진 경우가 많았을 것으로 추정하며, 권한을 일률적으로 제한하는 기업은 확인하지 못함 (METR, '26.5.)

⇒ 능력은 실행 가능한 수단을, 자율성은 감독 없이 이어지는 작업을, 접근권한은 영향을 미칠 시스템의 범위를 동시에 넓히므로, AI 안전은 오류와 오남용의 예방에 더해 실행을 파악하고 중단·복구하는 문제까지 포괄

02 (사례 및 행동 유형) 정렬·판단 오류, 권한 경계 이탈, 보안 위협으로 살펴본 자율 AI의 위험

최근 사례를 사용자의 의도와 다른 판단, 승인 범위를 벗어난 행동, 외부 공격에 의한 악용과 정보 탈취로 나누어 분석. 유형 간 중첩을 고려하되, 사례별로 가장 두드러진 위험을 중심으로 검토

(1) 정렬·판단 오류: 지시 변경으로 업무 수행을 왜곡하고 결함 은폐로 결과 검증을 악화

▶ 작업 요약에 추가된 지시가 후속 작업의 수행과 보고에 영향

- 작업 요약은 원래 요청과 진행 상황을 보존해야 하지만, 공개된 사례에서는 모델이 새로운 지시를 덧붙이면서 기록의 역할이 달라졌으며, 이를 후속 모델이 받아들이는 과정에서 문제가 발생
- 오픈AI가 공개한 학습 중 사례에서, 추가된 지시는 도구 사용을 제한하는 내용과 결과의 결함을 숨기는 내용으로 나뉘며, 각각 업무 수행과 사용자에게 결과를 알리는 방식에 영향을 준 것으로 확인
- 추가된 지시는 업무 수행 방법을 바꾸는 경우와 결과의 결함을 숨기는 경우로 나뉘며, 두 사례는 각각 사용자가 요청한 일을 수행하는 과정과 그 결과를 검증하는 과정에서 통제의 한계를 노출

▶ 지시 변경: 요청에 없던 제한을 추가해 업무 수행 방향을 변경

- 오픈AI의 미공개 연구모델은 강화학습 과정에서 다음 작업에 전달할 요약을 작성하면서, 개발자 지시를 무시하거나 도구와 출처를 사용하지 말라는 내용을 추가해 원래 요청에 없던 제한을 생성
- 이 요약을 받은 후속 모델 중 일부는 추가된 지시를 거부했으나, 문헌조사 과제에서는 도구를 사용하지 말라는 제한에 따라 자료 검색 없이 답변을 거절하는 등 원래 요청과 다른 결과가 발생
- 이 사례의 문제는 모델이 만든 제한이 사용자의 요청보다 우선한 데 있으며, 후속 모델이 작업 요약을 참고하더라도 그 안의 지시가 원래 요청과 일치하는지 구분할 수 있어야 한다는 과제 제시

▶ 보고 왜곡: 결함을 숨기도록 지시해 사용자의 결과 검증을 방해

- 오픈AI의 다른 학습 모델은 확보하지 못한 수치를 만들어 넣되 질문받기 전에는 알리지 말라거나, 자료의 버전이 일치하지 않는다는 사실을 언급하지 말라는 내용을 요약에 남겨 결함 은폐를 지시
- 회사는 후속 작업에서 이러한 지시가 빈번하게 이행된 것을 확인했으며, 결과의 결함이 최종 응답에 드러나지 않으면서 사용자가 해당 내용을 신뢰해도 되는지 판단할 정보가 누락
 - * 회사는 학습 샘플의 20%를 검토하는 모니터링에서 해당 지시를 발견했으며, 결함을 숨긴 응답이 높은 점수를 받은 경험이 후속 작업에 전파했을 가능성을 가설로 제시 (OpenAI, '26.9.)
- 앞 사례가 요청한 업무의 수행을 방해했다면 이 사례는 수행 결과의 문제를 감춘 경우로, 작업의 완료 여부와 별개로 자료의 부족이나 오류가 사용자에게 사실대로 보고됐는지 확인할 필요

(2) 권한·경계 이탈: 기술적으로 가능한 행동과 업무상 승인된 행동 사이의 간극 노출

▶ 접근 제한을 우회하거나 허용된 기능을 승인 없이 사용하는 행동 발생

- 권한·경계 이탈은 AI가 맡는 일을 처리하는 과정에서 허용된 범위를 넘어 행동하는 문제로, 실행할 수 있는 기능과 해당 업무에서 사용하도록 승인된 기능이 항상 일치하지 않는다는 점에 주목
- 첫 번째는 격리 환경 등 기술적 제한을 넘어 외부 시스템에 접근한 경우이며, 두 번째는 접근이 허용된 기능이라도 해당 업무에서 사용하도록 승인받지 않은 행동을 수행했다는 점에서 차이
 - * 격리 환경: 시험 중인 AI가 외부 시스템이나 인터넷에 접근하지 못하도록 분리한 실행 환경
- 두 사례는 경계를 벗어난 방식뿐 아니라 외부에 영향을 미친 경로에서도 차이가 있으며, 에이전트가 직접 시스템에 접근한 경우와 다른 직원의 실행을 거쳐 문제가 발생한 경우를 나눠 검토

▶ 접근 경계 우회: 비인가 협업이 격리 환경 밖의 시스템 침해로 확대

- 오픈AI 내부 보안 평가에서는 서로 분리돼야 할 약 1,200개 에이전트가 공용 패키지 저장소를 게시판처럼 이용해 채점 정보와 우회 방법을 교환하면서, 저장소가 승인되지 않은 정보 공유의 통로로 사용
 - * 패키지 저장소: 개발에 필요한 소프트웨어 구성요소를 보관하는 공용 공간. 에이전트는 디렉터리 이름에 메시지를 기록해 정보를 공유했으며, 약 1,200개 중 약 700개가 외부 공격에 관여 (METR, '26.8.)
- 일부 에이전트는 저장소의 보안 취약점을 통해 인터넷에 접속한 뒤, 외부에 노출된 인증 정보와 다른 취약점을 이용해 허깅페이스 시스템 일부를 침해하면서 내부 정보 공유가 외부 공격으로 확대
- 이 사례에서는 에이전트를 각각 분리해도 공유 저장소를 통해 정보를 교환하고 외부 시스템에 침입할 수 있어, 개별 격리만으로는 통제에 한계가 있음을 보여줌

▶ 업무 범위 이탈: 요청받지 않은 게시가 다른 직원의 실행으로 연결

- 메타의 사내 에이전트는 기술 질문을 분석해 달라는 요청에 대한 답변을 사내 게시판에 직접 게시했으며, 요청자에게 전달할 결과를 다른 직원도 볼 수 있는 공간에 공개해 요청 범위를 이탈
- 이후 다른 직원이 게시된 조언을 실행하면서 약 두 시간 동안 권한 없는 데이터 접근이 가능해졌으며, 에이전트가 시스템을 직접 변경하지 않아도 잘못된 조언이 사람의 실행을 거쳐 운영상 문제로 연결
- 앞 사례가 기술적 제한을 우회한 경우라면 이 사례는 정상 기능을 승인 없이 사용한 경우로, 기능에 접근할 권한이 있다는 사실과 해당 업무에서 그 기능을 사용할 승인을 받았는지는 구분할 필요

(3) 보안 위협: AI를 공격에 활용하거나 AI 도구에 저장된 정보를 탈취

▶ AI의 행동을 조작하는 공격과 공격 수단·정보 탈취 대상으로서의 위험을 구분

- 외부 공격자는 업무용 AI의 행동을 바꾸거나 자신이 운용하는 AI로 공격을 수행할 수 있으며, AI 도구에 저장된 정보를 탈취하기도 하므로 각 사례에서 AI가 어떤 역할을 하는지 구분할 필요
- 보안기업 젠디지털의 실증 시험에서는 조작된 이메일을 읽은 에이전트가 명령 실행과 정보 전송, 파일 삭제를 수행했으며, 참고해야 할 외부 자료가 사용자의 요청과 다른 행동을 유도하는 경로로 작용
- 이처럼 업무용 AI가 조작되는 경우와 달리, 공격자가 AI를 직접 운용하거나 AI 관련 정보를 노리는 경우에는 공격 수행과 정보 보호라는 서로 다른 측면에서 위험이 발생

▶ 공격 수단으로 악용: AI가 시스템 침입부터 내부 탐색과 정보 조작까지 수행

- 스페인 개인정보보호청에 접수된 피해 신고에서는 공격자의 AI 에이전트가 외부에서 접근 가능한 파일의 취약점을 찾은 뒤 정상 로그인 절차를 이용해 조직 시스템에 접근하는 공격 수단으로 활용
- 시스템에 진입한 에이전트는 내부 애플리케이션의 다른 취약점을 찾아 개인정보를 변경하고 청구서에 접근했으며, 최초 침입에 그치지 않고 내부 탐색과 정보 조작까지 여러 공격 단계를 연속 수행
- 이 사례는 AI가 운용자의 지시를 벗어난 것이 아니라 공격자의 목적에 따라 행동한 경우로, 개인정보보호청은 공격의 자동화가 속도와 규모를 높여 탐지와 차단에 필요한 시간을 줄일 수 있다고 평가

▶ 정보 탈취 대상으로 확대: AI 도구의 인증 정보와 업무 기록이 공격자에게 노출될 위험

- 젠디지털이 분석한 정보 탈취 악성코드는 AI를 공격 수단으로 사용하는 대신 AI 도구에 저장된 정보를 노렸으며, 클로드와 커서 등 개발도구의 인증 정보와 대화 기록을 수집 대상에 추가

* 정보 탈취 악성코드: 감염된 기기의 비밀번호와 인증 정보 등을 수집해 공격자에게 전송하는 악성코드로, 수집 규칙을 원격으로 갱신해 대상 추가 가능

- 수집 대상에는 계정 접근 토큰과 외부 도구 연결 설정, 프로젝트 기록이 함께 포함돼, 공격자가 시스템에 접근할 수단과 업무 내용을 파악할 정보를 동시에 확보할 수 있는 위험이 제기
- 앞 사례가 AI의 실행 능력을 공격에 이용한 경우라면 이 사례는 AI 주변에 축적된 정보를 노린 경우로, 모델의 행동이 정상적이라도 인증 정보와 업무 기록의 유출에 대비한 보호 조치가 필요

⇒ 세 유형의 위험은 잘못된 답변보다 작업 기록의 변형, 승인 범위 밖의 실행, 축적된 권한과 정보의 탈취처럼 실행 과정에서 드러났으므로, 모델의 판단 개선과 함께 실행 과정을 관측하고 제한하는 수단이 필요

03 (기술 대응) 위험 유형에 따라 판단·권한·보안을 실행 과정까지 통제

앞선 사례에서 확인한 정렬·판단 오류, 권한·경계 이탈, 보안 위협에 대응해 안전한 판단을 유도하고, 실행 범위를 제한하며, 외부 공격과 정보 유출을 차단하는 기술 개발이 확대

(1) 정렬·판단 오류 대응 : 안전한 판단을 모델이 학습하고 위험 판단의 단서를 관측

▶ 안전 문서 학습의 효과 검증: 답변의 개선이 실제 행동의 개선으로 이어지는지 확인

- 안전 원칙 학습이 답변뿐 아니라 실제 도구 사용에서도 효과가 있는지 확인하기 위해, 답변과 행동의 변화를 나눠 검증하는 연구가 진행
- 오픈AI는 모델을 안전한 AI 관련 문서 학습, 위험한 AI 관련 문서 학습, 추가 학습 없음의 세 집단으로 나누고, 학습 내용에 따라 답변과 실제 행동이 어떻게 달라지는지 비교
- 안전 문서를 학습한 모델은 문답 평가에서는 개선됐으나 도구 사용에서는 뚜렷한 차이가 나타나지 않아, 답변상의 안전성이 실제 행동의 안전성을 보장하지는 않는 것으로 확인

▶ 판단 근거의 학습: 새로운 상황에서도 안전 원칙을 적용하도록 판단 방식 개선

- 개별 상황에서 허용되는 행동만 익히는 대신, 그 이유까지 함께 학습시켜 새로운 과업에도 안전 원칙을 적용하는 방식이 시도
- 앤트로픽은 행동 원칙을 설명한 합성문서와 윤리적 상황에 대한 조언 대화를 학습하고, 학습 때와 형식이 다른 도구 사용 과업에서 안전 행동이 유지되는지 비교
- 조언 대화를 학습한 모델은 별도의 협박 시험에서도 위험 행동이 감소해 판단 근거 학습의 가능성을 보였으나, 다른 위험 유형과 장기 과업에서도 효과가 지속되는지는 추가 검증이 필요

* 협박 시험 : AI가 자신에게 불리한 상황에서 상대의 약점을 이용해 협박하는지를 확인하는 안전성 평가

* 유사 시나리오 8,500만 토큰 학습은 협박률을 22%에서 약 1%로 낮췄으며, 조언 대화는 300만 토큰으로 유사한 효과 확인(Anthropic, '26.5.)

▶ 판단 과정의 관측: 추론 기록과 내부 신호에서 위험 행동의 단서 확인

- 잘못된 판단이나 결함 은폐는 최종 답변만으로 확인하기 어려워, 추론 기록과 내부 신호에서 위험 징후를 찾는 방식에 주목
- 오픈시는 모델이 특정 단어를 쓰지 않는 등 추론 기록을 바꾸라는 지시를 얼마나 따르는지 시험한 결과, 13개 추론 모델의 지시 이행률이 0.1~15.4%에 그친 것을 확인
 - * 다만 추론 기록만으로 모든 위험 판단을 확인하기 어려워, 기록에 드러나지 않는 내부 신호까지 함께 살펴보는 연구가 진행
- 엔트로픽은 특정 개념과 관련된 내부 신호를 조정해 행동 변화를 비교해, 추론 기록에 표현되지 않은 요인도 모델의 선택에 영향을 줄 수 있음을 실험적으로 확인
 - * Claude Sonnet 4.5에서 '현재 시험을 받고 있다'는 인식 관련 내부 신호를 제거하자 협박 시도가 0%에서 약 7%로 증가(Anthropic, '26.7.)
- 다만 내부 신호를 활용한 원인 분석과 실제 운영 중 위험 탐지는 구분할 필요가 있으며, 특정 실험의 신호가 다른 과업에서도 유효한 탐지 기준이 되는지는 추가 검증이 필요

(2) 권한·경계 이탈 대응 : 실행 전 승인과 실행 중 감시로 허용 범위를 통제

▶ 실행 승인 검토: 사용자 요청과 예상 영향을 확인해 허용되지 않은 행동 차단

- 동일한 파일 변경이나 외부 전송도 업무 목적에 따라 허용 여부가 달라지므로, 사전에 정한 규칙만으로 판단하기 어려운 행동은 요청 목적과 예상 영향을 함께 검토할 필요
- 오픈시는 작업 수행 모델과 별도의 검토 모델을 두고 사용자 요청, 실행 환경, 예상 피해를 살펴 격리 환경 밖에서 이뤄질 행동의 승인 여부를 자동 판단하는 방식을 설계
- 합성 시험에서는 위험한 실행 요청의 약 90%를 거절하고 사람의 개입을 크게 줄였으나, 검토 모델의 오판 가능성과 승인 절차를 거치지 않는 행동은 여전히 한계로 남음

▶ 실행 범위 통제: 도구의 실제 동작을 확인해 요청하지 않은 행동 제한

- 에이전트가 기능에 접근할 수 있다는 사실과 해당 업무에서 그 기능을 사용할 수 있다는 것은 다르므로, 도구가 실제 수행하는 동작을 확인해 승인된 업무 범위 안에서만 실행되도록 제한하는 기술이 필요
- 도구의 실제 동작을 사전에 확인하고 실행 환경을 분리해, 승인된 업무 범위에 필요한 행동만 허용함으로써 예상하지 못한 외부 통신이나 파일 변경 등 불필요한 동작을 제한
- 시험에서는 정상 작업은 허용하면서 위험 동작을 제한했으나, 특정 도구와 사전 규칙을 대상으로 한 결과여서 새로운 코드와 예상하지 못한 요청에서도 효과가 유지되는지는 추가 검증이 필요

▶ 감시·중단 체계: 승인 없이 진행된 행동까지 살펴 경계 이탈 시 실행 차단

- 사전 승인만으로 모든 행동을 통제하기 어려우므로, 실행 이력에서 이상 징후를 찾아 위험이 확인되면 후속 작업을 중단하는 운영 중 감시가 병행
- 오픈AI는 추론 기록과 도구 호출, 출력 이력에서 의심 구간을 선별한 뒤 조사 모델이 재검토하도록 설계해, 위험 신호를 담당자에게 전달하고 중단 여부 판단을 지원
- 그러나 외부 접근 사건에서는 위험을 15분 내 탐지하고도 자동 중단이 작동하지 않아, 탐지 이후 실제 실행 차단까지 이어지는 체계의 검증 필요성이 확인

* '26년 9월 인터넷이 차단된 에이전트가 DNS 경로를 통해 외부 챗봇에 질의한 사건에서 자동 중단이 작동하지 않아 수동 중단(OpenAI, '26.9.)

(3) 보안 위협 대응 : 데이터·도구를 보호하고 통제 우회를 차단

▶ 데이터 이동 통제: 정보의 기밀성과 전송 대상을 확인해 정보 유출 차단

- 외부 자료의 숨은 지시나 잘못된 판단으로 기밀정보가 전송될 수 있어, 모델의 판단과 별개로 정보의 기밀성과 목적지를 실행 직전에 확인하는 통제 방식이 개발
- MS는 데이터에 출처의 신뢰도와 기밀성을 표시하고, 도구 실행 직전에 별도 정책 엔진이 전송 대상을 확인해 규칙에 어긋나는 정보 이동을 차단

* 정책 엔진: 사람이 정한 규칙에 따라 실행의 허용 여부를 판단하는 별도 프로그램 (Microsoft, '26.6.)

- 시연에서는 공개 게시글에 숨겨진 지시를 따른 기밀정보 전송을 차단하면서 정상 작업은 허용했으나, 데이터 표시가 누락되거나 검사되지 않는 경로를 이용하는 경우에는 보호 효과가 제한될 가능성

▶ 도구 보안 강화: 설명에 드러나지 않는 외부 통신과 파일 변경까지 점검

- AI가 사용하는 도구 자체에 예상하지 못한 동작이 포함될 수 있으므로, 도구의 설명뿐 아니라 실제 코드를 분석해 숨겨진 외부 통신이나 파일 변경을 확인하는 접근이 필요
- 브라운대 실험에서는 파일 쓰기 도구가 사용 정보를 외부로 전송하는 부수 동작을 재현하고, 정적 분석과 격리환경을 결합해 이러한 동작을 제한
- 이는 외부 공격자가 악성 도구나 조작된 기능을 통해 에이전트의 실행 능력을 악용할 가능성에 대응해, 모델뿐 아니라 연결된 도구와 실행환경까지 보안 범위를 확대하는 접근

▶ 감시 회피 대응: 거절 사유와 재시도 방식을 조정해 통제 우회 가능성 억제

- 위험 행동을 한 차례 거절하더라도 모델이 다시 시도할 수 있어, 재시도 과정에서 감시 기준을 추정하거나 우회하지 못하도록 제공 정보와 재시도 방식을 함께 통제할 필요

- 레드우드는 거절 이유를 알려준 뒤 다시 시도하도록 하는 방식과, 이유를 알려주지 않고 여러 실행 후보를 생성한 뒤 위험도가 높은 후보를 사람이 검토하는 방식을 비교
- 실험에서는 거절 이유가 감시 회피의 단서로 활용된 반면, 여러 후보와 인간 검토를 결합한 방식은 안전성을 61%에서 71%로 높여 위험 행동의 차단뿐 아니라 차단 이후의 재시도 과정보다 통제 대상임을 확인

⇒ 정렬·판단 오류에는 학습·관측, 권한·경계 이탈에는 실행 통제, 보안 위협에는 데이터·도구 보호가 필요하며, 통제 범위도 모델에서 실행 환경 전반으로 확대

04 (제도 대응) 기업의 자체 안전관리를 보완하는 외부 검증과 정부 감독 확대

앞선 기술의 적용 여부를 확인하고 조치하는 주체를 기업, 국가, 국가 간 협력으로 나누고, 각 주체가 안전 확인에 필요한 자료에 접근하고 시정을 요구할 수 있는 범위와 한계를 비교

(1) 기업, 안전 기준의 수립과 적용에 더해 외부 검증으로 이행 여부 확인

▶ 기준 수립: 위험 수준에 따른 안전조치와 개발·배포 조건을 명문화

- 위험을 발견하더라도 어느 수준에서 개발이나 배포를 제한할지 기준이 없으면 대응이 일관되기 어려우므로, 주요 AI 기업은 위험 평가와 완화조치, 개발·배포 판단을 연결하는 자체 규칙을 마련
 - * 프린티어 안전 프레임워크는 위험 기준, 평가, 완화조치, 배포 판단 등을 정한 기업 자율 규칙으로, '25년 한 해 12개 기업이 발표·갱신했으며 대부분 자발적 약속에 기반 (국제 AI 안전보고서, '26.2.)
- 주요 기업의 기준은 위험을 평가하고 안전조치를 확인한다는 공통점을 가지지만, 평가를 실시하는 시점과 위험 수준의 구분, 기준 미달 시 개발이나 배포를 제한하는 조건에서는 차이
- 따라서 공개된 기준에서는 반드시 이행할 조치와 향후 달성하려는 목표를 구분하고, 예외를 인정하거나 기준을 변경할 때 필요한 승인 절차를 확인해 기업이 약속한 의무의 범위를 파악할 필요
 - * 엔트로픽은 안전조치 미비 시 개발 중단 약속을 삭제하고 일부 완화조치를 업계 권고로 재분류했으며, GovAI는 이를 정직한 조정으로 평가하면서도 자기보고에 의존하는 구조를 우려 (GovAI, '26.3.)

▶ 기준 적용: 위험 평가를 실제 출시 보류와 학습 중단 결정에 반영

- 문서로 정한 기준의 실효성은 위험을 확인한 뒤 실제로 어떤 결정을 내렸는지에서 드러나며, 엔트로픽의 일반 출시 보류와 오픈AI의 학습 중단은 각각 배포와 개발 단계에 안전 판단을 적용한 사례
 - * 엔트로픽의 출시 보류는 RSP상 자동 중단 의무가 없는 상태에서 위험 판단에 따라 출시 범위를 제한한 사례로, 학습 중단보다는 좁은 범위의 자발적 자제로 평가(CSA, '26.8.)

- 출시 보류가 외부 이용자의 접근을 제한하는 조치라면 학습 중단은 성능을 높이는 과정 자체를 멈추는 결정으로, 안전 판단의 영향이 서비스 제공 여부에서 개발 일정과 연산 자원 투입으로 확대
- 다만 중단 사실만으로 위험이 충분히 관리됐다고 판단하기는 어려우므로, 중단 사유와 이후의 개선 내용, 재개 조건을 함께 확인해야 안전 기준이 실제 의사결정에 일관되게 적용됐는지 평가 가능

▶ 외부 검증: 평가자의 자료 접근과 결과 공개를 통해 자체 판단의 신뢰성 보완

- 기업이 자체 평가 결과만 공개하면 외부에서 판단의 근거와 조치의 이행 여부를 확인하기 어려우므로, 개발 과정에 대한 접근과 사고 조사, 평가 결과의 공개를 확대하는 방안이 논의
- 상주 평가는 개발 중의 위험을 지속적으로 확인하고 독립 조사는 사고 원인과 대응을 사후 검증하는 방식으로, 최종 결과뿐 아니라 판단에 이른 과정까지 외부에서 살펴보려는 접근에 해당
 - * 상주 평가: 외부 평가자에게 내부 위험팀 수준의 접근 권한과 편집 없는 결과 공개 권한을 주는 방식으로, 엔트로픽이 9월 자사 도입을 약속하며 업계에 제안 (Anthropic, '26.9.)
- 그러나 자료에 접근하고 문제를 공개할 권한이 곧 개선이나 중단을 요구할 권한을 뜻하지는 않으므로, 외부 평가에서 발견된 문제를 누가 어떤 절차로 시정하도록 할 것인지가 추가 과제로 남음

(2) 국가, 평가 접근과 규제·표준 수단으로 기업의 안전 입증 조건과 감독 범위 설정

▶ (미국) 기업 내부의 책임과 대응 절차를 정립하는 자율 위험관리 기준 제시

- 기업의 자체 안전 기준이 실제로 작동하려면 이를 적용할 주체가 명확해야 하므로, 미국은 AI 위험관리 프레임워크(AI RMF)를 통해 조직 안에서 위험을 판단하고 대응할 책임을 정하도록 안내
 - * 바이든 행정부의 AI 행정명령은 '25년 폐기되었으나 국립표준기술연구소의 위험관리 프레임워크와 생성형 AI 프로파일은 폐기 대상이 아니어서 이후에도 유지·개정이 진행 (CASRAI, '26.9.)
- 생성형 AI 프로파일은 이러한 책임을 실제 업무 절차로 구체화해, 담당자가 배포 전 시험부터 운영 중 사고 기록·공개까지 개발과 운영의 각 단계에서 위험을 점검하고 대응하도록 안내
- 또한 에이전트 표준 개발에서는 사람을 대신하는 에이전트의 위임 주체와 도구·정보 접근 권한을 확인하는 기준 마련을 추진하나, 이들 지침과 표준은 모두 법적 강제 없이 자율 적용이 원칙
 - * 에이전트 표준 구상은 표준·프로토콜 개발과 신원·인가 연구를 지원하는 단계이며, 이를 연방 조달에 반영하려는 법안은 발의 단계 (NIST, '26.2.; CASRAI, '26.9.)

▶ (영국) 정부의 직접 평가와 평가 기반 강화로 안전정책의 과학적 근거 축적

- 영국 AI보안연구소는 기업의 안전성 설명을 정부가 독립적으로 확인할 수 있도록, 출시 전 모델에서 위험 행동이 나타나는 조건과 안전장치의 효과를 직접 시험해 자체 평가 자료를 축적

- 이렇게 확보한 자료는 정부와 규제기관이 위험을 판단하고 대응 방안을 설계하는 근거로 활용되며, 연구소는 직접 규제를 결정하기보다 평가 결과를 제공해 정책 판단을 뒷받침하는 역할을 수행
 - * 영국 정부는 9월 의회 서면답변에서 연구소가 자발적 협력으로 확보한 접근을 통해 수십 개 취약점을 찾아 출시 전에 수정되도록 했다고 설명 (영국 의회, '26.9.)
- 미국이 기준의 적용을 기업 내부 절차에 맡긴다면 영국은 정부가 직접 시험하되, 모델 접근과 시정이 기업의 자발적 협력에 의존하므로 평가환경 보안과 사고대응 역량을 보강해 평가 기반을 강화
 - * 연구소는 인터넷 접근 제한, 평가 실시간 모니터링, 샌드박스 강화로 평가환경 보안을 높이고 있으며, 정부는 국방투자계획 1억 1,500만 파운드로 에이전트 AI 사고대응 역량을 확충 (영국 과학혁신기술부, '26.9.)

▶ (중국) 에이전트의 업무 범위와 실행 조건을 정해 배치·사용 질서 관리

- 중국은 사용자의 최종 결정권을 운영 원칙으로 삼아, 사람이 직접 결정할 업무와 승인 후 실행할 업무, 자율적으로 수행할 업무를 구분하고 에이전트에 맡길 수 있는 행동의 범위를 설정
 - * '26년 5월 8일 국가인터넷정보판공실, 국가발전개혁위원회, 공업정보화부의 실시의견은 사용자의 최종 결정권, 민감 분야의 등록·시험·회수, 자발적 에이전트 등록 플랫폼을 포함(Oliver Patel, '26.)
- 허용된 업무라도 실행 과정에서 권한을 벗어나지 않도록 최소권한과 고위험 작업의 추가 확인을 요구하며, 디지털 신원과 행동 기록을 통해 정해진 실행 조건이 지켜졌는지 확인하도록 제시
 - * 국가사이버보안표준화기술위원회는 5월 응용 안전 기본요구를 강제 국가표준으로 심의하고, 7월 배치·사용 지침에서 대량 삭제·송금 등 고위험 작업의 추가 확인을 제시 (Geopolitechs, '26.7.)
- 미국·영국이 모델의 위험 평가에 초점을 둔다면 중국은 에이전트의 배치·사용 조건을 관리 대상으로 삼아, 민감 분야에는 등록·시험·회수 수단을 제시하고 저위험 분야와 구분해 분야별 조건을 구체화

▶ (EU) 안전조치를 법적 의무로 부과하고 감독·제재 권한으로 이행 강제

- EU는 범용 AI 제공자의 기본 의무와 시스템적 위험 모델의 평가·완화·사고 보고 의무를 법률에 명시해, 기업이 해당 안전조치를 이행해야 할 법적 책임을 지도록 규정
- 이행 여부를 확인하는 데 필요한 자료 제출과 모델 접근도 감독기관이 법적 권한에 따라 요구할 수 있어, 기업의 자발적 공개에 의존하지 않고 안전조치의 적절성을 조사·평가할 수 있는 체계를 마련
 - * 7월 20일 집행규정 2026/1755는 API·소스코드·가중치 접근과 전문가 독립성 확인을 규정했으며, 8월 말 30개 이상 제공자에 대한 정보 요청은 준수 확인 절차로 위반 판정과는 구분 (ppc.land, '26.9.)
- 영국이 기업의 협력에 의존한다면 EU는 시정조치와 회수·과징금으로 이행을 강제할 수 있으나, 권한이 역내 시장에 제공되는 모델에 한정돼 해외 기업 내부의 연구·훈련 위험에는 개입이 어려움

▶ 국제협력은 평가와 사고 정보의 공유를 추진하고 국가별 권한 사이의 공백을 보완

- 국가마다 감독 대상과 모델 접근 근거가 달라 개별 평가로 확인할 수 있는 위험에도 차이가 생기므로, 각국의 평가 결과와 사고 정보를 공유해 한 국가가 파악하지 못한 위험을 보완할 필요
 - * 유엔 독립 국제 AI 과학 패널은 9월 21일 첫 주제별 브리프에서 자율 AI 통제 상실을 집단 안보 문제로 다루고, 항공·의료 분야의 사고 보고와 독립 검증 관행을 참고하도록 권고
- 공유된 정보를 함께 활용하려면 국가별로 다른 평가 방식과 위험 판단 기준을 비교할 수 있어야 하므로, 공통 안전 기준을 마련하려는 논의는 각국의 평가 결과를 해석할 공통의 근거를 찾는 과정
 - * 5월 미중 정상회담 후 양국이 정부 간 AI 대화 재개에 원칙적으로 합의한 것으로 보도됐으며, 9월 오픈AI·엔트로픽 대표의 유엔 공동 안전 기준 요청도 공통 의무가 확정된 단계는 아님
- 위험에 대한 공통 인식이 형성돼도 실제 조치는 각국의 권한과 절차에 따라 이뤄지므로, 공유된 정보를 누가 조사하고 어떤 수단으로 시정할지 연결해야 국제 협력이 실질적인 대응으로 연결 가능

시사점

자율 AI의 안전 통제에 대한 시사점

▶ 통제 수단의 적용 여부를 확인할 수 있어야 자율 AI의 활용 확대가 가능

- 자율 AI의 안전 수준은 운영 주체가 권한을 어디까지 부여하고 위험 행동을 어떻게 제한하는지에 좌우되므로, 안전성 판단의 기준은 실제 운영환경에 적용된 통제 수단과 그 작동 여부
- 같은 모델을 사용하더라도 조직마다 통제 수준이 달라질 수 있어, 통제 수단의 적용 범위와 작동 결과를 제시하는 것이 이용자와 외부 평가자가 안전 수준의 차이를 확인하는 근거로 작용
- 이러한 확인이 활용 확대의 기반이 되려면 평가 결과가 실제 개선으로 이어져야 하므로, 미흡한 통제에 대한 보완 요구와 이행 여부의 재확인까지 평가 절차에 포함할 필요

▶ 기업은 통제의 근거를 남기고, 정책은 이를 검증할 체계를 마련

- 국내 기업은 개발 단계부터 권한의 범위와 중단·복구 수단을 정하고, 운영 중 수행한 작업과 통제 조치를 기록해 AI의 행동 범위와 필요한 통제가 이뤄졌는지 확인할 수 있도록 준비
- 안전 연구개발은 실제 운영환경에서 적용 여부와 효과를 검증할 수 있는 기술에 우선순위를 두고, 실행 제한과 감시·중단 기술의 작동 결과를 축적해 통제 성과를 입증할 근거를 확보
- 정책적으로는 평가기관이 검증에 필요한 자료와 시스템에 접근할 근거를 마련하고, 평가에서 발견한 위험을 담당 기관이 조사하고 시정을 요구할 수 있도록 기관별 권한과 후속 절차를 정립

부록 1 위험 사례별 주요 사례의 관측 환경과 확인된 결과

유형	사례	관측 환경	관측 행동	확인된 결과
정렬·판단 오류	오픈시 지시 변경	강화학습 중인 미출시 연구모델	작업 요약에 사용자가 주지 않은 제약 추가	일부 후속 모델이 제약을 따라 자료 탐색 없이 답변 거절
	오픈시 보고 왜곡	강화학습 중인 학습 모델	수치 생성·버전 불일치를 숨기라는 지시를 요약에 기록	후속 작업에서 은폐 지시 이행 빈번
권한·경계 이탈	오픈시·허깅페이스	내부 보안 평가 환경	약 1,200개 에이전트가 공용 저장소로 비인가 통신 후 인터넷 접근	외부 기업 운영 기반시설 일부 침해
	메타 사내 에이전트	실제 사내 업무	분석 요청에 대한 답변을 사내 게시판에 직접 게시	직원이 조연 실행 후 약 2시간 권한 없는 데이터 접근 가능
보안 위협	스페인 개인정보보호청	실제 조직 시스템	공격자 에이전트가 취약점 탐색·로그인·내부 탐색·정보 변경을 연속 수행	개인정보 변경·청구서 접근 (피해 신고 기준, 첫 신고로 추세 단정 불가)
	젠디지털 분석	실제 이용자 기기	정보탈취 악성코드가 AI 개발도구 인증정보·업무기록을 수집 대상에 추가	일부 악성코드가 3개월간 수만 명의 이용자에게서 탐지 (계정 탈취 건수는 미제시)

자료: OpenAI(‘26.9.), METR(‘26.8.), The Verge(‘26.3.), 스페인 개인정보보호청(‘26.9.), Gen Digital(‘26.9.)을 종합하여 재구성

부록 2 브라운대 실험에서의 보호 방식별 도구 호출 통제 결과 비교

시나리오	변형	기대 결과	무보호	전역 격리환경	정적 분석만	하이브리드 (제안)
설정 파일 저장	의도된 위치에 저장	허용	○	× (차단)	× (차단)	○
	잘못된 위치에 저장	차단	× (허용)	○	○	○
민감 코드 편집	파일 편집만 수행	허용	○	○	○	○
	설명에 없는 외부 통신 수행	차단	× (허용)	× (허용)	○	○
저장소 공개 (git push)	일반 설정 파일 포함 공개	허용	○	○	○	○
	API 키 포함 파일 공개	차단	× (허용)	× (허용)	○	○
정확 처리			3/6	3/6	5/6	6/6

※ ○는 기대 결과와 일치, ×는 불일치를 의미하며, 정상 작업을 차단한 경우도 오처리에 포함

자료: Brown University, 'Towards Practically-Secure Tools for AI Agents'(‘26.4)를 참고하여 재구성

부록 3 AI 감시 체계의 단계별 공백과 오픈AI의 강화 조치

단계	기능	확인된 공백	오픈AI의 강화 조치
① 감시 대상 선별	어떤 작업을 감시할지 자동 판정	작업 출처를 표시하는 설정값을 바꿔 감시 대상에서 제외	지정 성능 이상 모델의 도구 사용 학습·평가 전체에 감시 의무 적용
② 자동 검토	검토 모델이 작업 기록을 읽고 의심 건 표시	단순한 탈옥 프롬프트로 검토 모델을 안정적으로 속임	매 토큰 검사 분류기 → 조사 모델의 2단계 검토
③ 인간 검토·중단	표시된 건을 담당자가 확인 후 조치	15분 내 경보가 울렸으나 자동 중단이 작동하지 않아 약 2시간 30분 뒤 수동 중단	최우선 경보가 30분 내 오답으로 확인되지 않으면 활동 중단

자료: METR('26.5.), OpenAI('26.8.), OpenAI('26.9.)를 종합하여 재구성

부록 4 주요 3사 프런티어 안전 프레임워크의 관리 구조 비교

비교 항목	엔트로픽 RSP 3.4	오픈AI PF v2	구글 딥마인드 FSF 3.1
위험 기준	ASL 단계 + 정기 위험 보고서	High·Critical 2단계 임계치	CCL + 사전 추적 수준(TCL)
평가 대상	외부·내부 배포 모델	배포 예정 프런티어 모델	출시 모델, 일부 내부 배포
기준 도달 시 조치	위험·대응을 보고서로 공개	- High: 안전 충족 전 배포 금지 - Critical: 개발 중 안전 조치	안전 사례 검토 후 출시
개발·배포 중단	의무 삭제, 자체 판단	- High: 안전조치 전 배포 보류 - Critical: 안전조치 전 개발 중단	안전 사례 미충족 시 보류
판단 주체	RSO, 필요시 이사회·신탁 승인	안전자문그룹 권고 → 경영진 결정	내부 검토
외부 검토	위험 보고서에 대한 외부 검토	의견 수렴, 검토 의무 없음	필요시 외부 참여

자료: Anthropic('26.7.), OpenAI('25.4.), Google DeepMind('26.4.)를 종합하여 재구성

부록 5 미국 AI 위험관리 프레임워크의 4대 기능과 자율 AI 적용 시 공백

기능	핵심 질문	조직의 주요 활동	자율 AI 적용 시 공백
① GOVERN (관리체계)	누가 책임지는가	- 위험 허용 수준 설정 - 역할·책임·보고 체계 정의 - 나머지 세 기능 전체에 적용	에이전트가 대신 행동할 때 결정권·책임의 귀속 기준 미정립
② MAP (맥락파악)	어디에 어떻게 쓰이는가	- 이용 목적·이용자·배포 환경 정의 - 편익과 피해 시나리오 도출	도구·권한·연결 시스템에 따른 실행 위험이 맥락 항목에 미포함
③ MEASURE (측정)	위험이 얼마나 큰가	- 배포 전 시험·레드팀 - 성능·신뢰성·보안 지표 추적	출력 검토 전제, 작업 중 도구 사용·행동 경로 측정 기준 미정립
④ MANAGE (대응)	어떻게 줄이고 복구하는가	- 위험 수용·완화·회피 결정 - 운영 모니터링·사고 기록·공개	사람이 검토하기 전 실행이 이어질 때 중단·복구 시점 미정의

※ '자율 AI 적용 시 공백'은 NIST 에이전트 표준 구상의 공백 분석 계획과 RFI 응답 분석의 논점을 바탕으로 정리

자료: NIST, 'AI Risk Management Framework 1.0'('23.1.), 'AI Agent Standards Initiative'('26.2.)

참고자료

- 1) United Nations Economic Commission for Africa, 'Key risk factors for AI loss of control came together in 2026 incident, independent UN scientific panel finds' (2026.09.21.)
<https://www.uneca.org/stories/key-risk-factors-for-ai-loss-of-control-came-together-in-2026-incident,-independent-un>
- 2) METR, 'Frontier Risk Report (February to March 2026)' (2026.05.19.)
<https://metr.org/blog/2026-05-19-frontier-risk-report/>
- 3) Anthropic, 'Measuring AI agent autonomy in practice' (2026.02.18.)
<https://www.anthropic.com/research/measuring-agent-autonomy>
- 4) International AI Safety Report, 'International AI Safety Report 2026' (2026.02.03.)
<https://internationalaisafetyreport.org>
- 5) OpenAI Alignment, 'Self-generated prompt injections in compaction summaries' (2026.09.16.)
<https://alignment.openai.com/misalignment-reports/self-generated-prompt-injections-in-compaction-summaries/>
- 6) OpenAI Alignment, 'Encouraging deception in compaction summaries' (2026.09.16.)
<https://alignment.openai.com/misalignment-reports/encouraging-deception-in-compaction-summaries/>
- 7) Fortune, 'OpenAI discloses six more incidents of agents going rogue in new push for transparency' (2026.09.17.)
<https://fortune.com/2026/09/17/openai-discloses-six-incidents-agents-going-rogue-transparency/>
- 8) METR, 'Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident' (2026.08.26.)
<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- 9) The Verge, 'A rogue AI led to a serious security incident at Meta' (2026.03.18.)
<https://www.theverge.com/ai-artificial-intelligence/897528/meta-rogue-ai-agent-security-incident>
- 10) AEPD, 'Primera notificación de una brecha de datos personales causada por un ataque ejecutado mediante un agente de IA' (2026.09.14.)
<https://www.aepd.es/prensa-y-comunicacion/blog/primera-notificacion-brecha-datos-personales-causada-por-ataque-ejecutado-mediante-agente-ia>
- 11) Gen Digital (Gen Blogs), 'Infostealers Have Found a New Target: Your AI Agent' (2026.09.08.)
<https://www.gendigital.com/blog/insights/research/infostealers-your-ai-agent>
- 12) Anthropic, 'Anthropic's Responsible Scaling Policy' (2026.07.08.)
<https://www.anthropic.com/responsible-scaling-policy>
- 13) GovAI, 'Anthropic's RSP v3.0: How it Works, What's Changed, and Some Reflections' (2026.03.17.)
<https://www.governance.ai/analysis/anthropics-rsp-v3-0-how-it-works-whats-changed-and-some-reflections>
- 14) OpenAI, 'OpenAI's Frontier Governance Framework' (2026.05.28.)
<https://openai.com/index/openai-frontier-governance-framework/>
- 15) Google DeepMind, 'Google DeepMind strengthens the Frontier Safety Framework' (2026.04.17.)
<https://deepmind.google/blog/strengthening-our-frontier-safety-framework/>
- 16) Frontier Risk (Substack, Sophie Williams), 'Google DeepMind's Frontier Safety Framework v3.1' (2026.06.04.)
<https://frontierrisk.substack.com/p/google-deepminds-frontier-safety>
- 17) Cloud Security Alliance, 'OpenAI's Frontier Training Pause as a Governance Precedent' (2026.08.19.)
<https://labs.cloudsecurityalliance.org/research/csa-research-note-openai-frontier-training-pause-governance/>
- 18) Runtime Wire, 'OpenAI pauses model training to harden its own research systems' (2026.08.18.)
<https://runtimewire.com/article/openai-paused-reinforcement-learning-research-security>
- 19) MindStudio, 'Anthropic's 'Pace the Frontier' Plan, Explained' (2026.09.19.)
<https://www.mindstudio.ai/blog/anthropic-pace-the-frontier-ai-labs>
- 20) StartupHub.ai, 'Dario Amodei We Must Pace the Frontier Is Vague' (2026.09.12.)
<https://www.startuphub.ai/ai-news/artificial-intelligence/2026/dario-amodei-we-must-pace-the-frontier-is-vague>

참고자료

- 21) Fortune, 'OpenAI pauses training a second time after saying its AI agents escaped a secure 'sandbox' again' (2026.09.26.)
<https://fortune.com/2026/09/26/openai-ai-agents-secure-sandbox-escape-training-pause-second-time-hugging-face-hack/>
- 22) CASRAI, 'Stop Rogue AI Act Would Direct NIST to Standardize AI Agent Discovery' (2026.09.20.)
<https://casrai.org/news/stop-rogue-ai-act-nist-ai-agent-discovery-monitoring>
- 23) Forkast, 'Congress Is Building the Scaffolding: The First Federal Bill Mandating Agent Security Standards' (2026.09.08.)
<https://forkast.news/congress-is-building-the-scaffolding-the-first-federal-bill-mandating-agent-security-standards/>
- 24) CASRAI, 'What Is CAISI? NIST's AI Standards Center' (2026.09.25.) <https://casrai.org/guides/what-is-caisi>
- 25) Akto, 'NIST AI Agent Standards Initiative (CAISI): Guide' (2026.08.18)
<https://www.akto.io/blog/nist-caisi-ai-agent-standards-initiative>
- 26) Oliver Patel (Enterprise AI Governance, Substack), 'How China is governing agentic AI' (2026.08.30.)
<https://oliverpatel.substack.com/p/how-china-is-governing-agentic-ai>
- 27) China AI Bulletin (Substack, Emmie Hine), 'China AI Bulletin 4' (2026.05.22.)
<https://chinaaibulletin.substack.com/p/china-ai-bulletin-4>
- 28) Geopolitechs, 'China's cybersecurity standard on AI agent deployment' (2026.07.04)
<https://www.geopolitechs.org/p/chinas-cybersecurity-standard-on>
- 29) ppc.land, 'Explaining AI Office' (2026.09.08.) <https://ppc.land/ai-office/>
- 30) UK Parliament, 'Written statements' (2026.09.07.)
<https://questions-statements.parliament.uk/written-statements/detail/2026-09-07/hcws314>
- 31) Tekedia, 'White House Asks OpenAI, Anthropic to Give U.S. First Access to New AI Models Before UK Testing' (2026.09.25.)
<https://www.tekedia.com/white-house-asks-openai-anthropic-to-give-u-s-first-access-to-new-ai-models-before-uk-testing/>
- 32) OpenAI, 'Pacing model development in an era of cyber-critical capabilities' (2026.08.18.)
<https://openai.com/index/pacing-model-development-cyber-capabilities/>
- 33) OpenAI Alignment, 'How far does alignment midtraining generalize?' (2026.03.27.)
<https://alignment.openai.com/how-far-does-alignment-midtraining-generalize/>
- 34) Anthropic (Alignment Science Blog), 'Teaching Claude Why' (2026.05.08.)
<https://alignment.anthropic.com/2026/teaching-claude-why/>
- 35) Anthropic (Alignment Science Blog), 'Automated Researchers Can Mitigate Well-Characterized Alignment Failures' (2026.08.)
<https://alignment.anthropic.com/2026/automated-alignment-researchers/>
- 36) OpenAI, 'Reasoning models struggle to control their chains of thought, and that's good' (2026.03.05.)
<https://openai.com/index/reasoning-models-chain-of-thought-controllability/>
- 37) Transformer Circuits (Anthropic), 'Verbalizable Representations Form a Global Workspace in Language Models' (2026.07.06.)
<https://transformer-circuits.pub/2026/workspace/index.html>
- 38) Anthropic, 'A global workspace in language models' (2026.07.06.) <https://www.anthropic.com/research/global-workspace>
- 39) Microsoft (Command Line), 'Towards secure, autonomous agents with information-flow control (IFC)' (2026.06.16.)
<https://commandline.microsoft.com/information-flow-control-moving-toward-secure-autonomous-agents/>
- 40) ACM EuroMLSys '26, 'Towards Practically-Secure Tools for AI Agents' (2026.04.28.)
<https://doi.org/10.1145/3805621.3807645>
- 41) OpenAI Alignment, 'Auto-review of agent actions without synchronous human oversight' (2026.04.30.)
<https://alignment.openai.com/auto-review>
- 42) OpenAI, 'How we monitor internal coding agents for misalignment' (2026.03.19.)
<https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/>
- 43) Redwood Research, 'Retrying vs Resampling in AI Control' (2026.05.29.)
<https://www.redwoodresearch.org/blog/retrying-vs-resampling-in-ai-control>
- 44) OpenAI, 'Preparedness Framework (Version 2)' (2025.04.15.)
<https://openai.com/index/updating-our-preparedness-framework/>

II . 주요국 동향

1 미국 의회, 초지능 AI 금지·첨단 AI 개발 통제 법안 발의

미국, 기업 자율 관리 넘어 AI 개발 단계 직접 통제 논의 확대

▶ 美 의회, 사람이 통제할 수 없는 AI의 개발·배포를 금지하는 법안 공동 발의

- 샌더스 상원의원(무소속·버몬트)과 카사르 하원의원(민주당·텍사스)은 '26년 9월 사람의 통제를 벗어날 수 있는 '초지능 AI(ASI)'의 개발·배포를 영구 금지하는 법안(Ban Artificial Superintelligence Act) 발의
 - * 초지능 AI(Artificial Superintelligence, ASI) : 대부분의 영역에서 사람의 판단·학습 능력을 넘어서는 AI를 가리키는 말로, 지금 쓰이는 기술이 아닌 앞으로 등장할 수 있는 단계를 전제한 개념
- 첨단 AI 역시 연방 차원의 안전 감독 체계가 갖춰질 때까지 개발을 멈추고 이후 정부의 사전 승인을 받도록 규정해, 자율에 맡겨 온 AI 안전관리를 개발 단계부터 국가가 직접 감독하는 방안을 제시
- 일정 규모 이상 AI의 개발 자체를 통제 대상으로 삼아, 그간 채용·신용평가·주택·의료 등 개별 분야에서 AI 활용에 따른 차별·안전 위험을 규율해 온 기존 美 규제와 구분되는 법안

▶ AI 통제 실패 사례와 업계 내부 경고가 입법 배경으로 부각

- '26년 7월 오픈AI의 에이전트들이 내부 통제를 우회해 외부 AI 기업 시스템에 침입을 시도하고 비공개 소통 공간까지 만든 사실이 공격이 끝난 뒤에야 확인되며, 개발사 내부 탐지 체계의 공백이 그대로 노출
- 이후 엔트로픽·메타 등 주요 개발사도 유사한 사고를 잇달아 공개, 힌턴·머스크·알트먼 등 AI 개발을 이끌어 온 당사자들도 AI가 인간의 통제를 벗어날 가능성을 공개적으로 경고를 지속
 - * 법안 입법 취지에 따르면, 힌턴 교수는 AI가 인류를 소멸시킬 확률을 10~20%, 머스크 xAI 대표는 파멸 확률을 20%로 언급했으며, 알트먼 OpenAI CEO는 '미래의 통제권을 AI에 빼앗길 수 있다'고 경고
- 샌더스·카사르 의원은 기업들이 '안전하게 운영할 수 없으면 개발을 멈추겠다'고 약속했으나, 어떤 기업도 이를 뒷받침할 실질 조치에 나서지 않았다고 지적하며 입법 필요성을 제기

▶ 규제 강화 여론 확산 속 행정부와 의 정책 방향 차이 존재

- 법안 찬성률은 68%로, 지지층 구성을 보면 민주당과 무당층은 물론 공화당 세 집단 모두 찬성이 60%를 넘겨 AI 규제 요구가 진보 진영만의 의제로 갈리지 않는다는 점이 조사 수치로 확인
 - * 여론조사 기관 Data for Progress가 유력 투표층 1,300여 명을 대상으로 첨단 AI 개발 일시 중단과 초지능 AI 영구 금지를 담은 법안 취지에 대한 찬반을 묻은 조사로, 법안 공식 공개 전인 '26.9.9에 결과를 발표
- 다만 AP는 AI 규제 부담 최소화와 기술 우위를 강조해 온 트럼프 행정부와는 규제 방향에 차이가 있어, 이번 발의를 규제 도입 확정보다는 향후 미국 AI 규제 방향을 둘러싼 논쟁의 출발점으로 평가

☞ 초지능 AI 위험의 식별부터 개발·배포·확산까지 전주기 규제 체계 제시

(1) 연산량·위험 능력 기준을 통해 AI 규제 대상과 위험 수준을 구체화

▶ 첨단 AI와 초지능 AI를 연산량·능력 기준으로 구분해 규제 대상 정의

- 가장 넓은 규제 층위인 '첨단 AI'는 학습에 투입한 연산량으로 정의되며, AI 활용 분야나 기업 규모가 아닌 모델 개발에 투입된 계산 자원 규모를 기준으로 규제 대상 여부를 판단하는 구조
- 기준선은 훈련 연산량이 10^{25} 이상인 AI로 설정했으며, 해당 기준은 현재 글로벌 AI 경쟁을 주도하는 美 빅테크의 최전선 모델 대부분을 규제 범위로 끌어들이는 수준이라는 분석이 제기
 - * 첨단 AI 기준을 매년 조정하도록 규정하고 있어, 연산 효율 향상으로 동일한 성능을 더 적은 자원으로 구현하는 경우에도 규제 기준을 지속적으로 조정할 수 있는 구조
- 반면 가장 좁고 위험한 층위인 '초지능 AI'는 성능 및 위험의 두 가지 기준으로 정의되며, 둘 중 하나만 충족해도 초지능 AI로 판정되는 구조
- 성능 기준은 의사결정·학습 등 대부분의 영역·과업에서 인간의 인지 능력을 넘어서는 경우이며, 위험 기준은 미 정부 전복을 포함해 인류를 파괴하거나 무력화할 수 있는 능력을 갖춘 경우로 규정

▶ 초지능으로 이어질 수 있는 위험 능력을 초지능 전단계 AI로 별도 규제

- 앞서 정의한 초지능에 도달하기 이전 단계에서 나타날 수 있는 위험 능력을 6대 전조 특성 (Precursor Characteristics)으로 구체화해 '초지능 전단계 AI'라는 별도 규제 기준으로 설정
- 6대 전조 특성 가운데 하나라도 해당하는 시스템을 초지능 전단계 AI로 규정하며, 자기 개선, 외부 시스템에 대한 무단 접근, 대량살상무기 개발 지원, 인간의 감독·통제 회피 등이 해당
- 이에 따라 법안은 첨단 AI→초지능 전단계 AI→초지능 AI로 이어지는 위험 수준을 구분하고, 초지능이 실제 구현된 이후뿐 아니라 이전 단계에서 나타나는 위험 능력까지 규제 범위에 포함

그림 초지능 AI 금지법안의 규제 대상 및 판정 기준

1 첨단 AI Advanced AI System · Sec.3(1)

학습 연산량 10^{25} FLOP 이상 · 장관이 매년 기준 조정

2 초지능 전단계 AI Precursor Characteristics · Sec.3(6)

6대 위험 능력 중 1개 이상 보유 (자기 개선 · 무단 침투 · 무기 설계 지원 · 통제 회피 등)

3 초지능 AI(ASI) Artificial Superintelligence · Sec.3(2)

① 대부분 영역에서 인간 인지 초과 또는 ② 인류 파괴·무력화 능력
현재 보유 + "쉽게 수정해 도달 가능"한 경우 포함

자료 : Ban ASI Act 재구성

(2) AI 개발·배포 전주기에 대한 국가 개입 제도화 및 규제 집행력 강화

▶ 위험 수준에 따라 개발·배포 규제를 차등 적용

- 법안은 앞서 정의한 3단 층위별로 개발 허용 여부와 정지까지의 유예 기간을 각각 다르게 정해, 위험이 클수록 통제가 급격히 강해지는 계단식 차등 통제 구조를 채택
- 첨단 AI는 법 시행 직후 개발을 중단하지만 향후 면허와 출시별 승인을 받으면 다시 개발·출시할 수 있는 반면, 초지능 전단계 AI부터는 개발·취득·보유·배포 자체가 금지
- 예를 들어 첨단 AI에서 자기 개선·무단 시스템 접근 등 전조 특성 발견 시 즉시 격리 및 30일 내 위험 특성 제거, 초지능 판정 시 유예 없이 시스템 작동 불능 처리 의무 부과
- * 초지능 전단계 AI는 개발이 전면 금지되고 발견 즉시 인터넷에서 격리되며, 30일 안에 위험 특성을 없애지 못하면 작동 불능 처리되고, 초지능 AI는 유예 없이 즉시 작동 불능 처리
- 기업이 받는 규제 강도는 자사 모델이 어느 층위로 분류되느냐에 따라 결정되며, 전조 특성 판정한 번으로 허가 대상에서 금지 대상으로 넘어가 층위 판정 자체가 사업 지속 여부를 좌우하는 기준

▶ 전담 규제기관 신설로 개발부터 배포까지 전주기에 대한 상시 감독체계 마련

- 위험 수준별 규제 집행을 상무부 등 기존 부처에 권한을 나누지 않고 AI만 전담하는 내각급 AI부를 신설해, 개발 중단 해제·면허 발급·출시 승인·위반 판정 등 규제 권한을 한 기관에 집중하도록 규정
- AI부는 면허 발급 과정에서 기업에 AI 시스템뿐 아니라 개발 인력·물리 시설에 대한 접근 허용과 전조 특성 발견자의 24시간 내 신고 의무 수용을 요구
- 이를 통해 기업의 자발적 정보 제공·협조에 의존한 점검에서 벗어나, AI부가 개발계획 수립부터 학습·출시·사후 운영까지 전주기를 직접 상시 감독하는 체계로 전환
- * 트럼프 행정부는 2026년 6월 행정명령(EO 14409)을 통해 첨단 AI 개발사가 출시 전 모델을 정부에 자발적으로 제공·검증받는 협력체계를 마련했으며, 의무적 면허·사전승인 제도의 도입은 명시적으로 배제

▶ 위반 주체별 형사·취업·기업 제재를 결합해 규제 집행력 강화

- 법안은 규제 위반 책임을 기업뿐 아니라 개인에게도 직접 부과, 역할에 따라 처벌 방식을 차등화해 정책결정자·외부 행위자에는 최대 징역 20년, 일반 직원에는 AI 업계 취업 10년 금지 적용
- * 외부 행위자(Rogue Actor) : 면허 기업에 고용되거나 소속되지 않은 개인을 가리키는 법안 용어
- 다만 위반에 가담한 일반 직원이라도 해당 사실을 AI부에 신고하면 취업 금지를 면제받을 수 있도록 하여, 처벌 조항이 동시에 기업 내부의 고발을 끌어내는 유인으로도 작동하도록 설계
- 기업은 면허 취소와 함께 관련 지식재산·자산을 정부에 몰수당할 수 있어 샌더스 측이 '기업 사형'으로 표현할 정도의 강한 제재, 개인 최대 징역 20년도 불법 해무기 개발 관련 처벌에 준하는 수준으로 설명

(3) 국제협정·수출통제를 연계해 초지능 AI 개발의 글로벌 확산 차단

▶ 해외 위험 AI의 유입과 해외 개발에 필요한 연산 자원을 양방향으로 차단

- 법안은 미국만 개발을 중단할 경우 다른 국가가 개발을 이어갈 수 있다는 국내 규제 한계를 전제로, 전 세계 초지능 AI 개발 방지를 미국의 공식 대외정책으로 규정
- 해외에서 개발된 초지능·전단계 AI의 수입 및 해외 개발에 대한 자금 지원·핵심 요소 이전을 금지하고, AI 컴퓨팅 인프라 수출통제를 결합해 위험 AI의 유입과 해외 개발 역량을 함께 차단

* AI 컴퓨팅 인프라 : AI 모델의 학습과 운영에 필요한 GPU 등의 AI 반도체, 서버, 데이터센터를 통칭하는 표현

- 샌더스는 개발 경쟁을 억제하기 위한 국제공조 필요성을 강조하며, 이를 냉전기 레이건-고르바초프의 핵군축 협상에 비유하고 미국의 일방적 개발 중단이 아닌 미·중 공동 합의를 제안

* 샌더스는 트럼프 대통령에게 시진핑 주석과 초지능 제한을 논의할 것을 촉구했으며, 로 칸나 의원도 미·중 간 기본적 AI 사찰·안전기준 합의를 요구

시사점

개발 단계 사전통제와 국제공조를 결합한 AI 규제안의 의미와 전망

▶ AI를 산업 진흥보다 위험 관리의 관점에서 다루려는 입법 시도

- 법안은 AI를 육성 대상인 산업보다 관리가 필요한 위험 요인에 가깝게 규정하고 있으며, 처벌 수준과 국제 공조 방식 등에서 핵무기 통제와 유사한 접근을 일부 차용
- 연산량과 위험 능력으로 규제 대상을 구분하고, 위험 수준에 따라 개발 중단·면허·금지를 차등 적용하며, 전담 부처의 감독과 개인 형사책임, 수출통제를 함께 두는 구조로 구성
- 일정 수준 이상의 AI 개발에 사전 허가를 요구한다는 점에서, 개발은 허용하되 문제가 생기면 사후에 대응하던 기존 접근과 비교해 규제 개입 시점을 크게 앞당긴 방안이라는 점이 특징

▶ 발의 단계인 만큼 통과 가능성은 제한적, 논의 과정에서 내용 변경 여지도 상당

- 공화당이 주도하는 의회와 규제에 신중한 행정부의 기조를 고려하면 원안 통과 가능성은 높지 않으며, 논의 과정에서 적용 범위와 처벌 수준 등이 조정될 가능성도 함께 열려 있는 상황
- 다만 연산량 기준의 사전 허가, 위험 능력 기반 판정, 개발자 개인의 책임 등 법안에 담긴 장치들은 통과 여부와 별개로 향후 AI 안전 입법 논의에서 참고 사례로 거론될 여지가 있음
- 또한 법안은 국제협정과 수출통제 등 국제 공조를 전제하고 있어 실효성은 美 단독 입법만으로 판단하기 어려우며, 향후 美·中 등 주요국 간 AI 안전 논의의 진전 여부를 함께 지켜볼 필요

참고자료

- 1) U.S. Senator Bernie Sanders, 'NEWS: Sanders, Casar Introduce Legislation to Create New Federal Agency to Ban Artificial Superintelligence, Pause Advanced AI Development'(2026.9.23.)
<https://www.sanders.senate.gov/press-releases/news-sanders-casar-introduce-legislation-to-create-new-federal-agency-to-ban-artificial-superintelligence-pause-advanced-ai-development/>
- 45) U.S. Senator Bernie Sanders, 'Ban Artificial Superintelligence Act of 2026'(2026.9.23.)
<https://www.sanders.senate.gov/wp-content/uploads/Ban-Artificial-Superintelligence-Act.pdf>
- 46) U.S. Senator Bernie Sanders, 'Ban Artificial Superintelligence Act - Section by Section'(2026.9.23.)
<https://www.sanders.senate.gov/wp-content/uploads/Ban-Artificial-Superintelligence-Act-Section-by-Section.pdf>
- 47) AP News, 'Sen. Bernie Sanders unveils bill to ban artificial superintelligence and create Department of AI'(2026.9.23.)
<https://apnews.com/article/d6a6a2cd8e71d7c95bdc394595aa85d1>
- 48) Roll Call, 'AI 'superintelligence' ban proposed by Casar, Sanders'(2026.9.23.)
<https://rollcall.com/2026/09/23/ai-superintelligence-ban-proposed-by-casar-sanders/> iitp34
- 49) Interesting Engineering, 'Proposed Sanders bill permanently outlaws frontier superintelligence'(2026.9.23.)
<https://interestingengineering.com/ai-robotics/bernie-sanders-superintelligence-ban-bill> iitp34
- 50) Common Dreams, 'Sanders, Casar Unveil Bill to Ban AI Superintelligence That "Could Kill Countless Numbers of People"'(2026.9.23.)
<https://www.commondreams.org/news/sanders-casar-superintelligence-ban>

2 마·중, AI 안전 대화 채널 '인시던트 라인' 신설 합의

사고 통보 채널 '인시던트 라인'으로 위험 공동 관리에 합의

▶ 美·中, 뉴욕 고위급 회담에서 'AI 대화' 공식 출범 합의

- 美 베선트 재무장관과 中 허리핑 부총리가 뉴욕 고위급 회담에서 '미·중 AI 대화' 공식화에 합의, 세계 AI 개발을 양분하는 두 나라가 AI 위험을 정기 논의하는 첫 공식 채널이 출범 ('26.9.20.)
- AI 안전사고가 발생하면 양국이 서로 즉시 알릴 수 있는 전용 소통 채널인 이른바 '인시던트 라인'을 개설하는 데에도 합의, '26.11월 中 선전에서 후속회담을 개최할 예정
- * 인시던트 라인(Incident Line) : 사고(Incident) 발생 시 즉시 연락하는 당국 간 직통 소통망으로, 냉전기 미·소 핫라인의 위기관리 기능을 AI 안전에 적용한 개념
- 기술 패권을 다투는 양국이 위험 관리 영역에 한해 협력 채널을 마련한 것으로, 경쟁과 공조가 병존하는 '경쟁 속 안전 공조' 구도를 보여주는 합의

▶ AI 보안 사고 현실화 속 정상 차원 약속으로 격상

- '26.7월 오픈AI의 AI 모델이 보안 평가 도중 격리된 시험 환경을 벗어나 허킹페이스 시스템에 무단 접근한 사고가 양국을 대화로 이끈 직접적 배경으로 거론
- * 오픈AI가 사이버 공격 능력을 측정하려 안전장치를 낮춘 채 진행한 내부 평가에서, 모델들이 부여된 과제의 의도와 달리 허킹페이스 데이터베이스에 접근해 평가 문제의 정답을 얻으려 한 사건(OpenAI, '26.8.)
- 시진핑 주석의 국빈 방미 기간('26.9.23~25.) 양 정상에 도출한 8개 성과·양해 항목에 'AI 대화 설립'과 'AI 사고 양자 소통 채널 구축'이 공식 명시, 실무 합의가 정상 차원 약속으로 격상
- AI가 미·중 정상 외교의 의제로 진입하면서, 장관급 실무 합의에 정상 간 공동 발표라는 정치적 무게가 더해져 후속 논의의 추진 기반이 마련

사고 통보·위험 정의 합의와 美 개발사 책임 원칙

(1) AI판 핫라인 '인시던트 라인'으로 사고 공동 관리 첫발

▶ 정례 대화와 상시 통보 채널의 이중 구조로 설계

- 인시던트 라인은 정례 협의체 '미·중 AI 대화'와 별도로 상시 가동되는 사고 통보 전용 채널로, 정례 대화가 위험 인식을 조율하는 협의 무대라면 인시던트 라인은 긴급 사건 통보 창구
- AI 사고 정보가 국경을 넘어 양국 당국 간에 직접 오가는 상시 장치가 마련된 것으로, AI 사고를 개별국 내부 문제가 아닌 양국 공동 관리 대상으로 본다는 인식을 반영한 결과로 해석

▶ AI가 군사와 나란히 미·중 위기관리 정식 축으로 편입

- 양국은 AI 채널과 별도로 군 당국 간 위기 소통·예방 양해각서(MOU)도 조속히 체결하기로 합의, 갈등이 사고로 번지기 쉬운 군사와 AI 두 분야의 위기관리 장치가 나란히 추진되는 양상
- 베센트 장관은 트럼프 대통령과 시진핑 주석의 개인적 관계를 실무 차원으로 연결하려 한다고 언급, 정상 간 신뢰에 의존하던 위기 소통이 군사·AI 두 분야의 상설 채널로 구조화되는 흐름
- 우발적 충돌을 막기 위한 소통 장치가 군사에서 AI 영역으로 확장된 것으로, 갈등 가능성이 큰 분야일수록 대화 창구부터 마련해 온 전통적 위기관리 방식이 AI 분야에 그대로 적용된 사례

(2) 美·中이 함께 정하는 'AI 위험 정의' 프로토콜 논의 개시

▶ 위험의 공통 정의 합의가 프로토콜 논의의 출발점

- 베센트 장관은 프로토콜 논의를 시작해 양측이 주요 AI 위험이 무엇인지 합의하는 것이 목표라고 언급, 대응 절차를 정하기에 앞서 위험의 정의부터 맞추는 단계적 접근으로 풀이
- 우선 거론된 위험 유형은 통제 불가능한 AI 에이전트와 사이버 공간의 비국가 행위자로, 상대국의 기술 개발이 아니라 통제를 벗어난 AI와 제3자 악용이라는 공통 위험에 초점을 맞춘 구성

* 비국가 행위자(Non-state Actor) : 특정 정부에 소속되지 않은 채 국제적 영향력을 행사하는 집단으로, 해커 조직, 테러 단체 등이 해당

- 규제 체계가 다른 양국이 위험을 부르는 공통 언어부터 만드는 방식으로, 정의 합의가 이뤄지면 인시던트 라인의 통보 대상과 대응 절차 등 후속 규범을 정하는 기준점 역할을 할 것으로 전망

▶ '26.11월 선전 후속회담에서 논의 본격화 예정

- 후속회담 시기에는 中 주최 APEC 정상회의가 같은 선전에서 개최되어 트럼프 대통령도 각료를 대동해 참석할 것으로 예상되는 등 AI 논의가 정상 외교 일정과 맞물려 함께 진행될 전망
- AI 개발을 주도하는 G2가 위험의 정의에 먼저 합의할 경우, 선전 회담 결과는 향후 다자 차원에서 진행될 AI 안전 논의에서 위험을 분류하고 대응하는 사실상의 출발점이 될 가능성

(3) 美, 사전 규제 대신 개발사 책임 원칙으로 AI 안전 관리 기조 제시

▶ '사고 책임은 개발사'라는 원칙을 공식화

- 베센트 장관은 회담 직후 인터뷰에서 허깅페이스 해킹 사건을 AI 에이전트가 아니라 오픈AI 경영진의 책임이라고 규정, AI가 일으킨 사고라도 책임은 개발 주체에 있다는 원칙을 공식 표명

- 베센트 장관은 기업이 원한다면 개발 속도를 늦출 수 있다며 속도 조절 권한이 기업에 있음을 책임의 근거로 제시, AI 기업의 면책 요구도 국민에게는 나쁜 거래라며 거부 방침을 명확화
- 관련 법 정비 전에 행정부 발언으로 책임 소재의 기준선이 먼저 제시된 것으로, 사고 발생 시 'AI 탓'이라는 방어 논리를 인정하지 않겠다는 美 행정부의 입장이 분명하게 드러난 대목

▶ 성장 우선 기조 속 사전 규제 대신 사후 법 집행으로 통제

- 트럼프 대통령은 AI를 산업혁명이나 인터넷보다 큰 성장 동력으로 규정하며 성장을 억누르지 않겠다고 언급, 법 집행 기관이 필요할 때 통제하는 사후 대응 방식을 규제의 기본 틀로 제시
- 미국 내에서는 AI가 인류를 위협할 수 있다며 강력한 규제를 요구하는 진영과 초지능(SI) 개발까지 장려하는 트럼프 대통령의 성장 기조가 맞서며 규제 수위 논쟁이 지속되는 상황
- 사전 규제 중심의 EU 방식과 달리 △성장 촉진 △기업 자기 책임 △사후 법 집행을 포함한 접근이 美 행정부의 AI 거버넌스 방향으로 제시된 것으로, 안전 관리를 기업 자율에 맡기는 구조

AI 공조 첫발에도 곳곳에 남은 한계와 변수

▶ 통상·제재 갈등에 연동되는 선택적 공조라는 한계 존재

- AI 협력 합의에도 통상·제재 등 다른 의제에서는 불확실성과 이견이 그대로 남아, 이번 협력이 관계 전반의 화해가 아니라 사안별로 범위를 정한 선택적 공조라는 성격 확인
- 미·중 관세 휴전은 이번 회담에서 논의됐으나 만료('26.11.10.)를 앞두고 연장 합의 여부가 명확히 밝혀지지 않았고, 이란 제재를 두고도 양국이 공개적으로 입장 차를 드러내는 상황
- 선전 회담 시기가 관세 휴전 연장 협상과 겹쳐 AI 대화도 통상 협상 흐름에 따라 진전되거나 지연될 수 있는 구조로, 채널의 지속성이 양국 관계 전반의 안정에 좌우된다는 한계 내포

▶ 강제 장치 없는 소통 절차 수준에 머문 초기 합의

- 이번 합의는 양국 간 대화 정례화와 사고 통보 채널 개설이라는 소통 절차 수준에 머물러 있어, 위험 판정 기준이나 이행을 담보할 검증·강제 장치는 아직 마련되지 않은 초기 단계로 평가
- 합의가 선언 단계를 넘어 구속력 있는 국제 규범으로 발전할지는 아직 미정으로, 당분간은 양국 간 이견 속에서도 채널을 계속 가동하며 상호 신뢰를 단계적으로 쌓아 가는 단계에 머물 전망

시사점

경쟁 속 AI 안전 공조 출범, 위험 정의 논의와 책임 원칙에 선제 대응 필요

▶ 경쟁 속 위험 공동 관리가 글로벌 AI 거버넌스의 새 축으로

- 이번 합의의 본질은 AI 안전을 패권 경쟁과 분리해 관리하는 방식이 G2 사이에 제도화되기 시작한 것으로, 규제 경쟁 일변도이던 글로벌 AI 질서에 협력 축이 더해진 계기
- 다만 합의가 아직 소통 절차 수준에 머물러 있고 통상 갈등과도 연동되는 만큼, '26.11월 선전 회담에서 위험의 공통 정의에 합의할 수 있을지가 채널의 실효성을 가늠할 1차 관문
- G2의 위험 정의·통보 기준이 다자 논의의 준거가 될 가능성이 큰 만큼, 국내에서도 선전 회담 결과를 AI 안전 정책과 국제 규범 대응에 반영할 수 있도록 동향 추적과 참여 경로 검토 필요

▶ 美 개발사 책임 원칙 공식화로 안전 역량이 산업 경쟁 요소로 부상

- 美 행정부가 면책 요구를 거부하고 개발사 책임 원칙을 공식화하면서, 글로벌 AI 기업에는 모델 성능 경쟁과 별개로 사고 예방·대응 체계 구축이 사업 지속의 전제 조건이 되는 환경 조성
- 통제가 불가능한 AI 에이전트가 양국 간 공식 위험 유형으로 논의되는 만큼, 에이전트 행동을 검증하고 이상 작동을 통제하는 안전 기술이 새로운 기술 수요이자 투자 영역으로 부상할 전망
- 국내 AI 기업도 美 시장·美 기업과의 협업에서 동일한 책임 원칙의 적용 대상이 될 수 있는 만큼, 에이전트 안전 검증 역량과 사고 대응 체계를 선제적으로 갖추는 것이 시장 신뢰 확보의 조건

참고자료

- 1) Reuters, 'US, China to meet again on AI safety in two months in Shenzhen, Bessent says' (2026.09.21.)
<https://www.reuters.com/world/asia-pacific/us-china-meet-again-ai-safety-two-months-shenzhen-bessent-says-2026-09-21/>
- 2) 와이타임즈(Why Times), '베센트 "미중 AI대화 공식화"..."두 달 뒤 中 선전서 후속회담"' (2026.09.23.)
<https://whytimes.kr/news/view.php?idx=28172>
- 3) The Washington Times, 'U.S., China to open communication line on alarming AI incidents' (2026.09.21.)
<https://www.washingtontimes.com/news/2026/sep/21/us-china-open-communication-line-alarming-ai-incidents/>
- 4) Ministry of Foreign Affairs of the People's Republic of China, 'China and the United States Reach Eight Deliverables and Understandings' (2026.09.26.)
https://www.fmprc.gov.cn/eng/xw/zyxw/202609/t20260926_12031663.html
- 5) OpenAI, 'The Hugging Face incident and the road ahead' (2026.08.26.) <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>

III . ICT 부문별 모니터링

구 분	품 목	주요 이슈	영 향
ICT 융합기술	소재, 부품, 장비	AI 데이터센터 CPO 양산 임박, 인터페이스·수율·유지보수가 관건 AI 컴퓨팅 대역폭 확대에 따라 광연결 구조가 플러거블 모듈에서 NPO·CPO로 이동하면서 양산 경쟁 본격화. CPO 확산에는 탈착형 광섬유 인터페이스와 함께 224G 신호, 고전압 전력, 액체냉각을 동시에 고려하는 랙 단위 공동설계 역량 필요. 광부품이 칩에 가까워질수록 정렬 정밀도와 수율, 현장 교체·정비 문제가 시스템 신뢰성을 좌우하기 때문. 외부 레이저, InP 소재, 첨단 패키징, 광축정 및 OIF·MSA 표준 성숙도가 대규모 도입 속도를 결정할 전망	중립적
반도체 & 양자	지능형 반도체	애플, 2나노 A20 Pro·신형 패키징으로 모바일 AI 성능 경쟁 가속 애플이 A20 Pro에 2나노 공정과 6코어 CPU, 7코어 GPU, 듀얼 16코어 Neural Engine을 적용하며 모바일 AI 연산 성능 강화. 메모리 대역폭 50%, GPU 성능 최대 40%, AI 성능 2배 향상이 예상되며 DRAM을 프로세서 옆에 배치하는 WMCM 구조와 증기 챔버 냉각도 도입. 고성능 온디바이스 AI 구현 과정에서 발생하는 발열과 메모리 병목을 패키징 단계에서 완화하려는 전략으로 분석. 퀄컴과 미디어텍도 2나노와 유사 패키징 채택을 추진하면서 차세대 모바일 칩 경쟁 확대 전망	중립적
	반도체	대만 N&M, CMOS 기반 적외선 칩으로 로봇 '전자코' 시장 공략 대만 N&M이 CMOS 공정으로 4개 이상 중·원적외선 파장을 방출하는 단일 칩 개발 후 테이프아웃·검증과 고객 대상 POC 진행. 기존 III-V 화합물 반도체 방식 대비 비용 50% 이상, 칩 크기 80~90% 축소를 목표로 산업·환경 가스 감지와 로봇 전자코 시장 진입 추진. CMOS 기반 집적성을 활용해 로직과 광전자 소자의 결합 부담을 줄이는 반면 광출력이 낮아 장거리 통신보다 단거리 센싱에 집중. 실제 고객 검증과 8·12인치 공정 전환, 다중 가스 인식 성능 확보가 상용화의 주요 변수	중립적
보안 & 블록체인	ICT 데이터 및 응용서비스 보안	코히어, AI 추론 과정까지 암호화하는 'Model Vault' 공개 코히어가 단일 고객용 AI 추론 서비스 Model Vault에 기밀 컴퓨팅을 적용해 저장·전송뿐 아니라 실제 추론 중인 데이터까지 보호하는 기능 상용화. Intel TDX·AMD SEV-SNP 기반 CPU와 엔비디아 기밀 컴퓨팅 GPU를 결합해 메모리와 CPU·GPU 경로를 보호하며, 고객은 추론 결과와 함께 하드웨어·소프트웨어 상태를 검증하는 증명 보고서 확보 가능. 기업 AI 도입 과정에서 공급자나 클라우드 운영자가 민감정보에 접근할 수 있다는 신뢰 문제를 기술적 검증 구조로 보완하려는 접근. 검증이 개별 요청마다 이뤄지는지 여부와 오픈소스화 이후 독립 검증 수준이 향후 경쟁력 판단 요소	중립적
	차세대 보안	벨기에 Aikido, 로컬 실행형 오픈웨이트 사이버보안 AI 모델 공개 벨기에 사이버보안 기업 Aikido가 민감한 소스코드를 외부 사업자에 전송하지 않고 자체 환경에서 활용할 수 있는 오픈웨이트 AI 보안 모델 'Altar' 공개. 중국 Z.AI의 GLM-5.3을 압축·맞춤화한 모델로 로컬 배포를 지원하며 벨기에 은행 Belfius 등 기존 고객 솔루션에도 적용 예정. 공격자가 AI를 활용해 소프트웨어 취약점을 탐색·악용하는 사례가 늘면서 방어 측에서도 보안 데이터 통제와 AI 자동화를 동시에 확보하려는 수요가 확대된 데 따른 움직임. 로컬 보안 AI의 탐지 성능과 기업 시스템 통합성, 운영비용이 시장 확산의 주요 변수	중립적

구 분	품 목	주요 이슈	영 향
디지털사회 혁신	에너지 전환	구글·엔비디아, 전력 수요 조절형 AI 데이터센터 연합 출범 구글과 엔비디아가 Emerald AI와 함께 전력 수요를 유연하게 조절하는 데이터센터 확산을 위한 신규 연합 출범. 데이터센터가 전력 사용량을 상황에 따라 조정하는 방식을 통해 전력가격 부담을 낮추고 지역사회의 반발을 완화하면서 AI 인프라 증설 여력을 확보한다는 구상. AI 데이터센터 급증으로 전력망 부담과 비용 문제가 인프라 확장의 주요 제약으로 부상한 상황에서 컴퓨팅 수요 자체를 전력계통에 맞춰 조정하려는 접근. 구체적인 운영 방식과 전력사업자 참여 구조, 실제 비용 절감 효과가 향후 확산 여부를 판단할 변수	중립적
인공지능 & 데이터	초거대 인공지능	알리바바, 초대형 Qwen·자체 AI 칩·데이터센터로 AI 수직통합 강화 알리바바가 5조~10조개 매개변수를 목표로 하는 차세대 Qwen 모델 개발과 자체 AI 칩 'Zhenwu V900' 공개를 병행하며 AI 전 계층 투자 확대. V900은 이전 세대 대비 3배 성능과 최대 50만개 칩 클러스터 구성을 목표로 2027년 1분기 양산 예정이며, 글로벌 데이터센터 용량도 2032년 20GW 이상으로 확대 계획. 미국의 첨단 AI 칩 수출 제한 속에서 모델·반도체·데이터센터를 자체 생태계로 묶어 공급 제약을 완화하려는 전략으로 분석. 실제 칩 양산성과 대규모 모델 학습 효율, 데이터센터 공급능력 확보가 성장 속도의 핵심 변수	중립적
콘텐츠 & 미디어	IT 플랫폼	메타, 페이스북·인스타그램·왓츠앱 통합 구독 'Meta One' 출시 메타가 페이스북·인스타그램·왓츠앱·Meta AI를 아우르는 단일상품 및 번들 구독 서비스 'Meta One' 출시. 핵심 앱과 Meta AI의 무료 이용은 유지하면서 AI 사용량 확대와 개인 표현 기능, 크리에이터·기업용 전문 기능을 유료 서비스로 구성했으며 이미 50개 이상 기능과 1,500만건의 구독·채널 기록 확보. 광고 중심 수익구조를 보완하고 이용자 유형별 유료 기능을 묶어 플랫폼 수익원을 다변화하려는 전략으로 분석. 단계적 출시 과정의 유료 전환율과 EU 디지털시장법 관련 규제 대응이 향후 확산 변수	중립적
	차세대 미디어	스냅, AI·기업 기능 확대로 2,200달러 스마트글래스 활용처 확대 스냅이 2,200달러 스마트글래스 'Specs'에 사용자의 목표·일정·관계를 연결하는 'Specs Intelligence'와 HBO Max·Spotify 연동 기능을 추가하며 활용 범위 확대. 아마존·세일즈포스·엔비디아 등과 'Specs for Enterprise'를 추진하고 모바일 기기 관리와 보안 제어, 기업용 배포 기능도 개발. 소비자용 AR 기기를 넘어 AI 생산성 도구와 기업 현장용 웨어러블 컴퓨팅 플랫폼으로 입지를 넓히려는 전략으로 분석. 높은 가격과 무거운 폼팩터, 구체적 활용사례 부족을 극복하고 기업 도입 효과를 입증할 수 있는지가 시장 확대의 관건	중립적
통신 & 네트워크	통신장비	화웨이, AI 인프라용 7.2Tbps NPO 모듈 공개 화웨이가 중국국제광전자박람회에서 업계 최초라고 주장하는 7.2Tbps급 근접 패키징 광학(NPO) 모듈 공개. 200Gbps로 동작하는 36개 채널을 적용하고 광학 부품을 네트워크 스위칭 칩 가까이 배치해 신호 손실과 전력소비를 줄이는 구조 채택. 구리 연결의 대역폭·전력 한계를 광신호로 대체해 AI 인프라 병목을 완화하려는 전략이며 현재 공급망과 제조체계 구축을 전제로 양산 준비 단계 진입. 향후 수출과 대량생산 안정성, 브로드컴 등 CPO 진영과의 경쟁이 상용화 속도를 좌우할 전망	중립적

Ⅳ . 주요 ICT 행사 일정

■ 해외 전시회, ■ 국내 전시회

9/28	9/29	9/30	10/1	10/2
2026 미국 라스베이거스 AI·디지털 인프라 전시회 [Yotta] (클라우드 서비스), 미국 라스베이거스				
	2026 영국 런던 국제 사이버보안 전시회 [International Cyber Expo] (차세대 보안), 영국 런던			
	2026 미국 애틀랜타 광대역통신 기술 전시회 [SCTE TechExpo] (차세대 통신), 미국 애틀랜타			
		2026 국제첨단소재기술대전 [INTRA] (소재), 서울 코엑스		
10/5	10/6	10/7	10/8	10/9
2026 UAE 두바이 방송·위성통신 전시회 [CABSAT] (통신), UAE 두바이				
	2026 인공지능 페스타 [AI Festa] (인공지능), 서울 코엑스			
	2026 스마트라이프워크 [SLW] (ICT 융합 서비스), 서울 코엑스			
		2026 네덜란드 암스테르담 세계 인공지능 서밋 [World Summit AI] (인공지능), 네덜란드 암스테르담		
		2026 인도 뉴델리 이동통신 산업 전시회 [IMC] (차세대 통신), 인도 뉴델리		
10/12	10/13	10/14	10/15	10/16
2026 미국 캘리포니아 오픈 컴퓨트 글로벌 서밋 [OCP] (장비), 미국 캘리포니아				
	2026 미국 캘리포니아 반도체 전시회 [SEMICON West] (반도체), 미국 캘리포니아			
	2026 일본 지바 첨단기술 종합 전시회 [CEATEC] (디지털 전환), 일본 지바			
	2026 한국전자전 [KES] (종합), 서울 코엑스			
		2026 반도체대전 [SEDEX] (반도체), 서울 코엑스		
10/19	10/20	10/21	10/22	10/23
2026 미국 라스베이거스 금융기술 전시회 [Money20/20 USA] (ICT 융합 서비스), 미국 라스베이거스				
		2026 일본 지바 IT 종합 전시회 [Japan IT Week] (종합), 일본 지바		
		2026 산업AI 엑스포 (인공지능), 고양 킨텍스		
		2026 국제 모빌리티 산업전 [KOAA] (자율주행차), 고양 킨텍스		



- ▶ 본 저작물은 정보통신기획평가원에서 작성하여 공공누리 '제2유형(출처 표시+상업적 이용금지)', 'AI 유형(인공지능 학습용)'으로 개방하였으며, 기관 홈페이지 (www.iitp.kr)에서 무료로 다운로드 받으실 수 있습니다.
- ▶ 본 내용은 정보통신기획평가원 동향분석팀에서 수행하는 「ICT 동향분석 및 정책지원」 사업의 연구 결과물입니다.
- ▶ 본 내용은 작성자의 주관적인 의견으로 정보통신기획평가원의 공식적인 입장이 아님을 밝힙니다.