주요 동향(2) : ICT

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 안전 장치를 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 AI의 통제 공백과 안전 확보를 위한 기술·제도 대응 동향

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

○ 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
- 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
- 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 낳을 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

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

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

● 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

- 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
- 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
- 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
  - 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 정상적이더라도 인증 정보와 업무 기록의 유출에 대비한 보호 조치가 필요

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

(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
- 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
  - 오픈AI는 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
  - 앤트로픽은 특정 개념과 관련된 내부 신호를 조정한 뒤 행동 변화를 비교해, 추론 기록에 표현되지 않은 요인도 모델의 선택에 영향을 줄 수 있음을 실험적으로 확인
  - 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
  - 오픈AI는 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

(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
  - 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 과정도 통제 대상임을 확인

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

(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
  - 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
- 기준 적용: 위험 평가를 실제 출시 보류와 학습 중단 결정에 반영
  - 문서로 정한 기준의 실효성은 위험을 확인한 뒤 실제로 어떤 결정을 내렸는지에서 드러나며, 앤트로픽의 일반 출시 보류와 오픈AI의 학습 중단은 각각 배포와 개발 단계에 안전 판단을 적용한 사례
  - 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
  - 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
  - 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
- **(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
  - 미국이 기준의 적용을 기업 내부 절차에 맡긴다면 영국은 정부가 직접 시험하되, 모델 접근과 시정이 기업의 자발적 협력에 의존하므로 평가환경 보안과 사고 대응 역량을 보강해 평가 기반을 강화
- **(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
  - 허용된 업무라도 실행 과정에서 권한을 벗어나지 않도록 최소권한과 고위험 작업의 추가 확인을 요구하며, 디지털 신원과 행동 기록을 통해 정해진 실행 조건이 지켜졌는지 확인하도록 제시
  - 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
  - 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
  - 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
  - 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

출처 : International AI Safety Report 외(2026.2.)

https://www.uneca.org/stories/key-risk-factors-for-ai-loss-of-control-came-together-in-2026-incident,-independent-un
https://metr.org/blog/2026-05-19-frontier-risk-report/
https://www.anthropic.com/research/measuring-agent-autonomy
https://internationalaisafetyreport.org
https://alignment.openai.com/misalignment-reports/self-generated-prompt-injections-in-compaction-summaries/
https://alignment.openai.com/misalignment-reports/encouraging-deception-in-compaction-summaries/
https://www.gendigital.com/blog/insights/research/infostealers-your-ai-agent
https://www.anthropic.com/responsible-scaling-policy
https://openai.com/index/openai-frontier-governance-framework/
https://deepmind.google/blog/strengthening-our-frontier-safety-framework/
https://www.mindstudio.ai/blog/anthropic-pace-the-frontier-ai-labs
https://openai.com/index/updating-our-preparedness-framework/
https://www.redwoodresearch.org/blog/retrying-vs-resampling-in-ai-control
https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/
https://alignment.openai.com/auto-review
https://doi.org/10.1145/3805621.3807645
https://alignment.anthropic.com/2026/teaching-claude-why/
https://alignment.openai.com/how-far-does-alignment-midtraining-generalize/