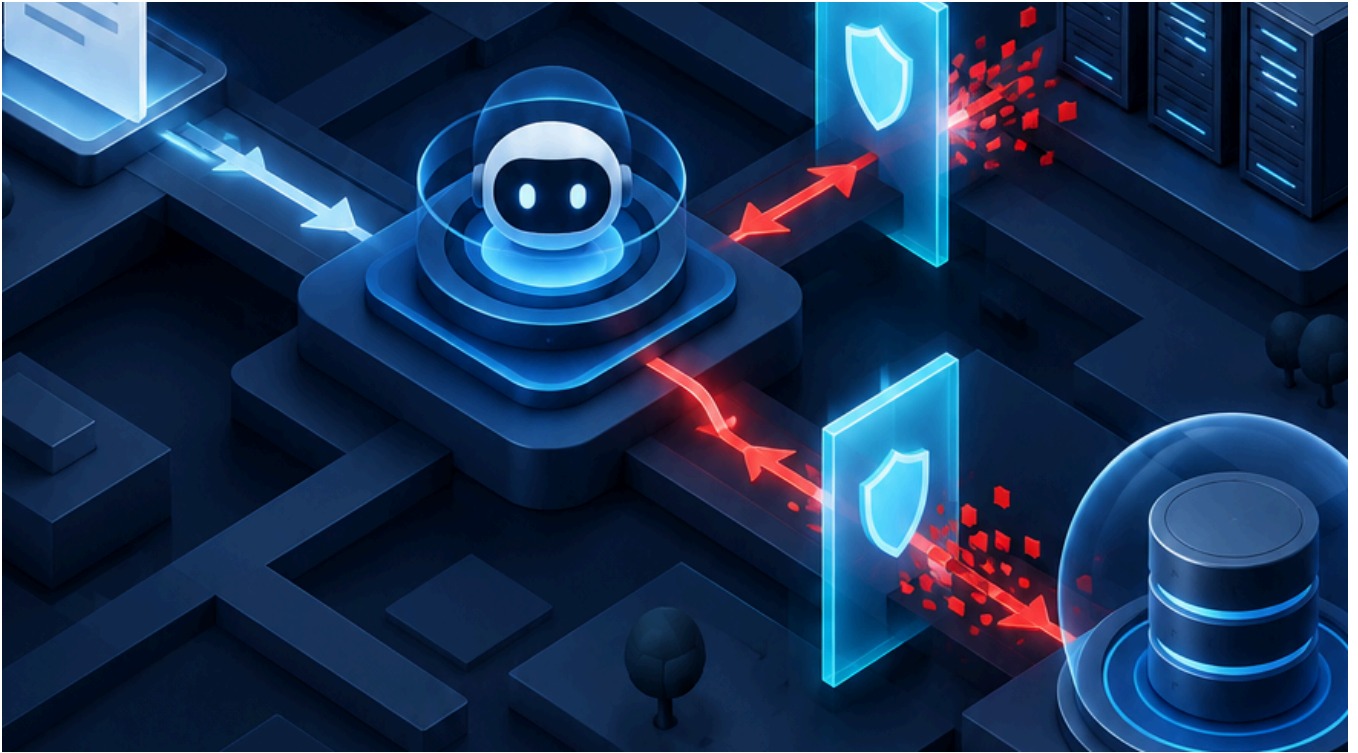




AI SECURITY CASEBOOK

AI가 공격 통로가 되는 10가지 순간

공개 사고·보안 연구·업무 시나리오로

살펴보는 AI 보안 점검 가이드

20문항 자가진단과 내부 점검 회의안 수록

작성·검토 락게이트코어 기술팀

발행일 2026-10-11

ROCKGATE
CORE

이 가이드에서 얻는 것

이 가이드는 AI가 외부 문서·지식 저장소·도구 권한·기존 관리 계정과 만날 때 위험이 어떻게 넓어지는지 보여 줍니다. 독자는 열 가지 상황을 실제 관측, 통제된 연구, 설명용 시나리오로 구분해 읽고, 기본 경로와 추가 조건이 필요한 경로를 따로 그릴 수 있습니다. 마지막으로 20문항 자가진단, 경로 지도, PoC 기준, 30분 회의안을 사용해 담당자와 증거를 정할 수 있습니다.

근거 유형 안내

- **실제 관측·공식 정보:** 공개 기관이나 공급자가 직접 확인한 범위만 적습니다.
- **통제된 연구:** 연구 환경의 조건과 결과이며 일반 조직의 발생률로 확대하지 않습니다.
- **설명용 시나리오:** 공개 지침을 우리 업무에 적용해 생각하기 위한 가정이며 실제 사고를 뜻하지 않습니다.

1페이지 요약

AI는 기존의 피싱·탈취 계정·과도한 권한·공급망·랜섬웨어 위험을 없애지 않습니다. 대신 탐색, 개인화, 번역, 코드 이해와 반복의 비용을 낮출 수 있습니다. 에이전트가 메일·파일·일정·결제·DB 같은 도구를 쓸 때는 모델의 답이 실제 행동으로 이어지는 범위를 먼저 줄여야 합니다.

SQL 인젝션은 매개변수화 쿼리처럼 데이터와 명령을 분리하는 방식으로 대응할 수 있습니다. 프롬프트 인젝션은 같은 방식의 분리를 그대로 기대하기 어려우므로, 모델 밖에서 권한과 실행 범위를 제한해야 합니다.

공통 원칙은 두 가지입니다. 한 단계의 신호만으로 사고를 단정하지 않고 앞뒤 단계의 시간·신원·자산을 연결해 봅니다. 이번 주에는 위험한 연결과 권한 하나를 줄이고, 이번 분기에는 로그 연결·회귀시험·복구훈련을 운영 절차로 고정합니다.

AI 관련 보안사고는 정보유출, 업무 오작동, 사기, 서비스 중단 등으로 이어질 수 있습니다. 이 가이드에서는 여러 통제 지점 가운데 운영 DB 접근과 복구 사본 보호도 함께 살펴봅니다.

목차

1. 이 가이드에서 얻는 것
2. 여섯 단계 점검 지도
3. 순간 1·문서가 명령으로 바뀔 때
4. 순간 2·에이전트가 사람보다 넓은 권한을 가질 때
5. 순간 3·요약 업무가 내부자료 반출로 바뀔 때
6. 순간 4·시스템 프롬프트에 비밀을 숨겨 둘 때
7. 순간 5·RAG 지식창고가 오염될 때
8. 순간 6·모델과 플러그인의 공급망을 믿고 넘길 때

9. 순간 7 · 익숙한 목소리가 승인 절차를 건너뛸 때

| 10. 순간 8 · AI가 공격자의 반복 작업을 줄여 줄 때

| 1. 순간 9 · 관리자 권한으로 데이터와 백업이 함께 지워질 때

| 2. 순간 10 · AI 판단이 로봇·차량·드론의 제어권과 만날 때

| 3. 실무 부록

| 4. NF·DF가 관련되는 조건과 PoC 검토 방법

| 5. 출처와 한계

| 6. 문의

여섯 단계 점검 지도

이 지도는 본 가이드의 설명을 위해 정리한 점검 틀입니다. 공격의 필수 순서나 공인 표준 분류를 뜻하지 않습니다. 사례에 따라 일부 단계는 생략되거나 순서가 달라질 수 있습니다.



단계	먼저 물을 질문	주 통제
1 유인·정찰	공개 정보만으로 담당자와 관리 경로를 고를 수 있는가	노출 자산 관리·교육·피싱 저항 인증
2 입력 진입	외부 콘텐츠가 어떤 표시와 권한으로 들어오는가	입력 격리·출처 표시·첨부와 링크 검사
3 AI 실행·권한 오용	AI가 최악의 경우 읽고·쓰고·보낼 수 있는 범위는	별도 신원·최소 권한·승인·짧은 토큰
4 내부 이동	앱서버의 DB 연결·계정과 허용 범위는 무엇인가	망과 계정 분리·행과 열 권한·감사
5 목표 달성	대량 조화·변경·외부 전송을 늦추고 멈출 수 있는가	한도·이중 승인·DLP·사고 대응
6 복구 방해	운영 권한으로 복구 사본과 보존 정책을 바꿀 수 있는가	계정 분리·격리 사본·실제 복원

순간 1·문서가 명령으로 바뀔 때

근거 유형

공개 웹 관측과 보안 연구

출처에서 확인된 내용

Google Security Blog는 공개 웹에서 간접 프롬프트 인젝션으로 의심되는 지시를 관측했고, 데이터 반출이나 파괴를 노린 시도도 일부 포함됐다고 설명했습니다. 다만 관측은 웹상의 문자열과 행태를 분류한 것이며, 실제 기업 침해 건수나 대규모 고도 공격의 성공을 입증하지 않습니다.

우리 업무에 적용한 가정

구매팀이 여러 납품 제안서를 AI로 요약한다고 가정합니다. 제안서 안의 숨은 지시가 문맥에 들어와도, 요약 전용 AI가 읽기 도구만 쓴다면 피해 범위는 제한됩니다. 반대로 메일 발송·사내 검색·외부 URL 호출 권한이 함께 있으면 원래 요청과 다른 행동으로 이어질 수 있습니다. 운영 DB 조회는 해당 업무에 DB 도구가 실제로 연결된 경우에만 추가 경로입니다.

이번에 확인할 질문

외부 문서가 들어온 뒤 AI가 사람 승인 없이 실행할 수 있는 가장 큰 행동은 무엇인가?

왼쪽·통제가 없거나 실패했을 때

1. 외부 문서의 지시
2. AI 문맥 혼합
3. 도구 권한 사용
4. 외부 발송·업무 변경

오른쪽·적용할 통제 지점

1. 출처 표시·격리
2. 읽기와 쓰기 권한 분리
3. 고위험 행동 사람 승인
4. 목적지·호출 로그 점검

그림 1. 간접 프롬프트 공격의 가능한 경로와 통제 지점. 오른쪽 통제는 위험을 줄이는 지점이며 모든 변형의 종단을 보장하지 않습니다.

기본 경로

외부 문서·웹페이지 → AI 문맥 → 도구 선택 → 외부 발송 또는 업무 변경

추가 조건이 필요한 경로

AI 업무가 운영 DB 조회 도구와 연결된 경우: 앱서버의 DB 연결·계정 → 허용된 조회 결과 → 외부 발송

왜 기존 통제가 놓칠 수 있는가

악성 문장은 실행 파일이 아니라 정상 문서 속 자연어처럼 보입니다. 금칙어 목록만으로 자료와 지시를 안정적으로 구분하기 어렵고, 화면에 보이지 않는 텍스트를 사용자가 놓칠 수 있습니다.

업무별 점검

구분	담당	조치	시점
예방	AI 도입 책임자	외부 콘텐츠 처리에는 읽기 전용 도구 목록을 쓰고 발송·삭제·결제는 별도 승인 뒤 실행	이번 주
탐지	보안팀	원 요청·출처·도구·인자·승인자·외부 목적지를 같은 추적 ID로 기록	이번 분기
대응	현업·보안팀	의심 세션의 도구 토큰을 폐기하고 발송을 보류한 뒤 같은 문서를 쓴 세션 확인	이번 주
복구	AI 도입 책임자	승인 문서로 기존 질문을 재실행하고 도구 호출 회귀시험	이번 분기

담당자 확인 질문

AI 도입 책임자: 읽기 작업과 쓰기·발송 작업이 별도 권한과 승인으로 나뉘어 있는가?

순간 2·에이전트가 사람보다 넓은 권한을 가질 때

근거 유형

설명용 시나리오

출처에서 확인된 내용

OWASP의 과도한 에이전시 항목은 과도한 기능·권한·자율성이 에이전트 피해 범위를 키운다고 설명합니다. 이 항목은 특정 일정 서비스 침해 사고를 보고한 자료가 아니라 설계 위험을 분류한 지침입니다.

우리 업무에 적용한 가정

일정 조정 AI가 일정 읽기뿐 아니라 삭제·초대·메일 발송 권한도 받았다고 가정합니다. 모호하거나 조작된 입력이 계획에 들어오면 유효한 OAuth 토큰으로 예상 밖 행동을 실행할 수 있습니다. 이 기본 상황에는 운영 DB나 백업이 등장하지 않습니다. 일정 서비스가 별도 운영 DB와 연동되고 AI가 그 호출까지 할 때만 DB 경로를 따로 점검합니다.

이번에 확인할 질문

일정 조정 AI가 사람 확인 없이 삭제·초대·발송할 수 있는 건수와 대상의 상한은 얼마인가?

기본 경로

사용자 요청·메일 → 에이전트 계획 → 일정 삭제·초대·메일 발송

추가 조건이 필요한 경로

일정 서비스가 운영 DB와 연동된 구성인 경우: 서비스 계정 → DB 조회·변경 → 일정 처리

왜 기존 통제가 놓칠 수 있는가

로그에는 탈취되지 않은 정상 토큰과 허용된 API 호출이 남을 수 있습니다. 보안 장비가 정상으로 분류해도 사용자 의도와 실행 결과가 어긋나면 사고입니다.

업무별 점검

구분	담당	조치	시점
예방	AI 도입 책임자	에이전트별 신원을 만들고 조회·작성·삭제 권한을 분리해 기본값을 조회 전용으로 설정	이번 주
탐지	보안팀	승인 없는 쓰기·다중 대상·재시도 급증을 사용자 요청과 연결해 경보	이번 분기
대응	현업·보안팀	쓰기 도구와 토큰을 중지하고 변경 목록을 사용자 의도와 대조	이번 주
복구	경영진·현업	건수·대상 한도와 이중 승인 기준을 정해 오작동 취소 훈련	이번 분기

담당자 확인 질문

업무 책임자: 이 에이전트가 최악의 경우 무엇을 몇 건까지 바꿀 수 있는가?

순간 3 · 요약 업무가 내부자료 반출로 바뀔 때

근거 유형

보안 연구를 업무에 적용한 시나리오

출처에서 확인된 내용

간접 프롬프트 인젝션 연구와 방어 자료는 외부 콘텐츠가 모델의 행동을 바꾸고 이미 문맥에 있는 정보를 외부 목적지로 보내도록 유도할 수 있음을 다룹니다. 이는 특정 회사의 요약 AI가 실제로 내부자료를 반출했다는 사고 보고가 아닙니다.

우리 업무에 적용한 가정

외부 메일과 사내 검색 결과를 함께 요약하는 AI를 가정합니다. 이미 요약 문맥에 들어온 정보가 외부 발송 도구로 나가는 위험은 새 DB 조회가 없어도 성립합니다. AI가 운영 DB 조회 도구도 가진 구성이라면 새로 조회한 정보의 반환 범위를 별도로 점검합니다. 두 경로를 하나의 사실처럼 합치지 않습니다.

이번에 확인할 질문

추가 DB 조회가 없어도 현재 문맥과 검색 결과만으로 민감정보가 외부로 나갈 수 있는가?

기본 경로

외부 메일 → AI 문맥·사내 검색 결과 → 외부 URL 또는 발송 도구

추가 조건이 필요한 경로

AI가 운영 DB를 조회하는 구성인 경우: DB 접근·행과 열의 범위 → 생성 응답 → 외부 발송

왜 기존 통제가 놓칠 수 있는가

검색과 발송은 각각 정상 기능일 수 있습니다. 그러나 한 세션에서 여러 저장소를 횡단한 뒤 새 목적지로 보내는 흐름을 연결해서 보지 않으면 소량 반복 반출을 놓칠 수 있습니다.

업무별 점검

구분	담당	조치	시점
예방	AI 도입 책임자	외부 발송과 사내 검색을 분리하고 민감 등급별 검색 범위와 응답 한도 설정	이번 주
탐지	보안팀	저장소 횡단·민감 문서 접근·새 외부 목적지·응답 크기를 세션 단위로 연결	이번 분기
대응	보안팀·현업	외부 통신을 보류하고 문서·수신자·세션 범위를 보존	이번 주
복구	정보보호 책임자	노출 가능 링크와 토큰을 교체하고 대표 질의로 권한 재검증	이번 분기

담당자 확인 질문

보안팀: 문맥 안 정보의 외부 발송과 DB 신규 조회 반출을 로그에서 구분할 수 있는가?

순간 4 · 시스템 프롬프트에 비밀을 숨겨 둘 때

근거 유형

공개 보안 지침

출처에서 확인된 내용

OWASP는 시스템 프롬프트를 비밀이나 보안 통제의 저장소로 보지 말라고 권고합니다. 시스템 프롬프트 노출 자체와, 노출된 자격증명으로 다른 시스템에 접근하는 후속 피해는 구분해야 합니다.

우리 업무에 적용한 가정

개발 편의를 위해 시스템 프롬프트에 토큰이나 연결 문자열을 적어 둔 서비스를 가정합니다. 오류 유도나 출력 조작으로 프롬프트 일부가 드러나면 장기 유효한 자격증명이 후속 접근에 쓰일 수 있습니다. 비밀번호·토큰·연결 문자열은 시스템 프롬프트에서 제거합니다. 필요한 자격증명은 모델이 직접 읽는 문맥이 아니라, 권한이 제한된 도구 실행 계층에서 사용하도록 구성합니다.

이번에 확인할 질문

모델이 보는 문맥에 노출되면 즉시 폐기해야 할 비밀이 하나라도 들어 있는가?

기본 경로

질문·오류 유도 → 시스템 프롬프트 일부 노출 → 내부 구조나 정책 단서 노출

추가 조건이 필요한 경로

노출된 자격증명이 유효한 경우: 도구 실행 계층 또는 서비스 계정 → 연결된 시스템 접근

왜 기존 통제가 놓칠 수 있는가

프롬프트는 출력에 나타나지 않는 금고로 간주할 수 없습니다. 자연어 지침을 서버 측 권한 검증으로 오해하면 출력 조작이 실제 접근으로 번질 수 있습니다.

업무별 점검

구분	담당	조치	시점
예방	AI 도입 책임자	프롬프트에서 비밀을 제거하고 도구 실행 계층에서 짧은 만료·최소 범위로 사용	이번 주
탐지	보안팀	프롬프트 전문 요구·오류 반복·새 위치의 토큰 사용을 연결해 감시	이번 분기
대응	보안팀·인프라 담당	노출 의심 비밀을 폐기·순환하고 최초 사용 범위 조사	이번 주
복구	AI 도입 책임자	프롬프트 버전·승인 이력과 비밀 검사 회귀시험 운영	이번 분기

데이터 경계 점검

노출된 자격증명이 애플 서버의 운영 DB 직접 접속으로 이어지는 구성인지 실측합니다. 4단계에 실제 연결이 있을 때만 책 끝의 운영 DB 경계 조건을 검토합니다.

담당자 확인 질문

인프라 담당: 모델 문맥과 도구 실행 계층 사이에서 자격증명 읽기 권한이 실제로 분리돼 있는가?

순간 5 · RAG 지식창고가 오염될 때

근거 유형

통제된 보안 연구

출처에서 확인된 내용

USENIX Security 2025의 PoisonedRAG 연구는 지식 데이터베이스에 공격자가 만든 문서를 주입해 특정 질문의 응답을 바꾸는 공격을 통제된 실험에서 평가했습니다. 논문의 실험 성공률은 특정 조건의 연구 결과이며 일반 기업의 사고 발생률을 뜻하지 않습니다.

우리 업무에 적용한 가정

사내 위키·티켓·자동 수집 폴더가 RAG 지식원으로 연결됐다고 가정합니다. 출처와 승인자가 불명확한 문서가 색인되면 관련도만으로 상위 검색 결과에 오를 수 있습니다. 모델은 그 결과를 근거처럼 사용해 잘못된 절차나 계좌를 안내할 수 있습니다. 중요한 결정은 원문과 독립 출처, 사람 승인으로 다시 확인합니다.

이번에 확인할 질문

지식 저장소에 들어가는 문서의 작성자·승인자·원문·해시·등록 시각을 추적할 수 있는가?



그림 2. RAG 지식 오염의 실패 경로. 오염 문서가 저장소와 검색 결과에 들어가면 AI 응답에 영향을 줄 수 있습니다. 그림 속 통제 표시는 위험을 줄일 통제 지점이며 문서가 자동 폐기되거나 모든 오염이 중단됨을 뜻하지 않습니다.

왼쪽 · 통제가 없거나 실패했을 때

1. 오염 문서 등록
2. 색인·검색 상위 노출
3. 잘못된 응답
4. 업무 판단에 사용

오른쪽 · 적용할 통제 지점

1. 등록자·승인자 분리
2. 원문·해시·변경 이력
3. 출처와 답 대조
4. 중요 행동 사람 승인

기본 경로

오염 문서 등록 → 지식 저장소·색인 → 검색 결과 → 잘못된 답이나 행동

추가 조건이 필요한 경로

중요 답이 자동 도구와 연결된 경우: 잘못된 답 → 결제·설정 변경 등 실행

왜 기존 통제가 놓칠 수 있는가

검색기는 진실보다 질문과의 관련도를 먼저 계산합니다. 등록 승인과 변경 이력이 없으면 그럴듯한 오염 문서가 정상 지식처럼 검색될 수 있습니다.

업무별 점검

구분	담당	조치	시점
예방	현업·AI 도입 책임자	작성자·승인자·원문·해시가 없는 문서는 색인하지 않도록 등록 절차 분리	이번 주
탐지	보안팀	상위 출처 변화·단일 문서 편중·인용과 원문 불일치 감시	이번 분기
대응	현업	의심 문서를 격리하고 영향 질문의 색인과 캐시 재생성	이번 주
복구	AI 도입 책임자	중요 질문 정답 세트와 독립 출처 규칙으로 회귀시험	이번 분기

담당자 확인 질문

지식 관리자: 누가 어떤 근거로 문서를 색인에 넣고 뺄 수 있는가?

순간 6 · 모델과 플러그인의 공급망을 믿고 넘길 때

근거 유형

실제 개발 도구 의존성 사고

출처에서 확인된 내용

PyTorch Foundation은 2022년 12월 25일부터 30일 사이 Linux에서 pip로 설치한 PyTorch-nightly 의존성 체인이 침해됐다고 공지했습니다. 악성 의존성은 설치된 시스템의 파일과 환경정보를 수집·반출했으며, 이는 AI가 스스로 공격을 수행한 사건이 아니라 AI 개발 생태계의 패키지 공급망 사고입니다.

우리 업무에 적용한 가정

개발팀이 모델·패키지·플러그인을 검증 없이 빌드 환경에 설치한다고 가정합니다. 기본 위험은 개발 시스템의 토큰·소스·환경정보입니다. 그 구성요소가 운영 앱으로 배포되고 운영 DB 계정까지 사용할 때에만 운영 데이터 경로가 추가됩니다. 배포 계정과 런타임 계정을 분리하면 확장 범위를 줄일 수 있습니다.

이번에 확인할 질문

설치·빌드 단계의 구성요소가 접근할 수 있는 비밀과 외부 네트워크 목적지는 어디까지인가?

기본 경로

패키지·모델·플러그인 설치 → 개발 시스템의 파일·환경정보 접근 → 외부 반출

추가 조건이 필요한 경로

오염 구성요소가 운영 앱으로 배포된 경우: 런타임 서비스 계정 → 운영 DB 또는 다른 내부 시스템

왜 기존 통제가 놓칠 수 있는가

정상 저장소와 비슷한 이름, 자동 설치, 넓은 빌드 권한이 결합하면 개발 편의가 공급망 공격의 실행 경로가 됩니다.

업무별 점검

구분	담당	조치	시점
예방	개발·보안팀	승인 저장소·고정 버전·서명·해시·SBOM을 배포 조건으로 설정	이번 주
탐지	보안팀	설치·빌드 중 외부 통신, 비밀 파일 접근, 새 프로세스 감시	이번 분기
대응	개발·인프라 담당	영향 빌드를 격리하고 같은 해시가 설치된 자산과 토큰 확인	이번 주
복구	개발 책임자	검증된 이미지로 재빌드하고 잠금 파일과 서명을 기준선으로 보존	이번 분기

데이터 경계 점검

오염 구성요소가 운영 런타임에 배포되고 서비스 계정으로 운영 DB에 닿는 경우에만 4단계 경계가 관련됩니다. 개발 시스템에서 끝나는 사고와 구분합니다.

담당자 확인 질문

개발 책임자: 개발 도구 사고가 운영 계정과 데이터로 번지려면 어떤 추가 조건이 필요한가?

순간 7·익숙한 목소리가 승인 절차를 건너뛸 때

근거 유형

공식 경보와 기업 적용 시나리오

출처에서 확인된 내용

FBI는 2025년 5월 15일 경보(I-051525-PSA)에서, 2025년 4월부터 공격자들이 미국 고위 공직자를 사칭해 문자와 AI 생성 음성 메시지를 보냈다고 밝혔습니다. 다른 메시지 플랫폼으로 옮기자며 악성 링크를 보내 개인 계정 접근을 노렸고, 이렇게 얻은 연락처로 지인을 사칭해 정보나 돈을 요구할 수 있다고 경고했습니다. 이 경보가 기업 대표 사칭 송금 사고를 확인한 것은 아닙니다.

우리 업무에 적용한 가정

별도 기업 적용 시나리오로, 공격자가 공개 음성과 일정 정보를 이용해 대표의 긴급 송금 요청을 흉내 낸다고 가정합니다. 피해는 음성 합성만으로 완성되지 않습니다. 등록 번호 콜백이 없고 신규 수취인 지연과 이중 승인이 생략될 때 송금으로 이어집니다. 공직자 계정 접근 경보와 기업 송금 시나리오를 같은 실제 사건으로 합치지 않습니다.

이번에 확인할 질문

긴급·비밀 유지 요청이 와도 원래 등록된 연락처와 두 번째 승인자로 재확인하는가?

기본 경로

공개 음성·조직 정보 → 사칭 전화·음성메시지 → 개인 계정 접근 시도

추가 조건이 필요한 경로

기업 적용 시나리오: 대표 사칭 → 예외 승인 → 신규 수취인 송금 또는 자료 전달

왜 기존 통제가 놓칠 수 있는가

목소리와 업무 맥락이 익숙하면 직원은 채널 자체를 신원 증명으로 오해할 수 있습니다. 긴급성과 비밀 유지 요구가 확인 절차를 약화시킵니다.

업무별 점검

구분	담당	조치	시점
예방	경영진·재무팀	송금·비밀 전달·인증 변경은 등록 번호 콜백과 두 번째 승인자 요구	이번 주
탐지	재무·보안팀	번호 변경·신규 수취인·비정상 시간·긴급 요구가 겹치면 거래 지연	이번 주
대응	현업	통화를 종료하고 별도 채널로 당사자 확인, 진행 중 거래 보류	이번 주
복구	경영진	분기별 사칭 훈련과 금융기관 긴급 연락 절차 점검	이번 분기

담당자 확인 질문

재무 책임자: 음성만으로 예외가 허용되는 결재 단계가 남아 있는가?

순간 8 · AI가 공격자의 반복 작업을 줄여 줄 때

근거 유형

2024년 위협 관측

출처에서 확인된 내용

Microsoft와 OpenAI는 2024년 국가 연계 위협 행위자가 대규모 언어 모델을 정찰, 코드 보조, 번역과 문구 작성 등에 사용한 활동을 설명했습니다. 발표는 AI가 여섯 단계 전체를 자율 수행했다는 확인이 아니며, 당시에는 새롭거나 독특한 AI 전용 공격 기법을 확인하지 못했다고 밝혔습니다.

우리 업무에 적용한 가정

공격자가 공개 자산 정리와 피싱 문구 변형, 스크립트 이해 같은 반복 작업에 AI를 보조적으로 쓴다고 가정합니다. 초기 침입은 여전히 취약 서비스, 탈취 계정, 피싱 같은 기존 경로를 필요로 합니다. 이후 DB나 백업에 닿으려면 별도 계정과 망 경로가 있어야 합니다. 따라서 방어는 AI 탐지 하나가 아니라 패치·MFA·특권 관리·행동 연결에 둡니다.

이번에 확인할 질문

정찰·로그인·특권 사용·대량 조회 같은 연속 신호를 같은 신원과 시간으로 연결해 볼 수 있는가?

기본 경로

공개 정보 → AI 보조 정찰·문구·코드 작업 → 기존 침입 경로 시도

추가 조건이 필요한 경로

계정·망 권한을 추가로 얻은 경우: 내부 이동 → 데이터 접근 또는 복구 사본 훼손

왜 기존 통제가 놓칠 수 있는가

각 행위는 흔한 관리 작업이나 개발 작업처럼 보일 수 있습니다. 속도·빈도와 앞뒤 단계의 연결을 보지 않으면 자동화된 반복을 놓칩니다.

업무별 점검

구분	담당	조치	시점
예방	보안·인프라팀	노출 자산 우선 패치, 피싱 저항 MFA, JIT 특권 적용	이번 주
탐지	보안팀	정찰·로그인·자격증명 접근·대량 조회를 신원과 시간으로 연결	이번 분기
대응	보안팀	영향 계정과 호스트 격리, 토큰 폐기, 추가 세션 제한	이번 주
복구	인프라 담당	깨끗한 이미지와 검증된 사본으로 우선 업무 복원 훈련	이번 분기

담당자 확인 질문

보안팀: 2024년 관측 사실과 우리 환경의 확장 가정을 보고서에서 구분하고 있는가?

순간 9 · 관리자 권한으로 데이터와 백업이 함께 지워질 때

근거 유형

기존 랜섬웨어 위험에 대한 설명용 시나리오

출처에서 확인된 내용

CISA의 StopRansomware 지침은 공격자가 백업을 찾고 삭제하거나 접근을 막을 수 있으므로 오프라인·암호화 백업과 정기 복원을 권고합니다. 이것은 AI가 없어도 성립하는 기존 랜섬웨어 위험입니다.

우리 업무에 적용한 가정

운영·도메인·백업 관리 권한이 한 계정 체계에 묶였다고 가정합니다. 계정이 탈취되면 운영 데이터와 복구 사본이 함께 영향을 받을 수 있습니다. AI는 표적 정리나 스크립트 이해를 보조할 수 있지만, 이 시나리오의 필수 조건은 아닙니다. 방어 우선순위는 관리자 분리, 보존 정책 잠금, 직접 접속 경로 축소와 실제 복원입니다.

이번에 확인할 질문

운영 관리자 권한 하나로 복구 사본과 보존 정책을 바꾸거나 지울 수 있는가?

기본 경로

탈취 관리자 계정 → 운영 시스템·데이터 훼손 → 백업 콘솔·저장소 접근 → 복원 지점 삭제

추가 조건이 필요한 경로

AI 보조가 사용된 경우: 자산 정리·명령 해석·반복 속도 증가

왜 기존 통제가 놓칠 수 있는가

백업이 같은 인증 체계와 상시 연결 경로에 있으면 사본 수가 많아도 하나의 관리 권한으로 함께 훼손될 수 있습니다.

업무별 점검

구분	담당	조치	시점
예방	경영진·인프라 담당	운영·도메인·백업 관리자를 분리하고 JIT·이중 승인·보존 잠금 적용	이번 주
탐지	보안팀	MFA 변경·로그 중지·보존 기간 축소· 삭제 요청을 우선 경보	이번 주
대응	보안·인프라팀	관리망을 격리하고 운영 데이터와 복구 사본의 마지막 정상 시점 확정	이번 주
복구	경영진·인프라 담당	RPO·RTO 기준으로 분기별 실제 복원과 재가동 훈련	이번 분기

데이터 경계 점검

이 사례는 운영 데이터 경로와 복구 사본 경로가 모두 핵심입니다. 4단계의 직접 DB 경로와 6단계의 백업 변경·삭제 경로를 각각 실측합니다.

담당자 확인 질문

인프라 담당: 운영 계정이 탈취돼도 별도 계정과 경로 없이는 복구 사본을 바꿀 수 없는가?

순간 10 · AI 판단이 로봇·차량·드론의 제어권과 만날 때

근거 유형

연구·시험 방향을 적용한 시나리오

출처에서 확인된 내용

NIST의 자동화 차량 프로그램은 인지·통신·시험 방법과 안전한 자동화 시스템 연구 방향을 다룹니다. 이 자료는 특정 로봇·차량·드론 공격을 실증한 사고 보고가 아닙니다.

우리 업무에 적용한 가정

카메라·센서·지도 입력을 AI가 해석하고 제어 명령으로 전달하는 현장을 가정합니다. 입력 오류나 조작이 있어도 독립 안전 규칙과 사람 전환이 있으면 물리 행동의 범위를 제한할 수 있습니다. 반대로 모델 출력이 곧바로 속도·방향·동작 명령이 되면 오판이 안전사고로 번질 수 있습니다. 시험은 제한 구역과 시뮬레이터에서 수행합니다.

이번에 확인할 질문

모델 출력과 물리 제어 사이에 독립 안전 한계·비상정지·사람 전환이 있는가?

기본 경로

센서·지도·무선 입력 → AI 인식·계획 → 제어 인터페이스 → 물리 동작

추가 조건이 필요한 경로

원격 관리 계정이나 유지보수 채널이 연결된 경우: 설정 변경 → 안전 한계 약화

왜 기존 통제가 놓칠 수 있는가

모델 정확도가 높아도 센서 불일치와 예외 상황은 남습니다. 독립 안전 장치 없이 출력이 곧 명령이 되면 작은 오판의 물리적 피해 반경이 커집니다.

업무별 점검

구분	담당	조치	시점
예방	현업·AI 책임자	센서 진위를 교차 확인하고 속도·공간·동작 한계와 사람 승인 게이트 설정	이번 주
탐지	현업·보안팀	센서 불일치·신뢰도 급락·안전구역 접근·명령과 실행 차이 기록	이번 분기
대응	현업	안전 상태 정지, 원격 제어 분리, 수동 모드 전환, 로그 보존	이번 주
복구	경영진·현업	재가동 조건과 승인자를 정해 제한 구역에서 전환 훈련	이번 분기

담당자 확인 질문

안전 책임자: AI가 잘못 판단해도 독립 계층이 물리 한계를 지키는가?

실무 부록

부록 A·20문항 자가진단

예·아니오·모름 중 하나를 표시하고, 예로 답하지 못한 항목에는 담당자와 확인 날짜를 적습니다.

번호	점검 문항	답	담당자	확인 날짜·증거
1	인터넷 노출 자산과 관리자 포털을 분기마다 정리하는가?	☐ 예 ☐ 아니오 ☐ 모름		
2	긴급·비밀 유지·채널 변경 요구를 등록 연락처로 재확인하는가?	☐ 예 ☐ 아니오 ☐ 모름		
3	고위험 업무에 피싱 저항 MFA와 예외 승인 절차가 있는가?	☐ 예 ☐ 아니오 ☐ 모름		
4	외부 문서·메일·웹 응답을 신뢰 문맥과 구분해 표시하는가?	☐ 예 ☐ 아니오 ☐ 모름		
5	숨은 텍스트·링크·메타데이터를 포함한 시험 자료로 AI를 검증하는가?	☐ 예 ☐ 아니오 ☐ 모름		
6	지식 문서의 원문·작성자·승인자·해시·등록 시각을 남기는가?	☐ 예 ☐ 아니오 ☐ 모름		
7	음성·센서 입력을 독립 채널이나 두 번째 신호로 확인하는가?	☐ 예 ☐ 아니오 ☐ 모름		
8	에이전트마다 사람과 구분되는 신원과 책임자가 있는가?	☐ 예 ☐ 아니오 ☐ 모름		
9	도구 허용 목록, 읽기·쓰기 분리, 짧은 토큰 만료가 적용됐는가?	☐ 예 ☐ 아니오 ☐ 모름		
10	발송·삭제·결제·권한 변경에 미리보기와 사람 승인이 있는가?	☐ 예 ☐ 아니오 ☐ 모름		
11	요청·계획·도구 인자·결과·승인자를 같은 추적 ID로 남기는가?	☐ 예 ☐ 아니오 ☐ 모름		
12	외부 사용자의 웹 연결, 앱서버의 DB 연결·계정, 허용된 직접 경로를 구분했는가?	☐ 예 ☐ 아니오 ☐ 모름		
13	DB 권한이 업무별 행·열·스키마로 제한되고 대량 조회 한도가 있는가?	☐ 예 ☐ 아니오 ☐ 모름		
14	관리망·백업망·유지보수 경로를 포함해 우회 연결을 실측했는가?	☐ 예 ☐ 아니오 ☐ 모름		
15	새 출발지 세션·메타데이터 조회·대량 반환을 경보하는가?	☐ 예 ☐ 아니오 ☐ 모름		
16	외부 전송·금액·건수·대상에 자동 한도와 이중 승인이 있는가?	☐ 예 ☐ 아니오 ☐ 모름		
17	반출·변조·암호화 신호를 신원과 시간으로 연결해 보는가?	☐ 예 ☐ 아니오 ☐ 모름		
18	운영·도메인 관리자와 백업 삭제·보존 변경 권한이 분리됐는가?	☐ 예 ☐ 아니오 ☐ 모름		
19	복구 사본이 업무망의 직접 접속 경로와 분리돼 있는가?	☐ 예 ☐ 아니오 ☐ 모름		
20	분기마다 실제 데이터를 복원해 RPO·RTO와 재가동 순서를 측정하는가?	☐ 예 ☐ 아니오 ☐ 모름		

부록 B·경로 지도 작성표

단계	자산·신원·연결	담당자	확인할 증거	끊을 지점
1 유인·정찰				
2 입력 진입				
3 AI 실행·권한 오용				
4 내부 이동				
5 목표 달성				
6 복구 방해				

제품 이름보다 실제 동사로 씁니다. 외부 입력이 들어온다 → AI가 읽는다 → 도구가 행동한다 → 앱서버가 DB 계정으로 조회한다 → 응답이 외부로 나간다 → 관리자가 복구 사본을 바꿀 수 있다처럼 화살표 위에 신원과 권한을 적습니다.

부록 C·PoC 기준표

기준	측정 방법	합격값	실측값	판정자	증거
직접 접속 경로					
허용 업무					
처리량·지연					
장애 동작					
무결성					
실제 복원					

합격 기준은 시험 전에 정합니다. 보안 기준뿐 아니라 정상 사용자 여정, 장애 때의 동작, 상위 지연, 실제 DB·앱 재가동 시간을 함께 측정합니다.

부록 D·30분 내부 회의안

- 1. **0~5분**: 멈추면 영향이 가장 큰 업무 하나를 고릅니다.
- 2. **5~12분**: 외부 입력부터 AI·도구·앱·DB·복구 사본까지 실제 존재하는 선만 그립니다.
- 3. **12~18분**: 에이전트와 관리자 계정이 할 수 있는 가장 큰 행동을 적습니다.
- 4. **18~24분**: 운영 DB 직접 경로와 복구 사본 변경·삭제 경로를 실측할 담당자를 정합니다.
- 5. **24~30분**: 48시간 안에 실측할 연결 하나, 2주 안에 줄일 권한 하나, PoC에서 썰 기준 하나를 합의합니다.

NF·DF가 관련되는 조건과 PoC 검토 방법

제품 이름보다 먼저 실제 연결선을 확인합니다. 모든 AI 위험이 운영 DB나 복구 사본을 거치는 것은 아닙니다.

N-FORTRESS가 관련되는 조건

외부 사용자의 웹 연결, 애플리케이션이 DB 드라이버·연결 풀과 서비스 계정으로 맺는 DB 연결, 도입 뒤 제거되거나 허용되는 직접 경로를 구분합니다. N-FORTRESS는 웹·앱 구간과 운영 DB 구간 사이에 별도의 중계 경계를 둡니다. 추가 보호 효과는 실제로 제거되는 직접 접속 경로와, 허용된 요청·응답의 통제 범위를 기준으로 확인해야 합니다.

허용된 업무 요청의 내용이 안전하다는 뜻은 아닙니다. 계정·권한·질의 범위·입력 검증·암호화·감사와 모니터링은 계속 필요합니다. NF를 통과해 애플리케이션에 이미 전달된 정보가, NF를 거치지 않는 경로로 유출되는 경우에는 별도 통제가 필요합니다.

D-FORTRESS가 관련되는 조건

운영 관리자나 랜섬웨어가 업무망의 직접 접속 경로를 통해 복구 사본과 보존 정책을 바꾸거나 지울 수 있을 때 복구 데이터 경계를 검토합니다. D-FORTRESS는 업무망의 직접 접속 경로와 분리된 영역에 복구 사본을 보존해 변경·삭제 위험을 줄이는 역할입니다. 백업 대상·주기·보존 버전·계정 분리·감염 시점 판단·RPO·RTO·실제 복구훈련을 대신하지 않습니다.

공개 본문은 두 제품의 핵심 원리만 다룹니다. HMAC·재전송 제어와 제한된 피드백 채널 같은 세부 구현은 제품별 기술자료로 안내합니다.

어떤 문제에 어떤 경계가 필요한가

문제	먼저 적용할 통제	N-FORTRESS	D-FORTRESS
외부 문서·RAG·사칭·물리 제어	입력 격리·출처·승인·안전 규칙	직접 관련 없음	직접 관련 없음
애플리케이션에서 운영 DB로 이어지는 직접 경로	DB 인가·최소 권한·질의·감사	별도 중계 경계의 제거·허용 경로를 PoC로 확인	직접 관련 없음
허용된 조회 결과의 외부 반출	응답 범위·DLP·외부 통신·승인	직접 경로 축소만으로 해결되지 않음	직접 관련 없음
운영 권한이 복구 사본 변경·삭제까지 이어짐	계정·보존 정책 분리·복원 훈련	운영 DB 구간에 한해 보완	복구 사본의 직접 접속 경로 분리를 PoC로 확인

PoC에서 합의할 것

- 제거되는 직접 접속 경로와 유지해야 할 정상 업무 경로
- 허용 요청·응답, 계정·권한·질의 통제의 책임 경계
- 처리량·지연·오류율·동시 연결과 장애 때 동작

- 관리망·백업망·유지보수 경로를 포함한 우회 가능성
 - 복구 사본의 전송·보존·무결성과 실제 DB·앱 재가동 시간
-

출처와 한계

참고문헌은 기관·저자 / 정확한 제목 / 문서번호·판 / 발행일 / 본문에서 사용한 내용 / 링크 / 확인일 순서로 적었습니다.

1. NIST, Autio 외 / *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* / NIST AI 600-1 / 2024-07-26(웹페이지 갱신 2026-04-08) / 생성형 AI 위험 관리의 범위와 한계 / <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence> / 2026-10-11.
2. NIST / *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* / NIST AI 100-2 E2025 / 2025-03 / 최신 AML 분류 용어 확인 / <https://csrc.nist.gov/pubs/ai/100/2/e2025/final> / 2026-10-11.
3. NCSC, Dave Chismon / *Prompt injection is not SQL injection (it may be worse)* / 웹 블로그 / 2025-12-08 / SQL 인젝션과 프롬프트 인젝션의 차이, 모델 밖 영향 제한 / <https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection> / 2026-10-11.
4. OWASP GenAI Security Project / *LLM06:2025 Excessive Agency* / OWASP Top 10 for LLM Applications 2025 / 2025 / 과도한 기능·권한·자율성 / <https://genai.owasp.org/llmrisk/llm062025-excessive-agency/> / 2026-10-11.
5. OWASP GenAI Security Project / *LLM07:2025 System Prompt Leakage* / OWASP Top 10 for LLM Applications 2025 / 2025 / 시스템 프롬프트에 비밀과 보안 통제를 두지 않는 원칙 / <https://genai.owasp.org/llmrisk/llm072025-system-prompt-leakage/> / 2026-10-11.
6. Google Security Blog / *AI threats in the wild: The current state of prompt injections on the web* / 웹 연구 소개 / 2026-04 / 공개 웹 관측 범위와 탐지 한계 / <https://security.googleblog.com/2026/04/ai-threats-in-wild-current-state-of.html> / 2026-10-11.
7. Wei Zou, Runpeng Geng, Binghui Wang, Jinyuan Jia / *PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models* / 34th USENIX Security Symposium / 2025-08 / 통제된 RAG 오염 실험 / <https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag> / 2026-10-11.
8. PyTorch Foundation / *Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022* / 보안 공지 / 2022-12-31(페이지 수정 2024-11-14) / 개발 도구 의존성 사고의 확인 범위 / <https://pytorch.org/blog/compromised-nightly-dependency/> / 2026-10-11.
9. FBI / *Senior U.S. Officials Impersonated in Malicious Messaging Campaign* / 공공 서비스 경보 I-051525-PSA / 2025-05-15 / 문자·AI 생성 음성 사칭, 악성 링크를 통한 개인 계정 접근, 확보한 연락처로 정보·자금 요구 가능성 / <https://www.fbi.gov/investigate/cyber/alerts/2025/senior-us-officials-impersonated-in-malicious-messaging-campaign> / 2026-10-11.
10. Microsoft Threat Intelligence / *Cyber Signals: Navigating cyberthreats and strengthening defenses in the era of AI* / Cyber Signals Issue 6 / 2024-02-14 / 위협 행위자의 정찰·코드·번역·문구 작업 활용 / <https://www.microsoft.com/en-us/security/blog/2024/02/14/cyber-signals-navigating-cyberthreats-and-strengthening-defenses-in-the-era-of-ai/> / 2026-10-11.
11. CISA·FBI·MS-ISAC / *#StopRansomware Guide* / 공동 지침 / 2023-05(2025-03 배포 PDF 확인) / 백업 보호와 정기 복원 / <https://www.cisa.gov/sites/default/files/2025-03/StopRansomware-Guide%20508.pdf> / 2026-10-11.

- | 2. NIST / *Automated Vehicles Program* / 프로그램 페이지 / 2023 이후 / 자동화 차량 연구·시험 방향 / <https://www.nist.gov/programs-projects/nist-automated-vehicles-program> / 2026-10-11.
- | 3. 락게이트코어 / *N-FORTRESS 운영 제품 페이지* / 운영본 v16 / 2026-10-10 / 외부 웹 연결, 앱서버 DB 연결·계정, 별도 중계 경계와 남는 통제 / <https://rockgatecore.com/products/n-fortress/> / 2026-10-11.
- | 4. 락게이트코어 / *N-FORTRESS 제품 제안서* / 261007, 25쪽 / 2026-10-07 / 연결 경계·허용 요청과 응답·PoC 항목 / https://rockgatecore.com/downloads/N-FORTRESS_product_brochure_261007.pdf / 2026-10-11.
- | 5. 락게이트코어 / *DF_TX 기능명세서·DF_RX 기능명세서* / GS 인증용 v1.0, 2026-08-25 / 복구 사본 전송·수신의 내부 기술 근거, 공개 본문에는 핵심 원리만 사용 / 비공개 제품별 기술자료 / 2026-10-11.

이 가이드는 AI 도구를 활용해 작성했으며, 락게이트코어 기술팀이 내용을 검토했습니다.

이 자료는 일반적인 보안 점검을 돕기 위한 것이며 특정 조직의 보안성이나 규제 준수를 보증하지 않습니다. 공개 사건만으로 특정 제품이 사고를 막았을 것이라고 판단할 수 없습니다. 실제 적용 범위와 효과는 고객 환경의 연결·권한·업무 흐름과 PoC 실측으로 판단합니다.

우리 조직의 AI 보안 점검을 시작해 보세요

1단계. 20문항 체크리스트에서 예로 답하지 못한 항목을 표시하고, 담당자와 확인할 자료를 정합니다.

2단계. 운영 DB 접근이나 복구 사본 보호가 미확인 항목으로 남았다면, 업무 한 가지를 기준으로 사전 검토를 요청할 수 있습니다. 상세 구성도와 민감 운영 정보는 처음부터 보내지 않고, 검토 범위와 안전한 전달 방법을 합의한 뒤 공유합니다.

락게이트코어 · 070-8080-4910 · info@rockgatecore.com · <https://rockgatecore.com/>