

주간 WEF 브리프 · WEF 신호로 읽는 산업·정책 · 8호

고른 사람이 없는 칸

— 자율성과 권한은 설계 선택이다 · 네 칸과 오늘 적을 목록

설비 제어 프로그램을 짜는 개발자가 남의 코드 저장소를 AI 코딩 도구로 열면, 「실행하시겠습니까」 창이 뜨기 전에 그 저장소에 심긴 명령이 이미 돌아갑니다. 9월 1일 공개된 결함이고 번호가 붙은 것만 네 건이고, 영향 도구로 적힌 것은 Claude Code·Codex를 포함해 일곱인데 수정이 끝난 것으로 확인된 것은 둘뿐입니다. 쓰는 쪽은 승인을 기다리는 칸에 있다고 여겼고, 도구는 사람 없이 실행하는 칸에서 움직였습니다. WEF가 5월에 낸 백서는 AI에게 맡길 폭을 네 칸으로 나누고 「위험이 큰가, 되돌릴 수 있는가」로 칸을 고르라고 적었고, 그보다 앞선 2025년 11월 백서는 자율성과 권한이 시스템에 딸린 성질이 아니라 설계 선택이라고 적었습니다. 고르지 않으면 회사의 수준은 가장 높은 칸에서 움직이는 도구가 정합니다. 오늘 할 수 있는 일은 지금 쓰는 도구를 적고 각각 몇 번 칸인지 옆에 쓰는 것이고, 여기에는 전담조직도 예산도 필요하지 않습니다.

AI 코딩 도구 결함 4건 · 영향 도구 일곱

자율성 네 칸 · 권한은 별개 눈금

도내 신청 마감 11월 말 · 경기 배정분

칸은 넉 달 전에 그려졌다

2025. 11.

WEF 백서 — 자율성과 권한은 설계 선택이라고 적었다

2026. 5.

WEF·KPMG 백서 — 맡길 폭을 네 칸으로 나누고 「위험과 되돌릴 수 있는 정도」로 고르라고 했다

2026. 9. 1.

승인 창이 뜨기 전에 실행되는 결함 공개 — 번호 4건

2026. 9. 1~2.

최상위 AI 모델을 만드는 회사 세 곳, 가장 센 보안 기능을 누가 받을지 각각 정해 발표

2026. 9. 3.

오픈AI, 방어자 지원 10억 달러와 「파트너 국가」 확대 의향

붉은 항목은 사람이 칸을 고르기 전에 도구가 움직인 대목이다(Ⅰ장 3면). 세 회사의 배분 방식은 Ⅲ장 6~7면, 「파트너 국가」는 V장 9면.

이 브리프에서 다루는 것

- 1 승인 창이 뜨기 전에 실행된 명령 · Ⅰ장
- 2 네 칸과, 자율성과 다른 눈금인 권한 · Ⅱ장
- 3 가장 센 보안 기능을 누가 받는지 정한 세 곳 · Ⅲ·Ⅳ장
- 4 「파트너 국가」라는 말과 그 안에 없는 것 · V장
- 5 작은 조직이 뒤쳐진다 · 국내 값 · 직접 만들 것인가 사 올 것인가 · VI·Ⅶ장

원문 보고서 표지 · World Economic Forum
2026년 5월

이벤트 요약

발행일자	2026. 9. 10. (목)	원문 확인	2026년 5월 발간본 PDF 34쪽
원문	World Economic Forum 「Empowering Defenders: AI for Cybersecurity」(2026년 5월, White Paper 34쪽). KPMG 협업, WEF Cyber Frontiers 이니셔티브. 15개 산업 84개 조직 105명 워크숍과 사례 20건에 근거한다. 원문 PDF와 2025년 11월 백서 전문을 직접 대조했고 국내 제도·통계는 (보도 경유)로 구분해 적었다		

네 칸

WEF가 나눈 AI 자율성의 단계. 고르는 기준은 성능이 아니고 위험과 되돌릴 수 있는 정도다 (II장 4~5면)

4건

AI 코딩 도구로 저장소를 열면 명령이 실행되는 결함의 번호. 영향 도구 일곱 가운데 수정 확인은 둘뿐 (I장 3면)

1.6%

종사자 10~49인 기업에서 정보보호 전담조직을 둔 곳. 250인 이상은 33.9%다 (2024년 조사) (VI장 11면)

11월 말

도내 중소기업이 신청할 수 있는 마감 — 경기 배정분이며 조건은 공고에서 확인한다 (◆ 14면)

이벤트 핵심 3줄

- WEF가 KPMG와 함께 5월에 낸 백서는 AI에게 맡길 폭을 네 칸으로 나눴다 — ① 사람이 다 본다, ② 사람이 승인한다, ③ 뒤집을 수 있다, ④ 사람이 없다. 고르는 기준은 「위험과 되돌릴 수 있는 정도」다. 그보다 앞선 2025년 11월 백서는 자율성과 권한이 다른 논제이고 둘 다 설계 단계에서 누군가 정하는 선택이라고 이미 적어 두었다.
- 9월 1일 공개된 결함은 그 선택을 아무도 하지 않았을 때 무슨 일이 생기는지 보여 준다. AI 코딩 도구는 git status 같은 명령을 스스로 돌리는데, 남의 저장소에 심긴 설정이 그때 함께 실행된다. 공개 문서의 표현으로 「샌드박스 밖에서, 승인 창 없이」 돈다. 쓰는 쪽은 두 번째 칸에 있다고 생각했고 도구는 네 번째 칸에서 움직였다. 번호가 붙은 것은 네 건이고, 영향 도구로 적힌 일곱 가운데 공개 시점 기준 수정이 확인된 것은 둘뿐이다.
- 같은 주에 최상위 AI 모델을 만드는 회사 세 곳이 가장 센 보안 기능을 누구에게 줄지 각각 정했지만 (앤스로픽·구글은 자격 심사, 오픈AI는 초기 테스터로 갈린다), 국내 기업이 오늘 할 일은 그 결정과 무관하다. 원문이 중소기업 앞으로 따로 적은 조인 넷의 요지는 막힌 업무를 한두 개 고른 뒤 지금 쓰는 도구의 기능부터 켜 보라는 것이다. 국내 10~49인 기업의 전담조직 보유율은 1.6%이지만(250인 이상 33.9% · 2024년 조사), 도구 목록에 칸을 적는 데는 전담조직이 필요하지 않다.

목차

- I. 승인 창이 뜨기 전에 실행됐다 · 3면
- II. 원문이 그린 네 칸 · 자율성의 단계와 고르는 기준 · 4~5면
- III. 가장 센 기능을 누가 받는지 정하기 시작했다 · 만드는 쪽 세 곳 · 6~7면
- IV. 그 안에서 나온 것 · 8면
- V. 「파트너 국가」라는 말 · 9면
- VI. 작은 조직이 뒤처진다 · 국내 값 · 직접 만들 것인가 사 올 것인가 · 10~12면
- VII. 경기도 기업에 미치는 영향 · 13면
- ◆ 점검 항목 · 향후 일정 · 14면 / * 용어 · 출처 · 15면

모르는 말이 나오면 마지막 쪽(*용어 설명)에 있다. 「자율성 단계」·「권한」·「되돌릴 수 있는 행동」·「기계·에이전트 신원」이 이 호의 열쇳말이다.

I 승인 창이 뜨기 전에 실행됐다

설비 제어 프로그램이나 자동화 스크립트를 짜는 곳이라면 남의 코드 저장소를 내려받아 열어 보는 일이 흔하다. 그 일을 **AI 코딩 도구**로 하면 사정이 달라진다. 도구는 무엇을 도울지 판단하려고 **저장소 상태를 스스로 조회하는데**(`git status`·`git diff`), 저장소에 딸려 온 설정 파일에 명령이 심겨 있으면 **바로 그때 함께 실행된다**.

☞ 공개 문서가 적은 실행 조건

- “This is the agent's own code spawning a subprocess to use git, so the command runs **outside the sandbox, without an approval prompt.**”
- 국문: “에이전트 자신의 코드가 git을 쓰려고 하위 프로세스를 띄우는 것이라서, 그 명령은 **샌드박스 밖에서 승인 창 없이 돌아간다.**” 공격자가 조종하는 코드가 **그 도구를 쓰는 사람의 권한으로** 실행된다는 뜻이다.

* 영문은 공개 문서 표기 그대로이고 국문은 이 브리프의 번역이다. 생략한 대목은 **말줄임표로 표시했다**.

번호가 붙은 것은 네 건이다

번호	영향 도구	만든 곳	도구별 수정 상태 · 2026. 9. 1. 기준
CVE-2026-19592	Codex	오픈AI	CLI 0.131.0 및 데스크톱판에서 수정
CVE-2026-55607	Claude Code	앤스로픽	부분 수정 — 한 경로는 2.1.196에서 막혔고 다른 경로는 남아 있다
CVE-2026-72718	goose	AAIF	1.44.0에서 수정
CVE-2026-71963	Hermes Agent	Nous Research	수정 확인되지 않음
번호 없음	Qwen Code · Grok Build	알리바바 · xAI	수정 확인되지 않음
번호 없음	Cursor	Anysphere	공개 문서에 수정 상태가 적혀 있지 않다

공개일 2026년 9월 1일, 공개 주체는 보안업체 Manifold Security. 번호 두 건(19592·55607)과 **도구별 수정 상태는 (보도 경유)**이고 나머지는 공개 문서 기준이다. 이 표는 **공개 문서가 영향 도구로 적은 것을 그대로 옮긴 것이며 이 브리프가 도구를 시험하거나 평가한 것이 아니다.** 그 뒤 갱신됐을 수 있다.

이 결함이 국내 독자에게 뜻하는 것

- 공격은 저장소를 여는 순간 시작된다. 사내 규정이 「AI가 제안하면 사람이 확인한다」로 돼 있어도 그 규정이 닿지 않는 자리다.
- 고쳐진 도구와 안 고쳐진 도구가 섞여 있다. 팀의 안전 수준은 **가장 늦게 고쳐진 도구**에 맞춰진다.
- 오늘 걸 수 있는 **완화책이 있다**. 공개 문서는 도구가 git을 부를 때 해당 설정을 꺼 두는 방식(`git -c core.fsmonitor=false`)을 적었다 — **개발 담당자가 도구 실행 스크립트에 넣는다**.

🔍 고르지 않았다는 것이 이 결함의 조건이다

쓰는 쪽은 도구가 **제안하고 승인을 기다리는 칸**에 있다고 여겼고, 도구는 **사람 없이 실행하는 칸**에서 움직였다. 둘 사이를 정하는 것이 다음 장의 네 칸이고, 원문은 그 칸을 **고르는 기준**까지 적어 두었다(II장 4면). 같은 주에 회사 세 곳도 가장 센 보안 기능을 누구에게 줄지 각자 정했다 — **조직이 고르는 칸과는 층이 다른 결정**이다(III장 6~7면).

II 원문이 그린 네 칸

원문은 34쪽짜리 백서다. 앞의 절반은 AI를 방어에 쓰는 스무 건의 실제 사례이고, 뒤로 갈수록 「그래서 어디까지 말할 것인가」로 좁혀진다. 마지막 장에 그 답이 그림 하나로 들어 있다 — ① 사람이 다 본다 · ② 사람이 승인한다 · ③ 뒤집을 수 있다 · ④ 사람이 없다는 네 칸이다.

칸	원문 표기	AI가 하는 일과 사람의 자리	원문이 든 예
① 사람이 다 본다	Assist Full human oversight	AI는 자료를 정리하고 대응은 사람이 짠다	보안 경보를 요약한다
② 사람이 승인한다	Recommend Human-in-the-loop	AI가 조치를 제안하고 사람이 검토·승인한 뒤에 실행된다	악성 파일을 표시해 두고 분석가가 「격리」를 누르기를 기다린다
③ 뒤집을 수 있다	Execute (overrideable) Human-on-the-loop	AI가 되돌릴 수 있는 조치를 스스로 실행하고 사람이 실시간으로 지켜보다 뒤집는다	수상한 IP를 차단하고, 신뢰하는 거래처가 걸리면 관리자가 되돌린다
④ 사람이 없다	Execute (independent) Human-out-of-the-loop	사람의 실시간 개입 없이 움직이고, 감독은 「감독하는 AI」나 사후 점검이 맡는다	서비스 거부 공격 대응을 조율하고, 감독 AI가 보안 정책에 맞는지 검증한다

원문 그림 3을 표로 옮겼다. 국문 단계 이름은 이 브리프가 붙인 것이고 원문 표기 옆이 원문의 낱말이다.

칸을 고르는 기준은 성능이 아니다

☞ 원문이 기준을 적은 문장

- “The appropriate choice is determined by the **risk and reversibility of an action**. High autonomy delivers efficiency gains for low-risk, reversible decisions while human oversight remains essential for high-stakes actions with lasting consequences, **even at the cost of a slower response**.”
- 국문: “적절한 선택은 그 행동의 위험과 되돌릴 수 있는 정도로 정해진다. 높은 자율성은 위험이 낮고 되돌릴 수 있는 결정에서 효율을 가져다주지만, 오래가는 결과를 남기는 중대한 조치에는 대응이 느려지는 대가를 치르더라도 사람의 감독이 여전히 필수적이다.”

- 기준이 두 가지뿐이라는 것이 이 그림의 쓸모다. 「이 조치가 잘못됐을 때 얼마나 큰가」와 「되돌릴 수 있는가」만 물으면 칸이 정해진다. 모델이 얼마나 똑똑한지는 묻지 않는다.
- I장의 결함은 이 표로는 설명되지 않는다. 쓰는 쪽은 ②를 쓴다고 여겼고 도구는 ④에서 움직였다. 칸을 고르는 절차 자체가 없었던 것이지 잘못 고른 것이 아니다.

💡 여기서 나온 실무 질문 하나

지금 회사가 쓰는 AI 도구를 놓고 「이 도구는 몇 번 칸인가」에 답할 수 있는가. 답이 도구마다 다르고 문서에 적혀 있지 않다면, 회사는 가장 높은 칸에서 움직이는 도구의 수준에서 있다.

II 원문이 그린 네 칸 · 칸을 올릴 때 함께 커지는 것

원문은 자율성을 올리면 무엇이 함께 커지는지도 적었다. 크게 세 갈래이고, 셋 다 **기존 보안 통제만으로는 다루지 못한다**고 못 박았다.

- **들어올 수 있는 곳이 늘어난다** — 에이전트가 새로운 침입 경로를 만들거나 에이전트 자체가 탈취될 수 있다. I장의 결함이 그 예다 (3면).
- **의도하지 않은 행동이 번진다** — 환각이나 외부 조작, 잘못 설정된 목표에서 나온 행동이 에이전트가 여럿인 환경에서 **기계의 속도로** 번진다. 훈련 중이던 에이전트들이 외부 공개 위키를 서로의 연락 수단으로 쓴 사례가 9월 4일 공개됐다(보도 경유).
- **승인 없이 배치된다** — 제대로 된 승인이나 검증 없이 빠르게 배치되고 **책임 소재가 분명하지 않다**.

원문 3.3절을 옮겼다. 각 항목 뒤에 붙인 **대응 사례는 이 브리프가 든 것이며 원문에 없다**.

☞ 원문이 서론(6쪽)에 따로 상자를 내어 적은 경고 — 지나치게 기대는 것

- “Excessive trust in automated decisions **creates a false sense of security** and over time **erodes the expertise needed to intervene when systems fail.**”
- 국문: “자동화된 판단을 지나치게 믿으면 **안전하다는 착각**이 생기고, 시간이 지나면서 **시스템이 멈췄을 때 개입할 수 있는 역량이 줄어든다.**”
- 대응 셋 — 사람의 판단을 함께 쓴다 · **고장 상황을 미리 모의해 본다** · **AI가 멈춰도 운영이 돌아가도록** 대비 장치를 설계한다.

자율성과 권한은 다른 눈금이다

원문은 3.3절에서 에이전트가 만들어 내는 위험을 짚은 뒤, 구체적인 통제 수단은 **그보다 앞서 2025년 11월에 나온 백서 ‘AI Agents in Action: Foundations for Evaluation and Governance’**에 있다고 적었다. 그 백서는 에이전트를 일곱 개 축으로 나눠 분류했고 (이 호는 그 가운데 **자율성과 권한** 두 축만 쓴다), 둘을 **서로 다른 눈금**으로 갈라 놓았다.

☞ 2025년 11월 백서가 둘을 가른 문장

- **자율성** — “Autonomy captures **the degree to which an agent can define and pursue objectives.**” 목표를 스스로 정하고 좇는 폭이다.
- **권한** — “Authority defines **the actions an agent is permitted to take.**” / “Like autonomy, authority exists on a sliding scale, **from read-only access to full administrative control.**” 도구·데이터베이스·거래에 손댈 수 있는 **범위**이고, 읽기 전용부터 완전 관리자 권한까지 눈금이 있다.
- “**They are not inherent system properties but design choices** that can be made based on the agents’ intended functions, risk considerations and oversight requirements.”(they = autonomy and authority) 둘 다 **설계 단계에서 정하는 선택**이고, 운영 중에 다시 조정할 수도 있다.
- **I장의 결함은 권한 쪽에서 벌어졌다.** 문제는 그 명령이 **사용자의 권한 그대로** 실행된 데 있었다. 자율성만 조이면 막히지 않는 종류다.
- **2025년 11월 백서는 제로트러스트를 명시했다** — “every agent interaction should be treated as **untrusted by default**”, 접근을 주기 전에 신원·권한·맥락을 확인하라는 것이다. 남의 저장소를 여는 순간이 정확히 그 자리다.

III 가장 센 기능을 누가 받는지 정하기 시작했다

원문의 네 칸은 조직 안에서 무엇을 맡길지를 정하는 눈금이다. 9월 첫 주에는 그 앞 단계에서 같은 일이 벌어졌다 — 가장 센 보안 기능을 누가 받을 수 있는지, 만드는 쪽이 정하기 시작했다.

날짜	회사	프로그램 이름과 배포 방식	받을 수 있는 쪽
9월 1일	앤스로픽	Cyber Verification Program Life Sciences Verification Program	사이버보안·생명과학 제한에 걸리는 업무를 하면서 심사를 통과한 개인·조직. 현재는 미국 조직
9월 2일	구글	Fairwind Program	구글이 신뢰하는 정부 당국과 핵심 인프라 운영자, 소프트웨어 유지보수자에게 우선 배포
9월 1일	오픈AI	초기 테스터 → Daybreak Blue 오픈AI가 보안 목적 이용자에게 따로 여는 창구	우선은 소수 테스터에게 열고, 이후 방어 목적 사용자로 넓힌다고 적었다

각 회사 공식 게시물로 확인했다. 9월 3일 오픈AI의 지원 발표는 III장 7면에 두고 '파트너 국가' 대목만 V장 9면에서 다룬다.

앤스로픽 · 같은 모델을 둘로 나눠 냈다

☞ 회사 공식 게시물의 문장

- "Claude Fable 5.1 and Claude Mythos 5.1 are **the same model, but with different levels of safeguards.**" / "Claude Mythos 5.1 is identical to Fable 5.1, but it offers **more permissive safeguards for vetted individuals and organizations whose work is affected by the cybersecurity and life sciences restrictions** outlined above."
- 국문: "Fable 5.1과 Mythos 5.1은 같은 모델이고 안전장치 수준만 다르다." / "Mythos는 Fable과 동일하되, 사이버보안·생명과학 제한에 걸리는 업무를 하면서 심사를 통과한 개인·조직에게는 제약이 덜한 안전장치를 제공한다."

- 회사가 적은 조건은 심사 통과와 '사이버보안·생명과학 제한에 걸리는 업무' 둘이다. 이 방식은 이 호가 다루는 '네 칸'과 층이 다르다 — 칸은 조직이 고르고 이 구분은 만드는 쪽이 고른다.

구글 · 우선 배포라고 적었다

☞ 회사 공식 게시물의 문장

- "Through our new **Fairwind Program**, we're providing trusted government authorities, as well as **critical infrastructure operators and software maintainers with prioritized access** to Gemini 3.8 Flash Cyber." — 우선 배포이지 배타적 제한이라고 적지는 않았다.
- 같은 글은 이 모델의 취약점 발견 성공률이 **70%를 넘는다**고 적었다(20개 프로그래밍 언어의 복잡한 코드베이스 기준, 회사 발표).

III 가장 센 기능을 누가 받는지 · 오픈AI의 두 걸음

오픈AI는 이틀에 걸쳐 두 번 적었다. **9월 1일**에는 새 모델의 능력을 자사 기준의 최고 등급으로 분류하면서 **그 능력을 누구에게 줄지를 정했고**, **9월 3일**에는 **그 이용료를 자기가 대겠다고 적었다**.

☞ 9월 1일 · 능력 등급과 배포 방식

- “Astra meets the **Critical** cybersecurity capability threshold under our Preparedness Framework.”
- 회사가 적어 둔 Critical의 정의 — “identify and develop functional zero-day exploits of all severity levels in many hardened real-world critical systems **without human intervention**”
- “Advanced cybersecurity work will **initially be available to a group of testers**, with access through Daybreak Blue following to expand defensive use.”
- 국문: “고급 사이버보안 작업은 **우선 일부 테스터에게만** 열리고, 이어 Daybreak Blue를 통해 방어 목적 사용으로 넓혀 간다.”

Critical은 오픈AI 자체 기준(Preparedness Framework)의 최고 등급이지 공적 인증이 아니다. Daybreak Blue는 보안 목적 이용자에게 따로 여는 창구이며 이전 세대 모델로 운영돼 왔다 — 새 모델의 최상급 기능이 곧바로 열린 것으로 읽지 않는다.

9월 3일 · 10억 달러와 첫 순번

☞ 자원의 크기와 대상

- “**\$1 billion** in subsidized Daybreak access to help **resource-constrained cyber defenders**” — 6개월에 걸쳐 쓰는 것을 목표로 적었다.
- 첫 순번으로 적힌 일곱 갈래 — 상하수도 운영자 · 전력망 운영자 · 주·지방정부 · 지역·커뮤니티 은행 · 비영리단체 · 오픈소스 유지보수자 · 그 밖에 보안 자원이 부족한 조직.
- “Following recent attacks on U.S. water systems, we offered affected states and utilities **up to \$1 million in no-cost API credits, Daybreak access, and technical assistance.**”
- 국문: 미국 수도 시설 공격 이후 피해 주·사업자에게 **무상 API 크레딧과 접근권, 기술 지원을 합쳐 최대 100만 달러 상당**을 제공했다는 것이다 — 현금 지원이 아니다.

세 곳이 같은 방향으로 간 대목

공통점	세 곳의 발표에서 확인되는 것
능력과 접근을 분리했다	모델의 성능은 공개하고, 가장 센 기능은 심사를 거쳐 따로 배분 한다
방어자를 이름으로 지목했다	정부 당국·핵심 인프라 운영자·소프트웨어 유지보수자·지역 금융처럼 업종과 역할로 대상을 적었다
요건 공개는 같린다	구글은 신청 양식과 「윤리적 운영 이력」이라는 요건 을 프로그램 페이지에 적었다. 앤스로픽과 오픈AI는 판정 절차와 요건 을 발표문에 적지 않았다

💡 이 층에서 고르는 쪽은 만드는 쪽이었다

이제까지 최상위 AI 모델을 쓰는 일은 **계약을 맺고 값을 내면 되는** 일이었다. 가장 센 보안 기능에 한해서는 **값을 낼 의사와 별개로 자격이 먼저 판정된다**. 국내 기업이 그 판정에 어떻게 접근하는지는 **이 브리프가 확인하지 못했다**(V장 9면). 다만 II장의 네 칸은 그 판정과 무관하게 오늘 쓸 수 있다 — 자기 회사가 쓰는 도구에 매기는 눈금이기 때문이다.

IV 그 안에서 나온 것

심사를 거쳐 들어간 쪽이 **무엇을 얻었다고 말하는지**는 9월 2일 미국 하원에서 한 차례, **참여 기업 자신의 입으로** 나왔다. 뉴욕증권거래소 사장이 금융서비스위원회 증언에서, 열두 개 회사 공동 심사 프로그램에 초기부터 참여해 자사 결함을 찾아냈다고 밝혔다 — **건수와 내용은 밝히지 않았다.**

☞ 하원 증언 (보도 경유)

- “We’re **an early participant** in Project Glasswing.”
- “We have been able to find a variety of items that we’ve been able to **very quickly address** because of the state-of-the-art technology.”
- 국문: “우리는 Glasswing의 초기 참여자다. 최신 기술 덕분에 여러 건을 찾아냈고 **매우 빠르게 조치할 수 있었다.**”
- 이 발언은 **보도로 확인했고 증언 전문은 대조하지 못했다.**

그 프로그램은 한 회사의 것이 아니다

- Glasswing은 2026년 4월 7일 출범한 **열두 개 회사의 공동 사업**이다 — AWS·앤스로픽·애플·브로드컴·시스코·클라우드스트라이크·구글·JP모건체이스·리눅스 재단·마이크로소프트·엔비디아·팔로알토네트웍스.
- 4월 공개 안내에 따르면 심각도 높은 취약점을 **수천 건**(5월 시점 값은 1만 6,000건 이상 — V장 9면) 찾아냈고 **그중 일부는 주요 운영체제와 웹 브라우저 모두에서** 나왔다. 크레딧 1억 달러와 오픈소스 보안 단체 기부금 400만 달러가 배정돼 있다.
- 뉴욕증권거래소는 **4월 출범 파트너 목록에 없다.** 언제 어떤 경로로 합류했는지는 확인하지 못했다.

원문의 사례에서 사람의 자리는 어디였다

원문이 실은 스무 건 가운데 자율성 수준이 드러난 사례를 모아 보면, **사람의 감독 자리를 낱말로 명시한 것은 IBM 하나이고 아람코·ING 사례에는 그 서술이 없다.**

제출 기관	무엇을 맡겼나	원문이 적은 성과	칸
IBM	경보를 스스로 조사하고 관련 정보를 붙이고 위험도를 매기며, 분석가는 감독과 상황 보고 에 집중한다	일상 조사의 약 95% 를 처리, 월 850시간 이상 자동화, 조사 시간 37% 단축	㉠ 사람이 승인한다 원문 표기 human-in-the-loop
아람코	공격 패턴을 찾아 차단 조치를 스스로 실행 한다. 특허를 받은 자체 해법 50건 이상	보안 해법의 약 40% 가 AI 기반이고 정확도는 99% 라고 회사가 밝혔다	㉢ 뒤집을 수 있다 ~ ㉣ 사람이 없다
ING	자료 유출 경보를 분류하고 급한 순서를 매겨 분석가에게 넘긴다	경보 500만 건 처리, 분석가 정확도 20% 향상	㉠ 사람이 다 본다

성과 수치는 **제출 기관이 스스로 밝힌 값**이다. 오른쪽 ‘칸’ 열은 **이 브리프가 붙인 것**이며 원문에 없다 — IBM은 원문이 **human-in-the-loop**라 적어 ㉠이고, ING는 승인 절차 서술이 없어 ㉠로 보았다.

💡 자동화 비율이 올라가도 칸은 그대로였다

IBM은 일상 조사의 약 95%를 맡기고도 **사람이 승인하는 자리를 유지**한다고 적었다. 설계는 **자동화 비율이 정하지 않는다** — 어느 칸에 두었는지가 정한다. 사람의 일은 **좁고 어려운 쪽으로 옮겨 간다.**

V 「파트너 국가」라는 말

9월 3일 글에서 가장 짧고 가장 멀리 가는 문장은 마지막 대목에 있다.

☞ 확대 계획을 적은 한 문장

- “In the coming weeks, we intend to **expand this model to partner countries.**”
- 국문: “앞으로 몇 주 안에 이 방식을 파트너 국가로 **넓힐 의향**이다.” 표현은 「의향(intend to)」이며 약속이나 일정이 아니다.

그 글에 적혀 있는 것과 적혀 있지 않은 것 · 2026. 9. 7. 확인 기준 · 각 회사 공개 게시물까지

항목	상태	확인된 내용
「파트너 국가」의 정의와 대상국	글에 없음	어떤 나라를 가리키는지 설명이 없고 특정 국가명도 나오지 않는다
요건과 판정 절차	글에 없음	기관용 안내 창구는 있으나 어떤 요건으로 누가 판정하는지 는 적혀 있지 않다
판정하는 주체	공개	세 프로그램 모두 회사가 직접 정하고 공적 인증을 거치지 않는다 — 앤스로픽·구글은 자격 심사, 오픈AI는 단계적 배포다
지리적 제한	일부 공개	앤스로픽 안내는 현재는 미국 조직이라 적으면서, 미국 정부와 협의해 미국 안팎의 파트너로 넓히겠다 고 함께 적었다. 구글·오픈AI 발표문에는 국가 제한이 없다
한국 포함 여부	확인하지 못함	포함되는지 아닌지를 판단할 공개 근거를 찾지 못했다

그사이 국내에서 정해진 것 · 올해 5~8월

- **2026년 6월 22일** — 금융분야 AI 가이드라인 개정본 시행. 일곱 원칙 가운데 하나가 「**AI의 보조적 수단성**」이라고 보도는 전했다. 보도자료 원문을 대조하지 못했으므로 이 브리프는 그 원칙의 범위를 원문의 칸으로 옮기지 않는다. 같은 보도는 **하반기 협의체에서 AI 에이전트 시범운영과 리스크 관리 방안을 검토한다고** 적었다.
- **2026년 8월 20일** — 개인정보 보호법 개정안 국회 본회의 통과(보도 경유). AI 개발 목적의 원본 개인정보 활용 특례가 신설됐고 공포 후 6개월 뒤 시행된다.
- **2026년 5월 29일** — 「**AI 기반 사이버위협 대응 민간 정보보호 추진계획**」(과학기술관계장관회의, 보도 경유). 이 계획은 **Glasswing의 1차 보고서(IV장 8면의 열두 개 회사 공동 사업)**를 인용했다 — 취약점 **1만 6,000건 이상** 가운데 **국내 영향이 확인된 것이 두 건**이라는 내용이다. 1만 6,000건은 국내 제품을 표적으로 모은 것이 아니므로 **두 건을 국내 노출의 총량으로 읽지 않는다**. 중소기업 뒤편으로는 진단 도구 무료 배포와 **소프트웨어 자재명세서(SBOM) 기술지원**을 담았다.

💡 기다릴 것과 오늘 정할 것이 다르다

AI에 무엇을 줄 것인가(데이터)는 8월에 국회를 통과한 개정 법률의 대상이 됐다. **가장 센 능력을 누가 받는가는** 해외 민간 기업이 정하고 있고 국내 접근 경로는 **이 브리프가 확인하지 못했다** — 기다려야 하는 축이다. 다만 **그 산출물은 이미 국내 정책 문서에 들어와 있고, 자기 회사의 도구를 어느 칸에 들지는 오늘 정한다**.

VI 작은 조직이 뒤처진다 · 원문이 먼저 적었다

9월 첫 주의 프로그램들이 **보안 자원이 부족한 방어자**를 첫 순번으로 지목한 것은 우연이 아니다. 원문은 넉 달 전에 그 격차를 적어 두었다.

☞ 원문 도입부의 서술

- “While **77% of organizations** report using AI in cybersecurity, deployment is **closely tied to organizational size and resources**. Larger companies, **supported by greater technical maturity and investment capacity**, report higher adoption rates, whereas **smaller entities, governments and non-governmental organizations (NGOs) tend to lag behind** due to **financial constraints, skills availability and data maturity**.”
- 국문: “조직의 **77%**가 사이버보안에 AI를 쓴다고 답했지만, 도입은 **조직의 규모·자원과 밀접하게 얽혀 있다**. 큰 기업은 기술 성숙도와 투자 여력이 받쳐 주어 도입률이 높고, 작은 조직·정부·비영리는 자금이 빠듯하고 숙련 인력을 구하기 어려우며 데이터도 쓸 만큼 정리돼 있지 않아 뒤처지는 경향이 있다.”

원문이 든 숫자들

값	무엇의 값인가	모수와 출처
77%	사이버보안에 AI를 쓰고 있다고 답한 조직	원문 각주 1의 외부 조사. 원문은 이 값에 규모별 편차 를 함께 적었다
54%	도입의 가장 큰 걸림돌로 숙련 인력 부족을 꼽은 조직	원문 각주 17

두 값 모두 원문이 WEF 「Global Cybersecurity Outlook 2026」에서 인용해 온 것이다(원문 각주 1·17). 원문은 **AI를 광범위하게 쓰는 조직에서 침해 지속기간이 약 80일, 평균 비용이 190만 달러 줄었다는 값도 인용했다 — 그 조직들이 모수다**.

원문이 중소기업 앞으로 따로 적어 둔 네 가지

원문은 도입 절차를 다룬 장(원문 차례의 2장)의 네 절마다 **「SME tip」** 상자를 하나씩 달았다. 큰 조직의 방법을 줄여 쓰라는 것이 아니라 **순서를 다르게 하라는 내용이다**.

단계	원문이 적은 조언	실무로 옮기면
첫째 맞추기	가장 중요한 사업 위험이나 애로에 직접 걸리는 용도 한두 개 에 집중하라	막힌 업무를 먼저 고른다
둘째 준비 확인	소수 인원을 교육 하고 이미 쓰는 도구의 AI 기능 을 활용하며, 기계·에이전트 신원 을 안전하게 다룰 수 있는지 확인하라	지금 계약된 제품의 기능부터 켜고 에이전트에게 사람 계정을 빌려 주지 않는다
셋째 검증	작게 돌려 보는 시험(개념검증)을 빠르고 값싸게 해 값어치를 확인하고, 못 미치면 접거나 방향을 바꿀 준비 를 하라	시작할 때 그만두는 조건 을 함께 적어 둔다
넷째 확장·유지	효과가 측정된 것만 확장하고 넓힐 때마다 성능·비용·위험 을 다시 살펴라	확장할 때마다 다시 쟀다

원문의 SME는 **중소·중견기업**이고, 네 조언은 원문 2.1~2.4절 상자다. 오른쪽 열은 이 브리프의 정리다.

VI 작은 조직이 뒤처진다 · 국내 값은 어떤가

원문이 인용한 값은 모두 국제 조사에서 온 것이다. 국제 값에는 **규모별로 나뉜 수치가 없고** 국내 조사에는 있다 — 아래 표는 항목과 모집단이 다르므로 **각 열을 따로 읽는다**.

항목	원문이 인용한 국제 값	국내 값
정보보호 예산을 편성한 곳	예산이 부족하다고 답한 보안팀 53% (2025년)	예산을 편성·집행한 기업 54.8% — 애로 1위가 예산 확보 49.1%
전담 인력·조직	인력이 부족하다고 답한 팀 55% , 숙련 인력 부족을 최대 걸림돌로 꼽은 곳 54%	250인 이상 33.9% 대 10~49인 1.6% — 전담조직 보유율이 스물한 배 차이 난다
인식과 실행의 간격	원문에 대응 항목 없음	중요하다고 답한 곳 80.6% , 실제 예산을 편성한 곳 54.8% — 25.8%포인트 차이
AI를 보안에 쓰는 비율	77% (작은 조직·정부·비영리는 뒤처진다고 원문이 함께 적었다)	이 값에 대응하는 국내 통계는 이 브리프가 확인하지 못했다

출처는 「2025 정보보호 실태조사」(과학기술정보통신부·한국정보보호산업협회, 종사자 10인 이상 5,500개 사업체, 2025년 9~12월 조사)이고 보도 경유다. 규모별 전담조직 보유율만 「2024 정보보호 실태조사」 값이며 재인용이다. 연도와 문항이 다르므로 두 조사의 값을 한 줄로 이어 읽지 않는다. 첫 행과 둘째 행은 국제 값과 국내 값이 묻는 것이 서로 반대다(부족 응답률 대 편성률·보유율) — 크기를 견주지 않고 각각 읽는다.

예산을 세우지 않은 이유 가운데 하나

- 예산을 편성하지 않은 기업이 든 사유에서 「어떤 활동이 필요한지 모르겠다」가 **33.4%**였다 — 「우리 사업은 정보보호와 무관하다」(37.0%) 다음으로 많다.
- 이 대목이 원문의 표가 쓰이는 자리다. 예산을 세우지 않은 기업의 **33.4%**가 무엇을 해야 할지 모른다고 답했다면, 자율성을 네 칸으로 나눈 표는 그 회사들이 오늘 바로 쓸 수 있는 도구가 된다(II장 4면).

같은 무렵 국내에서 시행된 시행령

2026년 7월 21일 시행된 인공지능 기본법 개정 시행령은 권리·안전에 큰 영향을 주는 AI(고영향 인공지능)를 들여올 때 영향을 미리 따져 보게 했고, 거기 반영할 「인공지능취약계층」 열한 가지에 중소기업 재직자를 넣었다고 보도는 전했다. 국제 조사가 「작은 조직이 뒤처진다」로 적은 자리를, 국내 법령은 다른 갈래(보호 대상)에서 다루고 있는 셈이다.

시행령 조문은 조사 환경의 접근 제약으로 원문 대조를 하지 못했다. 열한 가지의 전체 목록·조항 번호와 그 범주의 효과는 이 브리프가 확인한 범위 밖이다.

🔗 1.6%가 정하는 것은 순서다

세 프로그램 가운데 둘은 요건을 밝히지 않았고 국내 접근 경로도 이 브리프가 확인하지 못했다(III장 7면·V장 9면). 1.6%가 뜻하는 것은 자격의 유무가 아니라 순서다. 밖에서 정해지는 것은 기다리되, 도구를 적고 몇 번 칸인지 쓰는 일은 예산도 전담조직도 없이 오늘 시작된다(VII장 13면).

VI 작은 조직이 뒤처진다 · 직접 만들 것인가, 사 올 것인가

원문은 이 선택을 여덟 갈래로 나눠 비교한 표를 실었다. 두 길 가운데 하나를 고르기 전에 여덟 항목 중 어느 것이 그 회사에 무거운지부터 정하라는 쓰임이다. 원문은 실제로는 많은 조직이 둘을 섞어 쓴다고 함께 적었다.

항목	직접 만들 때	사 올 때
도입 속도	개발·학습·자원 배치가 필요해 느리다	완성된 해법이라 빠르다
자료 통제권	데이터·구조·모델을 더 많이 진다	철 수 있는 범위가 좁고 바깥 업체에 기대게 된다
인력	사내에 숙련 인력이 많이 필요하다	사내 수요가 적고 밖의 전문 인력을 쓸 수 있다
유연성	특수한 보안 업무에 맞춰 고칠 수 있다	고칠 여지가 적고 공급업체 종속 위험이 있다
위험	개발 실패 위험과 유지보수 책임을 진다	공급업체가 계속 남아 있을지, 자료를 다른 곳으로 옮길 수 있을지, 규제를 지킬 수 있을지가 위험이 된다
비용	초기 도입·유지비를 합친 부담이 크고 장기 비용은 낮다	초기 비용은 낮고 연계·확장에서 누적 비용이 커진다
확장성	인프라와 중단 없이 돌리는 일을 스스로 진다	공급업체가 정한 한도 안에서만 늘릴 수 있다
어디에 맞나	AI 자체가 그 회사의 차별점이고 경쟁력의 원천인 경우	AI가 보편 도구이고 빨리 값어치를 내야 하는 경우

원문 표 1을 옮겼다. 항목 이름은 이 브리프가 우리말로 바꾼 것이다. 원문은 이 표 앞에 「많은 보안 도구가 이미 AI 기능을 담고 있으므로 어느 쪽을 정하기 전에 지금 쓰는 도구부터 점검하라」고 적었다.

이 표를 도내 중소 제조기업에 견주어 보면

- 「자료 통제권」 항목이 설비·공정 자료에서 무겁게 걸린다. 도면·레시피·불량 이력은 밖으로 나가면 되돌릴 수 없는 자료다. II장의 기준을 그대로 옮기면 되돌릴 수 없는 쪽이고, 그래서 사 올 때는 자료가 어디에 남는가가 값보다 먼저 오는 항목이 된다.
- 「인력」 항목은 원문의 SME 조인과 짝을 이룬다. 사내에 보안 전담이 한두 명인 회사에서 직접 만드는 길은 사실상 닫혀 있다. 원문이 소수 인원 교육과 지금 쓰는 도구의 기능부터 보라고 적은 이유가 그 항목에 있다.
- 「어디에 맞나」 항목이 판정 문장이다. 보안은 대부분의 제조기업에서 보편 도구 쪽에 가깝다. 이 항목을 그대로 적용하면 판정이 사 오는 쪽으로 기운다 — 원문 자신은 많은 조직이 둘을 섞어 쓴다고 함께 적었다. 직접 만드는 것은 그 회사만의 공정 지식이 걸릴 때다.

이 장이 확인하지 못한 것

- 국내 제조기업이 AI에 실행 권한을 준 사례 — 생산관리·조달·품질·설비 운영에서 세 번째 칸 이상으로 운영한다고 공개된 국내 사례를 찾지 못했다.
- 도내 중소기업의 사이버보안 AI 도입률 — 원문의 77%에 대응하는 국내 값이나 경기도 단위 조사를 확인하지 못했다.
- 세 프로그램에 국내 기관이 접근한 사례와 AI 에이전트를 승인 없이 배치한 국내 사례 — 공개된 것을 찾지 못했다. 없다는 뜻이 아니라 이 브리프가 확인하지 못했다는 뜻이다.

VII 경기도 기업에 미치는 영향

AI 도구가 사람의 승인 없이 무언가를 실행하는 구조는 이미 개발·생산·조달 현장에 들어와 있고, 도내 기업은 서 있는 자리에 따라 걸리는 대목이 다르다.

☞ 설비 제어·자동화 소프트웨어를 만들거나 납품하는 곳

- 외부 저장소를 여는 절차가 곧 공격이 들어오는 통로다. 발주처·협력사가 보낸 코드, 오픈소스 라이브러리, 장비사 예제 코드를 AI 코딩 도구로 여는 때가 그 자리다.
- 도구별 버전이 관리 항목이 된다. 개발 책임자가 도구 목록과 버전을 사람 목록처럼 관리하고, 외주 개발사에도 어떤 도구를 어느 버전으로 쓰는지 묻는다(I장 3면 표가 그 점검표다).

☞ AI를 들여와 공정·품질·조달에 쓰는 곳

- 무엇을 어느 칸에 들지 먼저 적는다. 원문의 기준은 얼마나 크가와 되돌릴 수 있는가 둘이다. 설비 파라미터 변경·발주 확정·품질 판정은 되돌리기 어려운 쪽이므로 사람의 승인이 붙는 칸에 둔다. 적어 둔 칸이 발주처의 물음에 그대로 답이 된다.
- IT 담당자가 에이전트 전용 계정을 만든다. 원문은 준비 확인 단계에서 기계와 에이전트의 신원을 안전하게 다룰 수 있는지 확인하라고 적었다(VI장 10면). 사람 계정을 빌려 쓰면 사고 뒤에 누가 무엇을 했는지 가릴 수 없다. 에이전트가 여럿이면 서로 주고받는 경로도 함께 적어 둔다(II장 5면).
- 고장 상황을 미리 돌려 본다. 원문은 AI가 멈춰도 운영이 돌아가도록 대비 장치를 설계하라고 적었다. 공정 책임자가 자동 검사 장비가 멈췄을 때의 수동 절차를 문서로 갖추고 있는지가 그 항목이다.

도내 산업 구성에서 특히 걸리는 자리

자리	이번 호의 어느 대목과 만나나
반도체 장비·부품의 소프트웨어	장비 제어 코드를 외주·협력 구조로 주고받으면 남의 저장소를 여는 일이 일상이다 — I장의 결함이 작동하는 조건이 그 안에 있다(3면)
산업단지의 용수·전력 설비	9월 3일 지원의 첫 순번에 상하수도·전력망 운영자가 있으나 그 목록은 미국의 사업자·지방정부이고 다른 나라 확대는 「의향」 단계다(V장 9면). 도내 설비에 오늘 닿는 것은 설비 제어 계통 AI 도구의 칸과 11월 말 마감 지원사업이다(◆ 14면)

이 표는 도내 산업 구성과 이 호의 내용을 잇는 정리이며, 특정 기업이나 단지의 현황을 조사한 결과가 아니다.

국내 기관이 낸 대응 문서가 이미 있다

- 과학기술정보통신부·한국인터넷진흥원 「AI 보안 위협 대응 매뉴얼」(2026. 7. 8., 보도 경유)은 위협을 데이터·모델·에이전트·공급망·고성능 시스템 다섯 범주로 나누고 제조·에너지 등 여덟 개 분야별 대응책을 붙였다.
- 밖에서 나온 결과는 이미 국내 문서를 거쳐 들어와 있다(V장 9면). I장의 결함은 그 「공급망」에 들어간다 — 외부 코드가 실행 경로를 여는 형태다. 매뉴얼 원문은 이 브리프가 대조하지 못했다.

🔍 품이 가장 적게 드는 조치

지금 쓰는 AI 도구를 목록으로 적고 각각 몇 번 칸인지 옆에 쓰는 일이다. 비용이 들지 않고, 원문이 준비 확인 단계에 둔 항목이며, 다 적고 나면 어느 도구에 아무도 칸을 고르지 않았는지가 드러난다. 이 호의 표제 「고른 사람이 없는 칸」은 그 자리를 가리키며, 원문에 없는 이 브리프의 표현이다.

점검 항목

AI 도구를 들여와 쓰는 기업의 점검표

- Q1.** 지금 쓰는 AI 도구를 목록으로 적었을 때 각각이 **네 칸 가운데 몇 번인지** 답할 수 있는가? → 원문은 칸을 고르는 기준을 **위험과 되돌릴 수 있는 정도로** 적었다(II장 4면).
- Q2.** 그 가운데 **사람의 승인 없이 실행되는 동작**이 있는가? 그것을 회사가 **알고 고른 것인가**? → I장의 결함은 사용자가 고르지 않은 채 가장 높은 칸에 놓였던 경우다(3면).
- Q3.** 에이전트가 **사람 계정의 권한**으로 움직이고 있는가? → 원문은 준비 확인 단계에서 **기계와 에이전트의 신원**을 안전하게 다룰 수 있는지 확인하라고 적었다(VI장 10면 · 실무 적용은 VII장 13면).
- Q4.** 그 도구가 **멈췄을 때의 수동 절차**가 문서로 있는가? 마지막으로 돌려 본 것이 언제인가? → 원문은 **고장 상황을 미리 모의해 보고 대비 장치를 설계하**라고 적었다(II장 5면).

소프트웨어를 만들거나 납품하는 기업의 점검표

- S1.** 팀이 쓰는 AI 코딩 도구의 **이름과 버전**을 지금 댈 수 있는가? 그 가운데 **수정이 확인되지 않은 도구**가 있는가? → **9월 1일 기준(수정 상태는 보도 경유)** 미확인이 셋, 부분 수정이 하나, 상태 미기재가 하나다. 그 뒤 갱신했을 수 있으니 배포 기록을 오늘 확인한다(I장 3면).
- S2.** 외부에서 받은 코드 저장소를 **어떤 절차로 여는가**? 여는 행위 자체가 실행을 부를 수 있다는 점이 그 절차에 반영돼 있는가? → 공개 문서는 해당 git 설정을 꺼 두는 방식을 완화책으로 적었다(I장 3면).
- A1. 공통 —** 발주처가 **「AI 도구를 어떤 권한으로 쓰는가」**를 물었을 때 내놓은 문서가 있는가? → 도구별 칸을 적어 둔 목록이 그 문서다(VII장 13면). 직접 만들 것인가 사 올 것인가는 **여덟 항목 중 무엇이 무거운지**부터 적는다(VI장 12면).

향후 일정

시점	무슨 일	주목 포인트
2026. 11. 30. 접수 마감	「2026년 ICT 중소기업 정보보호 지원사업」 수요기업 모집 — 접수는 risc.kisa.or.kr (지역정보보호센터 온라인 접수), 주관은 과학기술정보통신부·한국인터넷진흥원이고 경기 배정분은 경기정보보호지원센터(판교) 가 받는다. 문의 02-405-5031	접수는 6월 30일~11월 30일 이고 1차 접수는 7월 31일에 끝났다. 대상은 중소기업기본법 제2조 중소기업 가운데 지역 전략산업 영위 기업과 침해사고 피해 기업 이며 중견·대기업은 제외다. 경기 배정분 우선순위와 예산 소진 시 조기 마감 은 (보도 경유)이므로 공고에서 확인한다
2026년 9~10월	오픈AI가 9월 3일 글 기준 「앞으로 몇 주 안에」 라고 적은 파트너 국가 확대 의 이행 여부	대상 국가와 선정 기준 이 함께 공개되는지. 9월 7일 기준 어느 것도 공개돼 있지 않다
상시	I장 표에서 수정이 확인되지 않은 도구 셋 과 부분 수정 하나의 갱신	각 도구의 배포 기록. 오늘 바로 확인할 수 있는 항목 이다

* 용어 설명

자율성 단계 — AI에게 어디까지 맡길지를 나눈 칸. 원문은 넷으로 나뉘고 이 브리프는 ④ **사람이 다 본다** · ⑤ **사람이 승인한다** · ⑥ **뒤집을 수 있다** · ⑦ **사람이 없다**로 부른다(II장 4면). ①은 사람이 대응을 짜고, ②는 사람이 승인한 뒤 실행되며, ③은 AI가 되돌릴 수 있는 조치를 하고 사람이 지켜보다 뒤집고, ④는 감독을 다른 AI나 사후 점검이 맡는다.

되돌릴 수 있는 행동(reversible action) — 잘못했을 때 원래대로 돌려놓을 수 있는 조치. 원문은 자율성 칸을 **행동의 위험과 되돌릴 수 있는 정도로** 고르라고 적었다. 수상한 접속을 잠시 막는 것은 되돌릴 수 있고, 설비 설정을 바꾸거나 발주를 확정하는 것은 되돌리기 어렵다.

자율성 대 권한 — 2025년 11월 백서가 가른 두 눈금. **자율성**은 목표를 세우고 언제 어떻게 움직일지 정하는 쪽이고, **권한**은 손댈 수 있는 범위(위키 전유부터 완전 관리자 권한까지)다. 백서는 둘 다 **설계 단계에서 정하는 선택**이며 운영 중에 다시 조정할 수 있다고 적었다.

소프트웨어 자재명세서(SBOM) — 소프트웨어가 어떤 부품으로 만들어졌는지 적어 둔 목록. 외부 코드에 결함이 생겼을 때 **우리 제품 어디에 그것이 들어 있는지를** 찾는 데 쓴다.

심사 접근(vetted access) — 만드는 쪽이 쓰는 쪽의 자격을 먼저 심사해 기능을 배분하는 방식. 값을 내면 쓸 수 있는 일반 판매와 다르다. 앤스로픽과 구글은 자격 심사를 명시했고 오픈AI는 초기 테스터에게 먼저 여는 단계적 배포다 — 셋 다 일반 판매와 다르지만 판정 방식은 같린다(III장 6~7면).

기계-에이전트 신원 — 사람이 아닌 프로그램에게 따로 부여하는 계정과 권한. 사람 계정을 빌려 쓰면 무엇을 누가 했는지 사후에 가릴 수 없다. 원문은 준비 확인 단계의 점점 항목으로 적었다.

샌드박스 — 프로그램을 가둬 놓고 돌리는 격리 공간. I장의 결함은 도구가 부르는 보조 프로그램이 그 **올타리 밖에서** 돈 경우가.

제로데이 · CVE — 제로데이는 아직 알려지지도 고쳐지지도 않은 보안 결함이다. CVE는 공개된 결함에 붙이는 국제 공통 번호로, **번호가 붙었다는 것은 확인했다는 뜻이지 고쳐졌다는 뜻이 아니다**.

개념검증(PoC) — 본격 도입 전에 작게 돌려 보는 시험. 원문의 중소기업 조연은 **빠르고 값싸게** 하되 **그만큼 조건을 미리 정해** 두라는 것이다.

* 출처 (Sources)

- World Economic Forum — 「Empowering Defenders: AI for Cybersecurity」, White Paper, 2026년 5월, 34쪽. KPMG 협업, Cyber Frontiers 이나서티비. 15개 산업 84개 조직 105명 참여 워크숍, 파트너 제출 사례 20건 (원문 PDF 전문 직접 대조)
weforum.org/publications/empowering-defenders-ai-for-cybersecurity/
- World Economic Forum — 「AI Agents in Action: Foundations for Evaluation and Governance」, White Paper, 2025년 11월, Capgemini 협업. 원문이 3.3절에서 구체적 통제 수단을 담은 문서로 지목한 **선행 백서**다. 분류 일곱 개 축, 자율성-권한의 구분, 제로트러스트 서울 (원문 PDF 전문 직접 대조)
weforum.org/publications/ai-agents-in-action-foundations-for-evaluation-and-governance/
- 원문이 다른 조사에서 인용해 온 값 — **77%-54%**는 WEF 「Global Cybersecurity Outlook 2026」(원문 각주 1-17), **80일-190만 달러**는 IBM 「Cost of a Data Breach 2025」(각주 3), **53%-55%**는 ISACA(각주 7). 원문이 각주로 밝힌 값을 그대로 옮겼고 이 브리프가 각 원지료를 따로 검증하지는 않았다. 모수는 본문 표에 함께 적었다
- Anthropic — 「Claude Fable 5.1 and Claude Mythos 5.1」 공식 게시글, 2026년 9월 1일. 같은 모델 · 안전장치 수준 차이, 검증 프로그램 두 갈래
anthropic.com/claude-fable-and-mythos-5-1
- Anthropic — 「Project Glasswing」, 안내, 2026년 4월 7일. 12개 참여사, Claude Mythos Preview, 크래쉬 1억 달러와 오픈소스 보안 단체 기부 400만 달러, 출범 이후 40여 개 조직 추가
anthropic.com/glasswing
- OpenAI — 「Path to Astra: critical capabilities and frontier safeguards」, 2026년 9월 1일. Preparedness Framework상 Critical 분류와 그 정의, 초기 테스터 제한 및 Daybreak Blue 확대 경로
openai.com/index/path-to-astra/
- OpenAI — 「Daybreak for Frontline Defenders: \$1B to protect essential services」, 2026년 9월 3일. 6개월 10억 달러, 첫 순번 일곱 갈래, 미국 수도 시설 공격 이후 최대 100만 달러 상당 지원, **“In the coming weeks, we intend to expand this model to partner countries.”**
openai.com/index/daybreak-for-frontline-defenders/
- Google — 「Gemini 3.8 Flash and 3.8 Flash Cyber」, 공식 블로그, 2026년 9월 2일. Fairwind Program의 대상, 취약점 발견 성공률 70% 초과(회사 발표)
blog.google
- Manifold Security — 「GitSpawn」 취약점 공개, 2026년 9월 1일. 실행 경로와 본문 인용문, **CVE-2026-72718-71963**, 완화책
manifold.security
- CVE-2026-19592-55607과 도구별 수정 버전-미수정 상태는 (보도 경유)**다
- 뉴욕증권거래소 사장의 미국 하원 금융서비스위원회 증언(2026년 9월 2일) 및 인용문 — **보도 경유**. 증언 전문과 청문회 정식 명칭은 조사 환경의 접근 제약으로 대조하지 못했다
- 훈련 중인 에이전트들이 외부 공개 위키를 연락 수단으로 쓴 사례(2026년 9월 4일 공개) — **복수 매체 보도 경유**. 이 브리프는 원 연구 보고서를 대조하지 못했다
- 정보보호 실태조사(과학기술정보통신부-한국정보보호산업협회) — **2025년판**: 총조사 10인 이상 **5,500개 사업체**, 2025년 9월 15일~12월 15일 조사, 2026년 3월 발표. 예산 편성 54.8%-인식 80.6%-어로 1위 예산 확보 49.1%-미편성 사유 「어떤 활동이 필요한지 모르겠다」 33.4%. **2024년판**: 전담조직 보유율 250인 이상 33.9% 대 10~49인 1.6%. **둘 다 보도 경유**이고 2024년판은 **재인용**이다. 보고서 원문 PDF는 확보하지 못했다
- 「AI 기반 사이버위협 대응 민간 정보보호 추진계획」(2026년 5월 29일 제9회 과학기술장관회의) — 앤스로픽 Glasswing 1차 보고서 인용(취약점 1만 6,000건 이상-국내 영향 2건), 중소기업 대상 SBOM 기술지원 — **보도 경유**
- 과학기술정보통신부-한국인터넷진흥원 「AI 보안 위협 대응 매뉴얼」, 「AI 보안 레드티어링 가이드」(2026년 7월 8일) — 위협 5범주-8개 분야, ISO/IEC 42119-7 반영 / 인공지능 기본법 개정 시행령(2026년 7월 21일 시행)의 「인공지능취약계층」, 11가지에 **중소기업 재직자 포함** — **둘 다 보도 경유**이고 원문-조문을 대조하지 못했다
- 같은 사업의 경기 배경분 — 우선순위(개인정보 보유 기업-AI-제조 등 전략산업), 침해사고 경험 기업의 신청 자격, 예산 소진 시 조기 마감 — **보도 경유**(2026년 8월 4일 발표, 복수 매체 일치). 경기정보보호지원센터는 **한국인터넷진흥원이 판교 정보보호클러스터에서** 운영하는 지역센터이며 이 브리프의 발행처와는 별개 기관이다
- 금융위원회 — 금융분야 AI 가이드라인 개정 관련 보도자료(2026년 6월 18일). 일곱 원칙과 「AI의 보조적 수단성」, 하반기 협의체의 AI 에이전트 시범운영-리스크 관리 검토 계획 — **보도 경유**이며 보도자료 원문은 대조하지 못했다
fsc.go.kr
- 개인정보 보호법 개정안 국회 본회의 통과(2026년 8월 20일), AI 개발 목적 원본 개인정보 활용 특례와 공포 후 6개월 시행 — **보도 경유**. 의원 원문은 접근 제약으로 대조하지 못했다
- 과학기술정보통신부-한국인터넷진흥원 — 「2026년 ICT 중소기업 정보보호 지원사업」, 수요기업 모집 공고(2026년 6월 26일 게시). **접수기간 2026년 6월 30일~11월 30일**, 1차 접수 7월 31일 마감, 대상은 중소기업기본법 제2조 중소기업 중 지역 전략산업 영위 기업과 침해사고 피해 지역 중소기업, 접수 risc.kisa.or.kr, 문의 02-405-5031 (**중소벤처기업부 기업마당** 게시 공고문으로 확인)
kisa.or.kr · risc.kisa.or.kr · bizinfo.go.kr

알려드립니다

검증. 적대적 검수(레드팀) · 한국어 교열·독자 시점 검수·정부 심의를 거쳐 발행합니다. WEF 원문 34쪽과 2025년 11월 백서 「AI Agents in Action」을 모두 원문 파일로 직접 대조했습니다. 9월 사건의 영문 인용은 각 회사 공식 게시글에서 옮겼고, **생략한 대목은 말줄임표로** 표시했습니다. 보도를 경유한 항목은 본문에 (보도 경유)로 구분했고, 단일 매체에서만 확인된 수치는 신지 않았습다.

모수. VI장 표의 값은 모두 원문이 다른 조사에서 인용해 온 것이며 원문의 자체 조사가 아닙니다. 80일-190만 달러는 「AI를 광범위하게 쓰는 조직」이 모수여서 전체 평균이 아닙니다. **77%와 54%**는 같은 조사(Global Cybersecurity Outlook 2026, 원문 각주 1-17)에서 왔고, 53%-55%는 ISACA 조사입니다.

이 브리프가 붙인 것. II장 표의 국문 단계 이름, IV장 표의 「칸」, 열, V장 표의 상태 구분, VI장 11면의 국제-국내 대조, VII장 표는 원문에 없는 **이 브리프의 정리**입니다.

국내 값의 등급. VI장 11면의 국내 통계와 V장 9면-VII장 13면의 국내 정책 문서는 **모두 보도를 경유했습니다**(◆ 14면의 지원사업 공고는 예외로 공고문을 확인했습니다) — 조사 환경의 접근 제약으로 보고서-조문-회의자료 원문을 대조하지 못했습니다. 「2025 정보보호 실태조사」와 경기 배경분 지원사업은 **복수 매체가 일치**하는 것을 확인했고, 「2024 정보보호 실태조사」의 규모별 값은 **재인용**입니다. 서로 다른 연도의 조사를 한 줄로 이어 읽지 않도록 본문에 연도를 함께 적었습니다.

확인하지 못한 것. 세 프로그램의 심사 기준과 절차, 「파트너 국가」의 정의-대상국-신청 경로, 국내 기관의 접근 사례, 국내 제조기업이 AI에 실행 권한을 준 사례, 원문의 **77%**에 대응하는 국내 통계는 확인하지 못했습니다. 공개 근거를 찾지 못했다는 뜻입니다.

원문 이용. 원문 판권장은 “All rights reserved”와 전제-전송 금지를 명시했습니다. 출처를 밝힌 인용-번역 범위에서만 썼습니다.

작성 도구. 자료 조사와 초안 정리에 **생성형 인공지능을 보조 도구로** 활용했고, 인용한 수치-문장은 모두 원문과 대조했습니다.

이 문서의 성격. 수치와 제도 서술은 **2026년 9월 7일 확인 기준**이며, 이 브리프의 시사점은 공개 자료에 근거한 분석으로 특정 기관의 확정 계획이나 정책 제안이 아닙니다.

본문에 나오는 기업·기관·제품은 사실 확인의 출처로 적은 것이며 **특정 제품이나 업체를** 권하는 것이 아닙니다.