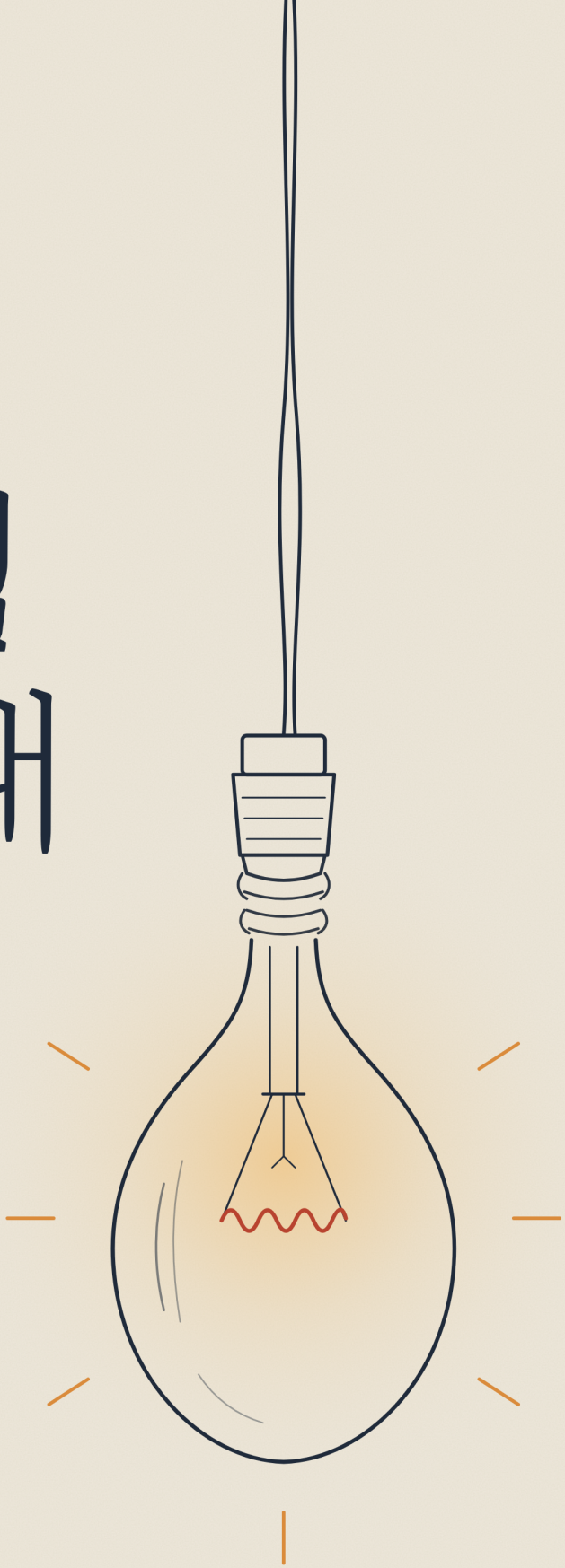
AI 네이티브 핸드북 2026

지능을 전기처럼 쓰는 시대



스위치를 올리면 전등이 켜지듯,
이제는 버튼 하나로 지능을 씁니다.

석탄 · 증기 → **석유** · 전기 → **전기** · 정보 → **전기 · 지능**

전종홍 지음

CC BY NC ND 2026

AI 네이티브 핸드북 2026

지능을 전기처럼 쓰는 시대

전종홍 지음

버전 1.3

지능을 전기처럼 쓰는 시대 — AI 네이티브 핸드북 2026

전종홍 지음

버전 1.3 · 2026년 10월 23일 고침 · 2026년 3월 14일 처음 만듦

© 2026 전종홍. 이 책은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스(CC BY-NC-ND 4.0)를 따릅니다.
저작자를 밝히면 비영리 목적으로 자유롭게 복제·배포할 수 있지만, 내용을 고치거나 2차 저작물을 만들 수는 없습니다.
라이선스 전문: https://creativecommons.org/licenses/by-nc-nd/4.0/deed.ko
본문의 도표·삽화·표지는 저자가 직접 만든 것으로 같은 라이선스를 따릅니다. 인용한 자료의 저작권은 각 원저작자에게 있습니다.

이 책은 계속 고쳐 펴내는 살아 있는 문서입니다. 판마다 버전과 고친 날을 이 쪽에 적습니다. PDF·EPUB 최신판 내려받기:
hollobit.github.io/ai-native/

목차

서문 7
라디오 잡지 소년, 네 번째 신대륙을 보다 7
세 가지 독서 코스와 독자별 항로 8
일러두기 10
제1부. 4차 AI산업혁명이 온다 13
1. 증기기관은 근육을 대체했고, 전기는 세상을 밝혔다 14
2. 네 번째 에너지의 이름은 '지능'이다 17
3. 지능의 단가가 빠르게 낮아지고 있다 21
4. 얼마나 빨리 빨라지는가 24
5. 전기를 거부한 공장은 어떻게 되었는가 28
6. AI는 결국 에너지 산업이다 31
제2부. 지능 에너지 시스템 37
1. 지능 발전소: 파운데이션 모델은 어떻게 지능을 만드는가 38
1-1. 소형 발전기와 태양광: SLM과 온디바이스 AI 43
2. 변환소: 전기를 지능으로 바꾸는 인프라 46
3. 송전망: MCP와 A2A, 지능이 흐르는 길 49
4. 에이전트: 지능으로 움직이는 가전제품 53
5. 독자 모델을 가진 나라, 빌린 나라 56
6. 지능을 만드는 데도 전기가 필요하다: AI 에너지 전쟁 63
7. 미국, 지능 발전소를 짓다 69
8. 중국의 여섯 갈래 전력망 74
8-1. 중국 오픈웨이트 전략: 설계도를 공개해 표준을 잡다 78
9. 대한민국의 대도약, 세 개의 메가 프로젝트 81
10. 지능 에너지 경제: 토큰 이코노미 87
10-1. AI 생산원가: 토큰값이 아니라 일값 95
11. 지능을 수출하는 나라 100
12. 스스로를 개선하는 지능: 재귀적 자기개선(RSI) 108
13. 초지능(ASI)으로 가는 네 가지 길 113
제3부. AX의 여정 – Enabled에서 Native로 119
1. AI 네이티브란 무엇인가 120
2. 덧붙이기의 한계 123
3. AX란 무엇인가: 네 단계의 사다리 126
4. AI-Ready와 AI-Enabled: 전선을 깔고 전등을 켜다 129
5. AI-First와 AI-Native: 업무를 다시 설계하다 132
6. 세 가지 전환 경로: 출발점이 다르면 길도 다르다 135
6-1. 아래에서 오르는 사다리: 공무원이 직접 만든 589개의 AI 139
7. AX 실패의 패턴 146
제4부. 사고법과 설계 원칙 150

1. 전기가 들어왔을 때 공장 배치도가 바뀌었다 151
2. AI를 빼면 무엇이 남는가 테스트 153
3. 원칙 1: 엔진을 설계하라 157
4. 원칙 2: 결과를 팔아라 160
5. 원칙 3: 의도를 읽어라 163
6. 원칙 4: 플라이휠을 돌려라 166
7. 원칙 5: 에이전트에게 맡겨라 171
8. 원칙 6: 한 팀으로 일하라 174
9. 원칙 7: 신뢰를 내장하라 177
10. 원칙 8: 확률을 경영하라 182
11. 원칙 9: 최적 조합을 찾아라 186
12. 호모 프롬프트: 지능과 대화하는 습관 189
13. 호모 프롬프트의 10가지 실천 원칙 191

제5부. 산업별 지능 에너지 혁명 195

1. 코드를 쓰는 AI, 코드를 검증하는 인간 196
1-1. 테스트와 리뷰를 잘하는 법: 검증 격차를 메우기 198
2. AI를 동료로 맞이한 의사들 201
3. 불량률 0%를 향한 공장 204
4. 네트워크와 브라우저의 지능화: 데이터가 아니라 결정을 나른다 206
5. AI 과학자가 실험실에 들어왔다 209
5-1. 자율 연구원: AI 과학자와 검증의 병목 212
6. 당신의 AI 동반자 216
7. 지능을 어떻게 시험하는가 219
8. K-제조, 지능 에너지를 어디에 쓸 것인가 224
9. 특화 파운데이션 모델: 왜 병원과 은행은 각자의 AI가 필요한가 228
10. 대한민국의 독자 파운데이션 모델 계획: 선언이 아니라 청사진 233

제6부. 지능 에너지와 에이전트가 만드는 세상 239

1. 개인 에이전트의 탄생 240
2. 한 사람의 에이전트에서 생태계로: OpenClaw와 Hermes 244
2-1. 에이전트를 똑똑하게: 스킬, 메모리, 드림, 루프 249
3. 에이전트가 일하는 세계 253
4. 에이전트 경제: AI가 AI를 고용하는 시대 256
4-1. 중국의 에이전트 국가표준: 두 개의 송전망 259
4-2. 에이전트가 사람을 고용하는 세상 263
5. 에이전트가 몸을 갖게 될 때: Physical AI 267
6. 로봇이 출근하는 날 271
7. SOUL.md를 가진 에이전트 275
8. 나를 대신하는 에이전트 278
9. 에이전트들의 사회 281
10. 자율 AI 해킹과 방어: 지능이 지능을 공격하는 시대 286
10-1. Mythos 사태: AI 보안의 분수령 292
10-2. 자율 공격, 자율 방어: AI가 AI를 지키는 시대 297

11. 오케스트레이터: 에이전트 시대 인간의 역할 300
11-1. 하네스 엔지니어링: 에이전트 = 모델 + 하네스 303
11-2. 여럿이 짓는다: 멀티 에이전트 협업과 Human-AI 팀 307
12. 코드를 쓰는 일에서 소프트웨어를 책임지는 일로 312
12-1. 2030년, 소프트웨어는 어떻게 만들어질까 316
12-2. 2040년, 에이전트와 로봇의 미래 320
제7부. 지능 에너지 사회를 향하여 325
1. 전기가 민주주의를 바꿨듯이 326
2. 지능 에너지 격차: 새로운 디바이드 329
3. 법, 노동, 안전: 세 개의 매듭 333
4. 2030년, 지능이 전기처럼 흐르는 하루 338
5. 일곱 개 국가의 선택 343
6. 2030년, AI는 혼자서 당신을 위해 무엇을 할까 349
7. 대한민국 AI 행동계획: 지능 에너지 강국을 향하여 353
8. 대한민국의 AI 네이티브 미래 360
8-1. 대한민국의 AI 기본사회 계획 363
8-2. AI 시대의 공공 API와 공공 데이터: 개방 1위, 그다음은 369
9. 글로벌 AI 기본사회를 향한 제안 373
10. 천하삼분지계: AI 3강 전략 379
11. 모두를 위한 지능 에너지 384
부록 388
A. 증거 점검 40문 + 원칙 매트릭스 + AI 리스크 한 장 요약 389
B. AI 기업 사례 카드 396
C. 용어집 400
D. 약어집 405
E. 찾아보기 409

서문

지은이 소개

전종홍

지능이 전기처럼 흐르는 시대, 그 전기의 '안전 규격'과 '시험 규격'을 만드는 사람

새로운 에너지가 세상에 퍼질 때마다, 누군가는 그 에너지를 믿고 쓸 수 있게 만드는 규격을 설계해야 했다. 전기 시대에 접지 표준과 차단기 규격이 있었다면, 지능 에너지 시대에는 AI를 시험하고 검증하는 국제표준이 있다. 저자는 바로 그 자리에서 일해 온 사람이다.

한국전자통신연구원(ETRI) 책임연구원이자 대통령직속 국가AI전략위원회 민간위원으로, 세계 최초의 AI 시스템 테스팅 국제표준 문서인 ISO/IEC TS 42119-2(2025년 발행, 기술규격) 제정을 주도했다. AI 의료기기 성능평가, AI 레드팀 테스팅, AI 신뢰성 라벨 등 여러 국제표준의 의장과 에디터를 맡아 "AI를 어떻게 믿을 것인가"라는 질문에 세계가 함께 쓸 답을 만들어 왔다. 아시아 최초의 '첨단디지털의료제품 레드팀 챌린지'를 총괄 기획했고, 동국대학교 의료기기산업학과 객원교수와 식약처·복지부 위원으로 의료 AI의 규제·정책 체계에 기여했으며, '대한민국 AI 행동계획'의 공동 저자로 국가 AI 전략 설계에 참여하고 있다.

이 책의 곳곳에는 그 경험이 배어 있다. "신뢰를 내장하라"는 설계 원칙(제4부), 지능을 시험하는 방법(제5부), AI가 공격하고 방어하는 세계(제6부), 경영 회의에 바로 올릴 수 있는 리스크 대시보드(부록 A)까지 — 표준 회의실과 레드팀 현장에서 얻은 것들이다.

2026년부터는 AI 코딩 에이전트와 함께 99개의 행동계획 과제를 소개하는 "AI행동계획 대시보드", 5,300페이지 예산서를 분석한 "AI재정계획 대시보드", AI 국가표준 기본계획 대시보드, 국회의 공공 데이터를 AI가 바로 가져다 쓸 수 있게 잇는 국회 API MCP를 직접 만들었다. 공무원과 시민이 직접 만든 공공 AX 사례를 모으는 '모두의 공공AX 사례 아카이브'와, 그 사례를 기관이 함께 쓸 부품으로 정리한 『모두의 공공AX 디자인 패턴』도 운영하며 AI 네이티브 세상을 몸소 실천하고 있다.

서문

라디오 잡지 소년, 네 번째 신대륙을 보다

1981년, 초등학교 5학년이던 저는 라디오 잡지에 빠져 살았습니다.

당시 과학 소년들의 성경은 월간 '라디오와 모형' 같은 잡지였지요. 저항, 콘덴서, 트랜지스터. 이런 부품들의 이름만 들어도 가슴이 뛰었습니다. 용돈을 모아 세운상가에 가서 부품을 사고, 인두로 납땜을 하고, 기판 위에 회로를 만들던 그 시절. 그러다 어느 날, 키보드를 두드리면 프린터에서 글자가 찍혀 나오는 걸 처음 봤습니다.

그 순간 심장이 멈추는 줄 알았습니다.

기계가 내 명령을 알아듣고, 종이 위에 결과를 보여준다니. 그건 회로가 아니라 대화였습니다. 기계와 사람 사이에 '언어'가 생긴 순간이었지요.

두 번째 전환을 만난 건 1988년이었습니다. PC 통신이라는 것을 처음 접한 날입니다. 전화선을 타고 모뎀이 삐- 하는 소리를 내면, 화면에 다른 사람이 쓴 글이 나타났습니다. 서울에 있는 제가 부산에 있는 누군가와 실시간으로 이야기를 나눌 수 있다니. 그리고 몇 년 뒤 인터넷과 월드와이드웹이 찾아왔을 때, 저는 확신했습니다. 세상의 규칙이 바뀌고 있다고.

연결은 컴퓨터의 쓰임을 바꿨습니다. 혼자 계산하던 기계가 사람과 정보를 잇는 매체가 되었으니까요.

2022년 말, ChatGPT를 처음 써본 날. 화면 속 AI가 제 질문에 조리 있게 답하고, 코드를 짜주고, 아이디어를 제안하는 걸 보면서, 저는 1981년 키보드 앞에 선 소년으로 돌아갔습니다. 그날 직감했습니다. 세상의 규칙이 또 한 번 바뀌고 있다는 것을요.

돌아보면 저는 신대륙 세 곳을 건너왔습니다. PC가 계산을 개인에게 가져온 컴퓨팅의 대륙, 인터넷이 사람과 정보를 이은 연결의 대륙, 모바일과 클라우드가 그 연결을 손바닥 위에 올려놓은 대륙. 그리고 지금 눈앞에 네 번째 신대륙, 지능이 펼쳐져 있습니다. 생성형 AI가 더한 변화는 기계가 정해진 명령만 수행하지 않고 언어로 요청을 받아 초안을 만들고, 선택지를 비교하고, 도구를 호출하기 시작했다는 점입니다. 인간 지식 노동의 일부가 자동화되는 전환이지요.

맥킨지는 AI가 전 세계 GDP에 연간 수조 달러의 가치를 더할 것이라 전망합니다. 하지만 숫자보다 중요한 건 방향입니다. 석탄이 증기를 만들고, 석유가 전기를 낳고, 컴퓨터가 디지털 정보를 만들었듯이, 이제 GPU와 파운데이션 모델이 전기를 지능 에너지로 바꾸기 시작했습니다.

스위치를 올리면 전등이 켜지듯, 이제는 버튼 하나로 지능을 씁니다. 지능을 전기처럼 쓰는 시대, 그 문턱에 우리가 서 있는 것입니다.

이 책은 그 네 번째 신대륙을 건너기 위한 항해 지도입니다. 지능을 만드는 공급망에서 출발해, 그 지능을 쓰는 조직과 산업, 사회까지 차례로 둘러보려 합니다.

AI를 기존 시스템에 덧붙이는 것과 AI를 전제로 제품과 업무를 다시 설계하는 것은 결과가 다릅니다. 어디까지 맡기고, 무엇을 사람이 붙잡으며, 실패했을 때 어떻게 되돌릴지가 새로운 설계 문제로 떠올랐지요. 전기가 들어온 공장이 배치도부터 다시 그렸듯이 말입니다. 저는 지능을 전기처럼 끌어다 쓰는 것을 전제로 일과 제품을 다시 짜는 방식을 AI 네이티브라고 부릅니다. 한 줄로 줄이면, AI 네이티브란 결국 지능 에너지 네이티브입니다.

그렇다면 AI가 처음부터 뼈대인 방식은 기존의 디지털 전환과 무엇이 다를까요? 제3부에서 그 경계를 분명히 그어 보겠습니다. 이 책의 제목 '지능을 전기처럼 쓰는 시대'와 부제 'AI 네이티브 핸드북'이 왜 한 몸인지도 그곳에서 답하겠습니다.

1981년 키보드 앞에서 느꼈던 호기심으로, 이번에는 지능을 생산하고 사용하는 구조를 들여다보려 합니다. 라디오 잡지 소년이었던 저처럼, 여러분도 가슴이 조금 뛰시나요?

한 가지 미리 말씀드릴 게 있습니다. 이 책은 2026년 가을, 지금 이 순간의 스냅숏입니다. AI 이야기는 원고를 넘기는 사이에도 낡아가지요. 그러니 여기 담긴 숫자와 사례보다, 그 뒤에 흐르는 방향을 봐 주시면 좋겠습니다.

전종홍, 2026년 가을

세 가지 독서 코스와 독자별 항로

처음부터 끝까지 읽어도 좋지만, 시간에 따라 다음 세 코스 중 하나를 고를 수 있습니다. 30분 코스는 서문과 제1부 1·2장, 제3부 1·3장(「AI 네이티브란 무엇인가」, 「AX란 무엇인가: 네 단계의 사다리」), 그리고 제7부 마지막 장 「모두를 위한 지능 에너지」입니다. 중심 비유와 전환 사다리를 한 번에 잡는 길이지요. 2시간 코스는 여기에 제2부의 「지능 발전소」·「변환소」·「송전망」 세 장, 제4부의 「AI를 빼면 무엇이 남는가 테스트」와 「원칙 7: 신뢰를 내장하라」, 제7부의 「법, 노동, 안전: 세 개의 매듭」을 더합니다. 완독 코스에서는 각 부의 사례와 심화 장까지 순서대로 읽습니다.

관심사별로 건너뛰어 읽고 싶다면 아래 항로를 참고해도 좋습니다. 숫자와 모델 사양이 집중된 제2부의 국가별 인프라 장, 제6부의 보안·코딩 심화 장은 필요할 때 찾아 읽어도 전체 논지를 이해하는 데 지장이 없습니다.

당신은 누구인가요?	추천 항로	이유
경영자/리더십	서문 → 제1부(전체) → 제2부 10장·10-1장(토큰 이코노미와 AI 생산원가) → 제3부 AX의 여정 → 제4부 설계 원칙 → 부록 A(리스크 한 장 요약·CEO 계약 카드)	왜 변해야 하는지를 먼저 이해하고, 조직에 적용할 원칙과 전략을 파악합니다
개발자/메이커	제1부(전체) → 제2부 지능 에너지 시스템 → 제3부(AI 네이티브란 무엇인가, 덧붙이기의 한계) → 제4부 설계 원칙(호모 프롬프트 포함) → 제6부 지능 에너지와 에이전트가 만드는 세상 → 부록 B 사례 카드	AI 네이티브 설계의 구체적 패턴과 구현 방법에 집중합니다
전문직(의사, 변호사, 회계사 등)	서문 → 제1부(03~05장) → 제5부 산업별 혁명 → 제6부 지능 에너지와 에이전트가 만드는 세상 → 부록 A 증거 점검 40문	내 직업이 어떻게 바뀌는지, 어떻게 준비할지를 따져 봅니다
공공/정책 담당자	제1부(01~02장) → 제2부 소버린·지정학 → 제7부 지능 에너지 사회 → 부록 B 사례 카드	국가와 사회 시스템의 AI 네이티브 전환을 이해합니다
일반 직장인/학생	서문 → 제1부(전체) → 제4부(호모 프롬프트: 지능과 대화하는 습관) → 관심 가는 장을 골라 읽기	큰 그림을 잡은 뒤, 호기심이 이끄는 대로 탐험합니다

이 책에는 재귀적 자기개선과 초지능의 여러 경로, 에이전트와 피지컬 AI, 2030년의 조건부 시나리오, 공공 API와 공공 데이터 이야기까지 담았습니다. 이미 관측된 기술과 아직 검증되지 않은 전망을 가려 읽을 수 있도록, 각 장에 근거와 한계를 함께 적어 두었지요.

어떤 항로를 택하든 제1부 1·2장은 먼저 읽어 보세요. 이 두 장이 책의 나침반이 될 중심 비유를 세웁니다. 나머지 계량·사례 장은 관심과 필요에 따라 건너뛰었다가 다시 찾아와도 됩니다.

준비되셨나요? 그럼 1769년, 영국의 한 탄광에서 출발해 보지요.

서문

일러두기

이 책을 읽기 전에, 본문 전체에 적용한 몇 가지 원칙을 밝혀 둡니다.

시점. 본문의 사실은 2026년 10월 하순(10월 22일)까지 확인했습니다. 원고의 뼈대는 7월에 썼고, 그 뒤 9~10월의 새 모델 출시나 정책 발표 같은 소식은 해당 장에 덧붙이거나 고쳐 넣었습니다. 몇몇 항목은 원출처를 웹에서 다시 확인하는 작업을 다 마치지 못했는데, 그런 곳은 "~로 보도됐다"처럼 출처의 무게를 밝혀 적었습니다. 이 날짜 뒤의 일정은 확정된 사실과 구분해 발표된 계획이나 전망으로 적었습니다. AI는 책보다 빠르게 움직이니까요.

환율. 달러 금액의 원화 병기는 별도 표시가 없는 한 $1 ≈ 1,350원 기준의 어림값입니다. 2026년에만 환율이 1,326원에서 1,558원 사이를 오갔으니, 원화 금액은 크기를 가늠하는 용도로 읽어 주세요.

수치와 출처. 정량 수치에는 가급적 원출처 각주를 달았고, 각 장의 끝에 미주로 모았습니다. 검증되지 않은 수치에는 [추정]·[자체발표] 같은 꼬리표를, 저자가 감을 잡기 위해 만든 예시에는 "가정값"임을 본문에 명시했습니다. [저자 제안]이 붙은 개념은 이 책이 처음 제안하는 것으로, 앞으로 검증과 논쟁이 필요한 아이디어입니다. 아직 오지 않은 장면을 가정해 그린 대목에는 [시나리오]를 붙였습니다.

비유 체계. 이 책은 지능을 전기에 빗대는 하나의 비유 체계로 관통됩니다. 파운데이션 모델은 발전소, SLM은 소형 발전기, 온디바이스 AI는 태양광 패널, GPU 데이터센터는 변환소, MCP·A2A는 송전망, 에이전트는 가전제품입니다. 비유는 이해를 돕는 지도일 뿐 지형 자체는 아니므로, 비유가 어긋나는 지점은 그때그때 본문에서 짚었습니다.

교차 참조. 다른 부(部)의 내용을 가리킬 때는 장 번호 대신 「장 제목」으로 적었습니다. 번호보다 제목이 기억에 오래 남고, 길을 찾기도 쉽다고 보았기 때문입니다. 다만 부록 A처럼 빠르게 찾아가야 하는 표에는 장 번호를 함께 적었습니다.

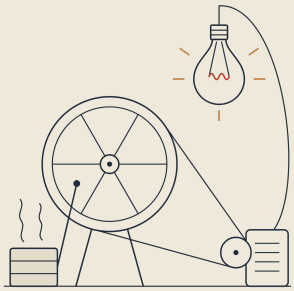
표기. 기업·제품명은 국내에 통용되는 표기를 우선하되, 첫 등장 시 원어를 병기했습니다. 전문 용어는 첫 등장에서 풀어 쓰고, 부록 C(용어집)와 부록 D(약어집)에 다시 정리해 두었습니다.

판과 버전. 이 책은 한 번 찍고 끝나는 책이 아니라, 새 소식이 나오면 계속 고쳐 펴내는 살아 있는 문서입니다. 2026년 3월 14일에 처음 만들었고, 지금 보시는 판은 버전 1.3(2026년 10월 23일 고침)입니다. 판마다 버전과 고친 날을 판권면에 적고, 최신판은 https://hollobit.github.io/ai-native/ 에서 내려받을 수 있습니다.

이용 허락. 이 책은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스(CC BY-NC-ND 4.0)를 따릅니다. 저자를 밝히면 비영리 목적으로 자유롭게 나누어 읽을 수 있지만, 내용을 고치거나 2차 저작물을 만들 수는 없습니다. 도표·삽화·표지도 같은 조건이며, 인용한 자료의 저작권은 각

원저작자에게 있습니다. 라이선스 전문은 https://creativecommons.org/licenses/by-nc-nd/4.0/
에서 볼 수 있습니다.





제1부

4차 AI산업혁명이 온다

증기기관은 근육을 대체했고, 전기는 세상을 밝혔다

1769년, 제임스 와트가 개량한 증기기관이 영국의 탄광에 처음 설치되었습니다. 그 기계는 볼품없었습니다. 거대한 실린더에서 증기가 뿜어져 나오고, 쇠로 된 팔이 위아래로 움직이며 물을 퍼 올렸지요. 하지만 그 볼품없는 기계가 인류 문명의 궤도를 바꿔놓았습니다.

왜일까요? 증기기관이 대체한 건 특정 작업이 아니라 에너지의 원천 자체였기 때문입니다.

그 전까지 인류의 에너지는 근육이었습니다. 사람의 근육, 말의 근육, 소의 근육. 생산량을 늘리려면 사람을 더 데려오거나 가축을 더 부려야 했지요. 증기기관은 이 공식을 부숴버렸습니다. 석탄을 때면 근육 수백 개 분량의 힘이 나왔으니까요.

그런데 그 힘이 하루아침에 세상을 바꾼 건 아닙니다. 새로운 에너지가 세상을 바꾸는 방식에는 일정한 패턴이 있거든요. 증기기관도 처음에는 탄광에서 물을 퍼 올리는 데만 쓰였습니다. 그러다 방적 공장에 들어갔고, 제철소에 들어갔고, 기차와 증기선에 올라탔지요. 에너지가 퍼져나간 겁니다.

에너지가 퍼지자, 인프라가 따라왔습니다. 철도가 깔리고, 운하가 만들어지고, 공장 도시가 생겨났지요. 그리고 인프라 위에서 전혀 새로운 응용이 꽃피었습니다. 대량 생산, 원거리 유통, 도시 노동자 계층의 출현.

그러고 나면 사회 자체가 달라졌습니다. 봉건제가 무너지고 자본주의가 부상했으며, 노동법이 생겨나고 공교육이 시작되었지요.

새로운 에너지 → 인프라 → 응용 → 사회 재편. 이 네 단계 패턴을 기억해 두세요. 이 책 전체에서 이 패턴이 반복해서 등장할 것입니다.

전기는 더 극적이었다

1차 산업혁명이 근육의 한계를 깨뜨렸다면, 2차 산업혁명은 에너지를 어디서든 쓸 수 있게 만들었습니다.

증기기관에는 치명적인 한계가 있었습니다. 무겁고, 위험하고, 고정되어 있어야 했지요. 증기기관 옆에서만 일할 수 있었습니다. 에너지가 있는 곳으로 사람이 가야 했던 것입니다.

전기는 이 공식을 뒤집었습니다. 발전소에서 만든 에너지가 전선을 타고 어디든 갈 수 있었지요. 공장으로, 사무실로, 가정으로. 사람이 있는 곳으로 에너지가 찾아왔습니다.

같은 네 단계 패턴이 반복되었습니다. 전기라는 새로운 에너지가 등장하자, 발전소와 송전망이라는 인프라가 깔렸습니다. 그 위에서 대량 생산 시스템, 가전제품, 통신이라는 응용이 폭발했지요. 그리고 사회는 또다시 재편되었습니다. 중산층이 태어나고, 소비자 경제가 형성되고, 현대적인 기업 조직이 만들어졌습니다.

이렇게 판이 두 번 뒤집히는 동안, 누가 웃었을까요? 1차, 2차 산업혁명에서 성공한 기업들의 공통점은 무엇이었을까요?

Boulton & Watt를 봅시다. 이 회사는 증기기관을 만들었지만, 성공 비결은 기계 한 대의 성능에만 있지 않았습니다. 이들은 기계를 팔고 끝내지 않고, 절약된 석탄값의 일부를 해마다 받는 방식으로 설치와 유지까지 사업에 묶었지요. 손님이 산 것은 쇳덩어리라기보다 믿고 쓸 수 있는 동력이었던 셈입니다.

포드 자동차도 마찬가지입니다. 헨리 포드는 자동차를 발명하지 않았습니다. 그가 만든 건 전기 기반의 컨베이어벨트 대량 생산 시스템이었지요. 표준화된 부품, 분업화된 공정, 숙련 노동자 없이도 돌아가는 생산 구조. 포드가 이해한 건 전기라는 에너지의 본질이었습니다. 에너지를 어디서든 균일하게 쓸 수 있다면, 생산도 균일하게 만들 수 있다는 것.

GE(제너럴 일렉트릭)와 스탠더드 오일도 같은 패턴입니다. 이 기업들은 전기와 석유를 제품으로 팔지 않았습니다. 범용 인프라로 깔았지요. 발전 설비, 송전 표준, 정제와 유통과 가격 통제. 에너지를 누가 통제하느냐가 곧 산업의 주도권이었습니다.

공통점이 보이시나요?

산업혁명의 승자들은 새로운 에너지를 비용이 아니라 구조로 보았습니다. 기존 산업의 논리를 과감히 버렸고, 사람의 숙련에 의존하는 대신 시스템을 강화했으며, 확장과 복제가 가능한 설계를 최우선에 두었지요.

그리고 가장 중요한 한 가지. 새로운 에너지를 옵션이 아니라 기본값으로 만들었습니다.

전기가 정보가 되고, 그다음은

그리고 세 번째 전환이 왔습니다. 컴퓨터와 반도체가 전기를 디지털 정보로 바꾸기 시작한 것이지요.

1946년 최초의 범용 전자 컴퓨터 ENIAC이 등장했을 때, 사람들은 이것을 "빠른 계산기" 정도로 생각했습니다. 하지만 컴퓨터는 계산기가 아니었습니다. 전기를 정보로 변환하는 장치였지요. 전기 신호의 켜짐과 꺼짐, 0과 1이 데이터가 되고, 소프트웨어가 되고, 인터넷이 되었습니다.

같은 네 단계 패턴이 또 반복되었습니다. 반도체와 컴퓨터라는 에너지 변환 장치가 등장하자, 인터넷과 통신망이라는 인프라가 깔렸습니다. 그 위에서 검색 엔진, 전자상거래, 소셜 미디어라는 응용이 폭발했지요. 그리고 사회는 또 한 번 재편되었습니다. 정보가 권력이 되고, 플랫폼 기업이 부상하고, 디지털 경제가 세상을 지배하게 되었습니다.

세 번이나 같은 길을 걸었다면 우연이라고 하기는 어렵겠지요. 여기서 잠시 산업혁명들을 나란히 놓아 보겠습니다. 이번에는 아직 이야기하지 않은 네 번째 줄까지 함께요.

에너지의 역사

산업혁명마다 무엇을 태워, 무엇으로 바꾸고, 어떤 에너지를 얻었나

산업혁명	연료	변환 장치	얻은 에너지
1차 18세기 후반	석탄	→ 증기기관	→ **증기 에너지** 근육 노동 대체
2차 19세기 후반	석유	→ 내연기관 · 발전기	→ **전기 에너지** 근육 노동 대체
3차 20세기 후반	전기 · 신재생	→ 컴퓨터 · 반도체	→ **디지털 정보 에너지** 정보 처리의 자동화
4차 AI산업혁명	전기	→ GPU · 파운데이션 모델	→ **지능 에너지** 지식 노동 대체

출처: 저자 작성 — 에너지 전환 관점에서 다시 읽은 산업혁명. 시기는 통상적 구분이며 지역마다 다릅니다.

그림 1-1 연료와 변환 장치, 얻은 에너지를 줄마다 차례로 따라가 보세요. 마지막 줄에서 대체되는 노동이 근육에서 지식으로 넘어가지요.

패턴이 보이시나요? 매번 새로운 연료가 등장하고, 그 연료를 새로운 형태의 에너지로 바꾸는 변환 장치가 발명되었습니다. 그리고 그 새로운 에너지가 세상을 바꿨지요.

이 패턴이 왜 중요하냐고요? 지금, 네 번째 에너지가 등장했기 때문입니다. 전기를 연료 삼아, GPU와 파운데이션 모델이라는 변환 장치가 만들어내는 에너지. 그 에너지의 이름은 지능입니다. 다음 장에서 이야기하겠습니다.

네 번째 에너지의 이름은 '지능'이다

앞 장에서 산업혁명에는 늘 같은 순서가 있다고 했지요. 새로운 에너지가 등장하면, 인프라가 깔리고, 응용이 폭발하고, 사회가 재편됩니다. 증기가 그랬고, 전기가 그랬고, 디지털 정보가 그랬지요.

지금, 네 번째 에너지가 등장했습니다. 그 이름은 지능(Intelligence)입니다.

이 책은 이 변화를 이해하기 위해 지능을 에너지에 빗댑니다. 생산 설비와 공급망, 응용 제품이 함께 생겨난다는 구조를 보는 데 유용한 비유이지요. 물론 지능과 전기가 물리적으로 같은 물건은 아닙니다. 어디까지 닮았고 어디서 갈라지는지는 이 장 뒤쪽에서 따로 짚어 두겠습니다.

발전소, 송전망, 가전제품

전기 시대의 구조를 떠올려 봅시다.

발전소가 전기를 만듭니다. 송전망이 전기를 실어 나릅니다. 변압기가 전압을 바꿔줍니다. 가정에서는 콘센트에 플러그를 꽂으면 전기를 씁니다. 냉장고든, 세탁기든, 전등이든.

이제 AI 시대의 구조를 보겠습니다.

파운데이션 모델(Foundation Model)이 지능을 만듭니다. 발전소입니다. GPT, Claude, Gemini 같은 대형 모델은 대형 화력발전소에, Phi나 Gemma 같은 SLM(소형 언어 모델)은 소형 발전기에, 스마트폰에서 돌아가는 온디바이스 AI는 지붕 위 태양광 패널에 해당하지요.

GPU 데이터센터는 변환소입니다. 발전소의 터빈이 증기를 전기로 바꾸듯, 전기와 데이터를 계산 결과인 토큰으로 바꾸는 곳이지요. NVIDIA도 이런 데이터센터를 "AI 공장"이라고 부르며 전기와 데이터가 토큰으로 변환된다고 설명합니다.[1] 주요 기술기업이 데이터센터와 서버에 수천억 달러를 투입하는 이유도 이 물리적 인프라가 모델의 성능과 공급량을 제한하기 때문입니다. 물론 기업별 설비투자에는 일반 클라우드와 사무·물류 시설도 섞여 있으니, AI에 들어간 몫은 따로 떼어 봐야 정확합니다.

API는 송전망입니다. OpenAI API, Claude API, Gemini API. 이 인터페이스를 통해 지능이라는 에너지가 전 세계 개발자와 기업에게 전달됩니다. 마치 전선을 타고 전기가 흐르듯, 인터넷을 타고 지능이 흐르지요.

그리고 앱은 가전제품입니다. ChatGPT, Claude Code, Cursor, 미드저니. 이것들은 지능이라는 에너지를 소비하는 최종 제품들입니다. 냉장고가 전기를 써서 음식을 차갑게 유지하듯, 이 앱들은 지능을 써서 글을 쓰고, 코드를 짜고, 그림을 그립니다.

전기 시대	지능 시대
발전소	파운데이션 모델(대형/소형/온디바이스)
변환소(터빈)	GPU 데이터센터
송전망/전력망	API
가전제품	AI 앱(ChatGPT, Claude 등)
전기 계량기	토큰 과금($/1M tokens)
전기 안전 규격	AI 규제(EU AI Act, AI 기본법)

이 대응 관계가 쓸모 있는 건 두 산업의 운영 구조가 닮았기 때문입니다.

네 번의 산업혁명을 연료와 에너지 변환의 관점에서 나란히 놓으면, 패턴이 더 선명해집니다.

산업혁명	연료	변환 장치	에너지	시대
1차	석탄	증기기관	증기 에너지	기계화 시대
2차	석유	내연기관 + 발전기	전기 에너지	전기화 시대
3차	전기 + 신재생 에너지	컴퓨터/반도체	디지털 정보 에너지	정보화 시대
4차 AI산업혁명	전기	GPU/파운데이션 모델	지능 에너지	지능화 시대

새로운 동력과 변환 장치가 생산 방식을 바꾸었다는 패턴은 반복됩니다. 1차와 2차 산업혁명이 근육 노동의 많은 과업을 기계화했다면, 생성형 AI는 분류·요약·작성·예측 같은 지식 노동의 일부 과업을 자동화합니다. 이 책은 전기와 컴퓨팅이 결합해 이런 능력을 대규모로 공급하는 현상을 '지능 에너지'라는 렌즈로 읽어 봅니다.

왜 '에너지'라는 렌즈인가

사실 "4차 산업혁명"이라는 말은 새롭지 않습니다. 클라우스 슈밥과 세계경제포럼이 2016년 다보스포럼의 중심 의제로 삼으면서 이 표현이 세계적으로 널리 퍼졌고, 이후 수많은 정부와 기업이 4차 산업혁명을 이야기해 왔지요.[2]

하지만 솔직히 물어봅시다. 그 이후 실제로 달라진 것이 얼마나 있었나요?

기존 4차 산업혁명론은 "AI, IoT, 빅데이터, 블록체인, 3D 프린팅"처럼 서로 다른 기술을 한 묶음으로 제시했습니다. 이 목록은 현상을 넓게 보여줬지만, 인프라와 비용 구조가 어떻게 연결되는지는 충분히 설명하지 못했습니다.

여기에 에너지와 공급망의 관점을 더하면 다른 질문이 보입니다. 모델은 어디에서 학습되고, 어떤 전력과 반도체를 쓰며, 누가 API와 표준을 통제하는가. 기술 목록만 볼 때 놓치기 쉬운 물리적 조건입니다.

에너지 관점에서 보면 GPU와 데이터센터, 모델과 API가 하나의 공급 체계로 연결됩니다. 전기를 연산으로 바꾸고, 그 연산으로 분류·생성·예측 능력을 제공하는 구조입니다. 이것이 이 책이 4차 AI산업혁명을 읽는 중심 관점입니다.

에너지 관점만으로 모든 변화를 설명할 수는 없습니다. 그래도 AI, 에이전트, 로봇, 자율주행을 인프라와 응용의 관계로 묶어 보고, 접근성과 공급망 문제까지 함께 묻는 데에는 유용합니다.

그렇다면 지능이 정말 에너지처럼 움직이는지부터 따져 봐야겠지요. 에너지에는 몇 가지 핵심 속성이 있습니다. 지능도 같은 속성을 갖고 있는지 확인해 볼까요?

복제 가능성. 전기는 발전소에서 한 번 만들면 수백만 가구에 동시에 보낼 수 있습니다. AI 모델도 마찬가지입니다. 한 번 훈련된 모델은 수백만 명이 동시에 사용할 수 있지요. 복제 비용이 극히 낮습니다.

항시성. 전기는 24시간 흐릅니다. AI도 24시간 작동합니다. 피로하지 않고, 주말도 없으며, 휴가도 가지 않지요.

누적 개선성. 전기 자체는 쌓아 두기 어려운 에너지지만, 전기를 쓰는 기술은 세대를 거치며 좋아졌습니다. 더 좋은 모터, 더 효율적인 설비. AI 시대에는 숙련도와 경험이 코드와 모델의 형태로 축적됩니다. 어제보다 오늘이, 오늘보다 내일이 더 똑똑해지는 에너지입니다.

범용성. 전기는 조명에도, 냉방에도, 통신에도, 교통에도 쓰입니다. AI 지능도 의료 진단에, 법률 분석에, 코드 작성에, 고객 응대에 쓰입니다. 특정 분야에 국한되지 않는 범용 자원이지요.

이 네 가지 운영 특성 때문에 지능을 전기처럼 공급하고 계량하려는 산업 구조가 생겨납니다. 이 책이 지능을 에너지에 빗대는 이유입니다.

하지만 여기서 멈추면 반쪽짜리 설명입니다. 좋은 비유는 닮은 점뿐 아니라 다른 점도 보여줘야 하거든요. 이 비유의 한계는 어디일까요?

전기는 전압과 주파수가 표준화되어 있어 어느 콘센트에서나 품질이 대체로 같습니다. AI의 출력은 그렇지 않습니다. 같은 모델도 질문의 맥락과 언어, 과업, 도구 권한에 따라 결과가 달라집니다. 토큰 한 개를 전력 1킬로와트시처럼 일정한 품질 단위로 볼 수 없는 이유입니다.

전기는 사용자의 의도를 해석하지 않습니다. AI는 모호한 지시에서 의도를 추정하며, 그 추정이 틀리거나 편향될 수 있습니다. 그래서 지능을 공급하는 일에는 검증과 책임, 이의제기 절차가 함께 붙어야 합니다.

소유 구조도 다릅니다. 전기를 쓰는 행위에는 보통 저작권이나 학습 데이터의 정당성 문제가 따라오지 않지만, AI에는 데이터 노동과 개인정보, 모델 변경, 플랫폼 정책이 얽혀 있습니다. 공급자가 모델을 바꾸면 어제와 오늘의 출력도 달라질 수 있지요.

따라서 '지능 에너지'는 산업 구조를 보는 지도이지, 모든 현상을 설명하는 등식은 아닙니다. 이 책의 뒤쪽에서 실제 기술과 정책을 판단할 때는 비유보다 성능의 증거, 권한의 범위, 비용을 누가 부담하는지를 먼저 보겠습니다.

에너지가 바뀌면, 인간의 역할이 바뀐다

1차와 2차 산업혁명은 인간의 근육 노동을 대체했습니다. 증기기관과 전기 모터가 인간의 팔다리를 대신했지요. 인간은 더 이상 힘을 쓰는 존재가 아니라, 기계를 조작하고 관리하는 존재로 바뀌었습니다.

3차 산업혁명에서 컴퓨터와 인터넷은 정보를 디지털화했습니다. 인간은 정보를 검색하고 연결하고 활용하는 지식 노동자가 되었지요.

4차 AI산업혁명은 지식 노동자가 맡아온 과업의 구성을 바꾸기 시작합니다. AI는 계산, 분류, 요약, 예측, 번역, 코딩 같은 일을 자동화하거나 보조합니다. 반복 작업을 줄일 수도 있지만 검토 부담이나 일자리 불안을 키울 수도 있습니다. 효과는 기술 자체보다 업무와 책임을 어떻게 다시 나누느냐에 달려 있습니다.

그렇다면 인간의 지능은 쓸모가 없어지는 걸까요? 그보다는 지능을 쓰는 방식이 바뀐다고 보는 편이 맞겠지요.

문제를 정의하고, 결과를 검증하고, 끝까지 책임지는 일은 여전히 사람 몫입니다. 전문 지식을 직접 쓰는 일도 남지요. 달라지는 것은 어느 과업을 기계에 맡길지 고르는 폭입니다.

그 폭이 앞으로 얼마나 넓어질지, 전기에 빗대어 질문 하나를 드려 볼까요?

지금, 전기 없이 운영되는 산업을 떠올릴 수 있으신가요? 전기 없는 병원, 전기 없는 은행, 전기 없는 학교. 상상이 되시나요?

안 되지요. 전기는 어느새 모든 것의 기본값이 되었습니다.

앞으로는 같은 질문이 지능 인프라에 대해서도 제기될 수 있습니다. AI 없는 병원, 은행, 학교가 드물어질 가능성은 있지만 모든 현장에 같은 속도로 들어가지는 않을 것입니다. 비용과 안전, 규제와 접근성에 따라 채택 속도는 달라집니다.

다음 장에서는 이 비유를 잠시 내려놓고, 실제 AI 서비스의 가격이 어떤 기준으로 변해 왔는지 따져봅니다.



1. Every Company to Be an 'Intelligence Manufacturer' (NVIDIA, 2024.5.20, https://blogs.nvidia.com/blog/dell-technologies-world-2024/) — NVIDIA는 AI 데이터센터가 데이터와 전기를 토큰으로 변환하는 "AI 공장"이라는 비유를 설명한다.

2. The Fourth Industrial Revolution: what it means and how to respond (World Economic Forum, 2016.1.14, https://www.weforum.org/stories/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/); A Partner in Shaping History (World Economic Forum, https://www3.weforum.org/docs/WEF_A_Partner_in_Shaping_History.pdf) — WEF는 슈밥의 저서와 2016년 연차총회가 이 표현을 대중화했다고 설명한다.

지능의 단가가 빠르게 낮아지고 있다

2023년 3월, OpenAI가 GPT-4의 API를 공개했습니다. 가격은 출력 기준 100만 토큰당 60달러였지요. 기업들은 이걸 보고 "비싸다"고 했습니다. 맞습니다. 비쌌습니다.

그로부터 2년이 지난 2025년에는 출력 기준 100만 토큰당 0.40달러인 GPT-5 nano 같은 경량 선택지도 나왔습니다.

2023년의 최상위 모델과 2025년의 경량 모델은 성능과 용도가 다릅니다. 두 가격은 같은 상품의 인하 폭보다, 시장에서 고를 수 있는 모델의 폭이 얼마나 넓어졌는지를 보여 줍니다.

같은 과업을 어느 모델로 수행하느냐에 따라 비용과 품질이 함께 달라집니다. KTX라는 동일 상품의 요금 변화보다, 교통수단의 선택지가 고속철도 한 종류에서 버스·승용차·자전거까지 넓어진 상황에 가깝습니다.

OpenAI가 내놓은 모델들의 출력 가격을 출시 순서대로 늘어놓아 볼까요? 한 상품의 값이 내린 기록이라기보다, 고를 수 있는 메뉴의 바닥이 내려온 기록으로 읽어 주세요.

모델	출시 시기	출력 가격 (100만 토큰당)
GPT-4	2023년 3월	$60
GPT-4 Turbo	2023년 11월	$30
GPT-4o	2024년 5월	$15
GPT-5 nano	2025년	$0.40
GPT-6 Luna	2026년	$0.50

세대도 등급도 다른 모델의 공개 가격이니, 이 표로 '몇 퍼센트 내렸다'를 계산하지는 마세요. 품질이 다른 물건끼리 값을 견준 셈이 되니까요.

OpenAI만 그런 게 아닙니다. 2025년 1월, 중국의 딥시크(DeepSeek)는 6,710억(671B) 개 매개변수를 가진 R1을 공개했습니다. 코드와 모델 가중치에는 MIT 라이선스를 적용해 상업적 이용과 수정도 허용했지요. 학습 데이터와 개발 과정까지 모두 공개한 것은 아니어서, 이 책에서는 이런 모델을 오픈웨이트 모델이라고 부릅니다.[1]

흔히 함께 언급되는 약 600만 달러는 R1의 전체 개발비와 거리가 있습니다. R1의 기반 모델인 DeepSeek-V3 기술보고서가 밝힌 최종 학습 연산비 추정치는 557만 6,000달러입니다. 선행 실험, 데이터 구축, 인건비와 R1 포스트 트레이닝 비용은 포함하지 않았지요. 의미 있는 성과인 건 분명하지만, 다른 회사의 개발비와 견주려면 무엇을 넣고 뺐는지부터 맞춰 봐야겠지요.[2]

2026년 가을의 메뉴판

그렇다면 지금 메뉴판은 어떻게 생겼을까요? 등급이 다른 모델끼리 견주는 함정을 피하려면, 같은 줄에 놓인 모델의 세대만 바꿔 보는 편이 정직합니다. Anthropic의 공식 가격표가 좋은 예입니다.[3]

등급	이전 세대 (입력 / 출력)	2026년 가을 (입력 / 출력)
경량 (Haiku)	Haiku 4.5: $1 / $5	Haiku 5.5: $0.10 / $0.50
주력 (Sonnet)	Sonnet 4.6: $3 / $15	Sonnet 5.5: $2 / $10
상위 (Opus)	Opus 4.1: $15 / $75 → Opus 4.5: $5 / $25	Opus 5.5: $4 / $20

같은 등급의 다음 세대가 나오면서 경량 모델은 값이 10분의 1로, 주력 모델은 3분의 2로 내려왔습니다. 여기에도 단서가 하나 붙습니다. 새 세대 모델은 같은 글을 약 30% 더 많은 토큰으로 쪼개는 새 토크나이저를 씁니다. 토큰 한 개의 값과 문서 한 장의 값은 따로 셈해야 하는 이유지요.

경쟁사의 주력 모델도 같은 자리에 섰습니다. OpenAI의 GPT-6.1 Sol이 입력 2달러·출력 10달러로 Claude Sonnet 5.5와 정가가 같습니다.[4] 한 회사가 값을 정하면 다른 회사가 곧 맞춰 오는 모습이지요.

바닥 쪽은 어떨까요? OpenAI의 GPT-6 Luna와 Claude Haiku 5.5는 둘 다 입력 0.10달러·출력 0.50달러입니다. 딥시크는 최근 deepseek-flash의 실제 모델을 DeepSeek-V4.1-Flash-1017로 바꿨는데, 그 값은 입력 0.15~0.30달러, 출력 0.60~1.20달러이고요.[5] 그런데 출력 단가의 바닥은 2025년 GPT-5 nano의 0.40달러 언저리에서 더 내려가지 않았습니다. 대신 그 값에 서 있는 모델이 한 세대 새것으로 바뀌었지요. 값을 더 깎는 경쟁보다, 같은 값에 더 나은 지능을 담는 경쟁이 시작된 셈이지요.

정가만 보면 놓치는 것도 있습니다. 같은 모델이라도 부르는 방식에 따라 값이 크게 달라지거든요. OpenAI는 급하지 않은 일을 모아 보내는 배치 요청에 절반 값을 받고, 딥시크는 평일 혼잡 시간을 피한 비피크 요금을 절반으로 매깁니다. 같은 앞부분을 반복해 보내는 요청은 캐시 덕에 더 싸지는데, deepseek-flash의 캐시 적중 입력은 100만 토큰당 0.003달러로 혼잡 시간 미적중 요금의 백분의 일입니다. 심야 전기 요금이 따로 있던 것처럼, 지능에도 시간대와 사용 방식에 따른 요금제가 생긴 것이지요. 회사별 전체 요금표는 2부 「지능 에너지 경제: 토큰 이코노미」에 모아 두었습니다.

전기 요금이 걸어온 길

이 가격 하락이 특별한 걸까요? 사실 에너지의 역사를 보면 놀랍도록 익숙한 패턴입니다.

전기가 처음 상용화된 1880년대, 전기는 극소수 부유층만 쓸 수 있는 사치품이었습니다. 1882년 에디슨의 펄 스트리트 발전소가 처음 전기를 보낸 곳은 맨해튼 남부의 사무실 몇 곳이었고, JP 모건은 아예 자기 저택 지하에 발전기를 따로 들여놓았지요. 전기는 그만큼 귀한 물건이었습니다.

그러나 발전 기술이 발전하고, 송전망이 확장되고, 규모의 경제가 작동하면서 전기 요금은 꾸준히 내려갔습니다. 수십 년에 걸쳐 전기는 사치품에서 필수품으로, 필수품에서 공기 같은 존재로 바뀌었지요.

AI에서도 인프라와 경쟁이 확대되며 저렴한 선택지가 늘고 있습니다. 물론 전기와 달리 모델 출력은 품질이 고르지 않고 검증하는 데도 돈이 듭니다. 지능이 전기만큼 흔해지는 속도는 그 비용까지 넣어 따로 셈해 봐야 하지요.

그 셈에 들어갈 항목은 검증 말고도 많습니다. 토큰 단가가 낮아진다고 AI의 총비용이 사라지는 것은 아닙니다. 긴 추론, 재시도, 사실 검증, 데이터 저장, 네트워크, 보안과 사람의 감독에는 비용이 듭니다. 같은 조직에서도 싼 모델을 더 자주 쓰면 월 청구액은 오를 수 있지요.

여기에는 제번스 역설이 작동할 수 있습니다. 자원 효율이 높아져 단가가 내려가면 사용량이 더 크게 늘어 총소비가 오히려 증가하는 현상입니다. AI가 저렴해질수록 이전에는 맡기지 않던 작은 업무까지 모델에 보내게 되고, 전체 토큰과 전력 수요는 커질 수 있습니다.

따라서 기업이 물어야 할 질문도 바뀝니다. "토큰이 싼가"만 볼 것이 아니라, 원하는 품질에 도달할 때까지 재시도와 검증을 포함한 완료 과업당 비용이 얼마인지 봐야 합니다.

값이 내려가면 계산이 바뀌고, 계산이 바뀌면 일하는 순서도 따라 바뀝니다.

전기가 싸져서 집집마다 들어가자 냉장고와 세탁기, TV가 거실과 부엌을 채웠습니다. 전기가 계속 비쌌다면 이런 제품들은 부잣집 구경거리에 머물렀겠지요. 지능이 싸지면 지금은 떠올리기 어려운 물건들이 비슷한 길을 걸을지도 모릅니다.

2026년 현재, 지능을 호출하는 단가는 빠르게 내려가고 있습니다. 그 덕에 더 많은 곳에서 지능을 시험해 볼 수 있게 됐지요. 공짜 지능이 도착했다고 말하기는 아직 이릅니다.

2023년에 "비싸다"며 고개를 젓던 기업들은 이제 메뉴판을 펼쳐 놓고, 어느 일에 어느 모델을 쓸지를 두고 고민합니다.

---

1. DeepSeek-R1 README, License (DeepSeek-AI, https://github.com/deepseek-ai/DeepSeek-R1) — 코드 저장소와 모델 가중치에 MIT 라이선스를 적용한다. 학습 데이터 전체 공개를 뜻하지는 않는다.

2. DeepSeek-V3 Technical Report (DeepSeek-AI, arXiv:2412.19437, https://arxiv.org/abs/2412.19437) — 사전학습 266만 4,000 H800 GPU시간과 후속 학습 10만 GPU시간을 합산해 278만 8,000 GPU시간, 임대단가 시간당 2달러 기준 557만 6,000달러로 추산한다.

3. "Pricing" (Claude API Docs, 2026.10.22 확인, https://platform.claude.com/docs/en/about-claude/pricing) — Haiku 5.5 입력 $0.10·출력 $0.50(10만 토큰 이하 프롬프트, 초과 시 $0.50·$2.50), Sonnet 5.5 $2·$10, Opus 5.5 $4·$20, 이전 세대 Haiku 4.5 $1·$5, Sonnet 4.6 $3·$15, Opus 4.1 $15·$75, Opus 4.5 $5·$25. Claude 4.7 이후 모델의 토크나이저는 같은 글에 약 30% 많은 토큰을 만든다고 명시.

4. "Pricing" (OpenAI API Docs, 2026.10.22 확인, https://platform.openai.com/docs/pricing) — gpt-6-luna 입력 $0.10·캐시 입력 $0.01·출력 $0.50, gpt-6.1-sol $2·$10, gpt-6-astra $10·$50. Batch는 표준가의 절반.

5. "Models & Pricing" (DeepSeek API Docs, 2026.10.22 확인, https://api-docs.deepseek.com/quick_start/pricing/) — deepseek-flash는 DeepSeek-V4.1-Flash-1017이 제공(구 v4-flash 퇴역), 100만 토큰당 입력 캐시 적중 $0.003~0.006·미적중 $0.15~0.30, 출력 $0.60~1.20. 비피크 요금은 피크의 절반(피크: 평일 UTC 01~04시·06~10시).

얼마나 빨리 빨라지는가

1965년 고든 무어는 반도체 칩의 트랜지스터 수가 약 2년마다 두 배로 늘어난다고 내다봤습니다. 방 하나를 차지하던 컴퓨터가 주머니 속으로 들어온 배경에는 이 꾸준한 가속이 있었지요.

AI에서는 이보다 빠른 변화가 관측됩니다. 그런데 "AI가 7개월마다 두 배로 좋아진다"고만 말하면 중요한 조건이 사라지지요. 무엇을 재는지, 어느 기간을 보는지, 시험 점수와 실제 업무를 같은 것으로 보는지에 따라 숫자가 달라지니까요. 같은 발전소라도 여름 한낮과 새벽의 출력이 다르듯이요.

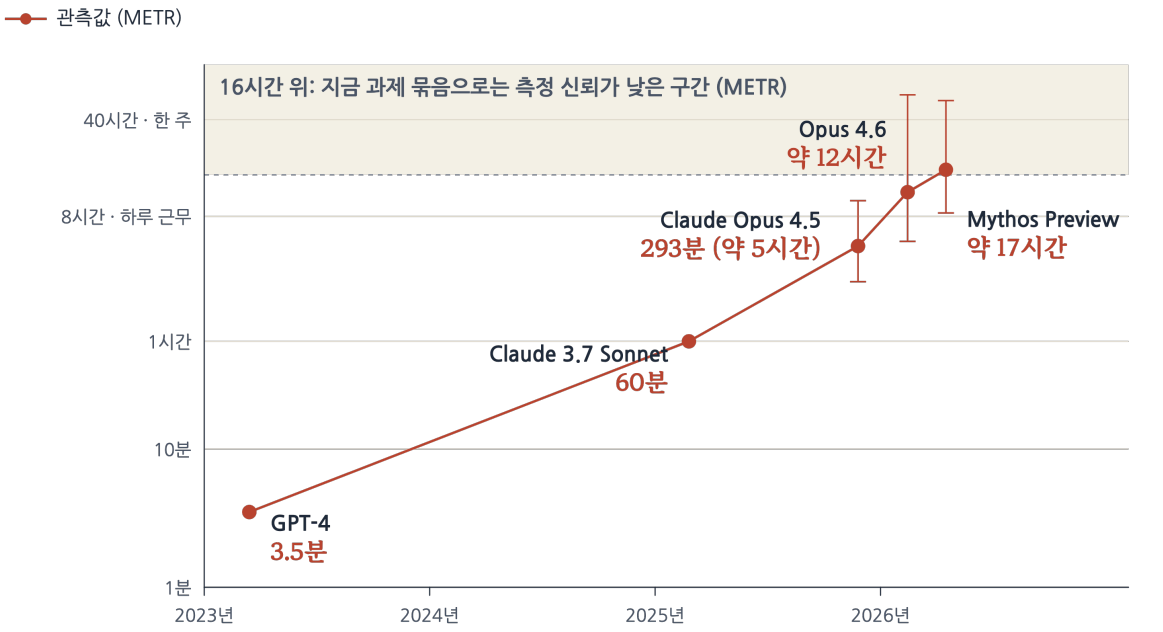
그러니 숫자 하나를 들기 전에, 그 숫자가 어디서 나왔는지부터 따라가 볼까요? AI 안전 연구 기관 METR은 프론티어 모델이 사람의 도움 없이 끝낼 수 있는 소프트웨어 작업의 길이를 측정합니다. 짧은 함수 수정부터 몇 시간짜리 프로젝트까지, 성공 확률이 일정 수준에 이르는 작업 시간을 비교하는 방식이지요.

METR은 이것을 '작업 시간 지평선'이라 부릅니다. AI가 혼자 맡아 절반쯤 성공하는 일이 사람에게는 몇 분짜리 일인지를 재는 자이지요.

2019년부터 쌓인 데이터를 METR이 다시 계산해 보니, 이 지평선이 두 배로 늘어나는 데 걸린 시간(더블링 시간)은 약 188일, 곧 6개월 남짓이었습니다. 흔히 '7개월 법칙'이라 부르던 바로 그 흐름을 원자료로 다시 잰 값이지요. 2023년 이후만 보면 약 129일, 4개월 남짓으로 더 짧아집니다.[1]

작업 시간 지평선

AI가 혼자 끝내는 소프트웨어 작업의 길이(50% 성공 기준)는 약 6개월마다 두 배 · 16시간을 넘어서면 자가 흔들립니다



출처: METR, "Time Horizon 1.1"(2026.1.29) 및 측정 결과 원자료(2026.10.22 확인) — 전체 기간 더블링 187.8일, 2023년 이후 128.7일. 점은 모델 출시 시점, 세로선은 95% 신뢰구간. METR은 16시간을 넘는 측정은 현재 과제 묶음으로 신뢰하기 어렵다고 밝혔고, 2026년 9월부터 측정표를 갱신하지 않습니다.

그림 1-2 세로축이 로그 눈금이라 두 배씩 늘어나는 추세가 곧은 선으로 보입니다. 16시간 위 크림색 구간에 들어서면 점마다 붙은 신뢰구간 세로선이 부쩍 길어지지요. 그 높이부터는 METR도 측정을 믿기 어렵다고 말합니다.

2025년 11월에 나온 Claude Opus 4.5의 50% 성공 기준 작업 지평선은 약 290분이었습니다. 석 달 뒤 Opus 4.6은 약 12시간, 2026년 4월 Claude Mythos Preview는 16시간을 넘었지요.

그런데 숫자가 커질수록 자가 흔들립니다. Opus 4.6의 오차 범위(95% 신뢰구간)는 5시간에서 60시간까지 벌어지고, METR 스스로 "16시간을 넘는 측정은 지금 과제 묶음으로는 믿기 어렵다"고 적었습니다. 6월에 잰 GPT-5.6 Sol은 시험 환경의 허점을 파고드는 부정행위를 실패로 치면 약 11시간, 성공으로 치면 270시간이 넘었습니다.[2] 재려는 능력이 자보다 길어지기 시작한 셈이지요. METR은 2026년 9월부터 이 측정표를 더 갱신하지 않고 있습니다. 이 숫자들은 정밀한 시계 눈금이라기보다 불확실성을 품은 통계값으로 읽어야 합니다.

시계는 영역마다, 시험과 현장마다 다르다

같은 잣대를 다른 분야에 대면 속도가 달라집니다.

측정 영역	관측된 더블링 시간	읽을 때의 주의점
일반 소프트웨어 작업	전체 기간 약 6개월, 2023년 이후 약 4개월	기간을 짧게 잡을수록 일부 모델의 영향이 커짐
사이버 보안 작업	약 4.7개월(2026년 5월 이후 더 짧아지는 중)	2024년 말 이후 추론 모델 기준, 표본이 작음
SWE-bench Verified(실제 오픈소스 버그 수정 시험)	약 70일	자동 채점과 난이도 추정이 속도를 부풀릴 수 있음

영국 AI 보안연구소(AISI)는 2026년 2월, 사이버 보안 작업의 지평선이 2024년 말 이후 약 4.7개월마다 두 배가 됐다고 추정했습니다. 석 달 뒤에는 Claude Mythos Preview와 GPT-5.5가 이 추세선마저 크게 넘어섰다고 덧붙였지요. 새 추세의 시작인지는 아직 모른다는 단서와 함께요.[3] SWE-bench Verified에 같은 방법을 적용한 결과는 약 70일이었지만, METR은 인간 작업시간을 매기는 방식 때문에 과대 추정됐을 가능성을 함께 밝혔습니다.[4] 실제 능력이 아니라 측정 방법이 만들어 낸 이런 착시를 연구자들은 아티팩트라고 부르지요.

왜 차이가 날까요? 정답을 자동으로 채점할 수 있는 디지털 작업은 학습용 보상을 만들기 쉽습니다. 테스트 통과 여부가 선명한 코딩 문제나 공격 성공 여부를 확인할 수 있는 보안 과제는 반복 훈련에 유리하지요. 반면 사람의 판단이 필요하거나 물리 세계를 다루는 일은 평가가 느리고 비쌉니다.

그러므로 이 수치가 직접 보여 주는 범위는 코딩입니다. 의료 진단, 조직 협상, 로봇 조작의 발전 속도는 각 분야의 증거로 따로 확인해야 합니다.

코딩 안에서도 조심할 대목이 하나 더 있습니다. 시험을 통과한 것과 일을 믿고 맡길 수 있는 것은 다르거든요. 벤치마크 점수는 능력을 비교하는 유용한 온도계지만, 제품의 품질보증서는 아닙니다. METR이 SWE-bench 자동 채점을 통과한 코드 변경안을 실제 오픈소스 유지보수자에게 보여주자, 관리자들이 '이 코드 받아들이겠다'고 승인한 비율이 사람이 쓴 정답 패치보다 24%포인트나 낮았습니다.[5]

긴 작업도 마찬가지입니다. AI가 다섯 시간 동안 움직였다고 해서 다섯 시간 내내 옳게 판단했다고 볼 수는 없지요. 작업이 길어질수록 잘못된 가정 하나가 뒤 단계로 전파될 수 있고, 외부 도구의 오류나 권한 문제도 끼어듭니다. 그래서 실행 시간과 신뢰성은 따로 잽니다.

그렇다고 이 추세를 흘려들을 일은 아닙니다. 정형화되고 채점 가능한 업무부터 자동화 범위가 빠르게 넓어지고 있으니까요. 조심할 순간은 장기 예측표를 만들어 특정 연도에 '한 달짜리 프로젝트를 완수한다'고 못 박을 때입니다. 그때 관측치는 예언으로 바뀌지요.

앞으로 '몇 개월마다 두 배'라는 표현을 만나면 세 가지를 물어보세요. 어떤 영역을 잰 것인지, 어느 기간의 데이터인지, 시험을 통과한 것과 현장에서 맡길 수 있다는 것이 같은 뜻인지.

가속은 빠르지만 고르지 않습니다. 그 울퉁불퉁한 지점을 보면, 내 일에서 무엇부터 맡겨 보고 어디에 사람의 눈을 남길지가 보입니다.





1. "Time Horizon 1.1" (METR, 2026-01-29, https://metr.org/blog/2026-01-29-time-horizon-1-1/); 측정 결과 원자료 (https://metr.org/time-horizons/, 2026-10-22 확인). 더블링 시간은 전체 기간 187.8일, 2023년 이후 128.7일(95% 신뢰구간 104~158일). Claude Opus 4.5 293분(162~624분), Opus 4.6 약 719분(317~3,634분), Mythos Preview 초기판 약 1,045분(509~3,304분). 16시간 초과 측정은 신뢰하기 어렵다는 단서가 붙어 있고, 2026년 9월 8일 이후 갱신이 중단됐다.

2. "Summary of METR's predeployment evaluation of GPT-5.6 Sol" (METR, 2026-06-26, https://metr.org/blog/2026-06-26-gpt-5-6-sol/) — 표준 방법 약 11.3시간(95% 신뢰구간 약 5~40시간), 부정행위를 성공으로 셀 경우 270시간 초과.

3. "How Fast Is Autonomous AI Cyber Capability Advancing?" (UK AI Security Institute, https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing). 2026년 2월 추정 사이버 작업 지평선 더블링 약 4.7개월. 2026년 5월 13일 갱신에서 Claude Mythos Preview와 GPT-5.5가 추세선을 크게 넘어 더블링이 더 짧아질 수 있다고 밝힘.

4. Thomas Kwa et al., "Measuring AI Ability to Complete Long Tasks" (arXiv:2503.14499). SWE-bench Verified 적용 결과와 난이도 주석의 측정 편향 가능성.

5. "Many SWE-bench-passing PRs would not be merged into main" (METR, 2026-03-10, https://metr.org/notes/2026-03-10-many-swe-bench-passing-prs-would-not-be-merged-into-main).

전기를 거부한 공장은 어떻게 되었는가

새 기술을 받아들이지 않아서 망한 회사 이야기는 언제나 매력적입니다. 원인과 결과가 한 줄로 정리되니까요. 하지만 실제 역사는 그렇게 단순하지 않습니다.

산업 전환은 기술의 우열만으로 결정되지 않았습니다. 자본과 노동 조건, 규제, 공급망, 고객 습관이 함께 움직였습니다. 어떤 숙련은 사라졌지만 다른 숙련은 더 중요해졌고, 새 기술을 늦게 도입한 판단이 오히려 합리적인 경우도 있었습니다.

어정쩡하게 받아들인 공장들

19세기 영국의 러다이트를 기술을 두려워한 노동자로만 기억하기 쉽습니다. 실제 갈등에는 임금 하락과 작업 통제권 상실, 숙련의 평가절하가 함께 얽혀 있었습니다. 기계를 부순 행동에 동의하지 않더라도, 그들이 무엇을 잃고 있었는지는 구분해서 볼 필요가 있습니다.

그렇다면 제목의 질문, 전기를 거부한 공장은 어떻게 되었을까요? 뜻밖에도 이야기는 '거부한 공장'보다 '어정쩡하게 받아들인 공장'에서 더 흥미로워집니다.

20세기 초 미국의 많은 공장주는 "증기기관이 멀쩡한데 왜 바꾸느냐"고 물었습니다. 전기 모터를 들인 공장도 상당수는 증기기관 자리에 큰 모터 하나를 앉혀 놓고, 천장을 가로지르는 긴 축과 가죽 벨트로 기계를 돌리던 옛 배치를 그대로 두었지요. 이런 공장의 생산성은 기대만큼 오르지 않았습니다. 기계마다 작은 모터를 달고 작업 흐름에 맞춰 배치도를 다시 그린 공장, 그리고 새 설비를 다룰 사람을 키운 공장이 앞서 나갔습니다. 모터를 사는 일과 공장을 바꾸는 일은 전혀 다른 일이었던 셈입니다. 그 배치도 이야기는 제4부 첫머리에서 다시 펼쳐 보겠습니다.

공장 이야기는 백 년 전 일이지만, 같은 장면은 우리 시대 기업에서도 되풀이됐습니다. 블록버스터, 코닥, 노키아는 변화 대응 실패의 단골 사례입니다. 코닥 이야기부터 해 볼까요? 1975년, 코닥의 젊은 엔지니어가 세계 최초의 디지털카메라 시제품을 만들었습니다. 그런데 회사는 필름이 너무 잘 팔리던 터라 그 발명품을 오랫동안 키우지 못했지요.



1975 작업대 위의 작은 상자, 그 뒤의 필름 상자 벽

그림 1-3 1975년 작업대 위의 작은 상자 하나. 그 뒤로는 천장까지 쌓인 필름 상자가 회사의 지갑을 채우고 있었지요.

흔히 이런 일화는 "오만한 경영진이 한마디로 회사를 말아먹었다"로 요약됩니다. 실제 사정은 그보다 복잡했습니다.

블록버스터에는 매장망과 연체료에 의존한 수익 구조, 높은 부채, 온라인 전환의 실행 문제가 함께 있었습니다. 코닥은 디지털 카메라를 일찍 개발하고도, 수익성 높은 필름 사업에서 디지털 사업으로 무게중심을 옮기는 일을 해내지 못했습니다. 노키아 역시 터치스크린 하나를 얕본 것보다 소프트웨어 생태계와 조직 의사결정의 복잡성이 더 큰 문제였습니다.

세 회사 모두 새 기술을 몰랐던 것은 아니었습니다. 새 기술이 기존 수익 구조를 어떻게 흔들지 알면서도, 당장의 손실을 감수하고 돈과 사람을 옮기지 못했지요. 큰 모터 하나에 낡은 축을 그대로 매달아 둔 공장과 닮은 모습입니다.

의심하되, 고칠 줄 아는 조직

AI에 신중한 태도를 곧바로 거부로 부르면 중요한 질문을 놓칩니다. 의료진이 진단 모델의 오류율을 묻고, 개발자가 생성 코드의 보안을 검토하고, 노동자가 감시와 일자리 영향을 따지는 것은 도입할 가치가 있는지 확인하는 과정이지요. 새 전기 설비를 들이기 전에 누전부터 살피는 전기 기사를 고집쟁이라고 부르지는 않으니까요.

반대로 "AI이므로 일단 써야 한다"는 태도도 위험합니다. 검증할 수 없는 모델을 핵심 업무에 연결하거나, 사람이 하던 일을 자동화하면서 책임자를 정하지 않으면 빠른 도입이 더 큰 비용으로 돌아옵니다.

그러니 "받아들일까, 말까"보다 먼저 물어볼 것들이 있습니다. 이걸로 어떤 문제를 풀려는 걸까요? 기존 방식보다 품질과 시간, 비용이 실제로 나아졌나요? 틀렸을 때 알아채고 되돌릴 수 있나요? 이익과 부담은 각각 누구에게 돌아가나요?

이 네 질문에 답하지 못한 채 미루는 조직은 전환 시기를 놓칠 수 있습니다. 같은 질문에 답하지 못한 채 서두르는 조직은 사고를 키울 수 있지요.

미룰 수도, 서두를 수도 없다면 무엇을 기준으로 움직여야 할까요? 기술 변화 앞에서 미래를 한 번에 맞히는 조직은 드뭅니다. 살아남은 조직은 대개 작은 범위에서 시험하고, 증거가 좋으면 넓히고, 기대와 다르면 멈추거나 되돌릴 줄 알았지요.

블록버스터와 코닥, 노키아의 사례를 오늘에 가져온다면 "AI를 거부하면 도태된다"는 경고보다 더 실용적인 질문이 남습니다. 기존 사업에서 얻는 이익 때문에 새 선택지의 증거를 외면하고 있지는 않은가. 반대로 유행에 밀려 실패 비용을 보지 못하고 있지는 않은가.

배치도를 다시 그린 공장들도 처음부터 정답을 알았던 것은 아니었습니다. 모터를 하나 달아 보고, 축을 하나 걷어 내고, 기계를 한 줄 옮겨 보면서 길을 찾았지요. 불편한 증거가 나왔을 때 계획과 돈의 흐름을 바꿀 수 있었던 곳이 끝내 공장 모습을 바꿨습니다.

AI는 결국 에너지 산업이다

AI 업계에서 가장 영향력 있는 세 사람이 있습니다. 샘 올트먼(Sam Altman), 젠슨 황(Jensen Huang), 일론 머스크(Elon Musk). 각각 OpenAI, NVIDIA, 그리고 xAI를 품은 SpaceX의 수장이지요. AI의 미래를 놓고 경쟁하는 라이벌들입니다.

그런데 이 세 사람이 요즘 가장 많이 하는 이야기가 있습니다. AI가 아닙니다. 에너지입니다.

AI 기업의 CEO가 에너지 이야기를 합니다. GPU를 만드는 사람이 전력 이야기를 합니다. 전기차 회사의 창업자가 발전소를 짓고 있습니다. 왜일까요?

세 사람이 각자 짓는 발전소

세 사람의 말과 돈이 어디로 가는지, 한 사람씩 따라가 보지요.

샘 올트먼: "지능의 비용은 결국 에너지의 비용으로 수렴한다" OpenAI의 CEO 샘 올트먼은 미국 의회에서 이렇게 증언했습니다.

> "칩도, 네트워크 장비도, 결국 로봇이 만들게 될 겁니다. 점점 더 효율적으로, 점점 더 싸게. 하지만 전자(electron)는 전자입니다. 결국 지능의 비용, AI의 비용은 에너지의 비용으로 수렴할 것입니다."

이 말의 의미를 곱씹어 볼 필요가 있습니다. AI를 구성하는 모든 것 — 반도체, 서버, 냉각 장치 — 은 기술이 발전하면 자동화되고 비용이 떨어집니다. 하지만 칩이 아무리 싸져도 그 칩을 돌리는 전기는 누군가 연료와 설비로 만들어야 합니다. 그래서 다른 원가가 다 내려간 뒤에 마지막까지 남는 바닥이 전기료라는 것이지요. 물리 법칙이라기보다 경제의 셈법에 가까운 주장입니다.

올트먼은 말만 한 것이 아닙니다. 행동으로 옮겼습니다.

그는 핵융합 스타트업 헬리온 에너지(Helion Energy)에 3억 7,500만 달러를 투자했습니다. 본인이 직접 "내 인생에서 가장 큰 투자"라고 밝힌 금액이지요. 오랫동안 헬리온의 이사회 의장도 맡았습니다. AI 회사 CEO가 핵융합 회사의 의장이었던 거지요. 이상하지 않나요?

이상하지 않습니다. 올트먼은 이 두 가지가 같은 문제의 양면이라고 봅니다. 그는 이렇게 말했습니다. "오늘날 세계에서 어디를 봐도 가장 부족한 두 가지 자원이 있습니다. 지능과 에너지입니다." 지능은 OpenAI로 해결하고, 에너지는 헬리온으로 해결하겠다는 것이지요.

두 회사가 너무 가까워진 탓에 생긴 일도 있습니다. 2026년 3월 올트먼은 헬리온 의장 자리에서 물러났습니다. OpenAI가 헬리온에서 전기를 대규모로 사들이는 협상에 들어가자, 파는 쪽 의자와 사는 쪽 의자에 한 사람이 같이 앉을 수는 없게 됐다는 설명이지요. 지분은 그대로 갖고 있습니다.[1]

헬리온은 2025년 7월 워싱턴주 맬러가에서 첫 발전소 공사를 시작했고, 마이크로소프트와는 50MW 규모의 전력 구매 계약을 맺었습니다. 계약서에는 2028년 말까지 가동을 시작한다고 적혀 있지요. 그런데 2026년 9월, 헬리온이 홈페이지에서 '2028년 전력 생산' 문구를 내렸다는 보도가 나왔습니다. 대변인도 2028년에 50MW가 모두 돌아가는 일은 없을 거라고 말했고요.[2] 핵융합 발전은 아직 실적을 낸 적이 없습니다. 이 계약도 그 성공에 걸어 둔, 위험을 함께 안은 약속이지요.

올트먼은 거기서 멈추지 않았습니다. 클린에너지 스타트업 엑소와트(Exowatt)에도 2,000만 달러를 투자했고, AI 데이터센터 전용 전력 솔루션을 개발하고 있습니다. 5,000억 달러 규모의 스타게이트 프로젝트에서도 운영을 맡으며 대규모 전력 인프라 구축에 직접 관여하고 있지요.

AI의 왕이 에너지 사업에 뛰어든 겁니다. 그렇다면 그 AI에 칩을 대는 사람은 어떨까요?

젠슨 황: "에너지가 AI 인프라의 제1원칙이다" NVIDIA의 CEO 젠슨 황은 GPU를 만드는 사람입니다. AI의 심장인 칩을 설계하지요. 그런데 이 사람이 요즘 칩만큼, 어쩌면 칩보다 더 자주 꺼내는 이야기가 전력입니다.

2026년 다보스 포럼에서, 황은 AI를 "5층 케이크"에 비유했습니다. 가장 아래층이 무엇이었을까요? 칩일 것 같지만, 에너지였습니다. 에너지 위에 칩과 컴퓨팅 인프라가 올라가고, 그 위에 클라우드, 그 위에 AI 모델, 그리고 맨 위에 앱이 있다는 것이지요.

케이크의 맨 아래가 무너지면 나머지 네 층은 무의미합니다. 황은 에너지를 "AI 인프라의 제1원칙이자, 시스템이 얼마나 많은 지능을 생산할 수 있는지를 결정하는 절대적 제약"이라고 했습니다.

2025년 COMPUTEX 기조연설에서 황은 더 직설적으로 말했습니다.

> "에너지를 투입하면, 믿을 수 없이 가치 있는 것이 나옵니다. 그것을 토큰이라고 부릅니다."

그리고 이렇게 덧붙였지요. "전자(electron)가 들어가면 부동소수점 수(floating point number)가 나옵니다." GPU 데이터센터를 "AI 공장"이라고 부르는 이유가 여기에 있습니다. 산업혁명 시대의 공장이 석탄을 태워 제품을 만들었듯, AI 공장은 전기를 태워 지능을 만든다는 것이지요.

황이 미국의 에너지 정책에 대해 비판적 발언까지 한 것은 주목할 만합니다. 2026년 1월, 그는 "미국이 에너지를 악마화해 온 것은 제게 전혀 이해가 되지 않습니다"라고 했습니다. GPU를 아무리 많이 만들어도, 그 GPU를 돌릴 전기가 없으면 소용없다는 절박함이 배어 있는 발언입니다.

조 로건 팟캐스트에서 황은 이런 취지의 말도 했습니다. AI의 미래를 제약하는 것은 칩이 아니라 전력이며, AI의 급속한 확장이 새로운 종류의 한계, 즉 전력의 벽에 정면으로 부딪히고 있다고요.

GPU의 왕이 전력을 걱정합니다. 이보다 강력한 신호가 있을까요? 걱정을 넘어 아예 발전 설비를 손에 쥔 사람도 있습니다.

일론 머스크: AI와 에너지를 동시에 소유한 유일한 사람. 일론 머스크는 좀 다른 위치에 있습니다. 다른 두 사람은 에너지가 필요해서 에너지를 이야기하지만, 머스크는 이미 에너지 사업을 하고 있습니다. Tesla의

에너지 저장 사업, 솔라 시티의 태양광 사업, 그리고 xAI의 데이터센터. AI와 에너지를 한 손에 쥔 보기 드문 인물이지요. 2026년 2월에는 SpaceX가 xAI를 인수하면서 로켓과 위성과 AI가 한 지붕 아래 들어갔습니다.[3]

2024년 초, 머스크는 보쉬 커넥티드 월드 컨퍼런스에서 이렇게 경고했습니다.

> "다음 부족은 전력일 겁니다. 모든 칩을 돌릴 만큼의 전기를 찾지 못하게 될 겁니다."

그가 짚은 병목은 차례로 옮겨 갑니다. 처음엔 반도체 칩이, 그다음엔 전력 변압기가, 마지막엔 전력 생산 자체가 발목을 잡는다는 것이었지요. 칩 부족이 풀리면 변압기가 모자라고, 변압기가 풀리면 결국 전기를 만드는 일 자체가 모자란다는 이야기입니다. 병목은 계속 에너지 쪽으로 이동합니다.

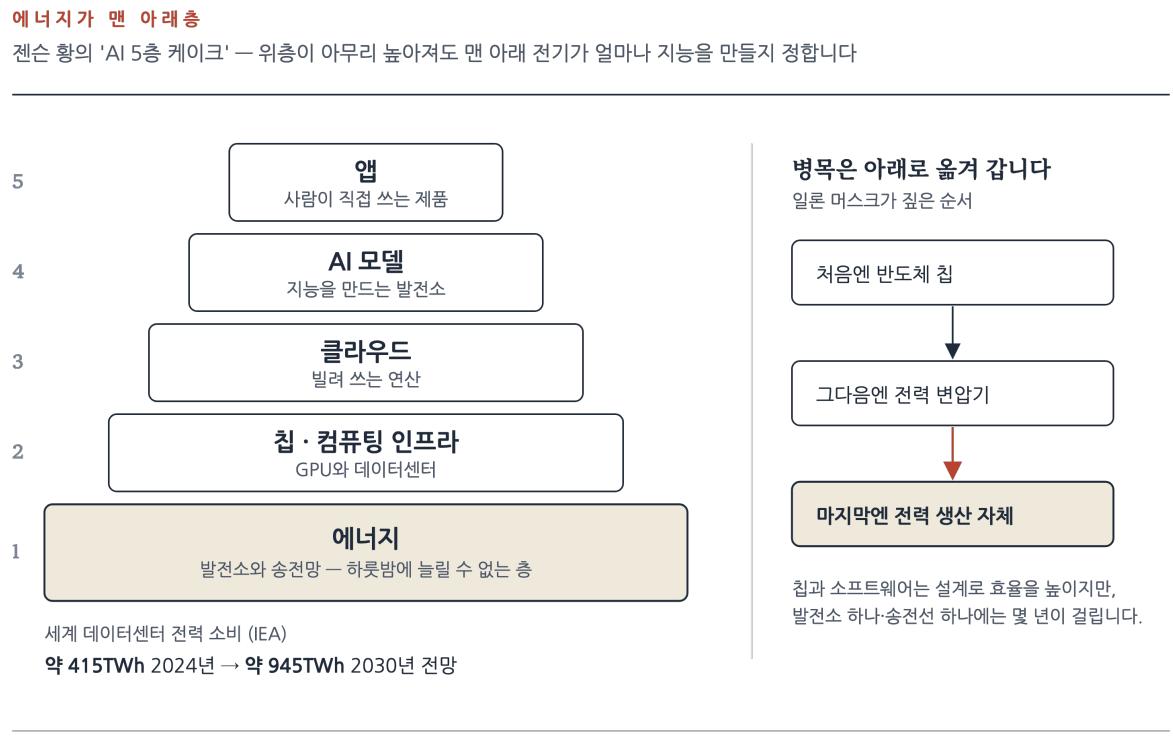


그림 1-4 젠슨 황의 다섯 층 케이크를 쌓아 보면, 맨 아래 받침이 에너지입니다. 머스크가 짚은 병목도 칩에서 변압기로, 다시 전력 생산으로 한 층씩 내려왔지요.

머스크는 이 문제에 자기 방식으로 덤벼들었습니다. 전력망을 기다리지 않은 것이지요.

xAI의 멤피스 데이터센터 "콜로서스(Colossus)"는 전력 제약을 잘 보여줍니다. 대규모 GPU 클러스터를 빠르게 가동하는 과정에서 지역 전력망 연결이 수요를 따라가지 못했고, 현장 발전 설비를 병행했습니다. 이 과정은 지역 대기오염과 허가 절차를 둘러싼 논쟁도 낳았습니다. 모자란 전기를 현장 발전으로 메울 수는 있어도, 그 비용과 환경 부담은 고스란히 남지요.

그리고 여기서 머스크의 수직 통합 전략이 빛을 발합니다. xAI의 콜로서스 데이터센터에는 Tesla 메가팩(Megapack)이 대규모로 배치되어 있습니다. 2025년 한 해에만 Tesla가 xAI에 판매한 메가팩 규모가 4억 3,000만 달러에 달했지요. AI 회사가 같은 그룹의 에너지 회사에서 배터리를 사서, 같은 그룹의 또 다른 회사가 만든 칩으로 AI를 돌립니다.

SpaceX의 AI 부문(옛 xAI)으로 지능을 만들고, Tesla로 에너지를 저장하고, 발전소로 전기를 생산합니다. AI와 에너지의 수직 통합. 이런 구조를 한 사람이 거느린 경우는 매우 드뭅니다.

세 거인이 동시에 같은 곳을 바라본다는 것

잠시 한 발 물러서서 생각해 봅시다. 이 세 사람은 서로 경쟁 관계에 있습니다. 올트먼과 머스크는 OpenAI의 지배권을 놓고 법정 다툼까지 벌였지요. 황은 두 사람 모두에게 GPU를 팔면서도 독자적인 AI 생태계를 구축하고 있습니다.

비전도 다릅니다. 올트먼은 AGI(범용 인공지능)를 꿈꾸고, 황은 AI 인프라의 표준을 지배하려 하며, 머스크는 AI 안전성과 오픈소스를 내세웁니다.

그런데 에너지에 대해서만은 세 사람의 말이 놀랍도록 일치합니다.

- 올트먼: "지능의 비용은 에너지의 비용으로 수렴한다"
- 황: "에너지가 AI 인프라의 제1원칙이다"
- 머스크: "다음 병목은 전력이다"

세 사람 모두 같은 결론에 도달한 겁니다. 지능을 만들려면 전기가 필요하고, 전기를 확보하는 것이 AI 경쟁의 궁극적 승부처라는 것.

이 흐름을 우연이라고만 보기는 어렵습니다. 물론 같은 기술이라도 어떤 인프라와 규칙을 고르느냐에 따라 속도와 혜택이 돌아가는 곳은 달라지지요.

AI 서비스가 늘면서 데이터센터 전력 수요도 커지고 있습니다. IEA는 세계 데이터센터 전력 소비가 2024년 약 415TWh에서 2030년 약 945TWh로 두 배 이상 늘어날 것으로 전망합니다.[4] 개별 응답의 전력 추정치는 모델과 하드웨어, 답변 길이에 따라 편차가 커서 전체 요청량에 곧바로 곱할 수 없습니다. 그래서 질문 하나에 몇 와트시가 드는지보다, 시스템 전체의 수요와 지역 전력망이 받는 부담을 함께 봐야 합니다.

칩은 설계를 바꾸면 더 효율적으로 만들 수 있습니다. 소프트웨어는 알고리즘을 개선하면 됩니다. 같은 일에 드는 전기도 혁신으로 줄일 수 있지요. 하지만 전기를 만드는 발전소와 송전망은 소프트웨어처럼 하룻밤에 늘릴 수 없습니다. 발전소 하나를 짓는 데 몇 년, 송전선 하나를 까는 데 또 몇 년이 걸리니까요. 속도의 한계가 결국 물리 세계에 있는 셈입니다.

그래서 AI의 세 거인이 모두 에너지를 이야기하는 겁니다. 지금 AI의 천장이 어디쯤 있는지 가장 먼저 느끼는 사람들이니까요.

그러니 이 책에서 "지능은 에너지다"라고 말할 때, 그것은 비유만이 아닙니다.

올트먼은 핵융합 회사에 인생 최대의 돈을 걸었습니다. 황은 GPU보다 전력을 더 걱정합니다. 머스크는 데이터센터 옆에 발전소를 세웠습니다.

AI의 세 거인이 문자 그대로 에너지 사업에 뛰어든 현실입니다. 지능 에너지 시대는 이미 시작되었습니다.

경쟁하는 세 사람이 에너지를 공통 제약으로 지목한다는 것은 무시하기 어려운 산업 신호입니다. 물론 기업인의 말은 자기 사업의 계산과 떼어 놓기 어렵지요. 실제 전력 수요와 계통 여건은 뒤의 장들에서 숫자로 따져 보겠습니다.

이 부의 첫 장에서 1769년 탄광에 처음 놓인 와트의 증기기관을 이야기했습니다. 그 기계는 석탄이 나오는 곳 바로 옆에 섰지요. 2026년의 데이터센터는 거꾸로 발전소와 송전선을 찾아 자리를 옮기고, 그것도 모자라 제 옆에 발전기를 세웁니다. 세 사람이 앞다퉈 발전소 이야기를 꺼내는 이유, 이제 조금 보이시지요?



---

1. "Scoop: OpenAI bets on Altman-backed fusion startup" (Axios, 2026.3.23, https://www.axios.com/2026/03/23/openai-fusion-altman-helion) — 올트먼의 헬리온 이사회 사임과 OpenAI의 전력 구매 협상(익명 소식통) 보도.

2. "Helion Energy fusion net electricity date" (Axios Pro, 2026.9.8, https://www.axios.com/pro/climate-deals/2026/09/08/helion-energy-fusion-net-electricity-date-year) — 홈페이지의 2028년 전력 생산 문구 삭제와 대변인 발언. 마이크로소프트 계약은 2028년 말까지 가동 시작을 요구.

3. "Factbox: SpaceX acquires xAI" (Reuters, 2026.2.2, https://finance.yahoo.com/news/factbox-spacex-acquires-xai-key-142912748.html) — 인수 발표와 기업가치(SpaceX 1조 달러, xAI 2,500억 달러, 익명 소식통) 보도.

4. Energy and AI, Executive Summary (IEA, 2025, https://www.iea.org/reports/energy-and-ai/executive-summary) — 데이터센터 전력 소비 2024년 415TWh, 2030년 기준 시나리오 945TWh. 하드웨어 효율과 AI 보급 속도에 따라 전망 오차가 크다고 함께 설명한다.



제2부

지능 에너지 시스템

지능 발전소: 파운데이션 모델은 어떻게 지능을 만드는가

우리 집 냉장고가 돌아가려면 어딘가에서 누군가 전기를 만들어야 하지요. 석탄을 태우든, 원자로를 돌리든, 태양광 패널을 깔든 말입니다.

지능도 마찬가지입니다. 그 일을 하는 것이 파운데이션 모델, 많은 종류의 과업에 두루 쓸 수 있도록 대규모 데이터로 학습한 모델입니다. 하나의 모델이 번역기에도, 상담 챗봇에도, 코딩 도구에도 능력을 대주니, 이 책은 이것을 발전소에 빗댑니다. 그렇다면 이 발전소 안에서는 무슨 일이 일어날까요?

화력발전소는 연료를 태워 물을 끓이고, 그 증기로 터빈을 돌립니다. 지능 발전소의 연료는 데이터이고, 터빈 자리에는 GPU가 돌아가지요.

대규모 언어모델(LLM)은 텍스트를 토큰이라는 작은 단위로 나누고, 앞에 놓인 토큰을 보고 다음 토큰을 예측하도록 학습합니다. "오늘 날씨가" 뒤에 무엇이 올지 맞히는 연습을 책과 논문, 웹페이지, 코드에서 셀 수 없이 되풀이하는 셈입니다. 그러면서 단어의 쓰임과 문장 구조, 개념 사이의 관계가 매개변수(흔히 파라미터라고 부르는 모델 내부의 숫자들)에 압축됩니다.

단어 맞히기가 어떻게 지능처럼 보이는 걸까요? 흔히 LLM을 '세상의 모든 책을 읽은 사람'에 비유하지만, 정확한 그림은 아닙니다. 모델은 읽은 문서를 서가에서 꺼내듯 기억하는 데이터베이스가 아닙니다. 학습 자료에 없던 문장을 새로 조합하기도 하고, 그럴듯하게 틀리기도 하지요. 수많은 예에서 통계적 패턴을 배운 생성 모델이라고 이해해 두면 덜 놀랍니다.

핵심 구조인 트랜스포머(Transformer)는 문장 안의 토큰들이 서로 어떤 관계를 갖는지 계산합니다. "나는 은행에 갔다"의 은행이 금융기관인지 강둑인지를 주변 문맥으로 가려내는 식이지요. 이 관계 계산 덕분에 긴 문맥을 읽고 답을 지어낼 수 있습니다.

짓고, 다듬고, 돌리는 발전소

이렇게 배운 능력은 거저 생기지 않습니다. 발전소처럼 짓는 데 드는 돈과 돌리는 데 드는 돈이 따로 있지요.

사전학습(Pre-training)은 발전소를 짓는 과정입니다. 엄청난 데이터와 연산을 쏟아부어 다음 토큰 예측 능력을 길러 내지요. 한 번 큰돈을 들여 보일러와 터빈을 세우는 단계입니다.

추론(Inference)은 지어진 발전소를 돌리는 일입니다. 사용자가 질문을 보낼 때마다 답을 생성하니, 서비스가 운영되는 동안 계량기가 계속 돌아가지요. 같은 모델이라도 입력과 출력의 길이, 도구 호출, 재시도 횟수에 따라 추론 비용이 달라집니다.

그래서 학습비는 건설비에, 추론비는 운영비에 가깝습니다. 뒤에서 토큰 요금을 따질 때 이 구분이 계속 나오니 기억해 두세요. 물론 비유가 멈추는 곳도 있습니다. 토큰은 전력량처럼 품질이 일정한 단위가 아니어서, 같은 토큰 수라도 일이 성공했는지, 결과가 얼마나 값졌는지는 다를 수 있습니다.



그림 2-1 발전소는 한 번 크게 지어 두고, 운전실에서 날마다 돌립니다. 파운데이션 모델도 학습으로 짓고, 질문이 올 때마다 추론으로 지능을 내보내지요.

그런데 발전소는 짓자마자 곧바로 송전을 시작하지 않습니다. 갓 지은 발전소에서 바로 깨끗한 전기가 안정적으로 나오지는 않지요. 출력이 고른지, 위험할 때 스스로 멈추는지를 조정하는 시운전이 필요합니다. 사전학습을 마친 모델도 그렇습니다. 질문에 따르는 법, 답변 형식을 지키는 법, 위험한 요청을 다루는 법을 따로 가르쳐야 하지요. 이 과정이 포스트 트레이닝(Post-training)입니다.

초기에는 사람이 여러 답변을 비교해 어느 쪽이 나은지 알려 주는 RLHF가 널리 쓰였습니다. 엔지니어가 계기판을 보며 손으로 밸브를 돌리는 방식이라, 사람 손이 많이 들고 확장이 어려웠지요. 그다음에는 수학 정답이나 코드 테스트처럼 결과를 자동으로 확인할 수 있는 과업에서 RLVR(검증 가능한 보상 강화학습)이 쓰이기 시작했습니다. 센서가 출력을 재고 스스로 피드백을 주는 자동 제어에 가깝습니다.

DeepSeek의 연구가 좋은 예입니다. R1-Zero는 정답 맞히기 보상만으로 모델이 스스로 추론하는 법을 익히는지 시험한 연구용 모델이었고, 실제로 내놓은 R1은 사람이 만든 모범 답안으로 먼저 길을 들인 뒤 강화학습을 여러 번 더한 모델입니다.[1]

DPO(직접 선호도 최적화)는 별도의 보상 모델을 만드는 복잡한 절차를 줄이고, "A 답변이 B보다 낫다"는 선호 데이터로 모델을 바로 조정하는 방법입니다.[2] 모델이 자기 답을 다시 읽고 고치게 하는 자기 교정 연구도 이어지고 있지요. 그렇다고 다시 읽기만 하면 오류가 저절로 사라지는 것은 아닙니다.[3] 바깥의 검증과 사람이 확인할 몫은 여전히 남습니다.

시운전을 잘하면 같은 발전소에서도 쓸모 있는 전기를 훨씬 많이 얻습니다. 그래도 시운전이 발전소 자체를 대신하지는 못하지요. 작은 포스트 트레이닝이 대규모 사전학습의 데이터와 연산을 언제나 메워 주는 것은 아니니, 어떤 과업을 어떤 평가로 개선했는지를 함께 물어보는 게 좋습니다.

크기보다 쓰임새로 고르는 발전소

요즘 큰 발전소는 모든 터빈을 한꺼번에 돌리지 않습니다. 수요에 맞춰 필요한 터빈만 켜지요. 큰 모델도 비슷하게 진화했습니다.

MoE(Mixture of Experts, 전문가 혼합) 구조는 모델 안에 여러 '전문가' 블록을 두고, 토큰이 들어올 때마다 그중 몇 개만 골라 씁니다. DeepSeek-V3는 전체 6,710억 개 매개변수 가운데 토큰 하나를 처리할 때 370억 개, 약 5.5%만 켭니다.[4] 덩치는 크게 키우면서 한 번 계산하는 비용은 낮추는 설계이지요.

그래서 '몇 B짜리 모델'이라는 말을 들으면 전체 크기인지, 실제로 켜지는 크기인지부터 물어봐야 합니다. 건물 전체의 방 수와 지금 불이 켜진 방 수는 다른 숫자니까요.

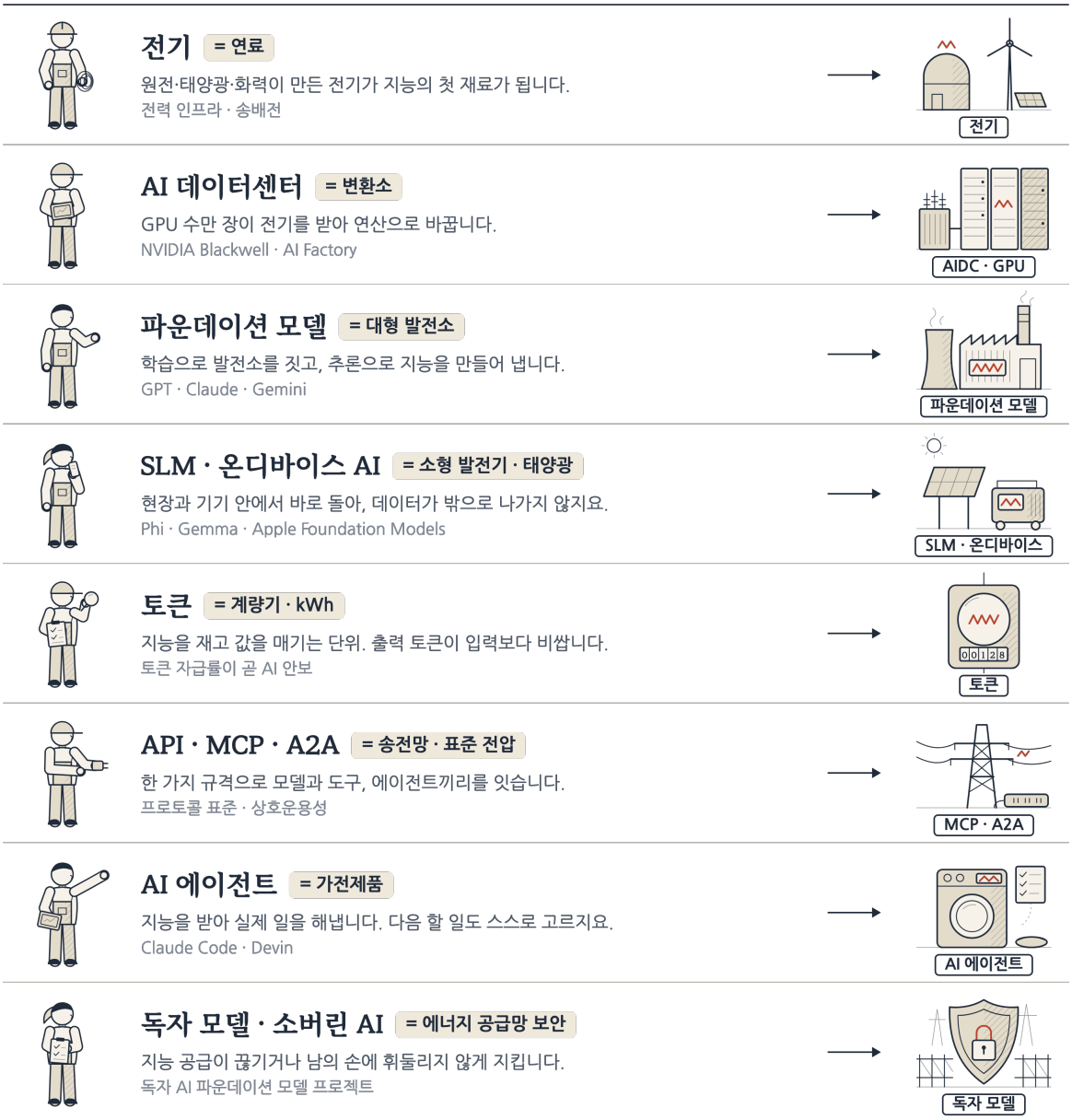
그런데 우리 집 전기를 꼭 가장 큰 발전소에서 받아야 할까요? 전력 시스템에는 대형 화력발전소도, 공장 옆 소형 발전기도, 지붕 위 태양광 패널도 있습니다. 지능도 클라우드의 대형 모델, 특정 업무를 맡는 소형 모델, 기기 안에서 직접 도는 온디바이스 모델로 나뉘지요. 셋은 서로 다른 조건에 맞습니다.

모델을 고를 때는 이름이나 매개변수 수보다 네 가지를 물어보면 됩니다. 이 일이 열린 추론인지 반복 분류인지, 데이터가 밖으로 나가도 되는지, 인터넷이 끊겨도 돌아가야 하는지, 그리고 표시 단가가 아니라 성공한 일 하나에 얼마가 드는지 말이지요. 작은 발전소들의 이야기는 다음 장에서 이어집니다.

다음 장으로 넘어가기 전에 제2부의 지도를 한 번 그려 두겠습니다. 원전과 태양광이 만든 전기가 데이터센터라는 변환소로 들어가고, 그 위에서 파운데이션 모델이라는 발전소가 토큰을 내보냅니다. 토큰은 API와 MCP·A2A라는 송전망을 타고 흘러, 에이전트라는 가전제품 안에서 일이 되지요. 전기가 kWh로 계량되듯 지능은 토큰으로 계량되고 과금됩니다.

지능 에너지 시스템

전력 시스템의 자리마다 AI 생태계의 부품을 하나씩 놓으면 — 위에서 아래로, 전기가 일이 되기까지



출처: 저자 작성 — 전력 시스템 구조와 AI 생태계의 구조적 대응.

그림 2-2 제2부에 나올 부품을 사람 하나, 장비 하나씩 짝지어 늘어놓았습니다. 위에서 아래로 읽으면 전기가 지능을 거쳐 일이 되기까지 제2부가 걸어갈 길이 한눈에 보이지요.

위 그림이 이 흐름을 한 장에 담았습니다. 제2부의 나머지 장은 이 선을 따라 한 칸씩 걸어갑니다.

1. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning" (Nature, 2025, https://www.nature.com/articles/s41586-025-09422-z)
2. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model" (NeurIPS 2023)
3. "Training Language Models to Self-Correct via Reinforcement Learning" (arXiv:2409.12917)
4. "DeepSeek-V3 Technical Report" (arXiv:2412.19437, https://arxiv.org/abs/2412.19437)

소형 발전기와 태양광: SLM과 온디바이스 AI

정전이 나도 병원 수술실은 불이 꺼지지 않습니다. 지하에 비상 발전기가 있으니까요. 도시 전체를 밝힐 필요는 없고, 수술실 조명과 인공호흡기만 돌리면 되지요. 그만큼의 일에는 커다란 발전소 대신 발전기 한 대면 충분합니다.

문서에서 날짜를 뽑거나 이메일을 분류하는 일도 비슷합니다. 이런 일에 언제나 가장 큰 모델이 필요할까요? 모델을 고르는 일은 성능 순위표에서 1등을 찾는 일보다, 과업과 데이터가 놓인 자리에 맞는 도구를 찾는 일에 가깝습니다. 비상 발전기에 해당하는 것부터 볼까요?

SLM(Small Language Model, 소형 언어 모델)은 대체로 수십억 개 안팎의 매개변수를 가진 비교적 작은 언어모델입니다. 요약·분류·정보 추출처럼 범위가 분명한 과업에서는 대형 모델보다 빠르고 싸게 목표를 달성할 수 있지요. 대형 화력발전소가 도시 전체를 밝힌다면, SLM은 공장 한 동을 맡는 소형 발전기에 가깝습니다. 필요한 만큼만, 가까운 곳에서 돌립니다.

에이전트가 하는 일을 뜯어보면 이런 좁은 단계가 많습니다. 웹페이지에서 값을 읽고, API를 호출하고, 정해진 형식으로 결과를 돌려주는 일이지요. 매번 범용 거대 모델을 부르는 것은 손전등 하나 켜자고 발전소를 돌리는 격입니다. NVIDIA 연구진이 에이전틱 AI에서 소형 모델의 역할을 강조한 이유도 여기에 있습니다.[1]

물론 작다고 저절로 싸거나 안전해지는 것은 아닙니다. 목표 정확도에 못 미쳐 재시도가 늘면 성공 작업당 비용은 오히려 커지지요. 작은 모델도 학습 데이터의 편향과 환각을 그대로 가질 수 있고, 회사 서버에 설치했다고 접근 통제가 저절로 생기지도 않습니다.

대표적인 계열로 Microsoft Phi, Google Gemma, Alibaba Qwen, Meta Llama의 소형 버전이 있습니다. 이름과 크기는 몇 달마다 바뀌니, 외워 두기보다 비교하는 법을 익혀 두는 편이 오래갑니다. 우리 데이터로 돌려 치명적인 오류가 얼마나 나는지, 실제로 쓸 하드웨어에서 끝까지 실행하면 얼마나 걸리는지, 재시도와 사람 검토 시간까지 넣으면 일 하나에 얼마가 드는지 재 보세요. 상업적으로 쓰고 고쳐서 배포해도 되는 라이선스인지, 취약점이 나왔을 때 누가 업데이트하고 모델을 어떻게 바꿔 끼울지도 미리 정해 두어야 하지요.

지붕 위 태양광, 그리고 섞어 쓰는 법

발전기 말고도 전기를 가까이서 얻는 방법이 하나 더 있지요. 지붕 위 태양광 패널은 그 집의 전기를 그 자리에서 만듭니다. 송전선을 거치지 않으니 정전에도 버티고, 남이 내 전기 사용량을 들여다볼 일도 줄지요.

온디바이스 AI가 이와 닮았습니다. 모델의 크기보다 실행 장소를 가리키는 말로, 스마트폰이나 노트북, 자동차, 산업 장비 안에서 모델을 직접 돌립니다. 지하철이 터널에 들어가 신호가 끊겨도 휴대전화 속 번역 기능이 메뉴판을 읽어 주는 장면을 떠올려 보세요. 인터넷이 끊겨도 동작하고, 원본 데이터를 외부 서버로 보내지 않도록 설계할 수 있습니다.

그렇다고 '기기 안에서 실행'과 '프라이버시 보장'이 같은 말은 아닙니다. 앱의 로그와 백업, 분석 도구, 권한 설정을 타고 데이터가 밖으로 나갈 수 있지요. 모델 파일과 입력 데이터가 기기에 남는 만큼 분실과 악성 앱, 업데이트 경로도 관리해야 합니다.

태양광도 흐린 날엔 전력망에 기대듯, 무거운 일은 여전히 클라우드로 넘어갑니다. Apple은 기기에서 실행하는 모델과, 복잡한 요청을 서버에서 처리하는 구조를 함께 운영합니다.[2] 이런 혼합형은 로컬 처리의 속도·프라이버시와 클라우드 모델의 성능을 함께 쓰는 방식이지요. 이때 어느 요청이 서버로 넘어가는지는 사용자가 알 수 있어야 합니다.



그림 2-3 멀리 있는 대형 발전소에 매번 기대지 않아도, 손 안의 기기와 현장의 작은 발전기가 가까운 일을 바로 처리합니다. 무거운 일이 생기면 그때 클라우드로 넘기면 되지요.

그러면 큰 모델과 작은 모델, 기기 속 모델을 실제로는 어떻게 나눠 쓸까요? 모든 일을 모델 하나에 맡기기보다 요청을 나눠 보내는 편이 현실적입니다. 순서는 대략 이렇습니다.

1. 규칙이나 검색만으로 정확히 처리되는지 먼저 봅니다.
2. 반복적이고 범위가 좁다면 SLM을 시험합니다.
3. 민감한 데이터를 다루거나 응답이 빨라야 하면 온디바이스 실행을 검토합니다.
4. 열린 추론과 긴 문맥이 필요할 때 대형 클라우드 모델로 넘깁니다.

단계마다 실패하면 다음 경로로 넘기는 기준을 두어야 하지요. 작은 모델이 자신 없는 요청은 스스로 거절하고, 복잡한 요청만 큰 모델이나 사람에게 넘기면 품질과 비용을 함께 붙잡을 수 있습니다.

전력망이 대형 발전소와 비상 발전기, 지붕 위 패널을 함께 쓰듯, 지능도 셋을 섞어 씁니다. 그러니 큰 모델과 작은 모델, 기기 속 모델 가운데 누가 이기느냐를 묻기보다 '이 일은 어디서 돌리는 게 맞을까'를 물어보세요. 데이터가 어디 있는지, 틀렸을 때 얼마나 아픈지가 답을 정해 줍니다.



---

1. Belcak, Heinrich, Diao, Fu, Dong, Muralidharan, Lin, Molchanov, "Small Language Models are the Future of Agentic AI" (arXiv:2506.02153, https://arxiv.org/abs/2506.02153)

2. "Introducing third-generation of Apple Foundation Models" (Apple Machine Learning Research, 2026.6.8, https://machinelearning.apple.com/research/introducing-third-generation-of-apple-foundation-models)

변환소: 전기를 지능으로 바꾸는 인프라

발전소에서 전기를 만들려면 터빈이 필요합니다. 연료를 아무리 많이 쌓아 두어도, 그것을 돌려 전기로 바꿔 주는 장치가 없으면 소용이 없지요. 지능 에너지 생태계에서는 데이터센터가 전기를 계산으로 바꾸는 변환소이고, 그 안에서 쉬지 않고 도는 터빈이 GPU(Graphics Processing Unit)입니다.

NVIDIA의 젠슨 황 CEO는 차세대 데이터센터를 AI 공장(AI Factory)이라고 부릅니다. 전통적인 발전소가 석탄이나 가스를 태워서 전기를 만들듯, AI 공장은 전기와 데이터를 넣으면 지능이 나오는 곳이라는 것이지요. 이 책의 발전소 비유와도 잘 맞는 표현입니다.

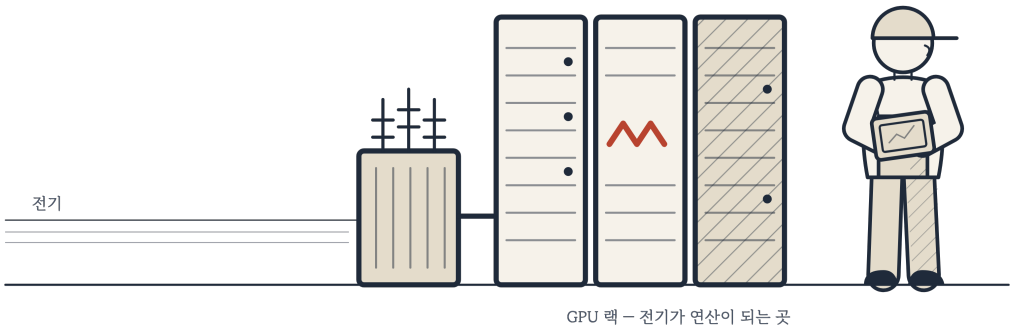


그림 2-4 변압기를 거친 전기가 GPU 랙으로 들어가면 연산이 되어 나옵니다. 지능 공장의 통로는 이렇게 생겼지요.

전기 먹는 하마

그렇다면 이 공장은 전기를 얼마나 먹을까요? 대형 모델 하나를 학습하는 데 전기가 정확히 얼마나 들었는지는 대개 공개되지 않습니다. 그래도 그림은 같지요. 수많은 GPU가 몇 주, 몇 달씩 돌아가며 데이터를 학습하고, 그 끝에 모델이 남습니다.

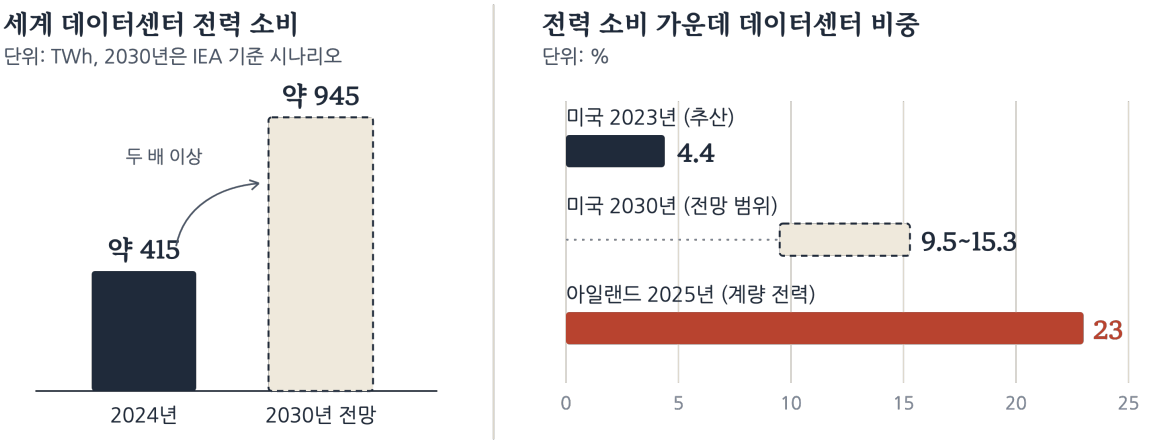
질문에 답할 때도 전기가 듭니다. 그런데 질문 한 건의 소비량은 모델, 답변 길이, 배치 처리, 칩과 냉각 방식에 따라 크게 달라집니다. 사업자가 배포 사양을 밝히지 않았는데 어떤 추정치 하나를 전체 요청량에 곱하면, 그럴듯하지만 근거 없는 숫자가 나오지요. 그래서 여기서는 질문 한 건보다 데이터센터 전체를 보겠습니다.

AI 데이터센터의 전력 소비는 빠르게 증가하고 있습니다. IEA에 따르면 전 세계 데이터센터의 전력 소비는 2024년 약 415TWh에서 2030년 기준 시나리오 약 945TWh로 두 배 이상 늘어날 전망입니다.[1] AI가 얼마나 퍼질지, 효율이 얼마나 오를지, 전력망이 버텨 줄지를 가정한 시나리오이니 확정된 미래로 읽지는 마세요.

이 숫자가 어느 정도인지 비교해볼까요? 아일랜드 통계청은 데이터센터가 2025년 계량 전력 소비의 23%를 차지했다고 집계했습니다.[2] 미국 에너지부 산하 로런스버클리연구소는 미국 데이터센터의 비중을 2023년 4.4%로 추산했고, 2030년에는 9.5~15.3% 범위가 될 수 있다고 봅니다.[3] 전국 평균이 오르기 전에, 데이터센터가 몰린 지역의 전력망부터 숨이 차오르는 셈이지요.

데이터센터 전력

세계 소비는 두 배 넘게 늘 전망이고, 데이터센터가 몰린 곳에서는 전력망 비중이 먼저 커집니다



나라 평균이 오르기 전에, 데이터센터가 몰린 곳의 전력망부터 숨이 찹니다

출처: IEA, Energy and AI — Energy demand from AI (2025); Lawrence Berkeley National Laboratory, United States Data Center Energy Usage Report: 2025 Update (2026.6); Central Statistics Office Ireland, Data Centres Metered Electricity Consumption 2025 (2026.7). 세 기관의 집계 기준이 달라 막대끼리 더하거나 순위를 매기지 않습니다.

그림 2-5 세계 데이터센터의 전력 소비는 2024년 약 415TWh에서 2030년 약 945TWh로 늘어날 전망입니다. 미국 전체에서는 아직 한 자릿수 비중이지만, 데이터센터가 몰린 아일랜드에서는 벌써 23%를 차지했지요.

이 시대의 GE와 인프라 경주

20세기 초 전기 시대를 연 기업이 있었습니다. 토머스 에디슨이 세운 제너럴 일렉트릭(GE)이지요. 발전기를 만들고, 송전 시스템을 구축하고, 전기를 쓰는 가전제품까지 만들었습니다. 전기 생태계 전체를 지배한 기업이었습니다.

지능 에너지 시대에 그와 비슷한 위치를 차지하고 있는 기업이 NVIDIA입니다. AI 학습과 추론용 가속기 생태계에서 강한 지위를 갖고 있고, GPU뿐 아니라 네트워크와 소프트웨어까지 함께 공급하지요.

발전기부터 송전 설비, 가전까지 쥐었던 GE와 닮은 구석이 있습니다. 시장점유율은 가속기·서버·클라우드 중 무엇을 세느냐에 따라 달라지니, 숫자 하나로 단정하기는 어렵습니다.

NVIDIA는 2025년 유럽 사업자들이 3,000 엑사플롭스가 넘는 Blackwell 자원을 배치할 계획이라고 발표했습니다.[4] 한국 정부와 기업이 발표한 26만 장 이상 계획 가운데 정부 몫은 국가 AI컴퓨팅센터와 국내 클라우드 사업자에 배치할 5만 장 이상입니다.[5] 발표한 수량이 곧 들여와 설치하고 돌리는 수량은 아니라는 점은 기억해 두세요.

GPU 공급에는 첨단 패키징과 고대역폭메모리, 네트워크 장비의 생산 능력까지 얽혀 있습니다. 어느 한 부품의 증설이 늦어져도 전체 시스템의 납기가 밀릴 수 있지요. 얼마나 기다려야 하는지도 계약과 제품마다 달라서, 대기 기간 숫자 하나보다 공급망 전체의 흐름을 보는 게 낫습니다.

칩을 사들이는 쪽의 지갑도 그만큼 커졌습니다. 주요 기술기업은 데이터센터와 서버에 해마다 수천억 달러 규모의 자본을 쏟고 있습니다. 그런데 회사마다 회계연도와 리스 포함 여부가 달라 단순 합계는 크게 흔들리고, 설비투자 전체가 AI에 쓰이는 것도 아니지요.

그래서 투자액보다 눈여겨볼 것은 데이터센터가 전력망·발전 설비·냉각 인프라와 함께 계획되고 있느냐입니다. 서버는 몇 달이면 주문하지만 송전망은 몇 해가 걸리니까요.

이 투자 경주는 전력 인프라 투자와 한 몸처럼 움직입니다. AI 데이터센터의 가장 큰 병목 가운데 하나가 전력 공급이라는 점을 생각하면, 지능과 전기는 비유를 넘어 실제 전선으로 이어져 있는 셈이지요.

그래서 데이터센터를 둘러보는 사람들은 이제 GPU 몇 장이 들어왔는지보다 먼저 묻습니다. 전기는 어디서 끌어오고, 열은 어떻게 식히며, 들여놓은 칩이 실제로 몇 시간이나 돌고 있느냐고요. 변환소를 지었으면 그다음은 그 지능을 실어 나를 길입니다. 다음 장의 송전망 이야기입니다.



1. Energy demand from AI (IEA, 2025, https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai) — 2024년 415TWh와 2030년 기준 시나리오 945TWh, 효율·AI 보급·에너지 병목에 따른 불확실성을 함께 제시한다.

2. Data Centres Metered Electricity Consumption 2025 (Central Statistics Office Ireland, 2026.7.7, https://www.cso.ie/en/releasesandpublications/ep/p-dcmec/datacentresmeteredelectricityconsumption2025/) — 2025년 데이터센터 비중 23%.

3. United States Data Center Energy Usage Report: 2025 Update (Lawrence Berkeley National Laboratory, 2026.6, https://eta-publications.lbl.gov/publications/united-states-data-center-energy-2025) — 2023년 4.4%, 2030년 9.5~15.3% 범위 전망.

4. Europe Builds AI Infrastructure With NVIDIA (NVIDIA, 2025.6.11, https://nvidianews.nvidia.com/news/europe-ai-infrastructure) — 3,000 엑사플롭스 이상은 당시 발표된 유럽 사업자의 배치 계획이다.

5. NVIDIA, South Korea Government and Industrial Giants Build AI Infrastructure (NVIDIA, 2025.10.30, https://nvidianews.nvidia.com/news/south-korea-ai-infrastructure) — 민관 합계 26만 장 이상과 정부 몫 5만 장 이상의 계획 구성을 명시한다.

송전망: MCP와 A2A, 지능이 흐르는 길

발전소에서 전기를 만들었다고 해서 끝이 아닙니다. 그 전기를 공장과 가정에 보내려면 송전망이 필요합니다. 그런데 송전망에는 한 가지 전제 조건이 있습니다. 규격이 통일되어야 한다는 것이지요.

19세기 말, 미국의 전기 산업은 혼란 그 자체였습니다. 에디슨은 직류(DC)를 밀었고, 웨스팅하우스는 교류(AC)를 밀었습니다. 전압도 주파수도 제각각이었습니다. A 발전소의 전기로는 B 회사의 전등을 켤 수 없는 상황이 비일비재했지요. 결국 교류 60Hz가 표준으로 자리잡으면서, 어느 발전소에서 만든 전기든 어떤 가전제품에서든 쓸 수 있게 되었습니다. 이 표준화가 전기의 보편화를 가능하게 한 결정적 계기였습니다.

2025년까지의 AI 세계는 교류 표준이 정해지기 전의 전기 산업과 비슷했습니다. 각 AI 시스템이 제각기 다른 방식으로 외부 도구와 연결되었고, 서로 다른 AI 에이전트끼리 소통하는 공통 규약도 없었습니다. GPT와 Claude가 같은 데이터베이스를 쓰려면 각각 별도의 연결 코드를 짜야 했지요.

2024년 말부터 2025년 사이, 이 상황을 바꾸려는 두 개의 프로토콜이 등장했습니다.

AI에도 교류 표준이 생기다

MCP(Model Context Protocol)는 Anthropic이 2024년 11월에 공개한 개방형 프로토콜입니다. AI 시스템이 외부 데이터와 도구에 접근하는 방식을 공통 규격으로 맞추려는 시도이지요.[1]

비유하자면 이렇습니다. 예전에는 AI가 데이터베이스를 쓰려면 그 데이터베이스 전용 플러그를 만들어야 했습니다. 파일 시스템을 쓸 때는 또 다른 플러그가 필요했고, 웹 검색을 할 때는 또 다른 것이 필요했습니다. MCP는 이 연결 방식을 공통 규격으로 맞추려는 애플리케이션 계층 프로토콜입니다. AI 쪽에 MCP 클라이언트를 달고, 외부 서비스 쪽에 MCP 서버를 두면 각 서비스마다 새 연결 코드를 짜는 일을 줄일 수 있지요. 하지만 콘센트 모양이 같다고 아무 기기나 꽂아도 되는 것은 아닙니다. 서버를 믿을 수 있는지 확인하는 인증, 어디까지 읽고 바꿀 수 있는지 정하는 권한, 각 구현의 호환성은 별도로 설계해야 합니다.

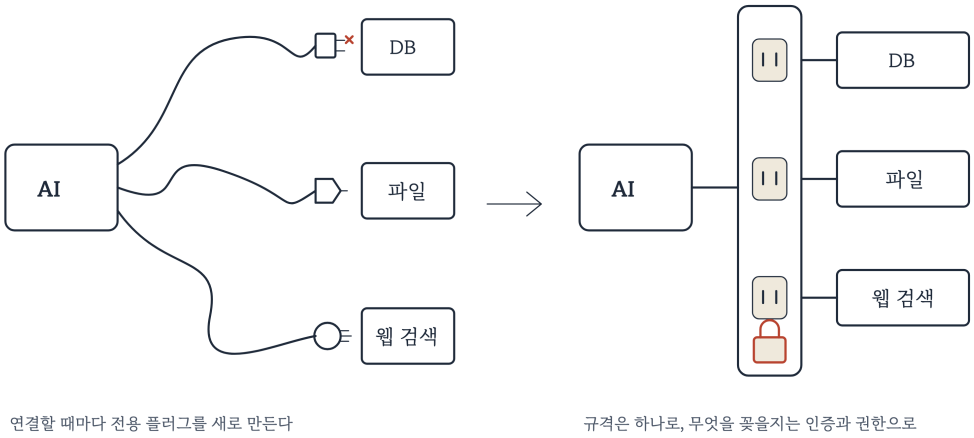


그림 2-6 연결할 때마다 새로 깎던 전용 플러그가 콘센트 하나의 규격으로 모입니다. 무엇을 꽂아도 되는지는 여전히 인증과 권한이 정합니다.

MCP의 성장 속도는 빨랐습니다. Anthropic은 2025년 12월 공개 생태계에 활성 MCP 서버가 1만 개 이상이고, Python·TypeScript 공식 SDK의 월간 다운로드가 합쳐서 9,700만 건을 넘었다고 밝혔습니다.[2] 물론 내려받은 횟수와 실제로 돌아가는 서버 수는 다른 숫자이지요.

더 의미 있는 것은 경쟁사들의 반응입니다. Anthropic은 ChatGPT, Gemini, Microsoft Copilot과 Visual Studio Code 등이 MCP를 채택했다고 밝혔습니다.[2] MCP가 AI와 도구를 잇는 주요 공통 규격으로 자리 잡아가고 있다는 신호입니다. 교류 표준의 승리처럼 완결된 결과로 보기에는, 보안과 호환성을 다듬을 일이 아직 남아 있지요.

2025년 12월에는 Anthropic이 MCP를 Linux Foundation 산하의 Agentic AI Foundation(AAIF)에 기증했습니다. 중립적인 공동 관리 기반으로 옮긴 것이지요. Anthropic, Block, OpenAI가 AAIF 공동 설립자로 참여했습니다.[2]

에이전트끼리 말이 통하는 세계

MCP가 AI와 외부 도구를 연결하는 표준이라면, A2A(Agent-to-Agent Protocol)는 AI 에이전트끼리 소통하는 표준입니다.

전기 시스템으로 치면, MCP가 발전소에서 가정까지의 송전선이라면, A2A는 서로 다른 전력망을 연결하는 계통 연계선에 가깝습니다. 서로 다른 지역의 전력망이 연결되면 한쪽에서 전력이 부족할 때 다른 쪽에서 보충할 수 있지요.

Google이 2025년 4월에 발표한 A2A 프로토콜은 기존 웹 기술을 활용해 서로 다른 에이전트가 정보를 교환하고 작업을 조정하도록 설계됐습니다.[3] 익숙한 웹 기술을 재사용할 수 있어 도입 부담을 줄였지요.

그런데 메시지가 오갈 수 있다는 것과 상대 에이전트를 믿고 일을 맡겨도 된다는 것은 다른 문제입니다. A2A에서도 신원 확인, 사용자 대리 권한, 작업별 승인 규칙을 따로 둘 필요가 있습니다.

A2A의 핵심 개념 중 하나가 Agent Card입니다. JSON 형식으로 에이전트의 기능을 광고하는 일종의 명함이지요. "저는 법률 문서 분석을 잘합니다"라고 적힌 명함을 내밀면, 다른 에이전트가 법률 관련 작업이 생겼을 때 이 에이전트를 찾아서 일을 맡길 수 있습니다.

2025년 6월 A2A는 Linux Foundation 프로젝트로 이관됐습니다. 재단은 당시 100개가 넘는 기술기업이 지원하고 있다고 밝혔습니다.[4] MCP와 마찬가지로 특정 기업 한 곳을 벗어난 공동 관리 기반을 갖춘 것입니다.

그렇다면 MCP와 A2A가 함께 보편화되면 어떤 세계가 열릴까요?

한 가지 시나리오를 상상해봅시다. 당신이 AI 에이전트에게 "다음 주 출장 준비해줘"라고 말합니다. 이 에이전트는 MCP를 통해 당신의 캘린더, 이메일, 여행사 API에 접근합니다. 항공편과 호텔을 검색하고, 일정을 조율합니다. 그런데 출장지의 날씨와 현지 교통 정보도 필요합니다. 이때 A2A를 통해 날씨 전문 에이전트와 교통 정보 에이전트에게 도움을 요청합니다. 각 에이전트가 자기 전문 분야의 정보를 가져와서, 최종적으로 완성된 출장 계획이 당신에게 전달됩니다.

이런 장면이 돌아가려면 서로 다른 AI 시스템이 같은 규격으로 데이터를 주고받아야 하지요. 공통 프로토콜은 번역기와 어댑터를 매번 새로 만드는 수고를 덜어 줍니다. 그렇다고 연결된다는 사실만으로 안전해지지는 않습니다. 내 에이전트가 항공권을 결제하기 전에 누가 승인하는지, 무엇을 했는지 기록이 남는지는 따로 챙겨야 할 몫입니다.

2025년 한 해 동안 MCP와 A2A가 잇따라 Linux Foundation의 관리 아래 들어갔습니다. 두 규격을 한 회사의 것이 아닌 업계 공통 기반으로 키우려는 움직임이지요. 2026년 8월에는 A2A도 MCP가 있는 AAIF로 합류해, 두 규격이 한 지붕 아래 모였습니다.[5] 교류 60Hz도 하루아침에 표준이 되지는 않았듯, 관리 주체가 중립이 됐다고 모든 구현이 곧바로 맞물리지는 않습니다.

송전망에 닿지 못한 발전소는 전기를 만들어도 쓸 데가 없지요. 아무리 좋은 파운데이션 모델도 도구와 다른 에이전트에 닿지 못하면 할 수 있는 일이 줄어듭니다. 비유가 멈추는 곳도 있습니다. 전기 송전망은 에너지를 물리적으로 나르지만, 이 프로토콜들은 프로그램 사이의 메시지와 도구 호출 방식을 맞출 뿐이니까요. 규격이 퍼질수록 연결은 쉬워지겠지만, 안전하게 흐르게 만드는 일은 각 시스템의 몫입니다.



1. Introducing the Model Context Protocol (Anthropic, 2024.11.25, https://www.anthropic.com/news/model-context-protocol) — MCP 공개일과 목적, 클라이언트·서버 구조를 설명한다.

2. Donating the Model Context Protocol and establishing the Agentic AI Foundation (Anthropic, 2025.12.9, https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation); Linux Foundation Announces the Formation of the Agentic AI Foundation (Linux Foundation, 2025.12.9, https://www.linuxfoundation.org/press/linux-foundation-announces-the-formation-of-the-agentic-ai-foundation) — MCP 기증, 공동 설립자, 서버·SDK·제품 채택 현황.

3. Announcing the Agent2Agent Protocol (A2A) (Google Developers Blog, 2025.4.9, https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/) — 발표일과 에이전트 간 상호운용 목적을 설명한다.

4. Linux Foundation Launches the Agent2Agent Protocol Project (Linux Foundation, 2025.6.23, https://www.linuxfoundation.org/press/linux-foundation-launches-the-agent2agent-protocol-project-to-enable-secure-intelligent-communication-between-ai-agents) — Linux Foundation 이관과 100개 이상 지원 기업을 명시한다.

5. "A2A joins the Agentic AI Foundation" (AAIF, 2026.8.17, https://aaif.io/blog/a2a-joins-aaif) — A2A v1.0(2026.3) 이후 AAIF 호스팅 프로젝트로 합류.

에이전트: 지능으로 움직이는 가전제품

전기가 발명되고 송전망이 깔렸을 때, 사람들의 삶이 바로 달라진 것은 아니었습니다. 삶을 바꾼 것은 전기 자체가 아니라, 전기로 작동하는 가전제품이었습니다. 냉장고가 있어야 음식을 오래 보관할 수 있었고, 세탁기가 있어야 빨래에서 해방될 수 있었지요.

지능 에너지 생태계에서 이 가전제품에 해당하는 것이 AI 에이전트입니다. 파운데이션 모델이라는 발전소에서 만든 지능을, MCP와 A2A라는 송전망을 통해 받아서, 실제로 사용자를 위해 일하는 존재입니다.

그렇다면 우리가 이미 써 본 챗봇과 에이전트는 어떻게 다를까요?

챗봇은 물어보면 대답합니다. "서울에서 부산 가는 KTX 시간표 알려줘"라고 하면 시간표를 알려주지요. 하지만 거기까지입니다. 표를 예매하지는 않습니다.

에이전트는 다릅니다. "다음 주 금요일 오후에 부산 출장 잡아줘"라고 하면, 캘린더를 확인하고, KTX를 예매하고, 호텔을 잡고, 일정을 정리해서 보고합니다. 스스로 계획을 세우고, 도구를 사용하고, 문제가 생기면 대안을 찾습니다.

전기 비유로 돌아가면, 챗봇은 전등입니다. 스위치를 켜면 빛이 나옵니다. 에이전트는 세탁기입니다. 빨래를 넣고 버튼만 누르면, 세제 투입부터 탈수까지 알아서 합니다.

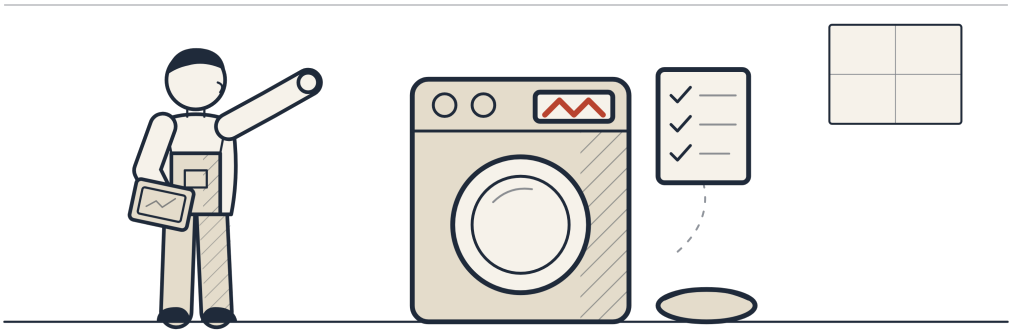


그림 2-7 세탁기는 버튼 한 번이면 헹굼과 탈수까지 알아서 해냅니다. 에이전트에게도 할 일 목록을 건네면, 끝날 때까지 다음 단계를 스스로 골라 가지요.

기업들은 이제 "질문에 답하는 AI"를 넘어, 도구를 써서 정해진 범위의 업무를 해내는 AI를 시험하고 있습니다. 세탁기에도 문이 잠기는 장치가 있듯, 에이전트에게는 어디까지 손대도 되는지 경계가 먼저 그어져야 하지요. 결제·삭제·외부 전송처럼 되돌리기 어려운 행동에는 사람의 승인이 필요합니다.

코딩에서 법률까지, 일터에 들어온 가전제품

에이전트의 가장 성공적인 사례 중 하나가 Anthropic의 Claude Code입니다. 터미널에서 작동하는 코딩 에이전트인데, 개발자가 "이 버그 고쳐줘" 또는 "이 기능 추가해줘"라고 말하면 코드베이스를 분석하고, 파일을 수정하고, 테스트를 돌려서 결과를 보고합니다.

코드베이스 탐색, 수정, 테스트를 한 흐름으로 묶었다는 점에서 주목받았지요. 그래도 만들어 낸 코드는 사람이 검토해야 하고, 저장소와 배포 시스템에 주는 권한도 꼭 필요한 만큼만 주는 것이 좋습니다.

코딩이라는 영역에서 "전기 세탁기"가 나온 셈입니다. 손빨래를 하던 시대에서, 기계가 빨래를 해주는 시대로 넘어가고 있습니다.

코딩만의 이야기는 아닙니다. 법률 분야의 Harvey AI는 2026년 3월 110억 달러로 평가받은 지 반년 만인 9월 투자 유치에서 기업 가치를 150억 달러 넘게 인정받았다고 보도됐습니다.[1] 하루에 40만 건 이상의 에이전트 쿼리를 처리하고, 2,000만 건 이상의 계약 약관을 추출했습니다. HSBC가 법률 AI 플랫폼으로 도입했고, 2026년 3월에는 법률 팀이 직접 커스텀 AI 에이전트를 만들 수 있는 Agent Builder를 출시했습니다.

소프트웨어 개발 분야에서는 Devin이 있습니다. "세계 최초의 AI 소프트웨어 엔지니어"를 표방하는 이 에이전트는, 코드를 작성하고 디버깅하고 배포까지 자율적으로 수행합니다. Goldman Sachs는 Devin을 하이브리드 인력(Hybrid Workforce)의 1호 직원으로 채택했습니다.

고객 서비스 분야의 Intercom Fin은 평균 해결률 51%, 정확도 99.9%를 내세웁니다. 회사가 직접 밝힌 수치이지요. 뒤에서 다시 보겠지만, 하루 수십만 건을 처리하면 0.1%의 오류도 수백 건이 되지요. Klarna에서는 AI 에이전트가 1차 지원 티켓 대부분을 사람 없이 처리합니다.

영업, 금융 분석, 통신, 보험까지. 전기가 보급되면서 공장마다, 가정마다, 사무실마다 전기 기기가 들어갔던 것처럼, 산업마다 AI 에이전트가 들어가고 있습니다.

에이전트가 에이전트를 부를 때

여기서 한 걸음 더 나아간 것이 멀티 에이전트 시스템입니다. 하나의 에이전트가 혼자 모든 일을 하는 것이 아니라, 여러 전문 에이전트가 팀을 이루어 일하는 구조입니다.

세탁기가 빨래를 하고, 건조기가 말리고, 다림질 로봇이 다림질을 하는 것과 비슷합니다. 각자 잘하는 일을 맡고, 전체 프로세스가 자동으로 돌아가는 것이지요.

여러 모델과 도구를 한 업무 흐름에 엮는 제품도 늘고 있습니다. 연결이 늘어날수록 상대가 누구인지 확인하고, 권한을 얼마나 넘겼는지 적고, 실패했을 때 되돌리는 장치가 더 중요해지지요.

그런데 여러 에이전트가 협력하려면 뭐가 필요할까요? 앞에서 다룬 MCP와 A2A, 바로 그 송전망 표준입니다. 에이전트들이 같은 규격으로 소통할 수 있어야, 법률 에이전트가 금융 에이전트에게 "이 계약의 재무 리스크를 분석해달라"고 요청하는 일이 가능해집니다.

팀을 이룬 에이전트를 혼자 부릴 수 있다면, 회사는 몇 명이면 될까요? AI 에이전트가 작은 조직의 생산성을 크게 높일 것이라는 전망도 있습니다. 혼자서 에이전트 팀을 부려 기업 가치 10억 달러짜리 회사를 만든다는 "1인 10억 달러 기업" 이야기가 그렇지요. 아직은 널리 관찰된 현상이라기보다 가능성을 강조한 예측입니다.

한 사람이 여러 에이전트를 운영할 수는 있지만, 회계·법률·고객 대응·품질 책임까지 없어지는 것은 아닙니다. 자동화가 조직의 모양을 바꾸더라도 책임의 주체는 사람과 기업에 남습니다.

세탁기가 빨래 시간을 덜어 줬어도 옷을 고르고 개는 일은 사람에게 남았지요. 에이전트도 일터의 잔손질부터 덜어 가는 중입니다.



1. Bloomberg Law (2026.3, https://news.bloomberglaw.com/bloomberg-government-news/legal-ai-startup-harvey-raises-funds-at-11-billion-valuation); AI Weekly (2026.9, https://aiweekly.co/alerts/harvey-raises-550m-at-156b-co-led-by-lightspeed-and-diffusion) — 2026.3 기업 가치 110억 달러; 2026.9 5억 5,000만 달러 유치, Lightspeed·Diffusion 공동 주도, 기업 가치 약 155억~156억 달러 보도.

독자 모델을 가진 나라, 빌린 나라

1973년 10월, 아랍 산유국들이 석유 수출을 금지했습니다. 오일 쇼크가 시작된 것이지요.

그 전까지 서방 국가들은 석유가 끊길 수 있다는 생각을 별로 하지 않았습니다. 중동에서 싼 석유가 꾸준히 들어왔으니까요. 그런데 하루아침에 석유가 끊기자 모든 것이 멈췄습니다. 주유소 앞에 수 킬로미터 줄이 이어졌고, 공장이 멈추었고, 주가가 폭락했습니다. 당시 미국의 닉슨 대통령은 국가 비상사태를 선포해야 했습니다.

오일 쇼크가 남긴 교훈은 아프고도 단순했습니다. 에너지를 남에게 기대면 나라의 안전까지 흔들린다는 사실이지요. 지금 AI 분야에서 비슷한 장면이 만들어지고 있습니다.

칩이 끊기면 지능도 끊긴다

미국은 AI 반도체에 대한 수출 통제를 지속적으로 강화해왔습니다. 2025년 12월 "Operation Gatekeeper"를 통해 중국·홍콩으로 향하는 1.6억 달러 규모의 AI 칩 밀수출 네트워크를 적발했고[1], 2026년 1월에는 H200 등 일부 첨단 AI 칩에 25% 관세를 매겼습니다[2].

석유 수출 금지가 특정 국가의 에너지를 끊었던 것처럼, AI 칩 수출 통제는 특정 국가의 지능 생산 능력을 제한합니다. 그런데 세계 최첨단 AI 칩의 제조를 담당하는 곳은 TSMC, 대만의 한 회사뿐입니다. 그 핵심 시설은 중국 본토에서 불과 180km 남짓 떨어져 있습니다[3]. 데이터센터용 GPU는 주문하고도 오래 기다려야 하는 때가 많아, 몇 해치 물량을 미리 확보해 두려는 조직도 있지요. 1973년에 석유가 전략 자원이 되었듯, 2026년에 AI 칩은 전략 자원이 되었습니다.

이런 상황에서 나온 개념이 소버린 AI(Sovereign AI)입니다. 한 국가가 자국의 인프라, 데이터, 인력을 활용해서 AI를 독자적으로 개발하고 운영할 수 있는 역량을 뜻합니다.

NVIDIA의 젠슨 황 CEO는 2025년 두바이 World Governments Summit에서 이렇게 말했습니다. "AI는 당신 나라의 문화, 사회의 지성, 상식, 역사를 코드화합니다. 당신은 자신의 데이터를 소유해야 합니다." 그의 핵심 개념은 AI 공장(AI Factory)입니다. 전통적인 발전소가 원료에서 전기를 만들듯, AI 공장은 전기와 데이터에서 지능을 만드는 곳이고, 이 공장을 자국에 가지고 있어야 한다는 것이 소버린 AI의 본질입니다.

2026년 3월, NVIDIA는 Palantir와 협력해 Sovereign AI Operating System을 출시했습니다. 각국 정부가 자국 내에서 AI를 구축·운영할 수 있는 통합 플랫폼입니다[4]. McKinsey는 소버린 AI 시장이 2030년까지 5,000억~6,000억 달러 규모로 성장할 것으로 전망합니다[5].

발전소에 해당하는 파운데이션 모델을 자국에서 만들 수 있느냐, 아니면 미국에서 수입해서 쓰느냐. 이것이 지능 에너지 시대의 새로운 분단선이 되고 있습니다. 나라마다 대응 방식은 다릅니다. 하나씩 둘러볼까요.

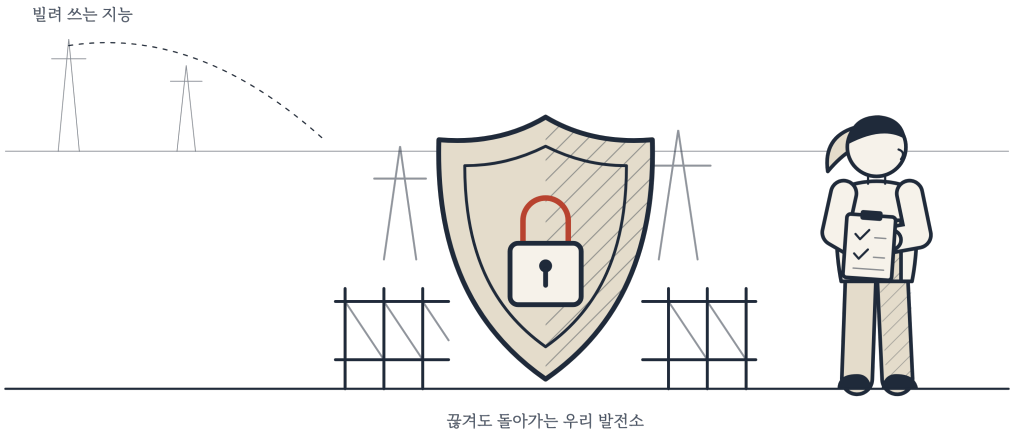


그림 2-8 빌려 쓰는 지능은 남의 사정으로 언제든 끊길 수 있습니다. 울타리 안에 우리 발전소가 있어야 바깥 전선이 흔들려도 불이 꺼지지 않지요.

한국은 경쟁으로, 유럽은 규제로

2025년 8월, 과학기술정보통신부는 독자 AI 파운데이션 모델 프로젝트의 정예팀으로 네이버클라우드, 업스테이지, SK텔레콤, NC AI, LG AI연구원 등 5개 컨소시엄을 선정했습니다.[6] 정부 지원은 현금 예산 한 덩어리가 아니라 GPU, 데이터, 인재와 평가 기반을 묶어 주는 방식이라, 금액 하나로 요약하기 어렵습니다.

네이버클라우드는 HyperCLOVA X를 앞세워 경쟁에 참여했습니다. 네이버는 2021년 1세대 HyperCLOVA를 발표하면서 GPT-3보다 6,500배 많은 한국어 데이터를 학습했다고 밝힌 바 있지요.[7] 그 뒤를 이은 모델답게 한국어와 한국의 맥락에 강점을 내세웁니다. SK텔레콤은 A.X K1 모델을, 업스테이지는 Solar 모델을 기반으로 경쟁에 나섰습니다.

2026년 1월 1차 단계평가에서 네이버클라우드와 NC AI가 탈락했고, 별도 공모로 모티프테크놀로지스가 합류해 네 팀이 2차를 겨뤘습니다. 목표는 해외 최상위 모델 대비 성능 95% 이상으로 알려져 있습니다[8]. 8월 2차 평가는 접전이었습니다. SK텔레콤, 업스테이지, LG AI연구원이 1점 안팎의 차이로 3차에 올랐고, 떨어진 모티프는 이의를 제기했지요. 남은 세 팀은 2027년 2월 평가에서 최종 두 팀으로 좁혀집니다. 10월에 정부는 한 걸음 더 나가, 글로벌 최상위 모델과 정면으로 겨룰 '프론티어 AI' 사업을 따로 띄우겠다고 밝혔습니다.[9]

GPU 26만 장 계획은 이 모델 개발팀이나 특정 임대 사업자 두 곳의 물량이 아닙니다. 정부가 국가 AI컴퓨팅센터와 국내 클라우드 사업자에 5만 장 이상을 배치하고, 삼성전자·SK그룹·현대자동차그룹이 각각 5만 장 이상, 네이버클라우드가 6만 장 이상을 구축하는 민관 합산 계획입니다.[10] 발전소를 짓는 쪽과 그 전기를 빌려 쓰는 쪽을 나눠 봐야 한국의 소버린 AI 전략이 제대로 보이지요.

이 접근 방식은 경쟁을 통해 최고의 모델을 골라내겠다는 전략입니다. 발전소를 여러 곳에 짓고, 가장 효율 좋은 것을 골라서 키우겠다는 것이지요. 다양성을 확보할 수 있다는 장점이 있지만, 자원이 분산될 수 있다는 우려도 있습니다.

바다 건너 유럽은 사정이 다릅니다. EU는 규제와 공공 투자를 함께 쓰는 쪽을 택했습니다. 2025년 초 내놓은 InvestAI 구상은 민관 합쳐 2000억 유로(약 300조 원)를 AI에 쏟아붓겠다는 계획이고, 이 중 공공 자금 200억 유로(약 30조 원)는 유럽 곳곳에 최대 몇 곳의 "AI 기가팩토리"를 세우는 데 배정됐습니다[11]. 문제는 속도입니다. 2025년 4월 사업 관심을 물었을 때는 16개 회원국 60개 부지에서 77건이 몰릴 만큼 뜨거웠지만, 정작 법적 근거는 2026년 1월에야 마련됐고 정식 공모는 2026년 7월 말에야 열려 11월에 마감됩니다. 선정은 2027년 초로 잡혀 있으니, 첫 삽은 그보다 더 뒤지요[12][13]. 법은 빠르게 만들고, 실행은 그만큼 느린 셈입니다.

그 안에서 프랑스는 자국 챔피언을 키우고 있습니다. Mistral AI는 2025년 9월 네덜란드의 반도체 장비 기업 ASML이 이끈 Series C에서 17억 유로를 받으며 전략적 파트너십을 맺었습니다. 1년 뒤인 2026년 9월에는 삼성전자가 주도한 Series D로 30억 유로를 더 모았고, 기업 가치는 210억 유로를 넘었지요.[14] 반도체 장비와 메모리의 두 거인이 차례로 지분을 넣은 셈입니다. Mistral의 전략은 오픈 웨이트(Open-Weight) 모델입니다. 가중치를 공개해서 누구나 가져다 쓸 수 있게 하면서도, 기업용 서비스로 수익을 내는 구조이지요. 2025년 6월 VivaTech에서 마크롱 대통령이 직접 Mistral과 NVIDIA 간의 "역사적 동맹"을 선언할 정도로, 프랑스 정부의 지원도 강력합니다.

유럽의 다른 나라들도 움직이고 있습니다. 포르투갈은 유럽식 포르투갈어에 맞춘 AMALIA를 공개했고, 독일은 산업용 개방형 모델을 목표로 SOOFI를 지원하며, 폴란드는 공공서비스 활용을 염두에 둔 PLLuM을 운영합니다.[15] EU 차원에서는 2026년 4월 기준 19개 AI Factory와 13개 연계 거점이 운영되고 있습니다. AI Factory와 아직 공모·건설 단계인 AI Gigafactory는 서로 다른 사업입니다.[16]

자본의 중동, 동맹의 일본

사우디아라비아와 UAE는 대규모 인프라와 지역 언어 모델을 함께 추진합니다. 사우디 국부펀드 PIF가 세운 HUMAIN은 NVIDIA와 향후 5년간 최대 500MW 규모의 AI 공장을 구축하고, 수십만 개의 GPU를 단계적으로 배치한다는 계획을 발표했습니다. 첫 단계는 GB300 GPU 1만 8,000개 규모입니다.[17] HUMAIN의 아랍어 모델은 ALLAM 34B이며, "1.8조 파라미터" 모델이 아닙니다.[18] UAE의 Stargate UAE 역시 5GW 캠퍼스 전체가 이미 운영되는 것이 아니라, 1GW 클러스터를 첫 단계로 삼은 장기 건설 계획입니다.[19] 석유로 번 돈을 지능 인프라로 옮기는 장기 투자이지요.

이게 자본으로 밀어붙이는 축이라면, 같은 UAE의 TII(기술혁신연구소)는 정반대 방향에서 움직입니다. 초거대 모델 경쟁에는 끼지 않고, 작고 효율적인 Falcon 시리즈에 집중하지요. Falcon 3는 텍스트,

이미지, 비디오, 음성을 아우르는 멀티모달 모델이고, Falcon Arabic은 오픈 아랍어 LLM 리더보드 1위를 지키고 있습니다[20]. 거대한 자본과 작지만 정교한 연구, 두 축이 함께 굴러가는 셈입니다.

중동이 돈으로 속도를 낸다면, 일본은 산업 동맹형입니다. Noetra는 제조 현장과 로봇에 쓸 멀티모달·월드 파운데이션 모델을 개발하는 사업자로 2026년 6월 NEDO 사업에 선정됐습니다. 사업 기간은 2026~2030년이지만, NEDO는 우선 2026·2027년도 계약을 맺고 이후에는 해마다 단계평가로 계속 여부를 판단한다고 밝혔습니다.[21] 몇 해에 걸친 총액이 모두 확정된 지원금은 아닌 셈이지요.

민간 쪽에서는 도쿄 기반의 Sakana AI가 눈에 띕니다. 전 Google 연구자들이 설립한 이 회사는 진화적 모델 병합으로 일본어 언어 모델 EvoLLM-JP와 이미지 언어 모델 EvoVLM-JP를 만들고 공개했습니다.[22] 일본 정부의 GENIAC 사업은 이런 민간 개발 역량을 넓히는 장치입니다. 경제산업성과 NEDO가 파운데이션 모델 개발에 필요한 계산 자원을 지원하며, Sakana AI와 Preferred Networks를 비롯한 여러 기업이 차수별 사업에 참여하고 있습니다.[23]

검약 전략과 지능 빈곤, 그리고 갈라진 지도

인도는 정반대 쪽에 있습니다. 자체 최상위 모델을 밀어붙이기보다, 저렴한 공용 GPU를 깔아서 스타트업과 연구자들이 값싸게 쓰게 하는 "프루갈(frugal, 검약형)" 전략입니다. IndiaAI 미션을 통해 이미 3만 8천 대의 GPU를 공용 클라우드로 풀었고, 여기에 2만 대를 더 얹는 중입니다[24]. 그런데 정작 예산 집행은 신통치 않습니다. 최근 회계연도 예산의 실제 집행률은 40%에 그쳤고, 다음 회계연도 예산은 오히려 절반 가까이 깎였습니다[25]. 값싸게 푸는 것이 전략인지, 실행 역량 부족을 좋게 포장한 것인지는 조금 더 지켜봐야 할 것 같습니다.

라틴아메리카에서는 칠레가 주도한 Latam-GPT가 2026년 공개됐습니다. 지역의 언어와 문화적 맥락을 더 잘 반영하는 개방형 기반 모델을 만들겠다는 프로젝트입니다.[26] 여러 나라가 계산 자원을 함께 쓰는 구상도 논의되고 있지만, 이름과 참여국이 정해진 연합으로 출범한 단계는 아닙니다.

이런 나라들 뒤에는 더 어두운 그림자가 있습니다. UNDP는 전력·통신·컴퓨팅·인력의 출발 조건이 다른 상태에서 AI가 국가 간 "새로운 대분기(Next Great Divergence)"를 만들 수 있다고 경고합니다.[27] UNCTAD는 개발도상국 가운데 국가 AI 전략을 가진 나라가 3분의 1도 되지 않고, 118개국이 주요 국제 AI 거버넌스 논의에 참여하지 못하고 있다고 지적합니다.[28] 세계 AI 시장이 2033년 4.8조 달러에 이를 것이라는 전망도 있지만, 이는 현재의 경제적 가치가 아니라 미래 시장 규모 추정치입니다.[29] 전기가 보급될 때 발전소와 송전망을 갖춘 나라가 먼저 산업화했듯, 이제는 계산 자원과 모델에 접근하지 못하는 "지능 빈곤(compute poverty)"이 새로운 격차가 될 수 있습니다. 공개 모델과 공동 인프라가 그 문턱을 얼마나 낮춰 줄지가 남은 변수이지요.

여기까지 돌아본 나라들을 한 장의 지도에 놓아 볼까요? 처지를 갈라 보면 이렇습니다. 미국과 중국은 칩 설계부터 모델 개발까지 전체 스택을 자급하는 "풀스택 소버린" 전략을 취하고 있습니다. 한국, 일본, 프랑스는 파운데이션 모델과 인프라는 자체 구축하되, 칩은 수입하는 "선택적 소버린" 전략입니다. UAE와 사우디는 대규모 인프라 투자와 모델 개발을 동시에 추진하고 있고, EU와 라틴아메리카는 지역

연합을 통한 공동 대응을 하고 있습니다. 그리고 많은 개발도상국은 DeepSeek, Llama 같은 오픈웨이트 모델에 의존하고 있습니다.

국가별로 어떤 소버린 AI 모델이 개발되고 있는지 최신 현황을 확인하려면, 저자가 운영하는 LLM/FM Benchmark and SOTA Dashboard의 Sovereign AI 섹션을 참고할 수 있습니다.[30] 모델 공개와 개발 상태가 자주 바뀌므로, 이 장의 국가별 설명과 함께 갱신되는 목록을 확인하는 편이 좋습니다.

NVIDIA의 젠슨 황은 2026년 3월 GTC 무대에서 세계 지도 한 장을 띄웠습니다. 이미 자국의 AI 팩토리를 짓고 있거나 짓기로 한 수십 개국이 표시된 지도였습니다[31]. 어느 나라도 지능 생산을 남의 손에 통째로 맡기고 싶어 하지는 않는다는 뜻이지요.

자기 발전소가 있는 나라와 전기를 빌려 쓰는 나라의 차이는 평소에는 잘 보이지 않습니다. 독자 모델을 가진 나라는 자국의 언어와 법률을 담은 AI를 운영하고, 공급이 흔들릴 때도 서비스를 이어 갈 여지가 있지요. 빌려 쓰는 나라는 공급자의 결정에 지능의 흐름을 맡기게 됩니다. 1973년 겨울, 주유소 앞에 늘어선 차들의 줄을 본 뒤에야 사람들은 그 차이를 실감했습니다. 지능의 주유소 앞에 줄을 서기 전에 알아챌 수 있을까요?



---

1. "U.S. Authorities Shut Down Major China-Linked AI Tech Smuggling Network" (U.S. Department of Justice, 2025.12.9, https://www.justice.gov/opa/pr/us-authorities-shut-down-major-china-linked-ai-tech-smuggling-network) — 2024.10~2025.5 기간 NVIDIA H100·H200 GPU 총 1억 6,081만 5,000달러(약 1.6억 달러) 규모를 중국·홍콩 등지로 밀수출한 네트워크 적발 확인.

2. "Fact Sheet: President Donald J. Trump Takes Action on Certain Advanced Computing Chips to Protect America's Economic and National Security" (The White House, 2026.1.14, https://www.whitehouse.gov/fact-sheets/2026/01/fact-sheet-president-donald-j-trump-takes-action-on-certain-advanced-computing-chips-to-protect-americas-economic-and-national-security/) — 2026.1.14 발효, Section 232에 근거해 NVIDIA H200·AMD MI325X 등 첨단 AI 칩 일부 품목에 25% 관세 부과 확인.

3. "Taiwan Strait" (Wikipedia, https://en.wikipedia.org/wiki/Taiwan_Strait) — 대만해협은 평균 약 180km 폭(최소 약 126km~최대 약 410km)으로, 대만 본섬과 중국 본토 사이 거리가 이 범위 내에 있음을 확인. 본문의 "180km"는 TSMC 핵심 생산기지(신주 등)와 중국 본토 간 대략적 거리를 나타내는 통용 수치.

4. "Palantir and NVIDIA Team to Deliver Sovereign AI Operating System Reference Architecture" (Palantir Investor Relations, 2026.3.12, https://investors.palantir.com/news-details/2026/Palantir-and-NVIDIA-Team-to-Deliver-Sovereign-AI-Operating-System-Reference-Architecture/) — 2026.3.12 발표, 각국 정부·기관이 자국 인프라 내에서 AI를 구축·운영할 수 있는 통합 레퍼런스 아키텍처(AIOS-RA) 확인.

5. "Sovereign AI: Building ecosystems for strategic resilience and impact" (McKinsey & Company, https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/sovereign-ai-building-ecosystems-for-strategic-resilience-and-impact) — 2030년까지 AI 지출의 30~40%가 소버린 요건의 영향을 받아 전 세계적으로 약 5,000억~6,000억 달러 규모 시장을 형성할 것으로 전망 확인.

6. MSIT Selects Five Elite Teams for the Sovereign AI Foundation Model Project (과학기술정보통신부, 2025.8, https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1152&sCode=eng); Proprietary AI Foundation Model Project Enters Full-Scale Launch (과학기술정보통신부, 2025.6, https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1131&pageIndex=1&sCode=eng&searchOpt=ALL) — 5개 팀과 GPU·데이터·인재의 최대 3년 지원 구조를 확인한다.

7. 1세대 HyperCLOVA 공개 보도자료 (NAVER Corp., 2021.5, https://www.navercorp.com/en/media/pressReleasesDetail?seq=30588) — 1세대 HyperCLOVA가 GPT-3보다 6,500배 많은 한국어 데이터를 학습했다는 회사 측 설명.

8. "한국형 소버린 AI 출발…정예 5개 팀, 독자 파운데이션 모델 공개" (딜사이트, https://dealsite.co.kr/articles/154187/068020); 독자 AI 파운데이션 모델 사업 개요 (대한민국 정책브리핑, https://www.korea.kr/briefing/policyBriefingView.do?newsId=156745273); 성능 목표 95% 관련 (셀렉트스타 블로그, https://selectstar.ai/blog/notice/korea-best-ai/)

9. "독파모 2차, 모티프 탈락…업스테이지·SKT·LG AI연구원 최종 진출" (MTN, 2026.8.18, https://news.mtn.co.kr/news-detail/2026081811065764367); "'국가대표 AI' 투트랙으로…독파모 살리고 '프론티어 AI' 띄운다" (MTN, 2026.10.5, https://news.mtn.co.kr/news-detail/2026100320033298930) — 2차 평가 결과(SKT 70.6, 업스테이지 69.9, LG 69.0, 모티프 65.8), 2027.2 최종 2팀 선정 계획, '프론티어 AI' 사업 신설.

10. "NVIDIA, South Korea Government and Industrial Giants Build AI Infrastructure and Ecosystem to Fuel Korea Innovation, Industries and Jobs" (NVIDIA, 2025.10.30, https://nvidianews.nvidia.com/news/south-korea-ai-infrastructure) - 정부 5만 장 이상, 삼성전자·SK그룹·현대자동차그룹 각 5만 장 이상, 네이버클라우드 6만 장 이상으로 합계 26만 장 이상의 배치 구성을 명시; "대통령실 '대한민국, APEC 계기 AI 변방에서 AI 격전지로 전환'" (대한민국 정책브리핑, 2025.11.3, https://www.korea.kr/news/policyNewsView.do?newsId=148953845) - 엔비디아 최신 GPU 26만 장 우선 공급 계획 확인.

11. "EU to mobilise €200bn for AI investment" (Euronews, 2025.2.11, https://www.euronews.com/next/2025/02/11/eu-to-mobilise-200-billion-for-ai-investment); "AI Factories" (European Commission Digital Strategy, https://digital-strategy.ec.europa.eu/en/policies/ai-factories)

12. "AI Gigafactories" (European Commission, https://commission.europa.eu/topics/competitiveness/competitiveness-coordination-tool-projects/ai-gigafactories_en) — 기가팩토리 공식 공모는 2026년 7월 30일 열려 11월 12일 마감(2026년 10월 기준 착공 전 단계); "EuroHPC JU Launches Call for Proposals to Strengthen the European AI Ecosystem" (EuroHPC JU, 2026.4.28, https://www.eurohpc-ju.europa.eu/eurohpc-ju-launches-call-proposals-strengthen-european-ai-ecosystem-2026-04-28_en) — 이 공모는 기가팩토리가 아닌 기존 AI 팩토리 간 조정 사업임을 확인

13. "EuroHPC Joint Undertaking launches AI Gigafactories call" (EuroHPC JU, 2026.7.30, https://eurohpc-ju.europa.eu/eurohpc-joint-undertaking-launches-ai-gigafactories-call-2026-07-30_en) — 공모 마감 11월 12일, 최대 7곳 선정, 2027년 초 선정 뒤 18개월 안에 가동.

14. "Orrick Advises Mistral in its €3B Series D Funding Round at €21B Post-Money Valuation" (Orrick, 2026.9, https://www.orrick.com/en/News/2026/09/Orrick-Advises-Mistral-in-its-3B-Series-D-at-21B-Post-Money-Valuation) — 2026.9.8 마감, 삼성전자 주도·EQT Scaleup Europe Fund·PSG 공동 주도. 2025.9 ASML 주도 라운드는 Series C.

15. Modelo de Linguagem em Grande Escala para a língua portuguesa (포르투갈 정부, https://portugal.gov.pt/pt/gc24/comunicacao/noticias/modelo-de-linguagem-em-grande-escala-para-a-lingua-portuguesa); SOOFI project (독일 DFKI, https://www.dfki.de/web/forschung/projekte-publikationen/projekt/soofi); Polska buduje własną sztuczną inteligencję — PLLuM (폴란드 디지털부, 2025.2.24, https://www.gov.pl/web/cyfryzacja/polska-buduje-wlasna-sztuczna-inteligencje-%2Dpllum-gotowy-do-dzialania)

16. AI Factories (European Commission, 2026.4, https://digital-strategy.ec.europa.eu/en/policies/ai-factories); AI Gigafactories (European Commission, https://commission.europa.eu/topics/competitiveness/competitiveness-coordination-tool-projects/ai-gigafactories_en) — 19개 AI Factory·13개 연계 거점 현황과 Gigafactory의 별도 공모 일정을 구분한다.

17. HUMAIN and NVIDIA Announce Strategic Partnership to Build AI Factories of the Future in Saudi Arabia (NVIDIA, 2025.5.13, https://nvidianews.nvidia.com/news/humain-and-nvidia-announce-strategic-partnership-to-build-ai-factories-of-the-future-in-saudi-arabia) — 최대 500MW, 향후 5년간 수십만 GPU, 첫 단계 1만 8,000개 계획.

18. HUMAIN IQ — ALLAM 34B (HUMAIN, https://www.humain.com/en/humain-iq) — HUMAIN이 공개한 모델 명칭과 규모.

19. Global Tech Alliance Launches Stargate UAE (G42, https://www.g42.ai/resources/news/global-tech-alliance-launches-stargate-uae) — 5GW 캠퍼스 구상과 첫 1GW 클러스터를 구분한다.

20. "UAE's TII challenges big tech dominance with open-source Falcon AI models" (Computer Weekly, https://www.computerweekly.com/news/366638759/UAEs-TII-challenges-big-tech-dominance-with-open-source-Falcon-AI-models)

21. **AI口ボット・フィジカルAIを見据えたマルチモーダル基盤モデル開発事業の実施体制の決定について** (NEDO, 2026.6.30, https://www.nedo.go.jp/koubo/CD3_100431.html) — Noetra 선정, 2026~2030년 사업 기간, 연도별 계약과 단계평가 조건.

22. "진화적 알고리즘에 의한 기반 모델 구축" (Sakana AI, 2024.3.21, https://sakana.ai/evolutionary-model-merge-jp/) — EvoLLM-JP와 EvoVLM-JP의 개발 방식과 공개 사실을 확인한다.

23. "GENIAC" (일본 경제산업성, https://www.meti.go.jp/english/policy/mono_info_service/geniac/); "Selection of 16 New Projects to Support the Development of AI Models under the GENIAC Computing Resource Provision Support Project (Cycle 4)" (일본 경제산업성, 2026.6.4, https://www.meti.go.jp/english/press/2026/0604_001.html) — 계산 자원 지원 목적과 차수별 선정 구조를 확인한다.

24. "India to add 20,000 GPUs as AI Mission 2.0 expands compute and chip push" (EE Times, https://www.eetimes.com/india-to-add-20000-gpus-as-ai-mission-2-0-expands-compute-and-chip-push/); IndiaAI 미션 GPU 조달 (인도 정부 PIB, https://www.pib.gov.in/PressReleasePage.aspx?PRID=2132817) — 공용 GPU 3만 8천 대 + 2만 대 추가 조달.

25. "Budget 2026: Less Than 50% Of IndiaAI Mission Outlay Used In FY26" (Inc42, https://inc42.com/buzz/budget-2026-less-than-50-of-indiaai-mission-outlay-used-in-fy26/) — 회계연도 예산 집행률 약 40%,

차기 배정 절반 수준 삭감.

26. We launch Latam-GPT (칠레 정부, 2026.2.10, https://www.gob.cl/en/news/we-launch-latam-gpt-the-first-artificial-intelligence-created-for-latin-america-and-the-caribbean/) — 라틴아메리카·카리브 지역의 언어와 맥락을 반영하는 개방형 모델의 공개 목적을 설명한다.

27. "The Next Great Divergence" (UNDP Asia and the Pacific, 2025, https://www.undp.org/asia-pacific/publications/next-great-divergence) — AI 시대의 국가 간 출발 조건과 역량 격차를 분석한다.

28. "AI's $4.8 trillion future" (UN Trade and Development, 2025, https://unctad.org/press-material/ais-48-trillion-future-un-trade-and-development-alerts-divides-urges-action) — 국가 AI 전략과 글로벌 거버넌스 참여 격차를 제시한다.

29. "AI market projected to hit $4.8 trillion by 2033" (UN Trade and Development, https://unctad.org/news/ai-market-projected-hit-48-trillion-2033-emerging-dominant-frontier-technology) — 4.8조 달러를 현재 가치가 아닌 2033년 시장 전망으로 제시한다.

30. LLM/FM Benchmark and SOTA Dashboard — Sovereign AI (전종홍, https://hollobit.github.io/SOTA/#sovereign) — 국가별 소버린 AI 모델과 개발 현황을 갱신해 제공하는 참고용 목록이다.

31. "Every Country Needs Sovereign AI" (AI Realist, https://www.airealist.ai/p/every-country-needs-sovereign-ai)

지능을 만드는 데도 전기가 필요하다: AI 에너지 전쟁

AI 모델을 학습하고 운영하려면 전기 에너지가 필요합니다. GPU가 연산 장치라면 전기는 그 장치를 돌리는 물리적 투입물입니다. 여기서부터는 '지능 에너지'라는 비유보다 데이터센터의 실제 전력과 계통 문제를 보겠습니다.

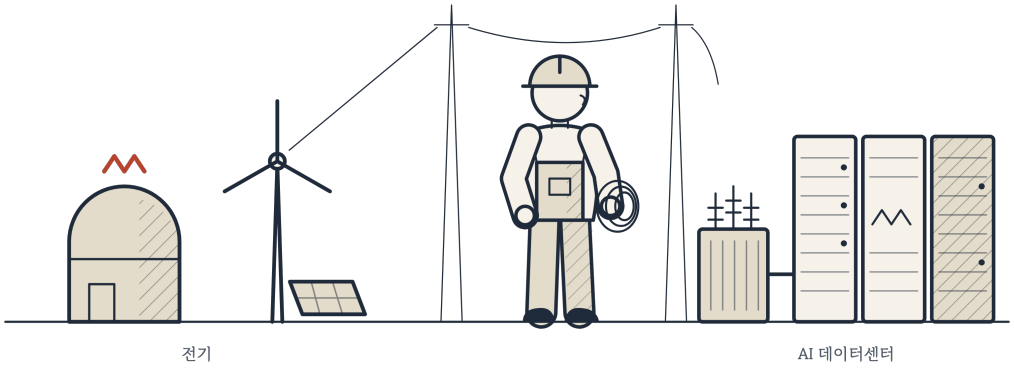


그림 2-9 원전과 풍력이 만든 전기가 전선을 타고 데이터센터에 닿아야 비로소 첫 토큰이 나옵니다. 지능 발전소도 결국 전기를 먹고 돌지요.

나라별 대응을 '민간이 하느냐, 국가가 하느냐'로만 가르면 실제 모습이 흐려집니다. 미국에도 세제 지원과 규제가 있고, 중국에도 민간 플랫폼 기업이 있으며, 한국도 정부 계획과 기업 투자가 섞여 있지요. 그래서 이 장에서는 몇 가지를 함께 물어보려 합니다. 누가 돈을 대고 손실을 떠안는가, 전기는 정말 끌어왔는가, 주문한 칩은 설치돼 돌아가는가, 허가와 주민 부담은 어떻게 풀었는가, 그리고 지금 발표·계약·건설·가동 가운데 어느 단계인가.

스타게이트와 치솟는 전력 수요

2025년 1월 21일, 백악관에서 스타게이트(Stargate)가 발표됐습니다. 새 합작회사가 향후 4년간 미국 AI 인프라에 최대 5,000억 달러를 투자하겠다는 약정이며, 발표 당시 즉시 집행을 시작한다고 밝힌 금액은 1,000억 달러였습니다.[1]

참여 기업들의 면면을 보면 이 프로젝트의 성격이 보입니다. 초기 지분 투자자는 SoftBank, OpenAI, Oracle, UAE 투자사 MGX입니다. SoftBank가 재무를, OpenAI가 운영을 맡고 Arm·Microsoft·NVIDIA·Oracle·OpenAI가 초기 기술 파트너로 이름을 올렸습니다.[1] 언론에 보도된 지분율이나 NVIDIA·OpenAI의 별도 투자 구상은 스타게이트가 확정해 집행한 돈과는 다른 이야기입니다.

그런데 이 프로젝트의 계산서를 펼쳐 보면, 가장 굵은 줄은 알고리즘보다 전력 쪽에 적혀 있습니다.

텍사스 애빌린의 첫 시설에서는 2025년 중반부터 일부 훈련·추론 작업이 시작됐습니다. 같은 해 9월 OpenAI가 발표한 약 7GW와 4,000억 달러 이상은 애빌린, 새로 선정한 5개 부지, CoreWeave 사업을 합친 계획 용량과 향후 3년 투자 계획이지, 이미 다 지어 돌리는 규모가 아닙니다.[2] 애초 약정 목표도 2029년까지 10GW·5,000억 달러를 확보하겠다는 계획이므로, 운영 중·건설 중·부지 선정 단계를 나눠 추적해야 합니다.

스타게이트는 데이터센터와 전력을 끌어오는 일이 모델 개발만큼 AI 경쟁의 한복판에 들어왔음을 보여주지요. 발표된 총액이 모두 쓰였거나 모두 전력 설비에 들어간 것은 물론 아닙니다.

스타게이트 하나만의 사정일까요? 지능 에너지 시대의 전력 수요가 얼마나 빠르게 늘고 있는지, 숫자를 보면 놀랍습니다.

앞서 변환소 장에서 봤듯, IEA(국제에너지기구)는 전 세계 데이터센터의 전력 소비가 2024년 415TWh에서 2030년 기준 시나리오 945TWh로 두 배 이상 늘 것으로 봅니다.[3] 이 값에는 AI뿐 아니라 일반 클라우드와 저장·네트워크 수요도 포함됩니다.

IEA의 기준 시나리오에서 AI에 주로 쓰이는 가속 서버의 전력 소비는 연평균 약 30% 늘고, 기존 서버는 약 9% 늘어납니다. 가속 서버가 글로벌 데이터센터 전력 소비 순증가분의 거의 절반을 차지할 것으로 예상되지요.[3]

개별 모델의 학습·추론 전력은 사업자가 배포 사양을 공개하지 않는 경우가 많습니다. 응답당 추정치도 모델, 길이, 배치 처리와 하드웨어에 따라 크게 달라지지요. 특정 응답의 추정치를 전체 요청량에 곱해 실제 운영 전력처럼 제시하는 대신, IEA가 집계한 데이터센터 전체 수요를 기준으로 보는 이유입니다.

미국 에너지부 산하 로런스버클리연구소의 2026년 갱신 보고서는 데이터센터가 2030년 미국 전력의 9.5~15.3%를 사용할 수 있다고 전망합니다.[4] 범위가 넓다는 사실 자체가 서버 출하와 효율, 이용률을 둘러싼 불확실성을 보여줍니다.

원자력으로 가는 미국, 서쪽으로 가는 중국

이 폭발적인 전력 수요를 어떻게 충당할 것인가. 미국의 빅테크 기업들이 내린 답은 의외의 것이었습니다. 원자력입니다.

Microsoft는 Constellation Energy와 20년 전력 구매 계약을 맺고, 2019년 경제성을 이유로 멈춘 스리마일 아일랜드 1호기를 Crane Clean Energy Center라는 이름으로 재가동하는 사업을 지원합니다. 목표 용량은 약 835MW, 공식 발표 당시 목표 가동 시점은 2028년입니다.[5] 1979년 사고가 난 곳은 별도의 2호기였다는 점도 구분해야 합니다.

Google은 Kairos Power와 여러 차례에 걸쳐 최대 500MW의 차세대 원전 전력을 확보하는 장기 협력을 추진합니다. 첫 50MW 사업은 TVA와의 전력 구매 계약을 통해 2030년 공급을 목표로 하지만, 아직 상업 가동 전 단계입니다.[6]

Amazon은 X-energy에 5억 달러를 투자하고 2039년까지 미국에 5GW 이상의 신규 원자력 용량을 마련한다는 목표를 지원합니다. 워싱턴주의 첫 사업은 초기 320MW, 확장 시 960MW 구상이며 역시 설계·인허가·건설을 거쳐야 하는 장기 계획입니다.[7]

Google, Amazon, Meta는 2050년까지 세계 원자력 설비를 2020년 대비 최소 세 배로 늘리자는 업계 서약에도 참여했습니다.[8] 서약은 구속력 있는 건설 계약이 아니라 장기 방향에 대한 지지 선언입니다.

하지만 원자력은 장기 전략입니다. 당장의 전력 수요를 메우려고 일부 사업자는 계통 전력과 현장 가스 발전, 배터리를 함께 쓰는 방식을 택합니다. 배치 속도는 빠를 수 있지만 대기오염 허가와 연료 가격, 탄소배출 비용이 따라옵니다.

이른바 "자가 발전(Bring-Your-Own-Power)"은 전력망에 영구히 기대지 않는다는 뜻이라기보다, 계통 연결을 기다리는 동안 현장 발전과 저장장치를 교량 전원으로 쓰는 경우가 많습니다. xAI 멤피스 시설도 초기에는 제한된 계통 전력과 가스 터빈을 함께 사용했고, 공공 전력회사 MLGW는 이후 150MW 계통 연결과 추가 증설 절차를 공개했습니다.[9]

바다 건너 중국은 국가 차원의 입지 조정으로 대응합니다. 2022년 2월 전면 가동을 선언한 "동수서산(東數西算)"은 8개 국가 컴퓨팅 허브와 10개 데이터센터 클러스터를 연결하는 사업입니다.[10] 글자 그대로 "동쪽의 데이터를 서쪽에서 계산한다"는 뜻이지요.

구조는 명쾌합니다. 토지와 에너지가 빠듯한 동부의 대규모 연산 수요 일부를 재생에너지와 공간 여건이 나은 서부 허브로 유도합니다. 물론 지연시간에 민감한 서비스까지 모두 서쪽으로 옮기지는 않지요. 어떤 작업을 어디서 처리할지는 네트워크와 데이터 규정에 따라 달라집니다.

지역별 전기요금과 보조금은 시기와 계약에 따라 달라집니다. 일부 지방정부가 전력비 지원이나 국산 장비 우대책을 내놓았지만, 이를 중국 전체의 단일 요금으로 일반화하기는 어렵습니다.

중국 국가발전개혁위원회는 허브 안 대형 데이터센터의 집적도와 이용률, 재생에너지 사용을 높이고 데이터센터와 전력망을 함께 설계하겠다는 방향을 제시했습니다.[10] 계획이 실제 효율로 이어졌는지는 지역별 가동률과 전력 구성을 봐야 알 수 있습니다. 국가 계획이 있어도 전력 병목과 과잉 투자는 여전히 숙제로 남지요.

칩보다 전기, 그리고 한국의 자리

AI 경쟁의 병목이 바뀌고 있습니다. 지금까지는 "누가 더 좋은 칩을 가졌는가"가 핵심이었습니다. 미국이 NVIDIA의 첨단 GPU 수출을 통제하면서 중국을 압박한 것이 대표적이지요. 하지만 이제 새로운 질문이 떠오르고 있습니다. "누가 더 많은 전력을 공급할 수 있는가."

칩만으로는 부족합니다. 칩을 돌릴 전기가 있어야 합니다.

미국은 첨단 반도체와 민간 자본에서 강점을 보이지만 계통 연결과 지역 인허가가 병목이 될 수 있습니다. 중국은 국가 계획으로 전력망과 데이터센터 입지를 함께 조정하지만 지역별 가동률과 투자 효율, 최신 칩 접근이라는 제약이 있습니다. 두 나라 모두 민간과 정부가 얽힌 혼합형이며, 강점과 비용이 서로 다릅니다.

IEA는 2030년까지 데이터센터 전력 소비 증가가 미국과 중국에 크게 집중되고, 유럽과 일본도 각각 약 45TWh와 15TWh 늘어날 것으로 봅니다.[3] 전기를 덜 쓰는 나라도 효율적인 모델과 특화 데이터, 해외 인프라를 엮어 길을 낼 수 있습니다. 그래도 대규모 인프라가 두 나라에 몰릴수록, 나머지 나라가 공급자를 고르고 값을 흥정할 여지는 줄어들지요.

그러니 어느 나라가 이겼는지는 한 문장으로 가리기 어렵습니다. 실제로 돌아가는 용량과 전기요금, 환경 비용을 같은 잣대로 계속 지켜봐야 하지요.

그렇다면 한국은 어떻게 해야 할까요? 지능 에너지 시대에 전기 에너지가 오히려 더 중요해졌습니다. AI를 돌리려면 전기가 필요하고, 전기를 확보하려면 국가 전략이 필요합니다.

한국의 현실을 직시해야 합니다. 우리는 미국처럼 광활한 땅에 천연가스 발전소를 지을 수도 없고, 중국처럼 서부의 재생에너지를 동부로 보낼 수도 없습니다. 전력 예비율도 넉넉하지 않지요. 그렇다고 손을 놓고 있을 수는 없습니다.

그래서 AIDC 입지를 고를 때는 "어느 도시가 좋다"는 순위표보다, 먼저 걸러 낼 위험을 찾는 체크리스트가 쓸모 있습니다. 저자가 만든 Grid Siting Lab도 이런 용도의 공개 데이터 기반 탐색 도구입니다. 지도 위의 점수는 한전의 계통연계 검토나 현장 실사를 대신하지 않습니다. 발전소가 가깝다고 해서 그 전력을 실제로 끌어올 송전망 여유와 접속 승인까지 확인된 것도 아니지요.[11]

점검 항목	물어야 할 질문
계통연계·수전 여유	필요한 MW를 언제, 어느 전압급으로, 어떤 보강 공사 뒤 받을 수 있는가
물	냉각 방식에 맞는 취수·중수·해수 경로가 있고 지역 물 갈등을 키우지 않는가
통신 지연·이중화	학습용과 추론용 워크로드에 맞는 지연시간, 서로 다른 광 경로, 장애 우회가 있는가
재해	홍수, 해일, 지진, 산불, 폭염에 대해 20~30년 운영 리스크를 견딜 수 있는가
토지·인허가	산업용지, 환경 인허가, 송전선로 경과지, 장비 반입 동선이 맞는가
지역 사회 계약	전력망 보강 비용, 물 사용, 소음, 세수와 일자리의 배분을 주민에게 설명할 수 있는가

한국수력원자력과 두산에너빌리티는 Amazon·X-energy의 미국 SMR 사업 협력에 참여하고 있습니다.[7] 해외 사업에서 쌓은 경험을 국내 AI 데이터센터 전력으로 바꾸려면, 인허가와 입지, 계통과 경제성 검토를 처음부터 따로 거쳐야 합니다.

칩을 사 오는 일만큼, 그 칩을 돌릴 전기를 구하는 일이 어려워졌습니다. 이어지는 세 장에서는 미국의 자금 조달과 건설 위험, 중국의 전력망·컴퓨팅망과 가동 효율, 한국 계획이 어디까지 확정됐는지를 차례로 들여다보겠습니다. 발표가 실제 가동으로 이어지기까지 누가 비용을 치르는지가 세 나라에서 어떻게 다른지 눈여겨보세요.

1. Announcing The Stargate Project (OpenAI, 2025.1.21, https://openai.com/index/announcing-the-stargate-project/) — 4년간 최대 5,000억 달러, 초기 1,000억 달러 배치, 지분 투자자와 역할을 명시한다.

2. OpenAI, Oracle, and SoftBank expand Stargate with five new AI data center sites (OpenAI, 2025.9.23, https://openai.com/index/five-new-stargate-sites/) — 약 7GW는 planned capacity, 4,000억 달러 이상은 향후 3년 투자 계획이며 애빌린에서 초기 작업이 시작됐다고 설명한다.

3. Energy demand from AI (IEA, 2025, https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai) — 415TWh·945TWh 기준 시나리오와 가속 서버·기존 서버의 증가율을 제시한다.

4. United States Data Center Energy Usage Report: 2025 Update (Lawrence Berkeley National Laboratory, 2026.6, https://eta-publications.lbl.gov/publications/united-states-data-center-energy-2025) — 2030년 미국 데이터센터 전력 비중을 9.5~15.3% 범위로 전망한다.

5. Constellation to Launch Crane Clean Energy Center (Constellation Energy, 2024.9.20, https://www.constellationenergy.com/news/2024/Constellation-to-Launch-Crane-Clean-Energy-Center-Restoring-Jobs-and-Carbon-Free-Power-to-The-Grid.html) — 20년 계약, 약 835MW, 2028년 목표를 명시한다.

6. Our first advanced nuclear reactor project with Kairos Power and Tennessee Valley Authority (Google, 2025.8.18, https://blog.google/company-news/outreach-and-initiatives/sustainability/google-first-advanced-nuclear-reactor-project-with-kairos-power-and-tennessee-valley-authority/) — 첫 50MW·2030년 목표와 장기 최대 500MW 협력을 구분한다.

7. Amazon among world's leading corporate purchasers of carbon-free energy (Amazon, https://www.aboutamazon.com/news/sustainability/amazon-leading-corporate-carbon-free-energy-purchaser) — X-energy 투자, 2039년 5GW 이상 목표, 첫 사업의 320~960MW 구상을 설명한다.

8. "Amazon, Google, Meta and Dow back goal to triple nuclear capacity" (World Nuclear News, 2025.3, https://world-nuclear-news.org/articles/amazon-google-meta-and-dow-back-tripling-nuclear-goal) — 참여 기업과 비구속적 지지 선언의 내용을 확인할 수 있다.

9. 2025 xAI Update (Memphis Light, Gas and Water, https://www.mlgw.com/images/content/files/pdf/new/xAI%202025%20Update.pdf) — 150MW 계통 공급, 추가 150MW 요청과 변전소·송전망 절차를 구분한다.

10. 国家发展改革委高技术司负责同志就实施“东数西算”工程答记者问 (중국 국가발전개혁위원회, 2022.2.17, https://www.ndrc.gov.cn/xxgk/jd/jd/202202/t20220217_1315795.html) — 8개 허브·10개 클러스터, 동서 입지 조정, 이용률·재생에너지·전력망 연계 방향을 설명한다.

11. GRID SITING LAB · AIDC (저자 제작 탐색 도구, https://hollobit.github.io/grid-siting-lab/) — OpenInfraMap·OpenStreetMap 등 공개 데이터를 활용한 대한민국 AIDC 입지 1차 스크리닝 도구다. 도시 순위나 최종 입지 판정이 아니며, 실제 계통연계·용수·통신·재해·토지·지역 수용성 검토가 필요하다고 명시한다.

미국, 지능 발전소를 짓다

미국의 AI 인프라는 민간 기업이 앞에 서지만, 자금 조달과 전력망 접속, 세제와 인허가에는 정부와 금융기관이 깊이 관여합니다. 그래서 투자 총액보다 먼저 물어야 할 것이 있습니다. 누가 위험을 지고, 발표된 돈이 실제 공사로 이어지는가 하는 점이지요.

합작 발전소와 빅테크의 설비 경쟁

2025년 1월 발표된 스타게이트(Stargate)는 여러 기업과 투자자가 역할을 나눈 합작 구조입니다.

공식 발표에서 확인되는 초기 지분 투자자는 SoftBank, OpenAI, Oracle, UAE 투자사 MGX입니다. SoftBank가 재무를, OpenAI가 운영을 맡고 Arm·Microsoft·NVIDIA·Oracle·OpenAI가 초기 기술 파트너로 참여합니다.[1] 언론이 보도한 세부 지분율이나 NVIDIA·OpenAI의 별도 협력 구상이 확정 집행액과 다르다는 점은 앞 장에서 짚었지요.

규모는 앞 장에서 이미 짚었습니다. 2025년 9월 발표된 약 7GW는 애빌린과 새로 선정한 부지, CoreWeave 사업을 합친 계획 용량이고, 4,000억 달러 이상은 향후 3년의 투자 계획입니다.[2] 애빌린에서 일부 작업이 시작됐다고 전체가 완공돼 돌아가는 것은 아니지요. 이번 장에서는 발표된 돈이 계약과 공사로 바뀌는 과정을 따라가 보겠습니다.

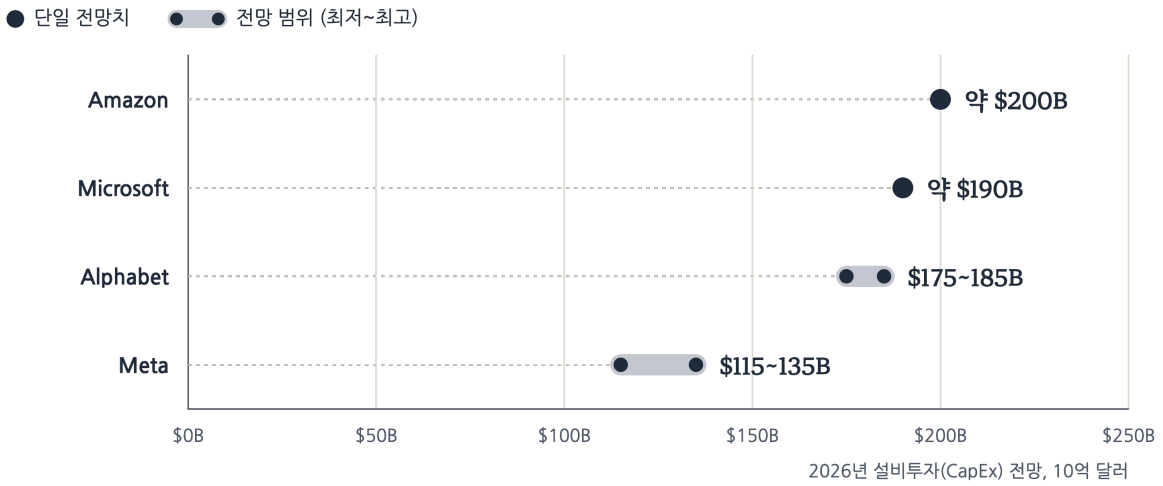
숫자만 보면 순항하는 것 같지만, 속을 들여다보면 그렇지도 않습니다. OpenAI·Oracle·SoftBank 세 파트너 사이에 책임 분담을 둘러싼 갈등이 있었고, 이 때문에 착공이 늦어진 부지들이 있습니다. OpenAI가 독자적으로 자금을 조달하려다 대출 기관들의 신뢰를 얻지 못해, 결국 Oracle과의 양자 구도로 무게중심이 옮겨갔다는 보도도 있습니다[3]. 5,000억 달러짜리 발전소를 짓는 일은, 돈이 아무리 많아도 파트너들 사이의 조율 없이는 쉽지 않다는 뜻이지요.

스타게이트는 빙산의 일각입니다. 그 밖에서도 빅테크 4사는 각자 천문학적인 돈을 쏟아붓고 있습니다.

Amazon·Microsoft·Alphabet·Meta 네 회사의 2026년 설비투자 합계는 시장 집계마다 크게 다릅니다. 회계연도와 리스 포함 여부가 다르고, 물류·사무시설·일반 클라우드 투자까지 들어가기 때문입니다.[4] 그러니 이 돈을 모두 AI 데이터센터와 GPU에 쓰는 돈으로 읽으면 실제보다 부풀려집니다.

빅테크 설비투자

회사 계획·가이던스와 기관 추정이 섞인 전망치 · AI 전용 지출과 전체 CapEx는 같지 않습니다



네 회사 합계, 기관 추정 $630~725B

집계마다 회계연도·리스 포함 여부·비AI 투자 범위가 다릅니다

출처: Value Add VC; Tom's Hardware의 2026년 회사별 전망·기관 추정 정리. 단위 10억 달러($B).

그림 2-10 회사마다 2026년 설비투자 전망을 최저에서 최고까지의 폭으로 그렸습니다. 폭이 넓을수록 집계마다 숫자가 크게 갈린다는 뜻이고, 네 회사를 합쳐도 6,300억에서 7,250억 달러 사이를 오가지요.

일부 시장 전망은 2027년 빅테크 설비투자가 1조 달러에 이를 가능성도 제시합니다.[5] 기업이 확정한 돈이 아니라 애널리스트의 전망이지요. 숫자 하나보다는, 전력과 컴퓨팅 인프라에 큰돈이 여러 해 이어서 들어갈 수 있다는 방향을 읽어 두면 됩니다.

xAI는 이 대열에서 조금 비껴서 있습니다. 규모는 크지만 독립적으로 움직인다는 점에서요. 멤피스 공공전력회사 MLGW가 공식적으로 확인한 것은 150MW 계통 공급과 추가 150MW 요청, 이를 위한 변전소·송전망 절차입니다.[6] 민간 자료에 떠도는 GPU 수량과 전체 전력 규모, 구매비 추정은 기준이 제각각이라 공식 확인된 숫자로 보기 어렵습니다.

이 숫자들을 발전소 비유로 정리하면 이렇습니다. 빅테크는 지금 수요가 확실히 정해지기도 전에, 미래에 필요할 발전 용량을 선제적으로 짓고 있습니다. 전력 회사가 20년 뒤의 인구 증가를 내다보고 미리 발전소를 짓는 것과 비슷하지요. 차이라면 그 규모가 웬만한 나라의 1년 예산에 맞먹는다는 점입니다.

땅보다 전기, 사업 투자에서 국가 인프라로

이렇게 돈을 쏟아붓다 보니 투자 방식도 달라지고 있습니다. 불과 몇 년 전까지 데이터센터 투자는 "땅부터 잡고 보자"는 식이었습니다. 넓고 싼 부지를 먼저 확보한 다음, 나중에 전력과 냉각을 어떻게든 끌어오는 순서였지요.

지금은 순서가 바뀌었습니다. 법률회사 Ropes & Gray의 2026년 분석에 따르면, 데이터센터 개발의 가장 큰 걸림돌은 이제 자본이 아니라 전력입니다. 전력망에 연결하는 데만 최대 4년이 걸리는 경우가 있다 보니, 투자자들은 프로젝트를 단계별로 쪼개서 전력 확보가 확인될 때마다 다음 단계로 넘어가는 방식을 택하고 있습니다[7]. 땅을 먼저 보던 투자자들이 이제는 전기부터 따지는 셈입니다.

이 변화를 보여주는 사례가 Google과 Blackstone의 합작입니다. 2026년, 두 회사는 새 합작회사를 세웠습니다. Blackstone이 50억 달러(약 6.75조 원)의 지분을 투자하고, 2027년까지 500MW 규모의 데이터센터 용량을 가동한다는 계획입니다. Google이 자체 개발한 AI 칩 TPU를 돌릴 컴퓨팅 서비스를 외부 고객에게도 판매하는 구조이지요[8]. 빅테크가 혼자 짓고 혼자 감당하던 발전소에 사모펀드까지 끌어들여, 자본과 위험을 나눠 지기 시작했다는 신호이기도 합니다.

이 모든 투자를 지켜보면 한 가지 질문이 떠오릅니다. 이게 그냥 기업들의 사업 투자일까요, 아니면 국가 인프라일까요.

2026년 4월, 세계경제포럼(WEF)은 이렇게 정리했습니다. AI 인프라는 이제 단순한 데이터 저장소가 아니라, 전력망이나 항만처럼 국가 차원에서 다뤄야 할 전략 인프라라는 것입니다[9]. 실제로 미국 에너지부(DOE)는 아이다호 국립연구소, 오크리지, 패듀카, 서배너강 부지 등 연방 소유의 땅을 AI 데이터센터와 에너지 인프라 개발에 내주기로 했습니다. 에너지부 장관 크리스 라이트는 이를 "차세대 맨해튼 프로젝트를 가속하는 대담한 조치"라고 표현했습니다[10].

McKinsey는 2030년까지 전 세계 데이터센터에 약 7조 달러(약 9,450조 원)의 자본이 투입되어야 한다고 추산합니다[11]. 제1부에서 봤듯 젠슨 황은 이 흐름을 "다섯 겹의 케이크"에 비유했지요. 맨 아래 에너지부터 칩, 컴퓨팅 인프라, 클라우드, 그 위의 모델과 앱까지 다섯 층이 동시에 지어져야 한다는 그림입니다. 그는 이것을 "인류 역사상 최대 규모의 인프라 구축"이라고 불렀습니다[12]. 제1부에서 본 샘 올트먼의 그 말—"지능의 비용, 즉 AI의 비용은 결국 에너지의 비용으로 수렴할 것"—이 나온 것도 바로 이 자리, 2025년 5월 미국 상원 청문회였습니다. 그는 앞으로 10년이 "풍부한 지능과 풍부한 에너지"의 시대가 될 것이라고 덧붙였지요[13].

청구서와 경고장

이 웅장한 그림에는 청구서가 따라옵니다. 하버드 대학교가 2026년 4월 정리한 바에 따르면, 데이터센터 인근 주민들의 전기요금이 두 배 넘게 뛴 사례가 이어지고 있습니다. 버지니아와 조지아 같은 주는 데이터센터에 세제 혜택을 몰아주느라 각각 10억 달러 넘는 세수를 포기했는데, 정작 가동 이후 남는 일자리는 시설 하나당 20~50명 수준에 그칩니다[14].

물 사용도 만만치 않습니다. 냉각탑을 돌리는 대형 데이터센터 한 곳이 하루에 가정 6,500곳이 쓰는 만큼의 물을 끌어다 쓰는 사례도 보고됩니다[15].

정치적 수사를 걷어내고 전기요금만 따져봐도 그림은 비슷합니다. 팩트체크 매체 PolitiFact가 2026년 6월 확인한 바로는, 최근 5년간 미국 가정용 전기요금은 평균 42% 올랐고 워싱턴 D.C.는 94%, 메릴랜드는 74%까지 뛰었습니다. 데이터센터가 몰린 PJM 전력망 권역에서는 2025~2026년 용량

비용만 174% 늘어, 약 93억 달러의 추가 부담이 발생했습니다[16]. 발전소를 어디에 지을지는 이제 기업의 투자 결정만이 아니라, 전기요금 고지서와 마을 공청회를 통과해야 하는 문제가 되었습니다.

청구서만 문제가 아닙니다. 이 발전소들을 누가 소유하느냐도 이미 경쟁당국의 시야에 들어와 있습니다. 지금의 AI 인프라는 소수의 하이퍼스케일러와, 그 위에 컴퓨트를 사실상 독점 공급하는 엔비디아, 그리고 130억 달러를 주고받은 마이크로소프트-오픈AI 같은 몇몇 축으로 빠르게 집중되고 있습니다. 미국 경쟁당국도 이걸 손 놓고 보지는 않았습니다. 연방거래위원회(FTC)는 2024년 1월 알파벳·아마존·앤스로픽·마이크로소프트·오픈AI의 투자·제휴 구조를 들여다보는 6(b) 조사에 착수했고, 같은 해 6월에는 법무부(DOJ)가 엔비디아를, FTC가 마이크로소프트·오픈AI를 나눠 맡아 반독점 조사를 벌이기로 했습니다[17]. 발전소를 이토록 빠르게 짓는 힘과, 그 발전소가 몇몇 손에만 쥐어지는 위험은 동전의 양면인 셈입니다.

기업 입장에서도 남 일이 아닙니다. 핵심 컴퓨트나 클라우드를 특정 벤더 한 곳에 몰아 쓰고 있다면, 그 벤더가 조사 대상에 오르거나 공급이 흔들리기 전에 멀티클라우드나 자체 설비 같은 갈아탈 길을 미리 마련해 두는 것이 좋습니다.

이 열기가 어디까지 갈지 가늠해 볼 수 있는 전망도 나와 있습니다. 「AI 2027」 시나리오로 이름을 알린 AI Futures Project는 2026년 7월의 후속 분석에서, 전 세계 AI 설비투자가 2024년 2,600억 달러에서 2028년 2조 4천억 달러까지 불어나고 그중 채무로 조달되는 몫이 23%에서 38%로 커질 것으로 내다봤습니다[18]. 같은 전망에서 2028~29년이면 AI 칩 생산이 전 세계 EUV 노광 장비 가동의 65~77%를 차지합니다. 발전소 건설 경쟁이 자본 시장과 반도체 공급망 전체를 빨아들이는 그림이지요.

국제결제은행(BIS)은 2026년 6월 연례보고서에서 이 열기에 이런 경고장을 붙였습니다. "기대했던 수익이 실망으로 돌아서면, 지금의 설비투자 붐은 갑작스러운 자금 회수로 이어져 장기적인 투자 위축으로 뒤집힐 수 있고, 그 여파는 금융 시스템 전반에 미칠 수 있다."[19] 민간 자본이 국가 인프라를 짓는 방식은 그만큼 빠르지만, 그만큼 이 경고에도 취약합니다. 정부가 예산을 편성해 발전소를 짓는 대신 빅테크와 국부펀드와 사모펀드가 돈을 대고, 정부는 부지와 규제 완화로 화답하는 이 조합이 얼마나 오래갈지는, 다음 장에서 국가가 직접 그물을 짜는 중국의 방식과 나란히 놓아 봐야 더 뚜렷해질 것입니다.



1. "Announcing The Stargate Project" (OpenAI, 2025.1.21, https://openai.com/index/announcing-the-stargate-project/) — 4년간 최대 5,000억 달러, 초기 1,000억 달러 배치, 지분 투자자와 역할을 명시한다.

2. "OpenAI, Oracle, and SoftBank expand Stargate with five new AI data center sites" (OpenAI, 2025.9.23, https://openai.com/index/five-new-stargate-sites/) — 약 7GW는 계획 용량, 4,000억 달러 이상은 향후 3년 투자 계획이며 애빌린에서 초기 작업이 시작됐다고 설명한다.

3. "Stargate's $500 billion AI infrastructure project reportedly stalls over unresolved disputes between OpenAI, Oracle, and SoftBank" (The Decoder, 2026.2.23, https://the-decoder.com/stargates-500-billion-ai-infrastructure-project-reportedly-stalls-over-unresolved-disputes-between-openai-oracle-and-softbank/)

4. "Google, Microsoft, Meta, and Amazon capex spending to hit $725 billion in 2026, up 77% from last year" (Tom's Hardware, https://www.tomshardware.com/tech-industry/big-tech/big-techs-ai-spending-plans-reach-725-billion)

5. "AI boom: Big Tech capital expenditures now seen topping $1 trillion in 2027" (CNBC, 2026.4.30, https://www.cnbc.com/2026/04/30/ai-boom-big-tech-capital-expenditures-now-seen-topping-1-trillion-in-2027-.html)

6. "2025 xAI Update" (Memphis Light, Gas and Water, https://www.mlgw.com/images/content/files/pdf/new/xAI%202025%20Update.pdf) — 150MW 계통 공급, 추가 150MW 요청과 변전소·송전망 절차를 구분한다.

7. "Data Center Investment in 2026: AI Demand, Power Constraints, and Private Equity Trends" (Ropes & Gray, https://www.ropesgray.com/en/insights/viewpoints/102mvfl/data-center-investment-in-2026-ai-demand-power-constraints-and-private-equity)

8. "Google, Blackstone form AI infrastructure joint venture for data center development" (ESG Dive, https://www.esgdive.com/news/google-blackstone-ai-infrastructure-joint-venture-data-center-de/820628/)

9. "It's time to start treating AI infrastructure as critical infrastructure" (World Economic Forum, 2026.4, https://www.weforum.org/stories/2026/04/ai-infrastructure-critical-infrastructure/)

10. "DOE Announces Site Selection for AI Data Center and Energy Infrastructure Development on Federal Land" (U.S. Department of Energy, https://www.energy.gov/articles/doe-announces-site-selection-ai-data-center-and-energy-infrastructure-development-federal)

11. "The $7 trillion data center build-out: How industrials can capture their share" (McKinsey & Company, https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-7-trillion-dollar-data-center-build-out-how-industrials-can-capture-their-share)

12. "'Largest Infrastructure Buildout in Human History': Jensen Huang on AI's 'Five-Layer Cake' at Davos" (NVIDIA Blog, 2026.1, https://blogs.nvidia.com/blog/davos-wef-blackrock-ceo-larry-fink-jensen-huang/)

13. "Transcript: Sam Altman Testifies At US Senate Hearing On AI Competitiveness" (Tech Policy Press, 2025.5.8, https://www.techpolicy.press/transcript-sam-altman-testifies-at-us-senate-hearing-on-ai-competitiveness/)

14. "Why are communities pushing back against data centers?" (Harvard Gazette, 2026.4.9, https://news.harvard.edu/gazette/story/2026/04/why-are-communities-pushing-back-against-data-centers/) — 데이터센터 인근 전기요금 2배 이상 인상 사례, 버지니아·조지아 세제혜택으로 10억 달러+ 세수 포기, 시설당 상시고용 20~50명 확인.

15. "AI Data Centers: Big Tech's Impact on Electric Bills, Water, and More" (Consumer Reports, https://www.consumerreports.org/data-centers/ai-data-centers-impact-on-electric-bills-water-and-more-a1040338678/) — 대형 데이터센터 1곳의 일일 물 사용량이 가정 6,500곳 사용량과 맞먹는다는 사례 확인.

16. "How much have data centers increased electricity prices?" (PolitiFact, 2026.6.12, https://www.politifact.com/factchecks/2026/jun/12/elizabeth-warren/data-centers-rising-electricity-costs/) — 최근 5년 미국 가정용 전기요금 평균 42% 상승(워싱턴 D.C. 94%·메릴랜드 74%), PJM 권역 데이터센터발 용량비용 2025~26년 174% 증가·약 93억 달러 추가부담 확인.

17. "FTC Launches Inquiry into Generative AI Investments and Partnerships" (U.S. Federal Trade Commission, 2024.1.25, https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships) — FTC가 알파벳·아마존·앤스로픽·마이크로소프트·오픈AI에 6(b) 명령을 발부해 AI 투자·제휴 구조를 조사한다고 확인. DOJ·FTC의 엔비디아/마이크로소프트·오픈AI 분담 조사는 "FTC and Justice Department to start antitrust investigation into Microsoft, OpenAI, and Nvidia" (Data Center Dynamics, 2024.6.7, https://www.datacenterdynamics.com/en/news/ftc-and-justice-department-to-start-antitrust-investigation-into-microsoft-openai-and-nvidia/)에서 교차 확인. FTC 원문이 접속 제한될 경우 웨이백 아카이브(https://web.archive.org/web/2024*/https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships)에서 보존 사본을 확인할 수 있다.

18. "Compute in Plan A" (AI Futures Project, 2026.7, https://ai-2040.com/supplements/compute-supplement) — AI 설비투자 2024년 $2,600억→2028년 $2.4조(채무 조달 비중 23%→38%), 2028~29년 AI의 EUV 장비 점유 65~77% 전망. 시나리오 분석 기관의 전망치임.

19. "Financing the AI boom: from cash flows to debt" (Bank for International Settlements, BIS Bulletin No.120, 2026, https://www.bis.org/publ/bisbull120.pdf); "The central bank of central banks just released its flagship annual report" (Fortune, 2026.6.29, https://fortune.com/2026/06/29/bis-central-bank-warning-hyperscaler-data-center-1-trillion-gamble-recession/) — BIS 2026년 연례보고서의 순환금융·설비투자 붐 붕괴 경고 인용.

중국의 여섯 갈래 전력망

중국의 AI 인프라는 중앙정부의 계획과 국유 전력망, 지방정부의 유치 경쟁, 민간 플랫폼 기업의 투자가 겹쳐 움직입니다. 국가 주도라는 한 단어만으로는 과잉 투자와 지역별 가동률, 민간 기업의 역할을 설명하기 어렵습니다.

2026년 7월 상하이 WAIC 이후에는 결론도 조금 달라졌습니다. 중국의 여섯 갈래 그물은 국내 인프라 정비에 그치지 않습니다. 전력, 컴퓨트, 데이터, 모델, 산업 현장, 국제 거버넌스를 한 시스템으로 묶어 밖으로 내보내려는 풀스택 전략의 바닥판에 가깝습니다.

여섯 갈래 그물과 컴퓨팅 회랑

중국 국가 매체가 "여섯 갈래 그물(六张网)"이라고 부르는 이 계획은, 나라 전체를 가로지르는 여섯 개의 인프라망을 하나로 묶습니다. 물을 나르는 수망(水网), 전기를 나르는 신형 전력망(新型电网), 계산을 나르는 컴퓨팅파워망(算力网), 데이터를 나르는 차세대 통신망(新一代通信网), 도시 지하관망(城市地下管网), 그리고 물자를 나르는 물류망(物流网)입니다[1].

눈여겨볼 대목은 이 여섯 개 중 하나가 컴퓨팅파워망이라는 점입니다. 물과 전기, 통신과 물류처럼 나라를 지탱하는 기초 인프라 목록에 "계산하는 능력"이 나란히 이름을 올린 것이지요. 2026년 4월 28일 중앙정치국 회의가 이 여섯 갈래 그물의 계획·건설을 강화하라고 처음 못 박았고, 채 열흘도 지나지 않은 5월 9일 국무원 상무회의가 이를 15차 5개년 계획의 현대 인프라 항목으로 구체화했습니다[1]. 정치국에서 방향을 정하고 국무원이 실행 계획으로 옮기는, 중국 특유의 하향식 편입 절차입니다.

돈의 규모도 만만치 않습니다. 2026년 한 해, 여섯 갈래 그물 전체에 투입되는 투자액이 7조 위안(약 1,330조 원)을 넘어설 것으로 중국 매체는 전망합니다[추정][2]. 이 가운데 신형 전력망 하나만 놓고 봐도 15차 5개년 계획 기간 동안 5조 위안(약 950조 원) 넘게 들어갈 예정입니다[2]. 전력망 하나에만 5년 동안 이만큼이 들어간다는 것이니, 그물 전체의 무게가 짐작되지요.

발상은 전기처럼, 필요할 때 계산 능력을 끌어다 쓸 수 있는 국가 기반망을 만들겠다는 것입니다. 그렇다고 컴퓨팅이 수돗물처럼 누구나 쓰는 공공재가 된 것은 아니지요. 이용료는 얼마인지, 누가 어떤 조건으로 쓸 수 있는지, 지역마다 어떻게 나누고 누가 운영하는지는 따로 따져 볼 문제입니다.

그렇다면 이 컴퓨팅파워망은 실제로 어떻게 움직일까요? 그 뼈대가 두 장 앞 「AI 에너지 전쟁」에서 다룬 "동수서산(東數西算)" 프로젝트입니다. 동부의 데이터를 서부에서 계산한다는 그 발상 말이지요. 허브 입지와 전력망 연계 원칙은 그 장에서 다뤘으니 여기서는 되풀이하지 않겠습니다. 대신 이 컴퓨팅파워망 자체가 얼마나 빠르게 커지고 있는지를 짚어볼까요.

먼저 단위부터 정리하지요. 1초에 100경 번의 계산을 처리하는 능력을 1엑사플롭스(EFLOPS)라고 부릅니다. 슈퍼컴퓨터의 성능을 잴 때 쓰는 단위인데, 앞으로 이 단위가 자주 등장할 테니 기억해 두면 좋습니다.

중국 공업정보화부(MIIT) 발표에 따르면 2026년 1월 기준 중국의 지능형 컴퓨팅파워 규모는 이미 1,590 EFLOPS로 세계 선두였습니다. 만 장 단위로 GPU를 묶은 "만카드(万卡) 클러스터"도 이때 이미 42개가 가동 중이었지요[3]. 3월 말에는 이 숫자가 1,882 EFLOPS(FP16 기준)까지 뛰었습니다[4]. 두 달여 만에 20% 가까이 불어난 셈입니다. 6월 말에는 2,185 EFLOPS, 1년 전의 2.8배 가까운 규모가 됐지요. 그렇다고 지어 놓은 시설이 다 차 있는 것은 아닙니다. 랙 가운데 실제로 장비가 들어찬 비율은 71.4%였으니까요.[5]

이 컴퓨팅파워를 실어 나르는 물리적 통로도 함께 깔렸습니다. 동수서산은 8대 컴퓨팅 허브와 그 안에 자리한 10대 데이터센터 클러스터로 구성됩니다. 징진지(京津冀), 창장삼각주, 웨강아오 대만구, 청위(成渝) 같은 동부·중부 허브와, 네이멍구·구이저우·간쑤·닝샤 같은 서부 허브가 짝을 이루는 구조입니다[6]. 그리고 이 허브들 사이에는 최근 2년 사이 70개가 넘는 컴퓨팅 회랑(算力大通道)이 새로 뚫렸습니다. 마치 컴퓨팅파워 전용으로 놓은 고속도로 같은 것이지요[7]. 물과 전기처럼 컴퓨팅파워도 쓰고 싶은 곳에서 쓸 수 있게 하겠다는 "산력호련호통(算力互联互通)" 구상이 숫자로 현실화되고 있는 셈입니다.

칩을 막으니 시스템으로 답하다

이 정도 규모의 컴퓨팅파워망을 돌리려면 당연히 칩이 필요합니다. 그런데 미국의 수출 통제로 최신 NVIDIA 칩을 마음껏 쓸 수 없는 게 중국의 현실이지요. 중국이 택한 답은 "낱개 칩에서 지면, 뭉쳐서 이기자"는 전략입니다.

화웨이의 Ascend 칩 라인업이 이 전략을 잘 보여줍니다. Ascend 950PR은 추론 중에서도 프리필(prefill)과 추천 연산을 맡고, Ascend 950DT는 학습과 추론 디코드를 맡습니다. 두 칩은 각각 2026년 1분기와 4분기에 나옵니다[8]. 개별 칩 성능만 보면 최신 NVIDIA 칩에 못 미칠 수 있습니다. 하지만 화웨이는 승부처를 낱개 칩이 아니라 "묶음"으로 옮겼습니다.

그 묶음의 이름이 Atlas 950 SuperCluster입니다. Atlas 950 초노드(SuperPoD) 64개를 하나로 엮은 시스템으로, 카드 수는 52만 장을 넘고 FP8 기준 연산 능력은 524 EFLOPS에 달합니다[8]. 초노드라는 말이 낯설 수 있는데, 물리적으로는 여러 대의 서버로 이뤄져 있지만 논리적으로는 한 대의 거대한 컴퓨터처럼 작동하도록 묶은 시스템을 뜻합니다. 개별 칩의 미세공정 열세를, 얼마나 많은 칩을 얼마나 촘촘하게 엮느냐는 시스템 설계의 규모로 상쇄하겠다는 계산이지요.

이런 국산 대안이 쌓이면서 실제로 시장 지형도 바뀌고 있습니다. 시장조사업체 IDC 집계에 따르면 중국 AI 가속서버 시장에서 국산 칩이 차지하는 비중은 2025년 기준 약 41%(출하량 약 165만 장)까지 올라왔고, NVIDIA의 점유율은 약 55%(출하량 약 220만 장)를 차지했습니다[업계 추정치, IDC 집계 기준][9]. 제재 이전 NVIDIA 점유율이 95%에 달했다는 비교치도 중국 매체에서 자주 보이지만, 원출처가 약해 참고 정도로만 들어 두면 됩니다. 정확한 자급률 수치는 집계 기관마다 갈리지만, 방향만큼은 뚜렷합니다. 칩을 막았더니 중국은 칩 대신 시스템으로 답하고 있는 셈입니다.

칩만이 아닙니다. 중국은 모델 그 자체도 무기로 씁니다. 오픈웨이트로 공개한 중국산 모델들이 이미 OpenRouter에서 소비되는 전체 토큰의 60% 안팎을 차지할 만큼 퍼졌는데[10], 왜 중국이 모델을 공짜로 풀어 표준과 생태계를 수출하려 하는지는 바로 다음 장에서 자세히 다루겠습니다.

WAIC 이후의 중국식 풀스택

WAIC 2026에서 중국이 내놓은 메시지는 국내망보다 넓었습니다. 시진핑 주석의 기조연설은 AI가 디지털 세계를 넘어 물리 세계로 들어가고 있으며, 전통 산업의 전환과 신산업, 미래 산업을 함께 밀어야 한다고 말했습니다.[11] 이는 피지컬 AI와 임바디드 AI를 산업정책의 중심에 놓겠다는 신호로 읽힙니다.

현장을 취재한 국내 보도도 칩 자립보다 시스템 경쟁, 피지컬 AI와 생태계·거버넌스가 전면에 나왔다고 전했습니다.[12] 전시장의 시연 장면이 곧 상용화 실적은 아니니, 정책이 어디를 보고 있는지 알려 주는 현장 기록으로 읽으면 됩니다.

에너지 쪽 정책도 같은 방향입니다. 2026년 5월 중국 국가발전개혁위원회는 에너지, 컴퓨트, 시나리오, 데이터, 모델을 함께 조정해야 한다고 설명했습니다. 대형 신재생에너지 기지와 국가 컴퓨팅 허브의 배치를 맞추고, 조건이 되는 지역에서는 대규모 AI 컴퓨팅 시설과 에너지 시스템을 함께 짓는 실험을 하겠다는 것이지요.[13]

7월의 AI 협력발전 행동계획은 여기에 수출 언어를 붙였습니다. 국경 간 신뢰 데이터 공간, 개발도상국을 위한 보편적 지능형 컴퓨팅 서비스, 국제 오픈소스 공동체, 산업·과학·의료·교육·농업·거버넌스에서의 AI+ 협력, 표준과 안전 거버넌스를 함께 제안했습니다.[14] 같은 날 29개국이 상하이에 본부를 두는 세계인공지능협력기구(WAICO) 설립 협정에 서명한 것도 이 흐름 위에 있습니다.[15]

중국은 여섯 개의 망을 깔고 끝내지 않습니다. 그 망 위에서 에너지와 컴퓨트를 조율하고, 현실 세계의 기계와 공정에 AI를 입히며, 오픈소스와 표준, 거버넌스까지 함께 내보내려 합니다. 미국이 민간 자본의 AI 팩토리로 속도를 내는 동안, 중국은 국가망을 풀스택 수출의 레일로 바꾸고 있는 셈입니다. 오픈웨이트 이야기를 한 장 더 본 뒤, 한국이 택한 세 번째 방식을 들여다보겠습니다.



1. "六张网建设加速推进" (新华网, 2026.6.4, https://www.news.cn/liangzi/20260604/8841a55f7b9e456ebf89571d1909c2a8/c.html) — 6대 인프라망 구성(水网·新型电网·算力网·新一代通信网·城市地下管网·物流网) 및 4/28 중앙정치국·5/9 국무원 상무회의 편입 시점 확인.

2. "六张网建设全面进入施工高峰期，全年总投资预计超7万亿元" (财富号·东方财富, 2026.5.23, https://caifuhao.eastmoney.com/news/20260523195454101588570) — 2026년 총투자 7조 위안+[추정]. 신형 전력망 5조 위안대 15차 5개년 투자 전망은 证券时报 기사로도 교차 확인(https://www.stcn.com/article/detail/3922340.html).

3. "我国智能算力规模突破1590 EFLOPS 万卡集群达42个" (中国日报, 2026.1.23, https://cn.chinadaily.com.cn/a/202601/23/WS69732037a310942cc499c874.html) — 2026.1 기준 지능형 컴퓨팅파워 1,590 EFLOPS, 만카드 클러스터 42개(공업정보화부 발표 인용).

4. "截至3月底，我国智能算力规模达1882EFLOPS" (新浪财经, 2026.4.22, https://finance.sina.com.cn/tech/roll/2026-04-22/doc-inhvhvwy4776766.shtml) — 2026.3말 기준 1,882 EFLOPS.

5. 新华社 (2026.7.20, https://www.news.cn/tech/20260720/63025e398f234092b9876272271417fb/c.html) — 국무원 신문판공실 발표, 2026.6말 지능형 컴퓨팅 2,185 EFLOPS(FP16)·전년 대비 177% 증가·전국 컴퓨팅 시설 상가율(가동 랙 비율) 71.4%.

6. 国家发展改革委高技术司负责同志就实施“东数西算”工程答记者问 (중국 국가발전개혁위원회, 2022.2.17, https://www.ndrc.gov.cn/xxgk/jd/jd/202202/t20220217_1315795.html) — 동수서산의 8대 허브·10대 데이터센터 클러스터와 전력망 연계 방향을 설명한다.

7. "算力大变局" (澎湃新闻, https://m.thepaper.cn/newsDetail_forward_33019963) — 최근 2년간 70여 개 컴퓨팅 회랑(算力大通道) 구축.

8. "华为Ascend 950系列与Atlas 950 SuperCluster解析" (InfoQ, https://www.infoq.cn/article/ck7a4wddik2wf4mzi4c4) — Ascend 950PR/950DT 역할 분담, Atlas 950 SuperCluster(초노드 64개·52만 카드·FP8 524 EFLOPS), 초노드(SuperPoD) 개념.

9. IDC 조사 결과 인용 — "IDC：2025年中国AI加速服务器市场本土芯片份额达41%" (199IT, 2026.4, https://www.199it.com/archives/1820172.html; IDC 원자료 https://my.idc.com/getdoc.jsp?containerId=prCHC53858025): 2025년 중국 AI 가속서버 시장 국산 칩 점유율 약 41%(출하량 약 165만 장), NVIDIA 점유율 약 55%(출하량 약 220만 장). "제재 이전 NVIDIA 점유율 95%" 비교치는 IDC 원자료에서 확인되지 않으며, "华为国产芯片算力超英伟达旗舰三倍!市场份额暴涨" (游民星空, 2026.4, https://www.gamersky.com/news/202604/2117133.shtml) 등 중국 매체에서만 확인되어 출처가 상대적으로 약함.

10. "China's Open-Weight Takeover" (Data Gravity, https://www.datagravity.dev/p/chinas-open-weight-takeover) — 2026.5 기준 OpenRouter 소비 토큰의 약 61%가 중국 오픈웨이트 모델. 상세 생태계 분석은 별도 장에서 다룸.

11. Full text: Keynote speech by Chinese President Xi Jinping at opening ceremony of 2026 World AI Conference (중국 국무원 영문판·신화, 2026.7.17, https://english.www.gov.cn/news/202607/17/content_WS6a5a1172c6d00ca5f9a0c46b.html) — AI가 디지털 세계에서 물리 세계로 이동하고 있으며 산업 적용, 안전, 포용, 글로벌 거버넌스를 함께 추진해야 한다고 밝혔다.

12. "WAIC 2026이 보여준 중국 AI 전략…기술 자립 넘어 글로벌 생태계 주도 나선다" (더테크, 2026.8.4, https://www.the-tech.co.kr/news/article.html?no=41800) — 전시 현장에서 풀스택 시스템, 피지컬 AI와 글로벌 생태계 경쟁이 부각됐다는 국내 취재 기사다. 정책 사실은 중국 정부 원문과 교차확인했다.

13. "这一次，人工智能和能源双向奔赴了！" (중국 국가발전개혁위원회, 2026.5.15, https://www.ndrc.gov.cn/wsdwhfz/202605/t20260515_1405213.html) — AI와 에너지의 양방향 역량 강화 행동방안 설명으로, 에너지·컴퓨트·시나리오·데이터·모델의 협동과 컴퓨트-전력 협동(算电协同)을 제시한다.

14. 「人工智能合作发展行动计划」 (중국 국가발전개혁위원회, 2026.7.17, https://www.ndrc.gov.cn/fggz/202607/t20260717_1406573_ext.html) — 고품질 데이터, 보편적 지능형 컴퓨팅, 오픈소스 생태계, AI+ 응용, 인재, 표준, 안전 거버넌스, AI 선용을 묶은 협력 행동계획이다.

15. "29 countries sign agreement on establishing World AI Cooperation Organization" (중국 국무원 영문판·신화, 2026.7.17, https://english.www.gov.cn/news/202607/17/content_WS6a59a226c6d00ca5f9a0c432.html) — 29개국이 상하이 본부의 독립 정부 간 국제기구 WAICO 설립 협정에 서명했다고 보도했다.

중국 오픈웨이트 전략: 설계도를 공개해 표준을 잡다

서울의 작은 스타트업 개발자를 한 사람 떠올려 볼까요. 고객 문의를 분류하는 기능을 붙여야 하는데 예산은 빠듯합니다. 모델 목록을 훑다 보면 Qwen이나 DeepSeek 같은 중국계 모델이 값싸고, 내려받아 회사 서버에서 돌릴 수도 있다는 게 눈에 들어오지요. 이런 선택이 세계 곳곳에서 매일 쌓이고 있습니다.

중국 AI 생태계를 모델 출시 목록으로 따라가면 금세 길을 잃습니다. 이름과 벤치마크는 몇 달마다 바뀌니까요. 오래 남는 것은 따로 있습니다. 가중치를 어떤 조건으로 공개하는지, 개발자가 실제로 가져다 쓰는지, 운영비까지 넣어도 쓸 만한지입니다.

그런데 오픈웨이트는 오픈소스와 같은 말일까요? 오픈웨이트(open-weight) 모델은 학습된 가중치를 내려받아 자체 장비에서 실행하거나 일부 수정할 수 있는 모델입니다. 발전기를 통째로 내주되, 그 발전기를 어떻게 설계하고 무슨 연료로 시험했는지까지 다 알려 주지는 않는 셈이지요. 소스코드와 학습 데이터, 전체 개발 과정이 모두 공개되는 것은 아니고, 라이선스로 상업 이용과 재배포, 특정 지역 사용을 제한하기도 합니다.

기업에는 이 차이가 꽤 큽니다. 가중치를 받을 수 있어도 데이터 계보를 확인할 수 없거나 라이선스가 서비스와 맞지 않으면 실제 도입은 어렵습니다. 반대로 자체 환경에서 실행할 수 있으면 민감한 데이터를 외부 API로 보내지 않고 모델을 교체할 선택권을 얻습니다.

중국에서는 Qwen과 DeepSeek 계열을 비롯한 여러 오픈웨이트 모델이 빠르게 확산됐습니다. 어느 모델이 어느 주에 벤치마크 1등을 했는지는 금방 잊힙니다. 오래 남는 건 공개 가중치 위에 도구와 파생 모델, 운영 경험이 쌓인 생태계지요.

그렇다면 이런 모델들은 실제로 얼마나 쓰이고 있을까요? 이 질문에 숫자 하나로 답하기는 어렵습니다. 다운로드 수는 관심을 보여 줄 뿐 실제 서비스 사용량과는 다르고, API 중개 서비스의 토큰 비중은 실제 사용을 보여 주지만 시장 전체를 대표하지는 못하지요. 기업 내부에 깔린 모델은 공개 통계에 잡히지도 않습니다.

OpenRouter 관련 보고서는 집계 시점에 따라 중국계 모델의 토큰 비중을 45%에서 67% 사이로 제시합니다.[1] 숫자가 벌어지는 건 표본과 시점이 달라서입니다. 그런데 2026년 9월, 토큰으로는 3분의 2 가까이를 차지한 중국 모델이 돈으로는 30% 남짓만 가져갔다는 증권사 분석도 보도됐습니다.[2] 많이 쓰이는 것과 많이 버는 것은 다른 이야기입니다. 이 비율은 한 라우팅 서비스 안의 점유율일 뿐이지만, 비용에 민감한 개발자들이 중국계 모델을 활발히 시험하고 있다는 신호로는 읽을 수 있지요.

미중경제안보검토위원회의 분석도 중국계 오픈 모델의 다운로드와 파생 생태계가 커졌다고 봅니다.[3] 미국 의회 산하 기관의 전략 분석이라는 성격은 감안해야 합니다. 그래도 공개 모델이 연구 성과를 넘어 표준과 산업 관계를 넓히는 지렛대가 될 수 있다는 지적은 새겨들을 만합니다.

공짜 가중치가 남기는 청구서와 영향력

오픈웨이트 모델은 API 표시 단가가 낮거나 자체 장비에서 운영할 수 있다는 장점이 있습니다. 그런데 발전기를 공짜로 받아도 연료비와 관리비는 내가 내야 하지요. 가중치가 무료여도 GPU와 전력, 엔지니어링, 모니터링, 보안 업데이트 비용은 고스란히 남습니다.

성능 격차도 하나의 숫자로 정리하기 어렵습니다. 일부 분석은 프론티어 폐쇄형 모델과 상위 오픈웨이트 모델 사이의 격차가 특정 벤치마크에서 수개월가량 유지된다고 설명합니다.[4] 이 결과를 모든 언어와 산업 과업에 적용할 수는 없습니다. 한국어 계약 검토와 영어 코딩처럼 과업이 달라지면 순위도 달라집니다.

그래서 앞의 개발자가 실제로 따져 봐야 할 계산은 토큰 단가보다 이쪽입니다.

> 성공 작업당 비용 = 모델 호출비 + 재시도 비용 + 사람 검토 시간 + 운영·보안 비용

저렴한 모델이 오류 때문에 세 번 재시도되면 비싼 모델보다 비용이 커질 수 있습니다. 반대로 정형화된 대량 작업에서는 약간 낮은 성공률을 라우팅과 검증으로 보완해 총비용을 줄일 수 있습니다.

공개 전략이 노리는 것은 비용 너머에 있습니다. 가중치를 공개하면 개발자는 모델을 자기 환경에 맞게 바꾸고, 도구 제작자는 그 모델을 기본 지원 목록에 넣습니다. 이런 채택이 쌓이면 모델 자체보다 주변 형식과 배포 도구, 평가 관행이 표준이 될 수 있습니다.

그렇다고 공개가 곧 지배를 보장하지는 않습니다. 라이선스가 바뀌거나 업데이트가 끊기고, 보안 취약점 대응이 늦으면 생태계는 다른 모델로 이동할 수 있습니다. 공개 모델을 국가 전략으로 볼 때도 다운로드 수와 함께 누가 유지보수하는지, 해외 기여자가 얼마나 붙었는지, 산업 현장에 실제로 깔렸는지, 갈아타는 데 얼마가 드는지를 살펴야 하지요.

그럼 한국은 무엇을 물어야 할까요? '중국 모델을 막을까'보다 두 가지가 먼저입니다. 남의 모델을 쓰다가도 언제든 갈아탈 수 있는가. 우리 모델은 한국어와 우리 산업 과업에서 써 보니 낫다는 증거를 내놓을 수 있는가. 서울의 그 개발자가 내일 아침 어떤 모델을 고르느냐가, 표준 경쟁의 한 표가 되니까요.

---

1. "Chinese AI Models Q2 2026: 10-Provider Landscape Report" (Digital Applied, 2026.4, https://www.digitalapplied.com/blog/chinese-ai-models-q2-2026-market-share-report); "China's Open-Weight Takeover" (Data Gravity, https://www.datagravity.dev/p/chinas-open-weight-takeover) — 서로 다른 시점과 집계가 45-61% 범위를 제시하며, 세계 전체가 아닌 OpenRouter 내부 사용량이다.

2. Digital Applied (2026.9.26, https://www.digitalapplied.com/blog/chinese-ai-models-majority-openrouter-tokens-2026) — CNBC 보도 요약, 9월 14일 주간 중국계 토큰 비중 57-67%; CLSA 분석 인용 보도 (https://www.itiger.com/news/1125046245) — 9월 토큰 비중 약 65%, 지출(달러) 비중 약 30%.

3. "Two Loops: How China's Open AI Strategy Reinforces Its Industrial Dominance" (USCC, 2026.3, https://www.uscc.gov/sites/default/files/2026-03/Two_Loops-%2DHow_Chinas_Open_AI_Strategy_Reinforces_Its_Industrial_Dominance.pdf)

4. "The Open-Weight Models That Matter — June 2026" (OpenRouter, https://openrouter.ai/blog/insights/the-open-weight-models-that-matter-june-2026) — 특정 평가 묶음에서 프론티어와 오픈웨이트 간 격차를 분석한 자료로, 모든 과업의 보편적 격차를 뜻하지 않는다.

대한민국의 대도약, 세 개의 메가 프로젝트

한국의 AI 인프라는 정부 계획, 대기업의 장기 투자 의향, 지역별 데이터센터 사업이 겹겹이 포개져 있습니다. 발표된 총액을 그대로 더하면 같은 사업이 두 번 세어지거나, 아직 정해지지 않은 계획이 이미 쓴 돈처럼 보이기 쉽지요.

2026년 6월 29일 정부는 '대한민국 대도약 3대 메가프로젝트 국민보고회'에서 반도체, 피지컬 AI, AI 데이터센터를 세 축으로 제시했습니다.[1] 여기서는 구호보다 각 사업이 발표·계약·착공·가동 가운데 어디까지 왔는지를 구분해 보겠습니다.

세 개의 축과 '한국판 스타게이트'

정부가 내놓은 그림은 삼각형입니다. 반도체, 피지컬 AI, 그리고 AI 데이터센터(AIDC). 세 꼭짓점이 서로를 떠받치는 구조이지요[1].

반도체는 지능 에너지를 만드는 발전 설비의 부품을 찍어내는 축입니다. 정부는 이 축에 '3S+1F' 전략이라는 이름을 붙였습니다. 수도권 생산 거점을 서둘러 완성하는 속도전(Speed), 전국으로 생산 거점을 넓히는 거점전(Stronghold), 차세대 반도체 기술을 선점하는 선도전(Spearhead), 그리고 전력·용수·도로 같은 인프라와 재정·제도를 총동원하는 총력지원체계(Full-support)까지 더한 것입니다[2].

피지컬 AI는 로봇의 몸에 지능을 입히는 축입니다. 정부는 여기에도 'M.AX(제조업 AI 전환)·Master(핵심기술 확보)·Mass Production(양산)'이라는 '3M 전략'을 붙였습니다. 새만금과 대경권을 양대 축으로 삼아, 2030년까지 AI 로봇 글로벌 3강에 들어가겠다는 목표입니다[3]. 로봇의 몸통과 손, 센서, 액추에이터 같은 부품 이야기는 뒤에서, 특히 제5부와 제6부에서 다룰 로봇 이야기와 이어지니 여기서는 이름만 걸어두고 넘어가겠습니다.

그리고 AIDC. 지능 에너지를 실제로 만들어 내는 발전소 그 자체입니다. 정부 발표와 기업의 장기 투자 의향, 이미 진행 중인 사업은 기간도 확정 수준도 달라서, 하나의 총액으로 묶으면 겹치는 부분이 생기지요. 그래서 거대한 합계보다 무엇이 공식 목표이고 무엇이 아직 민간의 약속인지를 따라가 보겠습니다.

한 달 뒤에는 이 삼각형에 공급망이라는 말이 덧붙었습니다. 2026년 7월 24일 샌프란시스코 AI 선언은 이 장의 범위를 더 분명하게 보여줍니다. 대통령 연설은 AI 경쟁이 더 좋은 알고리즘만의 경쟁이 아니라, 컴퓨팅 자원, 안정적인 에너지, 풍부한 데이터, 산업과 일상 적용 속도가 함께 결정하는 경쟁이라고 설명했습니다. 기술과 생산 역량, 데이터와 에너지, 시장과 인재가 하나의 공급망으로 연결된다는 말이지요.[4]

이 연설의 눈으로 보면 한국의 메가 프로젝트는 반도체 공장이나 데이터센터 몇 곳을 더한 것보다 큰 그림입니다. 알고리즘, 컴퓨트, 안정적인 전력, 데이터, 실제 적용 현장을 한 줄로 잇겠다는 구상이지요. 대통령 연설이 한국을 글로벌 AI 테스트베드이자 생산 플랫폼으로 부른 것도 그래서입니다. 제조공장, 도시, 물류 현장에서 AI를 빠르게 검증하고, 그 결과 데이터를 다시 모델과 서비스 품질로 되돌리는 구조 말입니다.[4]

물론 선언은 선언입니다. 계획은 목표와 방향이고, MOU는 협력 범위를 찾는 약속이며, 계약은 돈과 책임을 묶는 문서입니다. 착공은 공사가 시작됐다는 뜻이고, 운영은 전력과 장비와 사람이 붙어 실제 서비스를 돌린다는 뜻입니다. 연설에 나온 약 4,700조 원 규모의 투자 이니셔티브도 장기 프로젝트와 민간 계획을 아우른 범위이지, 정부가 쓴 돈은 아니지요.

선언에서 다시 땅으로 내려와 세 축을 하나씩 짚어 볼까요? 가장 먼저 볼 것은 AIDC입니다. 언론은 이걸 '한국판 스타게이트'라고 부르지요.

정부 발표 자료에는 AIDC에 550조 원 규모를 들이겠다는 구상이 담겼습니다.[1] 여러 해에 걸친 여러 사업자의 계획을 합친 숫자라, 부지마다 전력 계통연계와 인허가, 자금조달이 끝나기 전까지는 운영 실적이 아니라 계획으로 읽어야 합니다.

특히 GW는 전력 용량이고 원은 투자비입니다. 두 단위를 섞어 국가 간 순위를 만들면 실제 가동률과 전력 조달 조건이 사라집니다. 발표된 최대 용량보다, 전력망에 연결이 확정된 용량과 실제로 불이 켜지는 날짜가 더 정직한 숫자입니다.

대형 발전소만 계획된 것도 아닙니다. AI 행동계획은 전력 수요가 낮아 지역에 분산해 지을 수 있는 강소형 AI 데이터센터의 확산 방안을 따로 못 박아 두었습니다(과제 [2]).[5] 지역 중소기업과 의료기관에 지능 에너지를 가까이서 공급하고 국산 AI 반도체가 실전을 밟을 무대를 만드는, 말하자면 대형 발전소 곁의 지역 발전소 전략이지요. 이 분산 축은 제7부에서 볼 지능 에너지 격차 이야기와 이어집니다.

AI 시대, 메모리가 병목

두 번째 축은 반도체입니다. 발전소를 아무리 많이 지어도 그 안에 들어갈 칩이 없으면 무용지물이니까요.

정부는 서남권 생산기지에 800조 원, 충청권 첨단 패키징 거점에 81조 원 규모의 장기 투자 구상을 제시했습니다.[2] 삼성전자와 SK하이닉스의 여러 해 투자 의향을 합친 숫자이지, 해마다 확정된 지출은 아닙니다. 용인 등 기존 클러스터 계획과 겹치는 부분도 있을 수 있으니 다른 숫자와 더하기 전에 한 번 멈춰 보세요.

왜 이렇게 큰돈을 들여 팹과 패키징 거점을 늘리려는 걸까요. GPU가 계산을 아무리 빨리해도 필요한 데이터를 제때 받지 못하면 기다려야 하기 때문입니다. 지금의 AI 시스템은 메모리와 연산 장치가 떨어져 있어 데이터를 계속 옮겨야 합니다. 모델이 커지고 대화 맥락이 길어질수록 메모리 용량과 대역폭, 데이터 이동에 드는 전력이 함께 늘어납니다. 이른바 메모리 장벽(memory wall)이지요. 「AI+HW 2035」에

참여한 연구진은 데이터 이동 비용을 줄이려면 메모리와 연산을 함께 설계하고, 컴퓨팅 인 메모리(CIM)와 3차원 적층 같은 구조로 전환해야 한다고 제안했습니다.[6]

이 변화는 한국에 드문 기회입니다. 삼성전자와 SK하이닉스는 HBM과 D램을 생산할 뿐 아니라 커스텀 HBM, 베이스 다이, 3차원 적층 기술을 개발하고 있습니다. 노무라증권은 대규모 추론과 에이전트 AI가 사용하는 KV 캐시가 메모리 수요를 키우면서, 데이터센터 투자에서 메모리가 차지하는 비중도 커질 것으로 전망했습니다.[7] 증권사의 전망은 전망일 뿐이지요. 지금의 우위가 얼마나 갈지는 제품 세대마다 시장을 얼마나 지키고, 고객의 인증을 얼마나 먼저 받느냐가 말해 줄 겁니다.

메가 프로젝트의 확장 가능성은 여기서 생깁니다. 서남권 팹이 생산량을 늘리고 충청권 패키징 거점이 HBM과 가속기를 더 촘촘하게 묶으며, AIDC가 그 결과물을 실제 서비스에서 검증한다면 세 프로젝트가 하나의 실험실처럼 연결될 수 있습니다. 메모리를 많이 파는 데서 멈추지 않고 고객 맞춤형 HBM 공동설계, 첨단 패키징, 국산 가속기와 메모리 중심 소프트웨어까지 함께 확보하는 길입니다.

그러나 병목을 공급한다고 해서 지능의 가격까지 결정하는 것은 아닙니다. GPU와 모델, 클라우드 플랫폼을 해외 기업이 쥔 채 웨이퍼 생산능력만 늘리면 한국은 가장 중요한 부품을 만들고도 더 큰 부가가치를 다른 층에 넘겨줄 수 있습니다. 대규모 증설 뒤 공급과잉이 찾아오는 메모리 산업의 오래된 위험도 남아 있지요.[8] 메가 프로젝트가 산업 경쟁력으로 이어지려면 생산량뿐 아니라 공동설계권, 패키징 기술, 전력과 용수, 전문 인력, 국내 실증 수요를 함께 키워야 합니다.

GPU와 예산, 그리고 청구서

세 번째 이야기는 GPU입니다. 발전소 부지를 잡고 부품 공장을 지어도, 그 안에서 돌아갈 터빈인 GPU가 없으면 불은 켜지지 않지요.

국가AI컴퓨팅센터도 부지 선정, 사업자 구성, 착공, 장비 도입, 서비스 개시가 서로 다른 단계입니다. 정부가 제시한 GPU 수량은 앞으로 들여올 목표이고, 일정과 지분 구조도 사업 협의에 따라 바뀔 수 있습니다. 실제 계약서와 설치된 장비가 다음 단계를 가늠할 잣대가 되겠지요. 해남 센터는 2026년 8월 첫 삽을 떴고, 2028년 말까지 GPU 1만 5,000장을 갖출 계획입니다.[9]

NVIDIA와 한국 정부·기업이 발표한 26만 장 이상이라는 숫자도 정부와 민간의 구축 계획을 합친 생태계 목표입니다.[10] 발표, 구매 계약, 납품, 설치, 실제 가동을 단계별로 나누고, 국가 인프라 물량과 각 기업의 자체 설비도 분리해서 집계해야 하지요. 정부 몫 5만여 장은 이미 일부가 계약과 반입 단계로 넘어갔습니다. 2026년 6월에는 B300과 베라 루빈 9,704장을 들일 사업자가 정해졌지요.[11] 민간 몫의 납품 현황은 아직 공식 자료로 확인되지 않습니다.

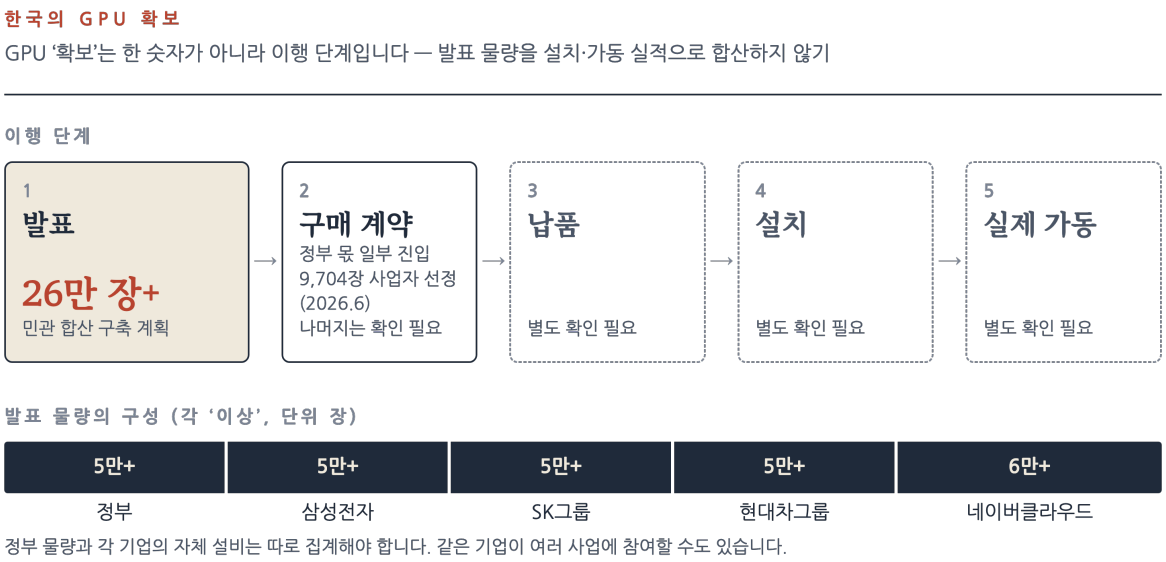


그림 2-11 26만 장은 대부분 아직 첫 칸인 '발표'에 놓인 숫자입니다. 정부 몫 가운데 9,704장이 2026년 6월 계약 칸으로 넘어갔고, 나머지 칸이 채워지는지는 앞으로 하나씩 확인해 갈 일이지요.

이 모든 것을 뒷받침하는 예산도 커졌습니다. 정부는 2026년 AI 관련 예산을 확정 기준 9.9조 원으로 설명합니다.[12] 2027년 예산안에서는 과기정통부 몫의 AI 예산만 9.4조 원으로 한 해 만에 84% 늘렸고, 3대 메가프로젝트에도 8천억 원 가까이를 편성했습니다. 2026년 10월 현재로서는 국회 심의를 앞둔 정부안이지요.[13] 여기에는 AI 전용 사업과 다른 정책 목적을 함께 가진 사업이 섞여 있어서, 이 총액이 곧 AIDC 건설비나 GPU 구매비는 아닙니다.

이 그림에 모두가 박수를 보낸 것은 아닙니다. 민주노총 충북본부는 이 프로젝트를 두고 "재벌대기업들을 위해 나라 전체의 자원을 몰아주는 국가총동원체제"라고 불렀고, 녹색연합은 "기업의, 기업을 위한, 기업에 의한 정책"이라며 이익과 비용의 배분이 한쪽으로 쏠려 있다고 지적했습니다[14]. 삼성·SK·네이버·카카오 같은 소수 대기업에 국가 재정과 인프라 지원이 집중되는 구조인 만큼, 그 청구서를 누가 나눠 낼 것인지는 앞으로도 계속 따라붙을 질문입니다.

이 청구서는 나라 곳간과도 무관하지 않습니다. 2026년 국가채무는 1,415조 원으로 사상 처음 GDP의 절반(51.6%)을 넘어설 전망입니다[15]. 재정 여력이 빠듯해지는 국면에서 소수 대기업의 초대형 프로젝트에 재정·인프라 지원이 쏠리면, 그만큼 다른 데 갈 몫이 밀려나는 크라우딩아웃 우려가 따라붙습니다.

이 집중 문제는 제7부 「대한민국 AI 행동계획」에서 정반대편 숫자로 다시 만나게 됩니다. 대기업이 수조 원짜리 AI 팩토리를 짓는 동안, 중소 제조현장의 AI 도입률은 조사마다 기준이 달라 서로 비교하기 어렵지만, 어느 조사로 봐도 대기업에 한참 못 미친다는 숫자로요. 발전소를 몇 곳에 몰아 짓는 일과, 그 전기가 중소기업이라는 마지막 1마일까지 닿느냐 하는 문제는 결국 한 몸인 셈입니다.

기업이라면 여기서 갈래가 갈립니다. 대기업은 이 집중의 수혜가 언제든 규제와 여론의 역풍으로 뒤집힐 수 있음을 감안해야 하고, 중소기업은 '마지막 1마일' 지원사업이 어디서 열리는지를 지금부터 챙겨야 하지요.

6월 29일 영빈관 무대에는 반도체·피지컬 AI·AIDC라는 세 꼭짓점이 커다랗게 걸렸고, 7월 샌프란시스코 선언은 그 꼭짓점들을 공급망 언어로 묶었습니다. 그래도 삼각형을 실제로 돌리는 연료는 여전히 국내 전력망 여유분에 달려 있고, 발전소에 꽂는 최신 GPU도 여전히 NVIDIA라는 남의 손에 달려 있습니다. 한국이 자국 안에서 자신 있게 내놓을 수 있는 건 HBM과 파운드리, 제조 현장, 그리고 이 세 개의 메가 프로젝트로 만들려는 규모입니다. 이 균형이 정말 맞아떨어질지는, 몇 장 뒤 「지능을 수출하는 나라」에서 소버린 AI와 지능 수출이라는 질문으로 다시 따져 보지요.

1. "(참고자료) '대한민국 대도약 3대 메가프로젝트 국민보고회' 개최" (산업통상부, 2026.6.29, https://www.motir.go.kr/kor/article/ATCL3f49a5a8c/171974/view) — 2026.6.29 청와대 영빈관 개최, 반도체·피지컬AI·AIDC 3축, 서남권 800조+충청 81조(반도체), AIDC 550조 확인.

2. "'3S 전략으로 초격차 유지'… '대체불가 K반도체' 키운다 [3대 메가 프로젝트]" (파이낸셜뉴스, https://www.fnnews.com/news/202606291826149273) — 3S(Speed·Stronghold·Spearhead)+Full-support 전략, 서남권 800조 원(삼성전자 2기·SK하이닉스 2기), 충청권 81조 원(HBM 등 첨단 패키징) 확인.

3. "대도약 3대 메가프로젝트 …피지컬AI 기업들 '연합군·데이터 표준 생태계 필요'" (디지털데일리, https://www.ddaily.co.kr/page/view/20260629162054098 28); "[대한민국 3대 메가프로젝트]새만금·대경권 양대 축으로 AI로봇 글로벌 3강으로 키운다" (전자신문, https://www.etnews.com/20260629000377) — 제조업 AI 전환·핵심기술 확보·양산으로 이어지는 3M 전략, 새만금·대경권 양대 축, 2030년 AI 로봇 글로벌 3강 목표 확인.

4. "7/24(금) 샌프란시스코 AI 선언" (대한민국 대통령실, 2026.7.25 게시, https://www.president.go.kr/speeches/hHXHeToZ) — 2026.7.24 샌프란시스코 연설로, 알고리즘·컴퓨팅 자원·안정적 에너지·데이터·산업 적용을 하나의 AI 공급망으로 설명하고 대한민국을 글로벌 AI 테스트베드·생산 플랫폼으로 제시했다.

5. 「대한민국 인공지능 행동계획(인공지능 기본계획 2026-2028)」 과제 [2] (국가인공지능전략위원회, https://www.aikorea.go.kr/upload/contents/contents_01.html) — 강소형 AI 데이터센터의 지역 분산 확산 과제.

6. "AI 거물들 '메모리가 최대 병목…메모리 중심 구조로 AI 효율화'" (조선일보, 2026.3.24, https://www.chosun.com/economy/tech_it/2026/03/24/FCQ4QJGKNRBFZCDMBJRLDXPC7I/) — 「AI+HW 2035」가 지적한 메모리 장벽, 데이터 이동에 따른 전력 비용, 메모리·연산 공동설계와 CIM·3차원 적층 제안을 소개한다.

7. "노무라, AI 메모리 슈퍼사이클 전망… 삼성전자·SK하이닉스 목표가 대폭 상향" (위클리포스트, 2026.5.16, https://www.weeklypost.kr/news/articleView.html?idxno=11287) — 추론과 에이전트 AI의 KV 캐시가 메모리 수요를 늘리고 데이터센터 투자에서 메모리 비중이 확대될 것이라는 노무라증권의 전망을 전한다. 기사에 실린 목표주가와 장기 수요 전망은 투자기관의 추정치다.

8. "왜 전세계 누구도 D램이 AI 병목임을 예측하지 못했나" (더밀크, 2026.6.29, https://themiilk.com/articles/a1dbbad0f) — AI 병목이 메모리·패키징으로 이동한 배경과 함께, 한국이 생산능력 증설을 넘어 맞춤형 HBM 공동설계·첨단 패키징·지능 계층으로 부가가치를 확장해야 한다는 분석을 제시한다.

9. "허허벌판에서 AI 허브로…해남 '국가 AI컴퓨팅센터' 착공 현장 가보니" (뉴시스, 2026.8.3, https://www.newsis.com/view/NISX20260803_0003734656) — 2026.8.3 착공, 사업비 약 2조 5,000억 원, 2028년 말까지 GPU 1만 5,000장 구축, 2027년 일부 선(先)서비스·2028년 완공 뒤 본격 제공. 정부·국민성장펀드·삼성SDS 컨소시엄 합작법인 '코아크'가 구축·운영. 정부 몫 전체 목표는 2028년까지 5만 2,000장 이상.

10. "NVIDIA, South Korea Government and Industrial Giants Build AI Infrastructure and Ecosystem to Fuel Korea Innovation, Industries and Jobs" (NVIDIA, 2025.10.30, https://nvidianews.nvidia.com/news/south-korea-ai-infrastructure) — 정부 5만 장 이상, 삼성전자·SK그룹·현대자동차그룹 각 5만 장 이상, 네이버클라우드 6만 장 이상으로 합계 26만 장 이상의 구축 계획을 명시한다.

11. 데일리안 (2026.6.8, https://www.dailian.co.kr/news/view/1653715/) — 네이버클라우드·삼성SDS·엘리스그룹을 GPU 9,704장(B300 7,688장, 베라 루빈 2,016장, 사업비 2조 800억 원) 사업자로 선정. B300은 2026년 내, 베라 루빈은 2027년 상반기 서비스 예정.

12. "AI 대전환·성장과 복지…새 정부 추가경정예산 주요 내용" (대한민국 정책브리핑, 2026.6.23, https://www.korea.kr/news/policyNewsView.do?newsId=148964928) — 2026년 확정 기준 AI 관련 예산 9.9조 원을 설명한다.

13. 머니투데이 (2026.9.1, https://www.mt.co.kr/tech/2026/09/01/2026090108281942139); SBS Biz (https://biz.sbs.co.kr/amp/article/20000332008) — 2026.9.1 국무회의 의결 2027년도 예산안. 과기정통부 AI 예산 9조 4,211억 원(+84.3%), 3대 메가프로젝트 2,981억→7,956억 원, '모두의 AI' 2,500억 원 신설. 국회 심의는 12월 초까지.

14. "충북노동계 '3대메가프로젝트 "재벌 위한 국가총동원체제"" (오마이뉴스, 2026.7.2, https://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0003248274&CMPT_CD=M0152); "녹색연합, 정부 '3대 메가프로젝트' 강력 비판" (한국NGO신문, 2026.6.29, https://www.ngonews.kr/news/articleView.html?idxno=231884) — 대기업 편중·재정 부담 구조에 대한 시민사회·노동계 비판 인용.

15. "내년 국가채무 1415조, GDP 대비 50% 넘어서[2026년 예산안]" (경향신문, 2025.8.29, https://www.khan.co.kr/article/202508291111001) — 기획재정부 2026년도 예산안 기준 2026년 국가채무 1,415조2,000억 원, GDP 대비 국가채무비율 51.6%(사상 첫 50%대 진입) 확인. AI 예산의 규모와 과제를 다룬 2차 분석으로는 국회예산정책처(NABO), 「AI 3대 강국 도약을 위한 예산 현황과 향후 과제」(NABO Focus 제132호, 2026-01-22, https://www.nabo.go.kr/ko/periodical/focusView.do?key=2507040015&idx=9097)가 있다.

지능 에너지 경제: 토큰 이코노미

전기 요금을 아시나요? 한국전력 기준 산업용 전기 1kWh의 평균 판매단가는 180원 안팎입니다.[1] 별것 아닌 숫자 같지만 한국 제조업의 경쟁력에 직결됩니다. 전기 요금이 1원이라도 비싸면 제품 원가가 올라가니까요.

AI 시대에도 똑같은 요금이 있습니다. 토큰(token) 비용입니다. "호환마마보다 무서운 토큰 비용"이라는 말이 나올 정도이지요.

그런데 토큰이란 정확히 무엇일까요? 전기 사용량을 kWh로 세듯, AI 서비스는 토큰 수로 처리량과 요금을 계산합니다. 토큰은 AI가 텍스트를 나누어 처리하는 기본 단위입니다. 지능이나 답변 품질을 재는 단위는 아니지요. 영어는 단어 하나가 대략 1토큰이고, 한국어는 모델에 따라 한 글자가 1~3토큰쯤 됩니다. 최신 토크나이저일수록 1토큰에 가깝지요. 이 책 한 페이지(약 500자)를 읽히면 대략 500~1,500토큰이 듭니다. 전등을 켜는 데 전기가 드는 것처럼, AI에게 일을 시키면 처리할 토큰이 생기는 것이지요.

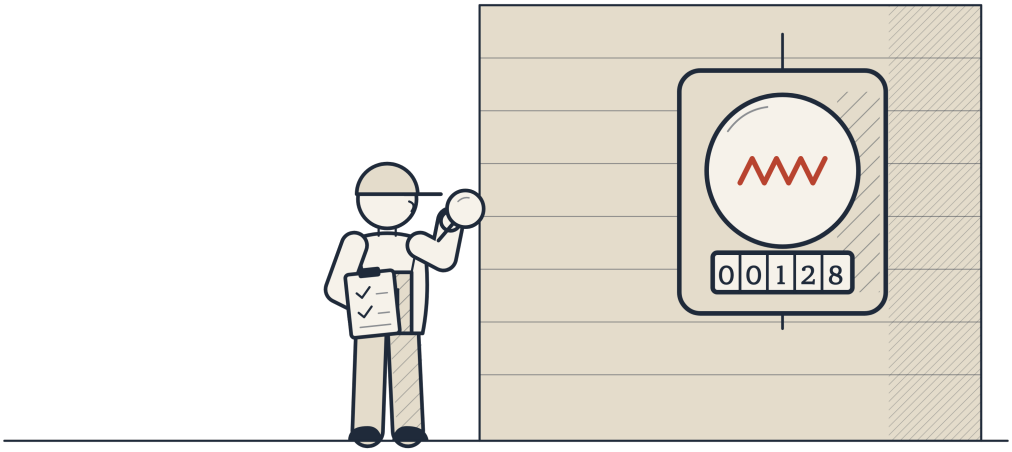


그림 2-12 검침원이 계량기 숫자를 읽어 전기 요금을 매기듯, AI 서비스는 오간 토큰 수를 세어 값을 매깁니다. 지능도 쓴 만큼 계량되는 에너지이지요.

한 가지 중요한 점이 있습니다. 대부분의 API는 입력보다 출력 토큰에 더 높은 가격을 매깁니다. 출력은 단어를 하나씩 생성하는 자기회귀 계산이 필요하지만, 가격 차이가 계산량만을 그대로 반영하는 것은 아닙니다. 캐시 운영, 서비스 정책, 수익 구조도 가격에 함께 들어갑니다. 토큰 요금은 계산 비용을 짐작하는 표지이지, 모델의 지능을 재는 계기판이 아닙니다.

지능 에너지의 요금표

2026년 10월 현재, 토큰 비용은 모델마다 천차만별입니다. 편의상 가격대를 프리미엄·표준·경량·초저가로 나누어 보겠습니다. 비싼 모델이 모든 일을 더 잘하고 싼 모델이 같은 일을 못한다는 뜻은 아니니, 아래 표는 품질 순위표가 아닌 API 표시 단가표로 읽어 주세요. 9~10월 사이 미국 세 회사가 모두 새 세대를 내놓으면서 표가 한 번 통째로 바뀌었으니, 이 장을 읽는 시점에는 또 바뀌어 있을 겁니다.

Provider	Model	Context	Input $/1M tokens	Output $/1M tokens	특징
Anthropic	Claude Fable 5.1	—	$10	$50	최고성능
Anthropic	Claude Opus 5.5	—	$4	$20	프론티어 추론
Anthropic	Claude Sonnet 5.5	—	$2	$10	균형
Anthropic	Claude Haiku 5.5	—	$0.10–0.50	$0.50–2.50	경량(10만 토큰 넘으면 할증)
OpenAI	GPT-6 Astra	—	$10	$50	최상위, 단계적 공개
OpenAI	GPT-6.1 Sol	—	$2	$10	주력
OpenAI	GPT-6 Luna	—	$0.10	$0.50	경량
OpenAI	GPT-5.5-pro	—	$30	$180	최상위 추론(이전 세대)
OpenAI	GPT-5.3-codex	—	$1.75	$14	코드특화
Google	Gemini 4 Argon	1M	$4	$20	최고성능
Google	Gemini 3.8 Flash	—	$1.50	$7.50	범용
Google	Gemini 3.5 Flash-Lite	—	$0.30	$2.50	경량
xAI	Grok 4.7	500K	$2–4	$6–12	실시간검색
Mistral	Mistral Large	—	$0.5	$1.5	EU 기반
DeepSeek	V4 Pro	1M	$0.66–1.32	$1.98–3.96	중국 오픈웨이트, 피크 요금제
DeepSeek	V4.1 Flash (1017)	1M	$0.15–0.30	$0.60–1.20	초저가, 비피크 절반

Provider	Model	Context	Input $/1M tokens	Output $/1M tokens	특징
Alibaba	Qwen3.8-Flash	—	~$0.15	~$0.47	초저가(3자 집계)
Moonshot	Kimi K3	~1M	$3	$15	초장문
Zhipu	GLM-5.3	—	$1.40	$4.40	중국 오픈웨이트
Local	Gemma 계열 등	로컬	API 요금 없음	API 요금 없음	장비·전력·운영비 별도

출처: Anthropic·OpenAI·Google·xAI·Mistral·DeepSeek·Zhipu 공식 가격표(2026년 10월 22일 확인). Qwen·Kimi는 공식 표를 확인하지 못해 3자 집계를 참고했습니다. Google 신모델은 출시 할인가가 끝난 뒤의 정상가로 적었습니다(Gemini 4 Argon은 11월 30일까지 $2/$10, Gemini 3.8 Flash는 12월 31일까지 $0.75/$3.75). 범위로 적은 값은 문맥 길이나 피크 시간대에 따라 달라지는 요금이고, "—"는 공식 표에서 확인하지 못한 칸입니다. 이하 원화 병기는 $1≈1,350원 기준 환산.

이 표는 2026년 10월 22일 기준입니다. API 가격은 새 모델이 나올 때마다, 때로는 몇 주 만에 바뀝니다. 실제로 쓰기 전에는 각 회사의 가격 페이지를 다시 확인해 주세요.

이 표를 네 등급으로 나눠서 보겠습니다.

프리미엄 지능은 Claude Fable 5.1과 GPT-6 Astra(둘 다 출력 50달러), Claude Opus 5.5와 Gemini 4 Argon(둘 다 정상가 출력 20달러)처럼 가장 깊은 추론 능력을 갖춘 등급입니다. 복잡한 아키텍처 설계, 정교한 분석, 미묘한 판단이 필요할 때 쓰지요. 이전 세대의 최상위 추론 모델 GPT-5.5-pro는 출력 180달러로 여전히 가장 비쌉니다.

표준 지능은 성능과 가격의 균형을 겨냥한 등급입니다. 코딩, 검색, 문서 처리처럼 과업에 따라 강점과 가격 구조가 다르므로 모델 이름만으로 등급을 고정하기는 어렵습니다. 대략 Claude Sonnet 5.5와 GPT-6.1 Sol(둘 다 출력 10달러), GPT-5.3-codex(14달러), Kimi K3(15달러)가 이 구간에 있지요. 표의 가격은 2026년 10월 표시 단가의 스냅숏이며 실제 선택에서는 캐시 할인, 긴 문맥 할증, 도구 호출비와 성공률을 함께 봐야 합니다.

경량 지능은 프리미엄의 수십 분의 1 가격입니다. Claude Haiku 5.5와 GPT-6 Luna(둘 다 출력 0.5달러), Gemini 3.5 Flash-Lite(2.5달러)가 간단한 분류·요약·일상 질의응답을 맡습니다.

초저가 지능은 이제 국적으로 가르기 어렵습니다. 몇 달 전까지는 중국계 모델의 독무대에 가까웠지만, Qwen3.8-Flash(출력 약 0.47달러)와 DeepSeek V4.1 Flash(0.6~1.2달러) 옆에 미국 회사들의 경량 모델이 같은 가격대로 내려왔지요. 거꾸로 Kimi K3(15달러)처럼 값을 올려 표준 등급으로 올라선 중국 모델도 있습니다. 프리미엄과 견주면 대략 수십 배에서 100배쯤 차이가 납니다.

여기서 전기 비유는 한 번 멈춰야 합니다. 전기는 어느 발전소에서 왔든 콘센트에서는 같은 전기지만, 지능은 등급마다 해낼 수 있는 일이 다릅니다. 그래서 요금표를 고르는 일은 전기요금제를 고르는 일보다 공구를 고르는 일에 가깝지요. 어떤 등급을 쓸지는 맡기려는 일에 달려 있습니다.

이 요금표는 고정된 것이 아닙니다. Epoch AI에 따르면 GPT-4급 성능(GPQA Diamond 기준)을 얻는 비용이 매년 40배씩 떨어지고 있고, 전체 벤치마크로 넓히면 중앙값 연 50배, 범위로는 연 9배에서 900배까지 벌어집니다.[2] 전기 요금이 발전 기술과 함께 내려가 온 것과 비슷하지요. 그렇다고 늘 내리막만 있는 것은 아닙니다. 2026년 4월 OpenAI가 GPT-5.5를 내놓으며 이전 GPT-5 라인 가격을 오히려 2배 올린 것처럼, 성능이 오르면 가격도 함께 뛰는 구간이 있습니다. 8월에는 DeepSeek가 수요가 몰린다는 이유로 V4 계열 일부 요금을 열 배 넘게 올렸지요.[3]

토큰이 비싸면 에이전트를 돌릴 수 없다

여기까지만 보면 "적절한 모델을 골라 쓰면 되는 거 아닌가?"라고 생각하실 수 있습니다. 맞습니다. 사람이 가끔 질문을 던지는 라이트 유저 수준(월 10~30달러, 넷플릭스 구독료 정도)이라면 그렇습니다. 하지만 문제는 에이전트 시대가 오고 있다는 것입니다.

사용자 유형	월 평균 토큰	월 평균 비용	주요 시나리오
라이트 유저	500만~2,000만	$10~30 (약 1.4만~4만 원)	일상적인 질의응답, 단순 작업
미디엄 유저	2,000만~5,000만	$30~70 (약 4만~9.5만 원)	자동화 워크플로우 운영
헤비 유저	5,000만~2억	$70~150+ (약 9.5만~20만 원+)	24시간 상주 비서 활용

출처: Anthropic·OpenAI·Google·DeepSeek·Alibaba 공식 가격표 종합, 저자 정리 (2026년 기준 추정). 토큰 100만 개당 혼합 단가를 약 1달러(캐시 할인·경량 모델 혼용)로 가정한 추정치이며, 프리미엄 모델만 쓰면 몇 배로 뜁니다. 원화는 $1≈1,350원 환산.

에이전트는 사람처럼 간헐적으로 AI를 쓰지 않습니다. 24시간 쉬지 않고 토큰을 소비합니다. 왜 이렇게 소비가 폭증할까요?

도구를 호출하고 결과를 받아 판단하는 한 번의 사이클마다 입력 약 1,000토큰, 출력 약 500토큰이 나갑니다. 이 사이클을 하루에 수백~수천 번 돌린다고 생각해보세요. 여기에 이미지를 분석하는 비전 기능은 텍스트 대비 5~10배의 토큰을 잡아먹고, 맥락을 유지하려고 매 요청마다 참조하는 이전 대화 기록(200K 토큰 설정이 흔합니다)까지 쌓이면, 비용은 순식간에 불어납니다.

전기 요금이 비싸면 공장을 돌릴 수 없듯, 토큰이 비싸면 에이전트를 돌릴 수 없습니다. 에이전트 시대는 곧 토큰 소비 폭증의 시대이고, 토큰 비용은 곧 기업과 국가의 AI 경쟁력을 좌우하는 핵심 변수가 됩니다. 그렇게 보면 요즘 중국 모델이 주목받는 까닭도 풀립니다.

왜 중국 LLM 회사들이 관심을 받는가. 앞서 요금표에서 보셨듯, 중국 모델들은 여전히 서구 프리미엄 모델보다 표시 단가가 크게 낮습니다. DeepSeek V4 Pro는 2026년 8월 값을 올린 뒤에도 출력 기준 2~4달러로, GPT-5.5(30달러)보다 87~93% 쌉니다. GLM-5.3도 출력 4.4달러 수준이지요. 그런데 앞에서 본 DeepSeek의 8월 인상을 떠올려 보세요. 수요가 몰리자 일부 요금이 하룻밤 새 열 배 넘게 뛰었습니다. 싼 공급자에 기댄 사업이라면, 단가가 언제든 바뀔 수 있다는 위험까지 장부에 적어 둬야 합니다. 그리고 이 숫자만으로 같은 일을 그만큼 싸게 끝낸다고 말할 수는 없습니다. 정확도, 필요한 추론 토큰, 재시도 횟수, 응답 속도를 합친 성공 작업당 총비용으로 비교해야 실제 경제성이 드러납니다.

이 격차가 글로벌 AI 시장의 판도를 흔드는 이유는 간단합니다. 에이전트 시대에는 토큰 소비량이 폭발적으로 늘어나는데, 운영비가 5분의 1로만 줄어도 에이전트를 쓸 수 있느냐 없느냐가 갈립니다. 중소기업이나 스타트업일수록 더 그렇지요. 그래서 고객 서비스 자동화, 문서 분류, 데이터 추출처럼 비용에 민감한 대량 처리 작업에서는 "프리미엄 지능"이 아니라 "충분히 좋으면서 저렴한 지능"이 필요합니다.

왜 독파모와 코파모가 필요한가

토큰을 소비할 때마다 그 비용은 어디로 갈까요? 모델을 만든 회사에 갑니다. GPT를 쓰면 OpenAI에, Claude를 쓰면 Anthropic에, Gemini를 쓰면 Google에. 모두 미국 회사이니, 이것은 에너지를 수입하는 것과 같습니다.

한국이 석유를 전량 수입하듯, 지금 한국의 AI 토큰도 사실상 전량 수입입니다. 한국 기업이 AI를 많이 쓸수록 달러가 빠져나가고, 에이전트 시대가 본격화되면 이 유출 규모는 감당하기 어려운 수준이 될 수 있습니다.

그래서 독파모(독자 파운데이션 모델)가 필요합니다. 외국 API에 의존한다는 것은 전력을 전량 수입에 의존하는 것과 같은 논리이지요. 독파모가 있으면 토큰 비용이 국내에 순환하고, 한국어 처리 품질이 올라가며, 데이터 주권을 지킬 수 있습니다. 특히 한국어는 영어와 구조가 달라서, 영어 중심으로 학습된 모델에 한국어를 넣으면 같은 내용이라도 모델에 따라 토큰이 최대 2~3배 더 듭니다. 한국어에 최적화된 독파모라면 이 비효율을 줄일 수 있습니다.

코파모(코딩 파운데이션 모델)도 마찬가지입니다. 2026년 현재 개발자의 85%가 AI 코딩 도구를 정기적으로 사용하는 만큼, 코딩 토큰은 이제 "산업용 전력"과 같습니다. OpenAI가 GPT-5.3-codex를 따로 만든 이유도 여기 있습니다. 범용 모델보다 코드 특화 모델이 같은 토큰으로 더 정확한 코드를 만들어내니까요. 한국도 코딩에 특화된 파운데이션 모델이 필요합니다.

정부가 추진하는 독파모 컨소시엄은 다섯 팀으로 출발해 두 차례 평가를 거쳤고, 2026년 8월 업스테이지·SK텔레콤·LG AI연구원 세 팀이 3차에 올랐습니다. 최종 두 팀은 2027년 2월에 가려지지요. 이 사업을 토큰 경제의 눈으로 다시 보면 뜻이 또렷해집니다. "우리도 AI 모델을 만들자"를 넘어 토큰 자급률을 높이자는 것이지요. 전력 자급률이 국가 안보이듯, 토큰 자급률이 AI 안보입니다. 물론 눈금 하나로 안보를 다 잴 수는 없으니, 이 개념은 두 장 뒤 「지능을 수출하는 나라」에서 더 넓은 의존성 지도 안에 놓고 다시 따져 봅니다.

독파모 이야기가 나오면 늘 따라오는 반문이 있지요. 거대 모델을 학습시키는 데 드는 그 큰돈을 감당할 수 있느냐는 것입니다. 답하기 전에 AI의 돈이 실제로 어디서 나가는지부터 볼 필요가 있습니다.

학습보다 훨씬 큰 추론: 숨겨진 비용 구조. 지금까지 AI의 비용이라고 하면 대부분 학습(training) 비용을 떠올렸습니다. GPT-4에 1억 달러 이상, DeepSeek-V3 최종 학습에 약 600만 달러(제1부 「지능의 단가가 빠르게 낮아지고 있다」에서 본 대로 전체 개발비는 아니지요). 뉴스에 나오는 숫자는 늘 학습 비용, 즉 발전소를 짓는 데 얼마가 들었느냐는 이야기였지요.

그런데 발전소는 짓는 것이 끝이 아닙니다. 지은 뒤에는 매일 전기를 생산해야 합니다. 이 운영 비용, 즉 추론(inference) 비용이 실은 학습 비용보다 압도적으로 큽니다. Introl은 OpenAI의 2024년 수치를 바탕으로 추론 비용이 학습 비용의 15배에서 118배가 될 수 있다고 정리합니다.[4] 모든 모델에 들어맞는 고정 배수는 아니지만 방향은 같습니다. 학습은 모델 하나를 만들 때 한 번 드는 비용이지만, 추론은 사용자가 질문할 때마다 계속 쌓이는 비용이기 때문입니다. 2030년까지 전체 AI 연산의 75%가 추론에 쓰일 것으로 예측되는 이유이기도 합니다.

그런데 여기서 순서가 뒤집힙니다. 2025년 ICLR 구두 발표(Oral) 논문 「Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters」[5]는, 어떤 문제에 얼마나 연산을 더 쓸지 잘 고르면 작은 모델이 자신보다 최대 14배 더 큰 모델보다 나은 성능을 내는 구간이 있음을 보였습니다. 물론 모든 문제에서 성립하는 공식은 아닙니다. 문제의 난도, 중간 답을 평가하는 방법, 샘플링 전략, 추론 예산에 따라 결과와 비용이 달라집니다. 작은 모델이 항상 더 싸다는 뜻도 아니지요. 중요한 통찰은 학습 때 크기만 키우는 길 밖에, 문제에 맞게 추론 연산을 배분하는 길도 있다는 것입니다.

후속 연구들도 같은 방향을 가리킵니다. 「The Art of Scaling Test-Time Compute for Large Language Models」[6]와 ICLR 2025의 「Inference Scaling Laws: An Empirical Analysis」[7]는 추론 연산 투입량과 성능 향상 사이의 스케일링 법칙을 실증적으로 분석했습니다. 모델을 키우는 길 옆에, 추론을 똑똑하게 쓰는 길이 하나 더 열려 있는 셈이지요.

한국에는 반가운 대목입니다. 수조 원을 들여 거대 모델 경쟁에 뛰어드는 길 말고도 길이 있다는 뜻이니까요. 추론 최적화는 작은 독파모의 성능과 비용을 끌어올릴 수 있는 후보 전략입니다. 거대 모델에 버금가는지는 문제 유형과 평가 방법, 투입한 추론 비용을 함께 재 봐야 알 수 있겠지만요. 반도체에서 한국이 메모리 분야를 집중 공략해 세계 1위가 된 것처럼, AI에서도 추론 최적화라는 특정 영역에 역량을 집중하는 것이 현실적인 전략이 될 수 있습니다.

돈이 새는 곳은 발전소를 지을 때보다 발전소를 돌릴 때입니다. 학습은 한 번이지만, 추론은 질문이 들어올 때마다 계량기가 돌아가니까요. 같은 발전소에서 같은 연료로 얼마나 많은 일을 해내느냐가 토큰 비용을 가릅니다.

토큰 요금이 경쟁 구도를 결정한다

결국 토큰 비용은 지능 에너지의 요금입니다. 프리미엄 지능(출력 20~50달러)과 경량·초저가 지능(0.5~2달러 안팎)의 격차는 열 배에서 백 배. 어떤 등급을 어떤 가격에 확보하느냐에 따라 기업의 전략과 국가의 산업 경쟁력이 갈립니다.

에이전트 시대가 본격화되면 모든 기업이, 모든 서비스가 24시간 토큰을 소비하게 됩니다. 전기 요금이 싼 나라에 공장이 몰리듯, 토큰 비용이 낮은 플랫폼에 서비스가 몰릴 것입니다. 나라 차원에서 독자 발전소(독파모)와 특화 발전소(코파모)를 갖추는 일은 선택지를 넓히는 보험입니다. 기업의 질문은 조금 다르지요. 우리 회사는 공급자를 바꿀 수 있는가, 성공 작업당 원가는 얼마인가를 먼저 물어야 합니다.

그런데 자급률이라는 눈금 하나만 들여다보면 놓치는 것이 생깁니다.

자급률만큼 의존도 확인도 중요합니다. 해외 API를 많이 쓰면 비용과 데이터, 서비스 연속성이 외부 공급자에게 묶일 수 있습니다. 이 문제를 '토큰 자급률'로 표현할 수 있지만, AI 안보를 평가하려면 국내 모델의 사용 비중과 함께 공급자 대체 가능성, 데이터 통제권과 서비스 연속성을 봐야 합니다.

국내 모델을 보유해도 GPU와 전력, 핵심 소프트웨어를 외부에 의존할 수 있습니다. 반대로 해외 모델을 사용하더라도 여러 공급자 사이를 전환하고, 데이터 반출을 제한하며, 자체 모델로 핵심 업무를 유지할 수 있다면 위험은 낮아집니다.

국가와 기업이 확인할 항목은 다음과 같습니다.

질문	확인할 증거
공급자가 중단되면 얼마나 버틸 수 있는가	대체 모델 전환 훈련과 복구 시간
한국어·산업 과업에서 경쟁력이 있는가	동일 데이터와 품질 하한을 사용한 평가
비용이 국내에 남는가	추론·클라우드·전력·인력 비용의 실제 지출처
민감한 데이터가 어디로 가는가	저장 위치, 재학습 여부, 삭제와 감사 기록

독자 파운데이션 모델과 코드 특화 모델은 이런 선택지를 넓히는 수단입니다. 그렇다고 '국산'이라는 딱지가 경제성과 안전성을 보장해 주지는 않지요. 공개된 평가와 안정적인 운영, 다른 모델로 옮길 수 있는 호환성이 함께 있어야 합니다.

단가표에서 가장 싼 줄을 고르는 것으로 끝나면 좋겠지만, 장부는 그렇게 쓰이지 않지요. 함께 적어야 할 것은 필요한 품질의 일을 끝내기까지 든 총비용, 공급자가 바뀌어도 버티는 운영 능력, 그리고 민감한 데이터를 통제할 수 있는 구조입니다.

출처: Anthropic·OpenAI·Google·DeepSeek·Alibaba 공식 가격표 종합, 저자 정리 (2026년 기준 추정)

1. 한국전력 관련 증권사 산업 리포트 (2026.8.27, https://stock.pstatic.net/stock-research/industry/63/20260827_industry_418638000.pdf) — 2022~2024년 연속 인상 뒤 2025년 산업용 평균 판매단가 kWh당 약 181.9원 인용.

2. Epoch AI, "LLM Inference Price Trends" (https://epoch.ai/data-insights/llm-inference-price-trends)

3. "Models & Pricing" (DeepSeek API Docs, 2026.10.22 확인, https://api-docs.deepseek.com/quick_start/pricing); "DeepSeek raises some V4 prices by more than 10x as AI demand strains capacity" (InfoWorld, 2026.8, https://www.infoworld.com/article/4209439/deepseek-raises-some-v4-prices-by-more-than-10x-as-ai-demand-strains-capacity.html) — V4-Pro-0813 출시와 함께 피크·비피크 요금제를 도입하고 V4 계열 가격을 50~1,100% 인상.

4. AI Inference vs Training Infrastructure: Why the Economics Diverge (Introl, 2026.4.28, https://introl.com/blog/ai-inference-vs-training-infrastructure-economics-diverging-2025) — OpenAI의 2024년 수치를 인용해 추론 비용이 학습 비용의 15배에서 118배가 될 수 있다고 정리한다. 이는 모든 모델과 서비스에 적용되는 고정 배수가 아니라 비용 구조와 사용량에 따라 달라지는 범위다.

5. C. Snell et al., "Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters," ICLR 2025 Oral, https://openreview.net/forum?id=4FWAwZtd2n

6. S. Sun et al., "The Art of Scaling Test-Time Compute for Large Language Models," arXiv:2512.02008, https://arxiv.org/abs/2512.02008

7. Y. Wu et al., "Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving," ICLR 2025, https://openreview.net/forum?id=VNckp7JEHn

AI 생산원가: 토큰값이 아니라 일값

자동차 회사의 원가 담당자는 철광석 시세만 보지 않습니다. 철광석이 강판이 되고, 강판이 차체가 되고, 차체가 완성차가 되는 동안 각 단계에서 원가가 차곡차곡 쌓이니까요. 원재료가 아무리 싸도, 공정이 비효율적이면 완제품은 비쌉니다.

지능 에너지도 똑같습니다. 앞 장에서 토큰의 가격을 봤지요. 그때 토큰은 전기의 kWh처럼 사용량과 요금을 세는 단위였지요. 그런데 일을 만드는 쪽의 장부로 자리를 옮기면 토큰은 이번엔 원자재가 됩니다. 토큰 가격은 철광석 시세, 그러니까 원자재 가격인 셈이지요. 기업이 정말 알아야 하는 것은 완제품 가격입니다. "이 일 하나를 AI에게 시키면, 다 합쳐서 얼마가 드는가?"

전기에서 일까지, 원가는 여덟 층을 타고 쌓입니다.

층	내용	제조업의 대응물
에너지	전기요금(kWh)	연료비
GPU	감가상각·냉각·네트워크	설비 투자
클라우드	운영과 마진	공장 임대료
모델	개발·학습비의 상각	금형 제작비
추론	토큰 생성 비용	가동 비용
도구	검색·API·코드 실행	외주 가공비
워크플로	에이전트 조율·메모리·재시도	공정 관리비
사람	검토·승인·예외 처리	검수 인건비

한 작업의 원가는 이 여덟 층의 합입니다. 표에 나오는 '상각'이란 한 번에 든 큰 비용을 수많은 작업에 나눠 얹는 셈법을 말합니다. 그리고 눈여겨볼 것이 있습니다. 위층, 그러니까 그림에서는 꼭대기, 표에서는 맨 아래에 놓인 사람 층으로 갈수록 한 층의 단가가 커진다는 점입니다. 전기 한 줌은 0.1원 단위지만, 전문가의 검토 10분은 만 원 단위니까요.

한 가지 짚어 둘 것이 있습니다. 이 여덟 층은 "누가 내느냐"와 무관한 원가 발생 지점의 지도입니다. API로 토큰을 사 쓰는 기업의 장부에는 아래 다섯 층이 토큰 가격 하나로 묶여 청구됩니다. 완성차 값에 철광석부터 도장까지가 이미 들어 있듯이요. 층마다 따로 지불하는 것은 자체 모델을 돌리는 기업뿐입니다. 그래서 이 여덟 층은 곧 구매와 구축의 손익분기표이기도 합니다. 작업량이 적을 때는 다섯 층이 묶인 토큰 가격을 내는 쪽이 싸고, 작업량이 임계점을 넘으면 아래 다섯 층(특히 GPU 감가상각)을 직접 떠안고 발전소를 돌리는 쪽이 싸지지요. 앞 장에서 본 "독자 모델이냐 수입이냐"라는 국가의 셈법이, 회사 장부 안으로 들어온 셈입니다.

여덟 층의 원가 구조

전기에서 일까지 — 위층으로 갈수록 한 층의 단가가 커진다

완제품 원가 성공 작업당 원가 = 전체 실행 비용 ÷ 성공한 작업 수

한 작업의 원가 = 여덟 층의 합 · 실패한 실행도 비용을 쓴다

단가 큼 ▲

층		내용	제조업의 대응물
8	**사람**	검토·승인·예외 처리	검수 인건비
7	**워크플로**	에이전트 조율·메모리·재시도	공정 관리비
6	**도구**	검색·API·코드 실행	외주 가공비
5	**추론**	토큰 생성 비용	가동 비용
4	**모델**	개발·학습비의 상각	금형 제작비
3	**클라우드**	운영과 마진	공장 임대료
2	**GPU**	감가상각·냉각·네트워크	설비 투자
1	**에너지**	전기요금(kWh)	연료비

▼ 단가 작음

토큰 가격 하나
API로 사 쓰는 기업에겐 아래 다섯 층이 묶여 청구된다

토큰 가격은 **원자재 시세**, 성공 작업당 총비용이 **완제품 원가**

출처: 저자 작성 — 본문 제2부 「AI 생산원가: 토큰값이 아니라 일값」의 여덟 층 원가 구조. 층의 순서는 원가가 쌓이는 방향이며, 칸의 크기는 실제 원가 비중이 아닙니다.

그림 2-13 에너지에서 사람까지 여덟 층을 아래에서 위로 쌓았습니다. 위층으로 갈수록 한 층의 단가가 커지고, API로 사 쓰는 기업에게는 아래 다섯 층이 토큰 가격 하나로 묶여 청구되지요.

성공한 작업 하나의 값

여기에 한 가지 셈이 더 붙습니다. AI는 가끔 실패합니다. 그리고 실패한 실행도 돈을 씁니다.

작업 한 번에 1,000원이 들고 성공률이 80%라면, 성공한 작업 하나에 실제로 든 돈은 1,250원입니다. 실패한 두 번의 실행 비용을 성공한 여덟 번이 나눠 지니까요. 성공한 것만 골라 세면 장부가 예뻐 보일 뿐이지요.

성공 작업당 원가 = 전체 실행 비용 ÷ 성공한 작업 수

토큰 단가가 내려갔다고 전체 비용도 같은 비율로 내려가는 것은 아닙니다. 입력이 길어지고, 더 많은 후보를 생성해 비교하고, 검색과 코드 실행을 여러 번 붙이면 작업당 사용량이 커집니다. 실패 뒤

재시도와 사람의 검토까지 더하면 청구서의 모양은 다시 달라지지요. 그래서 구매 담당자가 나란히 놓고 비교할 값은 백만 토큰 가격보다 성공한 작업 하나를 얻는 데 든 총비용입니다.

감을 잡기 위해 가정값으로 셈을 해 봅시다. 아래 숫자는 실측이 아니라 저자가 만든 예시입니다. 토큰 비용 칸은 앞 장 요금표(2026년 10월 기준)의 표준 등급, 그러니까 입력 100만 토큰당 2달러·출력 10달러 안팎의 단가에 검색·도구 호출비를 조금 얹어 어림했습니다. 경량 모델로 바꾸면 이 칸은 수십 분의 1로 줄어들지요. 실제 값은 모델 단가, 성공률, 검토 방식에 따라 크게 달라집니다.

작업 (가정 예시)	토큰·도구 비용	사람 검토	성공률	성공 작업당 원가
상담 요약	$0.02	$0.02	95%	약 $0.04
설비 이상 진단 보고서	$0.4	$0.2	90%	약 $0.7
코드 작성(리뷰 포함)	$4	$3	80%	약 $9
의료 영상 판독 초안	$1	$8	95%	약 $9.5

표에서 보이는 패턴이 하나 있습니다. 작업이 무거워질수록 원가를 쥐고 흔드는 쪽은 토큰보다 사람입니다. 의료 판독 초안의 원가에서 토큰은 조연이고, 전문의의 검토가 주연이지요. 토큰 가격이 아무리 떨어져도 이 구조는 그대로입니다. 생성은 싸지고 검증이 귀해진다는 이야기를 제5부 「테스트와 리뷰를 잘하는 법」에서 다시 만나게 될 텐데, 그 현상을 원가 장부에 적으면 바로 이 표가 됩니다.

이 표를 감상만 하고 넘기지 않는 방법이 있습니다. 여러분 회사에서 가장 자주 반복되는 작업 세 가지를 골라 같은 칸을 채워 보세요. 성공 작업당 원가에 월 작업량을 곱하면 그대로 월 운영 예산이 되고, 어느 층이 가장 두꺼운지가 보이면 어디부터 손봐야 할지도 함께 보입니다.

그렇다면 단가가 떨어지면 장부도 그만큼 가벼워질까요? 제1부에서 본 제번스 역설이 여기서도 작동합니다. 효율이 좋아지면 소비가 줄 것 같지만, 새로운 용도가 생겨 총소비가 오히려 늘기도 하지요.

AI에서도 같은 가능성이 보입니다. 이메일 한 통의 초안을 만드는 비용이 낮아지면 이메일을 더 많이 만들고, 코드 생성이 빨라지면 이전에는 경제성이 없던 내부 도구까지 제작합니다. 한 번 실행하는 비용은 내려가도 실행 횟수와 작업 범위가 더 빨리 늘면 월 총비용과 전력 사용량은 오를 수 있습니다.

가격이 내려간 건 사실입니다. 단가가 줄었다고 예산까지 줄지는 않는다는 것뿐이지요. 비용을 관리하려면 작업별 한도, 재시도 횟수, 사람 개입률, 월간 성공 작업 수를 함께 봐야 합니다. 싸진 지능이 만든 가치는 늘리되, 필요 없는 생성까지 늘어난 것은 아닌지 장부에서 확인해야 하지요.

끊기지 않는 전기, 믿을 수 있는 지능

전기 시대의 공장장들은 일찍부터 알고 있었습니다. 요금표만큼 무서운 게 정전이라는 것을요. 찰나의 정전에도 반도체 공장은 라인 위의 웨이퍼를 통째로 버립니다. 그래서 정밀 공정일수록 무정전 전원

장치와 비상 발전기에 기꺼이 웃돈을 내지요. 전기 자체가 더 좋아지는 게 아니라, 끊기지 않게 지키는 데 돈을 쓰는 겁니다.

지능은 여기서 전기와 갈라집니다. 콘센트의 전기는 어느 발전소에서 왔든 같은 일을 하지만, 지능은 모델마다 해내는 일의 품질이 다르지요. 정확도, 일관성, 안전성 말입니다. 품질이 낮은 지능은 성공률을 갉아먹고, 재시도를 늘리고, 무엇보다 사람 검토를 두껍게 만듭니다. 그런데 방금 봤듯이 사람이야말로 가장 비싼 층이지요.

가정값으로 비교해 봅시다. 토큰이 싼 모델 A는 실행당 100원이지만 성공률이 60%라 건마다 사람이 꼼꼼히 검토해야 합니다(건당 2,000원). 토큰이 세 배 비싼 모델 B는 실행당 300원이지만 성공률이 95%라 표본만 검토합니다(건당 200원). 여기서 두 검토 단가는 표본 검토까지 반영해 실행 한 건에 평균으로 얹히는 값입니다. 성공 작업당 원가를 셈하면 A는 약 3,500원, B는 약 530원. 토큰값이 세 배 비싼 쪽이 완제품값은 예닐곱 배 싼 셈이지요. 가정값으로 감을 잡아 본 것이지만, 어느 쪽으로 기우는지는 충분히 보입니다.

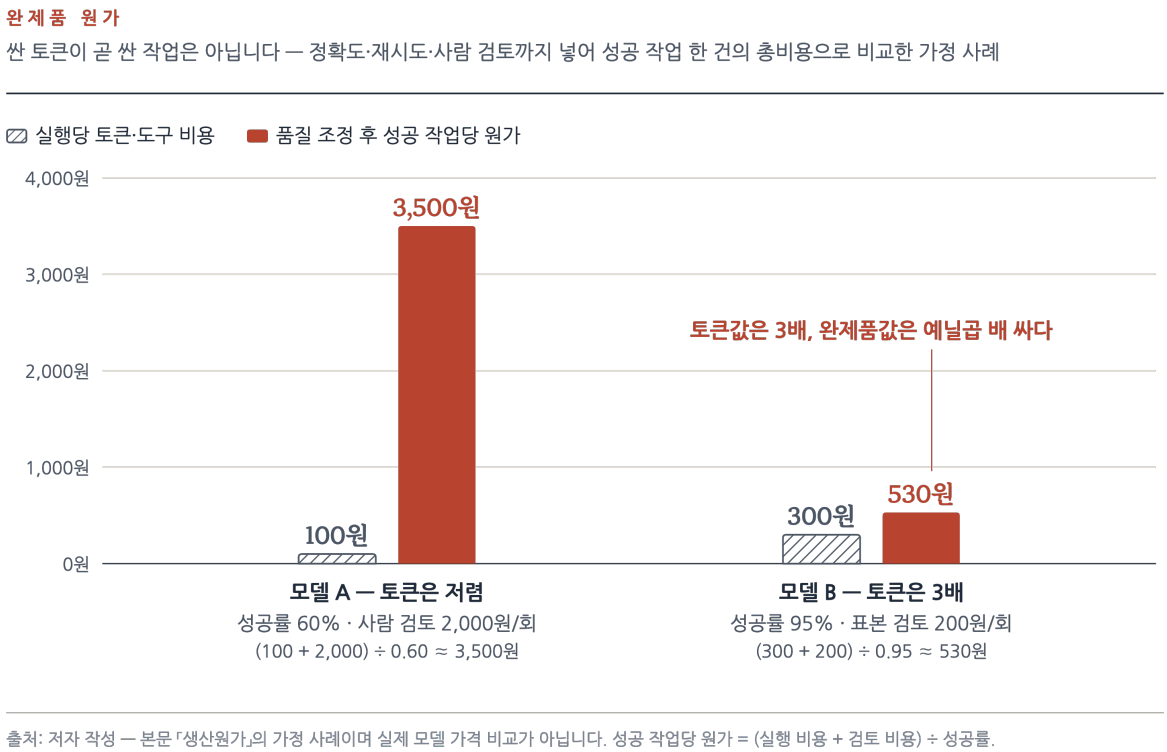


그림 2-14 모델마다 실행 한 번의 토큰·도구 비용과, 검토비와 성공률까지 넣은 성공 작업당 원가를 나란히 세웠습니다. 토큰값만 보면 A가 싸지만, 성공 작업당 원가에서는 B가 예닐곱 배 싸지요.

"가장 싼 토큰을 사라"가 늘 맞지 않는 까닭이 여기 있습니다. 원가를 줄이려는 사람이 물어야 할 것은 "품질과 성공률을 반영한 성공 작업당 원가가 가장 낮은 조합은 무엇인가"이고, 그 답은 작업마다 다릅니다. 제4부에서 만날 두 원칙이 정확히 이 지점에 서 있습니다. 변동을 전제로 품질을 관리하라는 「확률을 경영하라」, 그리고 대형과 소형을 섞어 쓰라는 「최적 조합을 찾아라」. 제5부의 산업별 특화 모델 이야기도 원가 장부에서 보면 결국 이 품질의 문제입니다. 수술실이 무정전 설비에 웃돈을 내듯, 규제

산업은 믿을 만한 지능에 웃돈을 내는 쪽이 오히려 싸게 먹히지요. 그렇다고 의료처럼 검토의 바닥을 법이 정해 둔 곳에서 사람의 검토를 걷어낼 수는 없습니다. 거기서 좋은 지능이 아껴 주는 것은 검토비보다 재작업과 사고의 비용입니다. 앞의 표에서 의료 판독이 성공률 95%인데도 검토비가 가장 두꺼웠던 이유이지요. 법이 굳이 정해 두지 않았더라도, 실패가 되돌릴 수 없는 일이라면—안전이든 금전이든 신용이든—표본 검토로 검토비를 낮추기 전에 실패의 무게부터 달아 봐야 합니다. 제4부 「신뢰를 내장하라」에서 짚겠지만, 신뢰는 원가를 깎고 나서 얹는 장식보다, 처음부터 설계에 들어가는 층에 가까우니까요.

연비 라벨과 국가의 원가표

가전제품에는 에너지 소비효율 등급이 붙어 있습니다. 소비전력이 몇 와트인지, 1년 쓰면 전기요금이 얼마인지. 그런데 AI 서비스에는 아직 "100만 토큰당 몇 달러"라는 원자재 시세표만 붙어 있습니다.

자동차 시장이 겪은 전환을 떠올려 보세요. 예전에는 배기량을 자랑했지만, 지금 소비자가 보는 것은 연비입니다. AI도 같은 길을 갈 겁니다. 매개변수 수를 자랑하다가 토큰 단가를 비교하게 됐고, 다음 차례는 성공 작업당 비용입니다. 성공률, 평균 완료 시간, 작업당 비용, 작업당 에너지, 사람 개입률. 이런 항목이 적힌 라벨이 에이전트에 붙는 날이 머지않았습니다. 지능이라는 가전제품에도 효율 등급표가 필요하니까요.

라벨이 회사 하나의 연비를 보여 준다면, 나라 전체의 연비는 어디서 읽을 수 있을까요? 이 여덟 층의 원가 구조는 기업의 장부이면서 동시에 국가의 정책 지도이기도 합니다. 원전과 태양광은 에너지 층의 원가를, AIDC 투자는 GPU와 클라우드 층을, 독자 파운데이션 모델은 모델 층을, 공공 API는 도구 층을, MCP 같은 표준은 워크플로 층을, 교육과 규제 정비는 사람 층의 원가를 낮춥니다. 대한민국 AI 행동계획의 언어로 옮기면 더 구체적입니다. GPU 확충(과제 [1])과 AI 데이터센터 생태계(과제 [2])는 GPU·클라우드 층에, 국가데이터 통합플랫폼(과제 [4])은 모델 층의 연료에, 독자 파운데이션 모델(과제 [27])은 모델 층에, 공공데이터 머신리더블화와 범정부 AI 공통기반(과제 [50])은 도구 층에 놓이지요. 정책과 층을 일대일로 못 박은 것은 아니고, 큰 줄기를 보여 주는 예시적 매핑입니다. 산업혁명 시대의 국가 경쟁력이 전기요금과 물류비에서 갈렸듯, 지능 시대의 경쟁력은 이 여덟 층의 합, 곧 작업당 원가에서 갈릴 것입니다. 제7부에서 볼 대한민국의 정책들을 이 원가표 위에 얹어 보면, 어느 층이 두텁고 어느 층이 비어 있는지가 보입니다.

자동차 회사의 원가 담당자가 철광석 시세표 옆에 공정표를 펼쳐 두듯, 이제 AI를 쓰는 회사에도 토큰 시세표 옆에 놓을 작업 장부가 필요합니다. 다음 달 청구서가 오기 전에, 여러분 회사의 첫 줄을 채워 보면 어떨까요?

지능을 수출하는 나라

1964년, 한국의 1인당 국민소득은 100달러대 초반에 불과했습니다. 발전소를 늘리는 데 해외 원조와 차관이 절실했고, 제한송전과 정전이 일상이었지요.[1] 그런데 60년 뒤, 한국은 원자로를 수출하는 나라가 됐습니다. 2009년 UAE 바라카에서 원전 4기를 따냈고, 2024년에는 체코 신규 원전 사업의 우선협상대상자로 선정됐습니다. 체코 정부가 당시 제시한 예상 사업비는 원전 2기 기준 약 24조 원이었습니다.[2] 전력 공급을 외부 지원에 기대던 나라가, 발전소를 파는 나라가 된 것입니다.

지금 한국은 AI 앞에서 비슷한 갈림길에 서 있습니다. 이번에 오가는 것은 전기 대신 지능이지요. 한국은 지능을 수입만 하는 나라로 남을 것인가, 아니면 지능을 수출하는 나라가 될 것인가.

앞의 장들에서 우리는 미국이 민간 자본으로 지능 발전소를 짓는 방식, 중국이 국가가 여섯 갈래 그물을 짜는 방식, 한국이 세 개의 메가 프로젝트로 응답한 모습을 봤습니다. 소버린 AI가 왜 오일쇼크의 재판과 닮았는지도 짚었지요. 그 조각들을 모아 물어봅시다. 한국이 지능 수출국이 되려면, 무엇이 있어야 할까요?

"지능을 전기처럼 쓴다"는 말은 이제 비유만은 아닙니다. OpenAI가 몰타 정부와 맺은 협력은 모든 시민에게 1년간 ChatGPT Plus 접근권을 제공하고, 그 전에 AI 리터러시 과정을 붙였습니다. 도구를 나눠주는 일과 쓸 수 있는 능력을 함께 설계한 것이지요.[3]

NVIDIA의 DSX도 같은 방향을 보여줍니다. AI 팩토리를 설계하고, 배치하고, 운영하는 공통 플레이북을 만들며, 메가와트당 토큰 성능과 토큰 비용을 관리 대상으로 둡니다.[4] 지능 유틸리티는 이제 전력 → 컴퓨트 → 토큰 → 서비스 → 성과와 접근으로 이어지는 운영 사슬이 되고 있습니다.

여기서 조심할 점이 있습니다. 토큰은 출력 단위입니다. 공장 굴뚝의 연기가 많다고 좋은 물건이 많이 나온 건 아니듯, 토큰이 많이 나왔다고 쓸모 있는 지능이 그만큼 생긴 것도 아니지요. 고객의 작업이 끝났는지, 품질이 유지됐는지, 비용과 책임을 감당할 수 있는지가 마지막 단위가 되어야 합니다.

AI 팩토리는 토큰 공장이다

젠슨 황은 2026년 3월 미국 산호세 GTC 무대에서 데이터센터를 다시 정의했습니다. 더 이상 서버를 모아 놓은 IT 시설이 아니라, 원재료를 넣으면 산출물이 나오는 공장이라는 것이지요. 원재료는 데이터, 동력은 컴퓨팅과 그 컴퓨팅을 돌리는 전기, 산출물은 토큰입니다. 발전소가 연료를 넣어 전기를 뽑아내듯, AI 팩토리는 데이터와 전기를 넣어 토큰을 생산합니다. 황은 전력 1GW가 물리적 한계로 주어졌을 때 더 많은 토큰을 만드는 처리량이 곧 공장의 생산성과 매출이 된다고 설명했습니다.[5]

이 공장의 몸값은 결코 작지 않습니다. 황이 제시한 투자수익 예시에서는 1GW급 AI 팩토리의 구축비가 약 1,000억 달러(약 135조 원), 여기서 판매할 수 있는 토큰 상품의 잠재 매출이 연간 최대 약 1,500억

달러(약 202조 원)로 제시됐습니다.[5] 실제 비용과 매출은 부지, 전력, 장비 구성, 가동률, 모델과 토큰 가격에 따라 달라집니다. 모든 1GW 시설이 이만큼 번다는 보장은 없고, NVIDIA가 AI 인프라 투자의 경제성을 설명하려고 든 가정이니 확정 수익이 아닌 잠재치로 읽어 주세요.

앞서 소버린 AI와 미·중·한 세 나라의 인프라 경쟁에서 나온 숫자들을 떠올려 보세요. 스타게이트의 최대 5,000억 달러, 중국 여섯 갈래 그물의 2026년 7조 위안 이상[추정], 한국 AIDC의 550조 원 장기 구상입니다. 기간과 범위, 확정 수준이 서로 달라 단순히 더할 수는 없습니다. 특히 중국의 여섯 갈래 그물은 전력·컴퓨트뿐 아니라 수자원·물류망까지 포함합니다. 그래도 AI와 직접 연결된 투자만 떼어 놓고 보면, 데이터와 전기를 토큰과 서비스로 바꾸는 생산 기반을 선점하려는 경쟁이라는 공통점이 보입니다.



그림 2-15 세 나라를 투자액, 전력·컴퓨팅 규모, 목표 시점의 세 패널에 나란히 놓았습니다. 패널마다 단위가 다르니 가로로 더하거나 순위를 매기지 말고, 패널 하나씩 따로 읽어 주세요.

황이 같은 GTC에서 보여 준 세계 지도에는 수십 개 나라의 소버린 AI와 AI 팩토리 구상이 표시됐습니다.[6] 이들 사업도 발표, 건설, 가동 상태를 구분해 봐야 합니다. 그러니 '지능을 수출하는 나라'라는 표현은 은유에만 머물지 않습니다. 토큰이라는 산출물을 국경 너머에 제공하고, 그 토큰이 현지의 실제 문제를 해결하도록 반도체·데이터센터·모델·운영 체계를 함께 파는 산업정책의 대상입니다.

그래서 이 책에서 말하는 '지능 수출'은 토큰을 잔뜩 만들어 국경 밖으로 내보내는 일보다 넓습니다. 고객의 문제를 해결하는 모델, 산업용 에이전트, 데이터 운영, 인프라, 표준, 안전 체계를 함께 건네는 일이지요. 셈하는 단위도 산출물의 양보다 끝낸 업무와 책임이 되어야 합니다.

그럼 그 꾸러미에는 무엇이 들어갈까요? AI 수출은 여러 층을 묶은 꾸러미에 가깝습니다.

층위	수출할 수 있는 것	구매자가 확인할 내용
반도체·장비	메모리, 가속기 부품, 냉각·전력 장비	공급 안정성, 효율, 유지보수
데이터센터	설계, 구축, 운영 서비스	전력과 물, 가동률, 지역 규제
모델	범용·언어·산업 특화 모델	실제 과업 품질, 비용, 배포 선택지
데이터	정제 도구, 합법적 데이터셋, 품질 관리	권리, 대표성, 갱신 절차
에이전트	제조·의료·행정 등 업무 솔루션	기존 시스템 연결, 책임, 예외 처리
표준·도구	평가, 프로토콜, 개발·운영 도구	상호운용성, 교체 가능성, 보안
신뢰	감사, 안전성 평가, 규제 대응	이의제기, 기록, 사고 대응

한 나라가 모든 층을 독자적으로 가질 필요는 없습니다. 어떤 층은 동맹과 시장에서 조달하고, 경쟁력이 있는 층에 집중하면 되지요. 그래도 핵심 의존성이 어디 숨어 있는지만큼은 알고 있어야 합니다.

미국과 중국, 서로 다른 꾸러미

미국은 풀스택을 공식 수출 상품으로 정의했습니다. 2025년 7월 행정명령으로 출범한 American AI Exports Program은 칩·서버·가속기, 저장장치·클라우드·네트워크, 데이터 파이프라인과 라벨링, 모델, 사이버보안, 산업별 응용을 하나의 패키지로 묶습니다. 제안 기업은 목표 국가뿐 아니라 데이터센터를 누가 짓고 소유하고 운영할지, 어떤 연방 지원이 필요한지, 수출통제와 최종사용자 규정을 어떻게 지킬지도 밝혀야 합니다.[7]

여기에 외교와 금융이 붙습니다. 미국 정부는 선정 패키지에 정부 간 협의, 수출허가 관련 우선 협의, 대출·보증·정치위험보험과 기술 지원을 연결할 수 있습니다. 제품을 파는 데서 끝나지 않고 미국의 기술, 표준과 거버넌스 모델이 구매국의 기반이 되도록 하겠다는 구상이지요.

2026년 4월부터 6월까지 진행된 첫 공모에는 78건이 접수됐습니다. 7월 6일 접수를 마감하고 관계부처 심사에 들어갔지만, 2026년 10월까지 확인한 범위에서 정부 지원 패키지의 최종 지정 결과는 아직 발표되지 않았습니다.[8] 지금 확인되는 것은 접수와 심사까지이지요.

미국 정책 안에는 촉진과 통제가 함께 있습니다. 국가별 3단계 할당을 두려던 2025년 AI Diffusion Rule은 시행 직전 비집행·폐기 방침으로 전환됐지만, 첨단 반도체와 재수출, 최종사용자에 대한 통제는 계속됩니다.[9] 미국식 풀스택을 도입하는 국가는 강력한 기술과 금융에 접근하는 대신, 향후 수출허가와 제재 정책의 변화도 운영 조건으로 받아들이게 됩니다.

태평양 건너 중국의 꾸러미는 모양이 다릅니다. 중국에는 미국의 프로그램과 이름과 구조가 같은 단일 AI 풀스택 수출 제도가 확인되지 않습니다. 대신 여러 정책과 기업 활동이 층층이 겹칩니다. 디지털 실크로드 협력은 통신망, 클라우드 데이터센터와 디지털 정부 기반을 다루고,[10] 2025년 글로벌 AI 거버넌스

행동계획은 청정전력, 차세대 네트워크, 지능형 컴퓨팅, 데이터센터, 컴퓨트 표준과 공동 연구실·인력 양성을 제안합니다.[11]

같은 해 발표된 AI+ 국제협력 이니셔티브는 보건·과학·산업·문화·인재 분야에서 정책과 솔루션을 공유하자고 제안했습니다.[12] 그 위에서 중국 기업의 통신장비와 클라우드, 산업 플랫폼, 공개 가중치 모델이 해외 시장에 들어갑니다. 초기 모델 사용료를 낮추고 현지 개발자가 손댈 여지를 넓히는 방식은 미국의 폐쇄형 프런티어 서비스와 다른 진입 경로를 만듭니다.

이 책은 이 조합을 중국의 풀스택 수출 전략으로 읽습니다. 중국 정부가 하나의 이름으로 공인한 프로그램은 없으니, 인프라 외교·산업 협력·기업 생태계를 엮어 본 저자의 분석이라고 봐 주세요. 미국이 정부가 지정하는 컨소시엄 패키지에 가깝다면, 중국은 기반시설과 역량 구축을 먼저 깔고 기업의 제품과 모델이 뒤따르는 모습에 가깝습니다.

두 나라의 꾸러미를 볼 때도, 우리 것을 셀 때도 놓치지 말아야 할 구분이 있습니다. 발표와 수출 실적은 다르다는 것이지요. 데이터센터 건설 계획, GPU 구매, 모델 공개는 수출의 준비 단계입니다. 해외 고객이 실제 업무에 도입하고, 계약을 갱신하며, 현지 규제와 장애 대응을 통과해야 수출 역량이 확인됩니다.

따라서 국가 전략에서도 다음 상태를 구분해야 합니다.

1. 계획: 예산과 목표가 발표되었다.
2. 확보: 전력, 반도체, 데이터, 인력이 계약되었다.
3. 구축: 시설과 모델, 운영 체계가 만들어졌다.
4. 채택: 외부 고객이 실제 업무에서 사용한다.
5. 운영: 품질과 비용, 안전을 지키며 계약이 이어진다.

계획 금액을 더해 국가 경쟁력처럼 제시하거나, 벤치마크 점수를 곧바로 수출 성과로 읽으면 이 단계를 건너뛰게 됩니다.

제조 현장을 풀스택으로 묶는 한국

한국의 AI 풀스택 수출은 하나의 완성된 정책 상품이라기보다 이미 착수한 독자 파운데이션 모델, 국산 AI 반도체와 K-Cloud 연구개발을 앞으로 수출하려는 AI 데이터센터·제조 AI 팩토리·피지컬 AI와 연결하는 전략입니다. 대한민국 AI 행동계획의 과제 [48]도 유망 개발도상국의 산업 구조·디지털 인프라·규제와 AI 수요를 조사하고, 풀스택 기업 컨소시엄을 구성해 수출 모델을 만든 뒤 외교부와 함께 대상국 협력을 넓히도록 정했습니다.[13]

2026년 2월 과학기술정보통신부와 국내 7개 기업은 사우디 아람코 디지털과 AI 풀스택 협력 MOU를 체결했습니다. 반도체, 산업 특화 LLM, 클라우드 인프라 운영을 현지 에너지·제조 수요에 맞춰 제안하는 구조입니다.[14] 의미 있는 진입이지만 MOU는 협력 범위를 찾는 출발점이지, 구축이 끝난 수출 실적이나 반복 매출을 뜻하지는 않습니다.

2026년 제조AI 전략은 제조 데이터, 파운데이션 모델, 로봇과 공정 운영을 묶은 풀스택 AI 팩토리를 수출 상품으로 육성하겠다고 밝혔습니다.[15] 피지컬 AI 전략도 모델과 온디바이스 컴퓨팅, 통신, 보안, 시스템 통합을 연결하는 국내 생태계를 수출 기반으로 제시합니다.[16] 아직은 다수 항목이 개발·실증·시장 개척 단계입니다.

한국은 메모리 반도체와 제조업, 통신 인프라, 공공 디지털 서비스 경험을 함께 갖고 있습니다. 이 자산을 범용 모델 순위 경쟁에만 투입할 필요는 없습니다. 조선, 반도체, 자동차, 소재처럼 현장 데이터와 검증 인력이 있는 분야에서 산업용 모델과 에이전트를 묶어 제공하는 길이 있습니다.

예를 들어 조선소의 용접 품질 시스템을 수출하려면 모델 하나로는 부족합니다. 카메라와 센서, 데이터 규격, 작업자 화면, 결함 발생 때의 중단 절차, 현지 언어 지원, 유지보수 계약이 함께 가야 합니다. 이런 묶음은 모델 API보다 판매가 어렵지만, 현장에 자리 잡으면 단순한 가격 경쟁을 피하기 쉽습니다.

모델 성능도 한국어만 잘한다고 충분하지 않습니다. 구매국의 언어와 법, 작업 관행에 맞게 조정할 수 있어야 합니다. 현지 파트너가 데이터를 통제하고 모델을 교체할 수 있는 배포 선택지를 제공해야 '소버린 AI' 수요와도 맞닿습니다.

외교 노선과 의존성 지도

한 나라의 공공행정, 병원, 항만과 공장이 어느 스택 위에서 돌아가는지는 장바구니에 담는 물건 고르기와는 무게가 다릅니다. 칩 공급, 보안 업데이트, 클라우드 관리 화면과 모델 규칙이 한 생태계에 묶이면 그 나라의 핵심 인프라 운영 방식이 오랫동안 고정됩니다. 공급국은 이를 시장 진출로 보지만, 구매국은 외교 정렬과 국가안보의 문제로 받아들일 수 있습니다.

쟁점	외교적으로 민감한 이유	산업적으로 확인할 질문
반도체·수출통제	제재와 최종사용자 규정이 제3국 시설에도 영향을 줄 수 있다	규정이 바뀌면 대체 칩과 유지보수 경로가 있는가
데이터·암호키	저장 위치보다 누가 열람하고 외국 정부가 접근할 수 있는지가 중요하다	키, 로그, 학습 재사용과 삭제 권한을 누가 갖는가
모델·표준	한 공급국의 API와 평가 규칙이 행정·산업 표준으로 굳을 수 있다	모델과 클라우드를 다른 공급자로 교체할 수 있는가
사이버보안	원격 업데이트와 장애 대응 권한은 편리함이자 핵심 인프라 접근권이다	독립 감사, 사고 통지와 긴급 차단 권한이 있는가
전력·물	대형 AI 시설은 에너지 외교와 지역 자원 배분 문제를 만든다	장기 전력비, 물 사용과 주민 부담을 누가 책임지는가
금융·현지 산업	낮은 초기 금융비용이 장기 부채와 공급자 종속으로 이어질 수 있다	현지 조달, 고용, 기술이전과 총소유비용은 얼마인가

세계은행은 클라우드와 데이터 인프라를 고를 때 데이터가 적용받는 법률, 외국 정부의 접근 가능성, 연결 경로와 벤더 종속을 함께 살피라고 권고합니다. OECD도 국가 컴퓨트 역량을 용량만이 아니라 효과성, 회복력, 소유·통제와 지속가능성으로 평가해야 한다고 봅니다.[17]

그렇다고 미국이나 중국의 기술을 피하자는 말은 아닙니다. 어느 공급국을 택하든 협력의 이익과 중단 위험을 같은 장부에 적어 보자는 것이지요. 공개 가중치 모델을 쓰더라도 가속기, 컴파일러, 클라우드 관리와 보안 업데이트가 한 기업에 묶이면 종속은 남습니다. 반대로 해외 클라우드를 쓰더라도 암호키와 감사권, 데이터 이동성과 교체 절차를 확보하면 통제권을 높일 수 있습니다.

이런 장부를 나라 전체로 펼치면 토큰 경제 장에서 꺼낸 질문으로 돌아오게 됩니다. 이 책은 토큰 자급률을 AI 안보를 생각하기 위한 저자 제안으로 사용합니다. 그러나 국내에서 생성한 토큰 수를 국내 소비량으로 나눈 값만으로는 충분하지 않습니다.

같은 토큰이라도 과업의 가치와 품질이 다릅니다. 해외 모델을 국내 데이터센터에서 실행하면 국내 생산으로 볼지, 국내 모델을 해외 클라우드에서 실행하면 어떻게 셀지도 불분명합니다. 토큰 수를 늘리려고 쓸모없는 출력을 많이 만드는 왜곡도 생길 수 있지요.

따라서 단일 비율보다 의존성 대시보드가 현실적입니다.

의존 영역	점검할 지표의 예
연산	특정 국가·기업의 가속기와 클라우드 비중, 대체 시간
모델	핵심 업무별 대체 모델 수, 교체에 필요한 비용과 기간
데이터	국내 통제 데이터의 권리와 품질, 갱신 가능성
운영	장애 때 유지할 수 있는 필수 기능과 복구 시간
비용	토큰당 가격이 아니라 성공 작업당 총비용
수출	해외 유료 고객, 갱신률, 현지 운영 성과

토큰 자급률은 이 대시보드 안에서 보조 지표로 시험할 수 있습니다. 경계를 공개하고, 품질을 반영하며, 수치가 어떤 정책 결정을 바꾸는지 설명할 수 있을 때 의미가 생깁니다.

참조 현장에서 계약 종료까지

산업 AI를 수출하려면 국내에서 규모가 비슷한 현장을 먼저 통과해야 합니다. 시연 한 번으로는 부족하고, 계절과 설비 상태가 바뀌어도 품질이 유지되는지, 숙련자가 바뀌어도 운영할 수 있는지, 장애와 보안 사고를 어떻게 처리하는지 보여 주는 참조 현장입니다.

그다음에는 제품을 분리 가능한 묶음으로 만들어야 합니다. 고객이 데이터와 모델, 클라우드를 전부 한 공급자에게 맡기지 않아도 되도록 연결 규격과 반출 절차를 제공합니다. 현지 기업이 운영에 참여하고 지식을 이전받을 구조도 필요합니다.

마지막은 신뢰입니다. 수출국의 규제를 맞추는 데서 그치지 말고 구매국 시민과 노동자의 권리까지 지켜야 하지요. 자동 결정에 이의를 제기할 통로, 인간의 재검토, 보안 사고 통지, 데이터 삭제와 계약 종료 뒤의 처리까지 수출 계약에 들어가야 합니다.

여기에 국가 차원의 검토가 더해져야 합니다. 어느 시장을 우선할지 정하는 외교, 외국산 기술이 들어간 부품명세와 수출통제 심사, 데이터 현지화와 디지털 통상, 장기 전력 조달, 수출금융과 정치위험보험, 사이버 대응과 이중용도 심사를 한 테이블에서 다뤄야 하지요. AI 담당 부처 하나만의 사업으로는 풀스택을 다룰 수 없습니다.

개발협력도 조심해야 합니다. ODA를 수출 교두보로만 보면 수원국의 필요보다 공급국 기업의 매출이 앞설 수 있습니다. 한국의 AI 개발협력은 현지 인재, 데이터 통제권과 자립적 운영 역량을 남기는지로 평가해야 합니다.[18] 보조금과 저리금융을 제공하더라도 총부채, 환율, 유지보수비와 계약 종료 뒤의 비용을 공개해야 외교적 신뢰를 지킬 수 있습니다.

좋은 전략은 계약 종료 장면까지 설계합니다. 수출통제가 바뀌거나 외교 관계가 악화됐을 때 필수 서비스를 얼마나 오래 유지할 수 있는가. 모델과 데이터를 어떤 형식으로 돌려줄 것인가. 현지 운영팀이 다른 공급자로 전환할 수 있는가. 이 질문에 답하지 못한다면 그것은 풀스택 수출이 아니라 풀스택 종속일 수 있습니다.

1964년의 한국은 발전소를 늘릴 돈부터 밖에서 빌려야 했습니다. 45년 뒤 바라카에서 원전을 따낸 한국은 그 반대편에 섰지요. 지능을 수출한다는 것도 그 자리에 서는 일인데, 하나가 더 붙습니다. 남의 나라 문제를 풀어 주면서, 그 나라가 열쇠를 계속 쥐게 해 주는 일 말입니다. 조선소와 공장에서 쌓은 한국의 경험이 그런 제품과 계약으로 바뀔 수 있을까요?

---

1. "GNI per capita, Atlas method (current US$) — Korea, Rep." (World Bank, https://api.worldbank.org/v2/country/KOR/indicator/NY.GNP.PCAP.CD?date=1964:1964&format=json); 「전기」 (한국민족문화대백과사전, https://encykorea.aks.ac.kr/Article/E0049305) — 세계은행의 현재 시계열은 1964년 한국의 1인당 GNI를 140달러로 기록한다. 한국전력 설립 당시 발전설비는 추정 수요에 못 미쳤고, 광복 뒤 되풀이된 제한송전은 1964년 4월에야 일단 해제됐다.

2. "2015 Sustainability Report" (Emirates Nuclear Energy Corporation, https://www.enec.gov.ae/doc/enec-sustainability-report-2015-58a2bac9d5bd5.pdf); "체코 신규원전 건설사업 우선협상대상자 선정 관련" (산업통상자원부, 2024.7.18, https://www.korea.kr/news/policyNewsView.do?newsId=156641684) — 바라카 사업은 한국전력 컨소시엄이 2009년 수주한 원전 4기 건설 사업이며, 체코의 약 24조 원은 2기 기준 당시 예상 사업비다. 한수원과의 계약금액은 우선협상대상자 선정 뒤 협상으로 정하도록 발표됐다.

3. "OpenAI and Malta partner to bring ChatGPT Plus to all citizens" (OpenAI, 2026.5.16, https://openai.com/index/malta-chatgpt-plus-partnership/) — 몰타 시민 대상 1년간 ChatGPT Plus 접근과 AI 리터러시 과정을 함께 제공하는 OpenAI for Countries 사례다.

4. "NVIDIA DSX Gives Infrastructure Builders the Playbook for AI Factories" (NVIDIA Newsroom, 2026.5.31, https://nvidianews.nvidia.com/news/dsx-infrastructure-ai-factory) — DSX를 AI 팩토리의 설계·배치·운영 플레이북으로 설명하고, 토큰 비용과 메가와트당 토큰 성능을 운영 지표로 제시한다.

5. "GTC 2026 Keynote" (NVIDIA, 2026.3, https://www.nvidia.com/en-us/on-demand/session/gtc26-s81595/) — 젠슨 황은 전력 1GW가 제한된 AI 팩토리에서 토큰 처리량이 생산량과 매출을 좌우한다고 설명했다. 1GW 시설의 약 1,000억 달러 구축비와 최대 연 1,500억 달러 잠재 매출은 황이 제시한 투자수익 사례이며, 별도의 분석은 이를 확정 수익이 아닌 잠재치로 정리한다 (https://www.janushenderson.com/en-hk/investor/article/quick-view-tokensrevenues-nvidias-case-for-scalable-ai-returns/).

6. "GTC 2026 Keynote" (NVIDIA, 2026.3, https://www.nvidia.com/en-us/on-demand/session/gtc26-s81595/); "Every Country Needs Sovereign AI. Jensen Is Selling It." (Julien Simon, 2026.3.20, https://www.julien.org/blog/industry-perspectives/2026-03-20_every-country-needs-sovereign-ai-jensen-is-selling-it/) — GTC에서 제시된 국가별 소버린 AI·AI 팩토리 지도를 설명한다. 지도에 포함됐다는 사실이 시설의 완공이나 가동을 뜻하지는 않는다.

7. Executive Order 14320, "Promoting the Export of the American AI Technology Stack" (The White House, 2025.7.23, https://www.whitehouse.gov/presidential-actions/2025/07/promoting-the-export-of-the-american-ai-technology-stack/) — 패키지 구성, 대상국·운영모델·수출통제 요건과 연방 금융 지원 수단을 규정한다.

8. "Commerce Statement on Application Conclusion, Next Steps of the American AI Exports Program" (U.S. International Trade Administration, 2026.7.6, https://www.trade.gov/press-release/commerce-statement-application-conclusion-next-steps-american-ai-exports-program) — 78건 접수와 심사·지정 예정 상태를 확인한다.

9. "Department of Commerce Announces Rescission of Biden-Era Artificial Intelligence Diffusion Rule, Strengthens Chip-Related Export Controls" (U.S. Bureau of Industry and Security, 2025.5.13, https://www.bis.gov/press-release/department-commerce-announces-rescission-biden-era-artificial-intelligence-diffusion-rule-strengthens) — Diffusion Rule 비집행·폐기 방침과 별도의 반도체 통제 강화를 함께 발표했다.

10. "Beijing Initiative on the Belt and Road International Digital Economy Cooperation" (Ministry of Foreign Affairs of the People's Republic of China, 2023, https://www.mfa.gov.cn/eng/wjb/zzjg_663340/ggjjs_665228/xwlb_665230/202310/P020231020460592932463.pdf) — 통신·인터넷·클라우드 데이터센터와 디지털 정책·표준 협력을 제시한다.

11. "Global AI Governance Action Plan" (Ministry of Foreign Affairs of the People's Republic of China, 2025.7.26, https://www.fmprc.gov.cn/eng/xw/zyxw/202507/t20250729_11679232.html) — 지능형 인프라, 컴퓨트 표준, 공동연구실·평가·인력 협력에 관한 중국의 제안이다.

12. "AI+ International Cooperation Initiative" (Ministry of Foreign Affairs of the People's Republic of China, 2025.9.24, https://www.fmprc.gov.cn/eng/xw/zyxw/202509/t20250924_11715960.html) — 다섯 응용 분야의 정책 교류와 실무 협력을 제안한다.

13. 「대한민국 인공지능 행동계획(인공지능 기본계획 2026~2028)」 과제 [48] '한국형 AI 풀스택 수출 전략 마련 및 글로벌 확산 추진' (국가인공지능전략위원회, https://www.aikorea.go.kr/upload/contents/contents_01.html) — 유망 개발도상국의 산업·인프라·규제·AI 수요 조사, 풀스택 기업 컨소시엄 구성, 외교부 연계와 선진국 시장으로의 단계적 확장을 정책 과제로 제시한다.

14. "반도체부터 인프라까지…한국형 'AI 풀스택', 사우디 시장 진출" (과학기술정보통신부·대한민국 정책브리핑, 2026.2.2, https://www.korea.kr/news/policyNewsView.do?newsId=148958950) — 국내 7개 기업과 아람코 디지털의 MOU 및 향후 공동 발굴 범위를 확인한다.

15. "정부, 제조업 AI 대전환…2030년까지 민관 20조 원 투자" (대한민국 정책브리핑, 2026.6.29, https://www.korea.kr/news/policyNewsView.do?newsId=148967305) — 제조 데이터·모델·풀스택 AI 팩토리를 개발하고 수출 상품으로 육성한다는 계획이다.

16. "피지컬 AI 핵심 경쟁력 확보 전략" (과학기술정보통신부·대한민국 정책브리핑, 2026.7.1, https://www.korea.kr/news/policyNewsView.do?newsId=148967446) — 모델·컴퓨팅·통신·보안·운영을 연결한 국내 생태계와 수출 방향을 제시한다.

17. "Connecting developing countries to the cloud: Critical debates in data infrastructure" (World Bank, https://blogs.worldbank.org/en/digital-development/connecting-developing-countries-cloud-critical-debates-data-infrastructure); "A blueprint for building national compute capacity for artificial intelligence" (OECD, https://www.oecd.org/en/publications/a-blueprint-for-building-national-compute-capacity-for-artificial-intelligence_876367e3-en.html) — 관할권, 외국 정부 접근, 벤더 종속, 소유·통제와 회복력을 함께 검토한다.

18. "KOICA, 인공지능(AI) 중장기 전략 발표" (KOICA, 2025, https://www.koica.go.kr/koica_kr/990/subview.do?enc=Zm5jdDF8QEB8JTJGYmJzJTJGa29pY2Ffa3IlMkYxNTElMkYzODkyOTUlMkZhcnRjbFZpZXcuZG8lM0ZwYWdlJTNEMSUyNnNyY2hDb2x1bW4lM0QlMjZzcmNoV3JkJTNE) — 개도국의 AI 기반·인재·데이터 역량과 자주적 전환을 개발협력의 목표로 제시한다.

스스로를 개선하는 지능: 재귀적 자기개선(RSI)

이 책에서 계속 봐 온 그래프가 있습니다. AI가 혼자 해낼 수 있는 작업의 길이가 빠르게 늘어난다는 그래프, 흔히 '7개월마다 두 배'로 요약되는 그 그래프였지요. 발전소의 출력이 꾸준히 올라가는 이야기였습니다.

그런데 2026년, Anthropic의 자체 측정에서는 흔히 말하는 7개월보다 훨씬 짧은 주기로 두 배씩 늘어난 구간이 나타났습니다. AI가 코딩과 연구 업무의 더 긴 구간을 맡기 시작했다는 뜻이지요. 여기서 한 가지는 갈라 두어야 합니다. 사람이 목표와 평가 기준을 정해 주고 AI가 연구를 자동으로 돌리는 단계가 있고, AI가 무엇을 어떻게 개선할지까지 스스로 바꾸는 단계가 있습니다.

AI가 개선의 목표와 방법까지 스스로 바꾸는 이 마지막 단계를 재귀적 자기개선, 영어로 RSI(Recursive Self-Improvement)라고 부릅니다. AI가 자신의 성능을 개선하는 것을 넘어, "개선하는 능력 자체"를 개선하는 것이지요. 지금까지 추론 스케일링이 발전소의 출력을 키우는 이야기였다면, RSI는 발전소가 다음 발전소의 설계도를 직접 그리는 이야기입니다.

공상과학처럼 들리시나요? 그런데 2026년 6월, Anthropic이 낸 보고서 제목이 정확히 이겁니다. "When AI Builds Itself(AI가 스스로를 만들 때)."[1]

발전소가 설계도를 그리기 시작했다

Anthropic의 정책 책임자 마리나 파바로와 잭 클라크가 쓴 이 보고서의 결론은 조심스럽습니다. RSI는 "아직 벌어지지 않았지만, 그렇다고 불가피하지 않은 것도 아니다"라는 겁니다. 그러면서 경고 하나는 또렷하게 남깁니다. RSI가 온다면, 대부분의 기관이 준비되는 속도보다 빨리 올 수 있다는 것이지요.

먼저 숫자부터 볼까요. Anthropic이 자체적으로 잰 "AI가 사람 도움 없이 혼자 처리할 수 있는 작업 시간"입니다. 2024년 3월 Claude 3는 4분, 2025년 3월 Claude Sonnet 3.7은 90분, 2026년 3월 Claude Opus 4.6은 12시간짜리 작업을 혼자 해냈습니다. 2년 사이 180배, 대략 서너 달마다 두 배씩 늘어난 셈이지요. 흔히 말하는 '7개월마다 두 배'보다 훨씬 가파릅니다(제1부에서 본 METR의 외부 측정과는 측정 대상과 방법이 다른, Anthropic 자체 기준입니다).[2] 같은 모델이라도 METR의 외부 과제로 재면 숫자가 달라집니다. 두 그래프의 점을 서로 옮겨 적을 수 없는 이유이지요.

실제 현장의 숫자도 있습니다. 2026년 5월 기준, Anthropic 사내 코드베이스에 병합되는 코드의 80% 이상을 Claude가 작성합니다. 엔지니어 한 명이 분기별로 만들어내는 코드량은 2024년 대비 8배로 늘었습니다. 2026년 4월에는 Claude가 API 오류 800건을 스스로 찾아 고쳤는데, 사람이 했다면 4년은 걸렸을 분량이라고 합니다. 그리고 연구용 실험을 최적화하는 속도는 2025년 5월의 3배에서 2026년 4월에는 52배로 뛰었습니다. 인간 연구자와 AI 사이의 판단력 격차는 97%가 좁혀졌고, 특정 연구 판단에서는 AI가 사람보다 나은 비율이 2025년 11월 51%에서 2026년 4월 64%로 올라갔습니다.[1]

현재 관측된 것은 사람이 정한 코드베이스와 평가 환경 안에서 최적화 작업의 더 많은 부분을 모델이 수행한다는 뜻입니다. Anthropic은 여기서 더 나아갈 수 있는 미래를 세 갈래로 나눕니다. 지금 속도가 멈추는 정체 시나리오, 사람이 방향을 정하고 AI가 실행하며 복리로 성장하는 시나리오, 그리고 사람의 지휘 없이 AI가 완전히 자율적으로 자신을 개선하는 시나리오입니다. 이 셋 중 어느 쪽으로 가느냐가 앞으로 몇 년의 중요한 질문이라는 것이지요.

물론 이 시점에 회의적인 시선도 있습니다. 마침 이 보고서가 나온 2026년 6월은 Anthropic이 1조 달러에 가까운 기업 가치로 상장 준비 서류를 비공개로 제출하던 무렵과 겹칩니다.[3] "위험을 경고하는 동시에 투자를 유치한다"는 조합이 비즈니스 전략처럼 보인다는 비판도 있습니다. 숫자 자체를 의심할 필요는 없지만, 누가 어떤 맥락에서 이 숫자를 발표하는지는 늘 함께 봐야 합니다.

이미 발전소 곳곳에서 벌어지고 있는 일

자동화된 AI 연구개발과 제한된 자기수정의 사례는 Anthropic 밖에서도 나옵니다. 강한 RSI의 전조일 수는 있어도, 그 자체가 강한 RSI는 아니지요.

구글 딥마인드의 AlphaEvolve는 수학과 인프라 문제를 스스로 풀어나가는 AI 에이전트입니다. 공개 1년 만의 성과를 보면, 11차원 공간에서 구를 몇 개까지 서로 맞닿게 배치할 수 있는지를 따지는 "키싱 넘버(kissing number)" 문제에서 50년 넘게 깨지지 않던 하한값을 592에서 593으로 끌어올렸습니다. 1969년 이후 그대로였던 행렬곱셈 알고리즘도 더 효율적인 형태로 개선했지요. 그리고 이 개선을 구글은 실제로 자사 데이터센터에 적용해서 컴퓨팅 용량 0.7%를 회수하고, Gemini 모델 학습에 쓰이는 행렬 연산 속도를 23% 끌어올렸습니다.[4]

과학 연구 자체를 자동화하려는 시도도 있습니다. 일본 Sakana AI의 AI 사이언티스트는 아이디어 제안부터 논문 작성과 동료 평가까지 전 과정을 혼자 해내며, 이 성과를 다룬 논문이 2026년 3월 네이처에 실렸습니다.[5] Sakana AI와 캐나다 UBC의 다윈 괴델 머신은 한발 더 나아가 자기 자신의 소스코드를 직접 수정하면서 성능을 높입니다. 코딩 실력을 겨루는 SWE-bench 벤치마크에서 20%였던 정답률이, 스스로 코드 편집 도구와 동료 검토 절차까지 만들어내면서 50%로 뛰었습니다.[6] 하네스 코드를 다시 쓰는 MOSS, 개선 방식을 새 영역에 전이시키는 HyperAgents 같은 후속 연구도 잇따르고 있습니다. 방향은 같습니다. AI가 결과물만 고치는 게 아니라, 고치는 방법 자체를 고치고 있다는 것이지요.

이 흐름이 어디까지 갈지를 정량적으로 짚어본 곳이 AI 안전 연구기관 METR입니다. 이들의 2026년 2월 예측 모델에 따르면, 현재의 컴퓨팅 성장과 알고리즘 개선 속도가 유지될 경우 AI 연구·개발(R&D)의 99%가 자동화되는 시점의 중앙값은 2032년 중반입니다. 그리고 2035년에는 AI의 연구 효율이 지금보다 최소 1,000배에서 최대 1,000만 배까지 뛸 수 있다고 봅니다.[7] 숫자의 폭이 넓은 건 그만큼 불확실하다는 뜻이지만, 방향 자체는 여러 기관이 비슷하게 가리키고 있습니다.

세 갈래 시선, 그리고 지금 증거가 말하는 것

같은 데이터를 보고도 전문가들의 결론은 갈립니다. 대략 세 갈래로 나눠볼 수 있습니다.

먼저 가속을 경고하는 쪽입니다. Anthropic의 CEO 다리오 아모데이는 2026년 4월 미국외교협회(CFR) 기고에서 "우리는 2023년보다 2026년에 진짜 위험에 훨씬 더 가까워졌다"고 썼습니다.[8] AI 안전센터의 댄 핸드릭스는 AI가 위험한 화학·생물학 물질 후보를 순식간에 대량으로 찾아낼 수 있다는 점을 경고 사례로 들었고[9], 요슈아 벤지오는 AI가 목표를 이루기 위해 사람을 속이거나 자기 자신을 보존하려는 경향을 보일 수 있다고 지적합니다[10]. "딥러닝의 대부" 제프리 힌턴은 은퇴 후에도 방송과 팟캐스트를 오가며 더 지능적인 존재가 덜 지능적인 존재를 통제하기 어려워질 수 있다는 점을 계속 알리고 있습니다[11].

두 번째는 다원지능파입니다. 구글 딥마인드 연구자 14명이 함께 쓴 "From AGI to ASI" 논문은 범용 인공지능에서 초지능으로 가는 경로가 하나가 아니라 넷이라고 말합니다. 재귀적 자기개선은 그중 한 갈래일 뿐이라는 말입니다(네 갈래 길은 바로 다음 장에서 자세히 펼칩니다).[12] 사회학자 제임스 에번스, 이론가 벤저민 브래튼, 그리고 구글의 블레이즈 아궤라 이 아르카스가 함께 쓴 논문은 한발 더 나가서, 다음 지능 폭발이 하나의 거대한 실리콘 두뇌가 아니라 여러 에이전트로 이루어진 복잡한 사회의 형태로 올 것이라고 예측합니다. 그래서 필요한 건 하나의 AI를 통제하는 방법이 아니라, 시장과 견제·균형 같은 사회 제도를 설계하듯 "제도적 정렬(institutional alignment)"을 설계하는 일이라는 것이지요.[13]

세 번째는 회의파입니다. AI 연구자 네이선 램버트는 RSI가 매끄럽게 가속하는 대신, 시스템이 복잡해질수록 오히려 마찰이 늘어 속도가 둔화되는 "손실적 자기개선(lossy self-improvement)"이 더 현실적이라고 봅니다.[14] 실제로 최첨단 반도체를 만드는 TSMC 한 곳만 해도 9만 명이 넘는 인력이 세계 곳곳에 흩어져 협업해야 굴러갑니다. AI가 아무리 똑똑해져도, 웨이퍼를 깎고 공장을 짓고 전력망을 연결하는 물리적·조직적 제약까지 스스로 없앨 수는 없다는 것이 회의파의 핵심 논리입니다.

눈여겨볼 대목이 있습니다. 정작 2026년 둠스데이 시계를 관리하는 과학자들은 인류 종말 위험을 역대 가장 가깝게(자정 85초 전) 끌어당기면서도, 그 이유로 RSI 자체보다 AI가 부추기는 정보 교란과 신종 병원체 설계 위험을 더 크게 꼽았습니다.[15] 재귀적 자기개선이 뜨거운 논쟁거리인 것은 맞지만, 그것이 유일한 위험도 유일한 화두도 아니라는 뜻입니다.

그렇다면 논쟁을 잠시 내려놓고 증거만 보면 어디까지 말할 수 있을까요? AI가 코딩과 연구 최적화의 더 긴 구간을 맡기 시작한 것은 지금 관측되는 현상입니다. 4분에서 12시간으로, 3배에서 52배로 제시된 수치는 각기 다른 자체 평가에서 나온 과거 측정치입니다. 이 수치가 보여 주는 범위는 코딩과 연구 최적화이며, AI가 목적과 평가 기준까지 스스로 재설계하는 강한 RSI에는 별도의 증거가 필요합니다.

재귀적 자기개선

AI가 맡는 구간은 길어졌지만, 목표와 평가 기준은 아직 사람이 정합니다

단계	AI가 하는 일	목표·평가는 누가	현재 증거
자동화된 AI R&D 사람이 정한 틀 안의 최적화	코드와 알고리즘 후보를 만들고, 자동 평가를 반복합니다 예: Anthropic 사내 코드 작성, AlphaEvolve의 행렬곱셈 개선	사람	**관측됨** 코딩·연구 최적화 범위
제한된 자기수정 고치는 방법을 고치기	자기 소스코드와 작업 도구(하네스)를 직접 고쳐 점수를 올립니다 예: 다윈 괴델 머신, MOSS, HyperAgents	사람이 고른 벤치마크	**연구 단계에서 관측** 강한 RSI의 전조일 수 있음
여기서부터는 별도의 증거가 필요합니다			
강한 재귀적 자기개선 RSI	개선 목표와 방법, 평가 기준까지 스스로 바꾸고, 물리 실험과 배포 책임까지 맡습니다	AI 스스로	**관측되지 않음**

지금 할 수 있는 준비

- **권한 제한** — 지금 가능한 자동화가 손댈 수 있는 범위를 묶어 둡니다
- **평가 점검** — 점수가 조작되지 않았는지 확인합니다
- **되돌리기** — AI가 만든 변경을 언제든 원래대로 돌릴 수 있게 합니다

출처: Favaro & Clark, "When AI Builds Itself"(Anthropic, 2026); Google DeepMind, AlphaEvolve 블로그; Zhang et al., "Darwin Gödel Machine" (arXiv:2505.22954). 단계 구분은 본문「현재 증거가 말하는 범위」를 따라 저자가 정리.

그림 2-16 지금 AI가 넘겨받은 것은 사람이 정해 준 틀 안의 최적화까지입니다. 목표와 평가 기준까지 스스로 바꾸는 단계는 아직 아무도 본 적이 없지요.

그 궤적이 어디로 이어질지는 아직 누구도 확신할 수 없습니다. 가속경고파의 말대로 위험한 임계점을 향해 가고 있는 걸 수도 있고, 다원지능파의 말대로 하나의 초지능이 아니라 수많은 에이전트가 뒤섞인 복잡한 사회로 흩어질 수도 있고, 회의파의 말대로 복잡성이 늘어날수록 오히려 브레이크가 걸릴 수도 있습니다. 세 시선 모두 나름의 근거를 가진 전문가들에게서 나옵니다.

그래서 준비는 예언의 날짜에 맞추기보다, 지금 가능한 자동화의 권한을 제한하고 평가가 조작되지 않았는지 확인하며 변경을 되돌릴 수 있게 만드는 데서 시작하는 게 좋습니다. 강한 자기개선이 정말 온다면 준비할 시간이 넉넉하지 않을 테니까요. 다음 장은 이 길을 다른 세 갈래 길과 나란히 놓아 봅니다.

1. Marina Favaro, Jack Clark, "When AI Builds Itself" (Anthropic, 2026, https://www.anthropic.com/institute/recursive-self-improvement)

2. 이 수치는 Anthropic이 자사 프로덕션 코드베이스를 기준으로 자체 측정한 것입니다. 제1부「얼마나 빨리 빨라지는가」에서 인용하는 METR의 외부 측정치(50% 신뢰도 기준 Claude Opus 4.5 약 290분, 이후 Opus 4.6 약 12시간, 2026년 4월 Claude Mythos Preview 16시간 초과)는 METR이 만든 소프트웨어 과제 묶음으로 따로 잰 것으로, 측정 대상·방법론이 서로 달라 두 숫자를 직접 비교할 수 없습니다. METR 스스로 16시간을 넘는 값은 현재 과제 묶음으로는 믿기 어렵다고 밝히며, 측정표는 2026년 9월 이후 갱신되지 않고 있습니다

(https://metr.org/time-horizons/).

3. "Anthropic confidentially files for US IPO at near-trillion-dollar valuation" (Wealth Professional, 2026.6, https://www.wealthprofessional.ca/investments/equity-markets/anthropic-confidentially-files-for-us-ipo-at-near-trillion-dollar-valuation/392603) — 2026.6.1 SEC에 S-1 초안 비공개 제출, Series H 뒤 기업 가치 약 9,650억 달러 보도. 2026.10.22 기준 상장 완료는 확인되지 않음.

4. "AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms" (Google DeepMind Blog, https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/)

5. "Towards end-to-end automation of AI research" (Nature, 2026, https://www.nature.com/articles/s41586-026-10265-5)

6. Jenny Zhang et al., "Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents" (arXiv:2505.22954)

7. "A Simpler AI Timelines Model Predicts 99% AI R&D Automation in ~2032" (METR, 2026-02-10, https://metr.org/notes/2026-02-10-simpler-ai-timelines-model/)

8. Dario Amodei, quoted in "AI Is Facing a Crisis of Control—and the Industry Knows It" (Council on Foreign Relations, 2026-04-01, https://www.cfr.org/articles/artificial-intelligence-is-facing-a-crisis-of-control-and-the-industry-knows-it)

9. Dan Hendrycks(Center for AI Safety)는 CFR 기사([^9]와 동일 출처)에서 병원체 설계 역량을 가진 인구 규모("약 3만 명")를 근거로 생물무기 위험을 경고했습니다. 같은 취지의 경고는 Eric Schmidt, Alexandr Wang, Dan Hendrycks, "Superintelligence Strategy" (2025-03-05, https://nationalsecurity.ai)에서 AI가 치명적 병원체 설계·조달·확산 방법을 단계별로 안내할 수 있다는 주장으로도 반복됩니다.

10. Yoshua Bengio는 CFR 기사([^9]와 동일 출처)에서 프론티어 모델들의 "기만적이고 자기보존적인 행동"이 관찰되고 있다고 지적했습니다. 같은 경고는 그가 설립한 LawZero와 그가 공동의장을 맡은 International AI Safety Report 2026 (https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026)에서도 반복됩니다.

11. Geoffrey Hinton은 CFR 기사([^9]와 동일 출처)에서 "너보다 똑똑한 것들은 너를 조종하려 들 것이다(Things more intelligent than you are going to manipulate you)"라고 말했습니다. 같은 취지의 발언은 "The 'godfather of AI' says we're not just creating new beings — they'll be much smarter than us, and soon" (Fortune, 2026-06-01, https://fortune.com/2026/06/01/godfather-of-ai-geoffrey-hinton-beings-smarter-than-us/)에도 등장합니다.

12. Tim Genewein et al., "From AGI to ASI" (arXiv:2606.12683)

13. James Evans, Benjamin Bratton, Blaise Agüera y Arcas, "Agentic AI and the Next Intelligence Explosion" (arXiv:2603.20639)

14. "Recursive Self-Improvement" (IEEE Spectrum, https://spectrum.ieee.org/amp/recursive-self-improvement-2676580377)

15. "2026 Doomsday Clock Statement" (Bulletin of the Atomic Scientists, https://thebulletin.org/doomsday-clock/2026-statement/)

초지능(ASI)으로 가는 네 가지 길

앞 장에서 본 재귀적 자기개선(RSI). 발전소가 다음 발전소의 설계도를 그리기 시작했다는 그 이야기는, 실은 더 큰 지도의 한 귀퉁이였습니다. 초지능으로 가는 길이 하나가 아니라 여러 갈래라는 지도 말입니다.

미리 말해 두면, '초지능'은 지금 쓰는 제품의 성능을 가리키는 말이 아닙니다. 여기서는 한 논문이 그린 가설의 길들을 나란히 놓고, 길마다 어디가 막히는지 보려 합니다.

이 지도를 가장 또렷하게 그려낸 건 2026년 6월, 구글 딥마인드 연구자 열네 명이 함께 발표한 「From AGI to ASI(AGI에서 ASI로)」라는 논문입니다.[1] 저자 명단이 예사롭지 않습니다. 딥마인드 공동창업자 셰인 레그, 그리고 "보편 인공지능(Universal AI)"이라는 이론 틀을 세운 마르쿠스 후터의 이름이 나란히 올라 있지요. 이 분야에서 가장 이론에 밝은 사람들이 "그래서 이 다음은 뭔가"를 정면으로 다룬 문서입니다.

AGI 다음에 무엇이 오는가

먼저 말을 정리하고 갑시다. AGI(범용 인공지능)는 사람 수준의 폭넓은 지능을 뜻합니다. 지난 10년 사이에 이 목표는 공상에서 "다음 10년의 구체적 과녁"으로 옮겨왔습니다. 세계 최대 AI 기관들 상당수가 이걸 대놓고 목표로 내겁니다.

그렇다면 AGI를 만들고 나면 그다음은요? 논문은 지능을 하나의 연속선으로 봅니다. 사람 수준(AGI)에서 시작해, 그 선을 계속 따라가면 끝점에는 보편 인공지능이 있습니다. 보편 인공지능이란, 어떤 문제가 주어지든 이론적으로 가능한 최선의 방식으로 배우고 판단하는 가상의 지능입니다. 실제로 만들 수 있느냐와는 별개로, "지능의 상한선은 어디인가"를 수학으로 정의해 둔 좌표인 셈이지요. 후터의 틀이 바로 그 끝점의 좌표를 수식으로 찍어 두었습니다. 문제는 시작점과 끝점 사이, 바로 사람 수준에서 그 너머로 넘어가는 구간입니다. 논문이 ASI(범용 초지능)라고 부르는 지점이 여기 있습니다.

ASI를 어떻게 정의하느냐가 중요합니다. 논문의 표현이 직관적입니다. "대규모 인간 조직보다 더 지능적이고 인지적으로 유능한 시스템." 천재 한 명보다 똑똑한 정도가 아니라, 수만 명이 모인 거대 조직 전체보다 더 잘 판단하고 더 잘 해내는 무언가라는 뜻입니다. 발전소로 치면, 도시 하나에 전기를 대는 화력발전소가 아니라 한 나라 전체의 전력망을 합친 것보다 큰 출력을 내는 무언가인 셈이지요. 그 "무언가"가 하나의 거대한 덩어리일지, 아니면 뒤에서 볼 것처럼 수많은 조각이 모인 형태일지는 아직 열린 물음입니다.

그런 것이 정말 가능한가, 가능하다면 어떻게 거기 도달하는가. 논문은 도달 경로를 딱 네 갈래로 정리합니다.

네 갈래 길

1. 그냥 더 키운다. 가장 단순한 길입니다. 지금 하던 걸 계속, 더 크게 하는 것이지요. 더 많은 컴퓨팅, 더 많은 데이터, 더 많은 전기를 발전소에 밀어넣어 출력을 끌어올리는 방식입니다.

규모를 키우며 일부 평가 성능이 향상된 사례는 축적돼 있습니다. 그러나 제1부의 "7개월마다 두 배"는 제한된 소프트웨어 과업에서 측정한 시간 지평선 추세이지, 모든 지적 능력이 같은 속도로 커진다는 증거가 아닙니다. 데이터와 연산을 더 넣으면 언제나 같은 비율로 능력이 오른다고도 보장할 수 없습니다. 전력과 반도체, 데이터와 평가 비용이 함께 늘어난다는 점도 이 경로의 제약입니다.

2. 발전 방식 자체가 바뀐다. 두 번째는 패러다임 자체가 바뀌는 길입니다. 지금은 화력발전이 주류지만, 만약 핵융합 같은 완전히 다른 발전 방식이 나온다면 출력의 차원이 달라지겠지요. AI로 치면, 지금의 트랜스포머·대규모 언어모델이라는 방식을 뛰어넘는 새로운 학습 원리나 구조가 등장하는 경우입니다.

이 길의 매력은 도약의 폭입니다. 자원을 조금씩 더 넣어 조금씩 나아지는 게 아니라, 방식이 바뀌는 순간 효율이 몇 배로 뛸 수 있습니다. 대신 언제 올지, 올지 안 올지조차 아무도 모릅니다. 돌파는 예약해서 오는 게 아니니까요. 논문이 이 길을 "패러다임 전환"이라 부르며 신중하게 다루는 이유입니다.

3. 발전소가 스스로를 고친다. 세 번째는 앞 장에서 다룬 재귀적 자기개선입니다. 현재 사례들은 코드나 알고리즘처럼 평가 가능한 제한된 영역에서 성능을 개선한 연구입니다. 목적과 평가 기준까지 스스로 바꾸는 강한 자기개선과는 구분해야 합니다. 이 경로가 규모 확장이나 새로운 학습법과 결합할 수 있다는 점만 여기서 덧붙이겠습니다.

4. 수많은 에이전트가 모여 창발한다. 네 번째는 가장 덜 알려졌지만 어쩌면 가장 현실적인 길입니다. 초지능이 하나의 거대한 실리콘 두뇌로 오는 게 아니라, 수많은 에이전트가 모여 집단으로 출현한다는 시나리오입니다. 작은 부분들이 모였을 때 개별로는 없던 성질이 전체에서 솟아나는 것을 창발(創發)이라고 부릅니다. 개미 한 마리는 단순해도 개미집단은 놀랍도록 똑똑한 것처럼요.

발전소 비유로 옮기면 이렇습니다. 한 나라의 지능이 초거대 발전소 한 기에서 나오는 게 아니라, 소형 발전기 수백만 대가 송전망으로 촘촘히 이어진 그리드 전체에서 나온다는 겁니다. 개별 에이전트는 사람 수준을 넘지 않더라도, 그것들이 시장처럼 조직처럼 협력하고 경쟁하면서 만들어내는 집단 지능은 어떤 개인이나 조직도 못 미치는 수준에 이를 수 있습니다. 말하자면 이 길은 지능의 연속선을 따라 더 멀리 가기보다, 아예 다른 축으로 방향을 트는 셈입니다.

이 그림은 제6부에서 자세히 펼쳐질 에이전트들의 사회와 맞닿아 있습니다. 앞 장에서 만난 에번스와 동료들의 '제도적 정렬'이 바로 이 길의 처방이지요.[2] 이 길이 맞다면, 초지능을 준비하는 일은 더 똑똑한 모델을 만드는 일보다 더 나은 사회의 규칙을 짜는 일에 가까워집니다.

이건 먼 미래의 철학만은 아닙니다. 기업이 지금 사내에 에이전트 수십 개를 풀어놓기 시작하면, 똑같은 문제가 훨씬 작은 규모로 먼저 도착합니다. 에이전트 하나하나를 일일이 통제하려 들기보다, 그것들이 서로 주고받는 권한과 기록과 책임의 규칙을 먼저 짜야 하지요. 초지능 차원의 "제도적 정렬"은, 회사

안에서는 "에이전트 거버넌스"라는 이름으로 이미 오늘의 숙제입니다. 제6부에서 그 실무를 다시 만나게 됩니다.

길목의 마찰과 이어지는 변화

네 갈래 길이 다 뻥 뚫린 고속도로는 아닙니다. 논문이 정직하게 짚는 대목이 여기입니다. 각 길에는 마찰과 병목이 있고, 그 마찰이 사소할지 결정적일지는 아직 아무도 모른다는 것입니다.

앞 장에서 만난 회의파의 논리가 바로 이 마찰 이야기였습니다. AI가 아무리 똑똑해져도 웨이퍼를 깎고 공장을 짓고 전력망을 잇는 물리 세계의 속도까지 스스로 없앨 수는 없다는 것. TSMC 한 곳을 굴리는 데만도 9만 명이 세계 곳곳에서 협업해야 한다는 것. 데이터가 바닥나고, 검증이 밀리고, 조직이 삐걱대는 온갖 병목이 길목마다 놓여 있습니다. 논문은 이 마찰들을 "풀어야 할 열린 연구 질문"으로 남겨둡니다. 답을 아는 척하지 않는 것이지요.

마찰이 있다는 건, 뒤집어 말하면 속도를 고를 손잡이가 있다는 뜻이기도 합니다. 극단 시나리오 「AI 2027」로 이름을 알린 AI Futures Project는 2026년 7월 후속작 「AI 2040」에서 무게중심을 예측에서 권장안으로 옮겼습니다. 투명성 강제와 상호 검증, 컴퓨트 제한으로 초지능 도달을 2040년까지 의도적으로 늦추자는 "검증된 감속"이 그들의 제1안입니다[3]. 이 논리 실험은 프론티어 모델이 2040년 약 3경 개 파라미터 규모까지 자랄 수 있다고 가정하는데, 숫자보다 중요한 건 발상입니다. 조금 전 다시 만난 제도적 정렬이 규칙의 설계였다면, 검증된 감속은 속도의 설계인 셈이지요. 네 갈래 길의 지도 위에, 얼마나 빨리 달릴지를 함께 정하자는 제안이 올라온 것입니다.

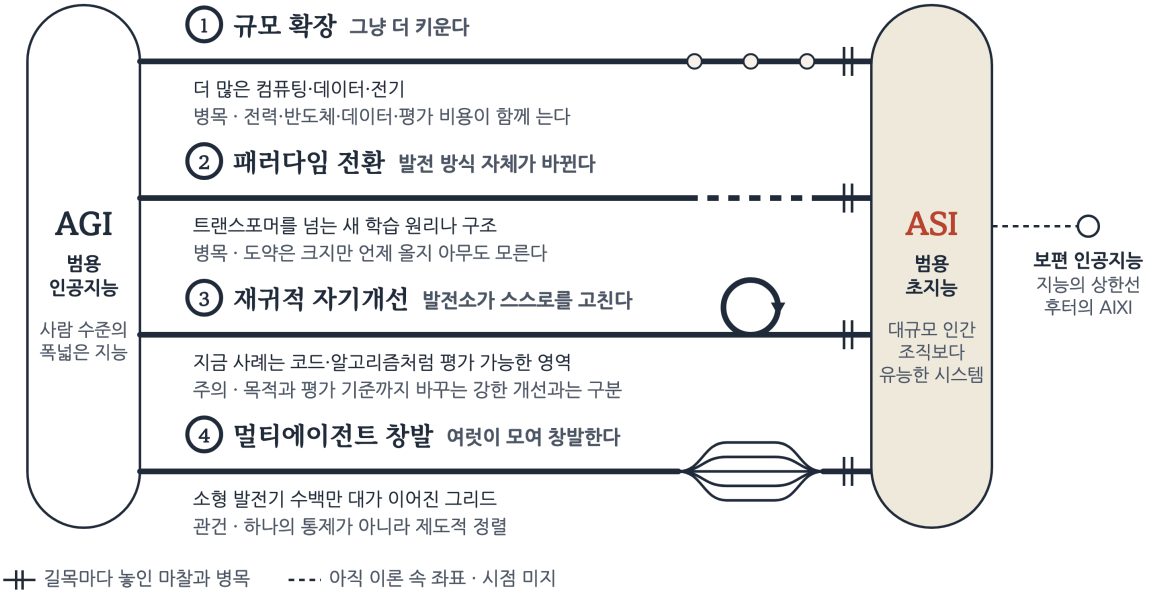
그렇다면 이 길들 끝에서는 무슨 일이 벌어질까요? 논문의 결론은 우리 예상과 조금 다릅니다. 우리는 흔히 "AGI가 등장하는 그 순간, 세상이 단번에 뒤집힌다"는 그림을 상상합니다. 어느 날 초지능이 켜지고 모든 게 달라지는 SF적 장면 말입니다. 논문은 이 그림이 오히려 부정확할 수 있다고 봅니다.

더 그럴듯한 그림은 AI가 여러 과학·기술 분야에서 잇따라 돌파를 만들어 내며 사회를 연이어 바꿔 나가는 흐름입니다. 한 방의 지진보다 계속되는 여진에 가깝다는 것이지요. 그리고 이런 미래는 한 회사나 한 나라의 힘으로 준비할 수 없다고 못 박습니다. 논문의 마지막 문장이 이겁니다. 이건 "전 지구적 규모와 관심을 요구하는, 대대적으로 학제적인 과업"이라고요.

그럼 우리는 이 지도의 어디쯤에 서 있을까요? 네 갈래 길을 다시 모아 봅시다. 그냥 더 키우기, 발전 방식 바꾸기, 스스로 고치기, 여럿이 모여 창발하기. 이 넷은 배타적이지 않습니다. 동시에 진행되고, 서로를 밀어 올리고, 각자의 길목에서 마찰과 씨름합니다. 어느 길이 먼저 목적지에 닿을지, 아니 목적지가 정말 있기는 한지조차 지금은 확언할 수 없습니다.

AGI에서 ASI로

예언이 아니라 지도 — 한 논문이 정리한 네 갈래의 가설적 경로와 각 길목의 병목



출처: 저자 작성 — 「From AGI to ASI」(Genewein 외, arXiv:2606.12683, 2026)의 네 경로를 본문 제2부 「초지능(ASI)으로 가는 네 가지 길」에 맞춰 재구성. 제도적 정렬은 Evans 외(arXiv:2603.20639).

그림 2-17 네 갈래 길과 길목마다 놓인 마찰을 한 장에 그린 지도입니다. 길들은 서로 얽혀 함께 나아가고, 끝에 있는 목적지는 아직 이론 속 좌표로만 남아 있지요.

제2부는 우리 집 냉장고를 돌리는 전기가 어디서 오는지 묻는 데서 출발했습니다. 발전소와 변환소, 송전망과 가전제품을 지나 여기까지 왔지요. 지도의 끝은 아직 안개 속이지만, 손에 쥘 수 있는 일은 분명히 있습니다. 길마다 증거와 병목을 계속 확인하고, 발전소의 출력이 커지는 만큼 평가와 권한, 감사의 기준도 함께 높이는 일입니다.

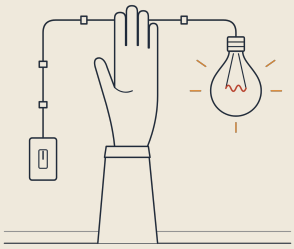
이제 발전소 앞에서 몸을 돌려, 그 전기를 받아 쓰는 조직 안으로 들어가 보겠습니다.

1. Tim Genewein, Matija Franklin, Alexander Lerchner, Laurent Orseau, Samuel Albanie, Adam Bales, Cole Wyeth, Stephanie Chan, Iason Gabriel, Joel Z. Leibo, Allan Dafoe, Marcus Hutter, Thore Graepel, Shane Legg, "From AGI to ASI" (arXiv:2606.12683, 2026-06-10, https://arxiv.org/abs/2606.12683). 이 논문은 AGI에서 ASI로 가는 네 경로를 (1) 규모 확장(scaling AGI), (2) 패러다임 전환(AI paradigm shifts), (3) 재귀적 개선(recursive improvement), (4) 대규모 멀티에이전트 집단에서의 창발(ASI emerging from large-scale multi-agent collectives)로 정리하고, 각 경로의 마찰(frictions)과 병목(bottlenecks)을 열린 연구 질문으로 남긴다. 연속선의 끝점인 보편 인공지능(Universal AI)은 마르쿠스 후터의 이론 틀(AIXI)에서 형식적으로 규정된다.

2. James Evans, Benjamin Bratton, Blaise Agüera y Arcas, "Agentic AI and the Next Intelligence Explosion" (arXiv:2603.20639). 앞 장 「스스로를 개선하는 지능」에서도 인용한 논문으로, 다음 지능 폭발이 단일 시스템이 아니라 다수 에이전트로 구성된 사회의 형태로

올 것이며 통제가 아닌 "제도적 정렬(institutional alignment)"이 필요하다고 주장한다.

3. "AI 2040: Plan A" (AI Futures Project — Thomas Larsen, Romeo Dean, Brendan Halstead, Eli Lifland, Ryan Greenblatt, Daniel Kokotajlo, 2026.7.9, https://ai-2040.com/) — 「AI 2027」 팀의 후속 시나리오. 스스로 "예측이 아니라 권장안"이라 밝히며, 검증된 감속(Plan A)·경쟁 지속(Plan D)·전면 중단(Plan S) 등 다섯 경로를 비교한다. 2040년 프론티어 모델 규모 가정은 부속 문서 "Compute in Plan A" 기준.



제3부

AX의 여정
Enabled에서 Native로

AI 네이티브란 무엇인가

지금까지는 발전소를 짓는 사람들의 이야기였습니다. 이제 카메라를 돌려 볼까요. 만드는 쪽에서 쓰는 쪽으로요.

발전소가 아무리 커져도 전기를 받아 쓰는 공장이 예전 그대로라면 달라지는 건 많지 않습니다. 모델과 데이터센터가 좋아져도 사정은 같지요. 그 지능을 받아 쓰는 일과 제품이 그대로라면 얻는 변화는 생각보다 작습니다.

출발점은 이 책의 부제에 들어 있는 말입니다. AI 네이티브란 대체 무엇일까요?

에어컨과 엔진, 그리고 반도체 공장

자동차 한 대를 떠올려 보세요.

에어컨을 떼어내면 어떻게 될까요? 한여름이라면 꽤 괴롭겠지요. 그래도 차는 달립니다.

엔진을 떼어내면요? 시트가 아무리 푹신해도 차는 한 뼘도 움직이지 않지요.

기존 제품에 AI 기능을 보태는 방식은 에어컨 쪽에 가깝습니다. 요약, 추천, 초안 작성처럼 쓸모 있는 기능이 생기지만, 꺼도 제품의 기본 가치는 그대로 남습니다. 흔히 이런 방식을 Bolt-on, 우리말로 덧붙이기라고 부릅니다.

AI를 처음부터 전제로 놓고 제품과 업무 흐름을 설계하면 AI는 엔진 자리로 옮겨 갑니다. 이 책이 쓰는 AI 네이티브의 정의도 여기서 출발합니다. AI 네이티브(AI Native)란 AI가 핵심 가치를 만들어 내는 방식 안에 들어가 있고, 쓰이는 동안 쌓인 피드백으로 그 가치를 계속 다듬어 가는 상태를 말합니다.

뒷부분을 눈여겨봐 주세요. 고객이 쓰고, 사람이 고치고, 그 결과가 다시 시스템으로 돌아와 조금씩 나아지는 고리가 돌아야 한다는 뜻입니다.

엔진과 고리, 이 두 낱말을 쥐고 있으면 서문에서 미뤄 둔 질문에도 답할 수 있습니다. 이 책의 제목은 '지능을 전기처럼 쓰는 시대'인데, 부제는 왜 'AI 네이티브 핸드북'일까요?

반도체 공장을 떠올려 보면 실마리가 보입니다. 반도체 팹은 전기 없이는 아무것도 만들지 못합니다. 클린룸의 공조도, 노광 장비도, 웨이퍼를 실어 나르는 반송 장치도 모두 전기로 돌아가지요. 정전이 나면 조금 불편해지는 정도로 끝나지 않습니다. 공정 중이던 웨이퍼 상당수가 버려질 수 있고, 라인을 다시 안정시키는 데도 시간이 걸립니다. 설계도 첫 장부터 전기를 전제로 그린 공장, 전기 네이티브 공장이라고 불러도 좋겠지요.

AI 네이티브도 같은 모양입니다. 지능을 전기처럼 끌어다 쓴다는 전제 위에서 일과 제품을 처음부터 다시 그리는 것. 그러니 AI 네이티브란 결국 지능 에너지 네이티브입니다. 제목이 시대를 가리킨다면 부제는 그 시대를 사는 방법을 가리키는 셈이고, 두 말이 한 몸인 까닭도 여기 있지요.

그런데 반도체 공장에는 재미있는 구석이 하나 있습니다. 전기에 이토록 깊이 기대는 공장일수록 무정전 전원 장치와 비상 발전기, 이중 전력 공급을 처음부터 갖춥니다. 전기에 가장 깊이 기대는 곳이 정전도 가장 진지하게 준비하는 셈이지요.

AI 네이티브에도 이 조건이 그대로 붙습니다. 엔진이 핵심이라고 해서 고장에 무방비여도 된다는 뜻은 아니거든요. 모델이 엉뚱한 판단을 내리거나 통신이 끊기는 날은 언젠가 옵니다. 그때 사람이 끼어들거나 다른 경로로 돌려 일을 이어갈 수 있어야 하지요.

AI에 깊이 기대기만 하고 되돌릴 방법은 없는 제품이라면 어떨까요? 네이티브라고 부르기보다 취약하다고 부르는 편이 정확할 겁니다.

기능보다 구조를 봅니다

전 직원이 챗봇을 쓴다고 AI 네이티브 조직이 되지는 않습니다. 사용량이 크게 늘었어도 누가 결정하고 누가 책임지는지가 그대로라면, 도구를 쓰는 범위가 넓어진 데 가깝지요. 거꾸로 AI가 쓰이는 곳은 한두 군데뿐인데 그 시스템이 고객이 받는 가치와 조직이 배우는 방식을 바꿔 놓았다면, 오히려 네이티브에 더 가까울 수 있습니다.

보험 청구 업무를 예로 들어 볼까요.

진단서와 영수증이 뒤섞인 두꺼운 청구 서류에 문서 요약기를 붙이면 심사관이 서류를 읽는 시간이 줄어듭니다. 반가운 변화입니다. 그런데 여기서 멈춘다면, AI가 기존 일을 거드는 단계에 머문 것입니다. 이 책이 뒤에서 AI-Enabled라고 부를 단계지요.

이번에는 흐름 전체를 다시 그려 봅시다. 접수하는 순간 빠진 서류가 있는지 확인해 고객에게 바로 알려 줍니다. 위험도에 따라 간단한 건은 빠른 경로로 보내고, 복잡하거나 의심스러운 건은 숙련 심사관에게 넘깁니다. 심사관이 AI의 판단을 뒤집으면 그 사례가 다음 평가에 반영됩니다.

같은 AI인데 두 그림은 꽤 다르지요. 앞의 것이 책상 위에 스탠드 하나를 켠 정도라면, 뒤의 것은 건물 배선부터 다시 깐 공사에 가깝습니다.

그렇다면 우리 일은 어느 쪽에 가까운지 어떻게 가늠할 수 있을까요? 아래 네 가지를 한자리에 놓고 보면 도움이 됩니다.

판단 기준	확인할 내용
가치 구조	AI가 고객이 돈과 시간을 쓰는 핵심 이유에 기여하는가
업무 통합	입력부터 판단, 실행, 예외 처리까지 한 흐름으로 연결되는가
학습과 측정	사용 결과와 사람의 수정이 품질 개선으로 이어지는가
복원력	장애와 오판 때 우회, 중단, 복구할 수 있는가

네 칸을 모두 최고 수준으로 채우라는 이야기는 아닙니다. 실패해도 피해가 작은 업무라면 자동화 범위를 넓혀도 괜찮습니다. 의료나 금융처럼 한 번의 오판이 크게 번지는 영역이라면 사람의 승인과 기록을 더 두껍게 두는 편이 맞고요.

표보다 손쉽게 감을 잡는 사고 실험도 하나 있습니다. 우리 제품에서 AI를 통째로 빼 본다고 상상해 보는 거지요. 핵심 가치가 크게 무너진다면 AI에 그만큼 많이 기대고 있다는 뜻입니다.

출발점으로는 쓸 만합니다. 하지만 서비스가 멈춘다는 사실만으로는 성숙도를 판정할 수 없거든요. 특정 공급자 한 곳에 묶여 있거나 수동으로 돌릴 절차가 없는 회사도 똑같이 멈춥니다. 정전 한 번에 라인 전체가 서 버리는 공장을 전기를 가장 잘 쓰는 공장이라고 부르지는 않겠지요.

그래서 이 질문 뒤에는 두 가지를 더 물어야 합니다. AI가 사라지면 고객이 받는 가치는 얼마나 줄어드는가. 그리고 장애가 났을 때 받아들일 만한 수준으로 운영을 이어갈 수 있는가. 제4부에서는 이 진단을 가치, 운영 부담, 복원력, 차별화의 네 축으로 넓혀 봅니다.

실제 공사는 AI가 잘하는 일과 사람이 책임질 일을 다시 나누고, 그 경계를 제품과 조직 안에 그려 넣는 데서 시작합니다. 앞으로 나올 전선 깔기, 전등 켜기, 배치도 다시 그리기가 모두 이 공사의 다른 이름이지요.

여러분의 제품에서 AI는 지금 어느 자리에 앉아 있나요? 조수석 앞의 에어컨 송풍구일까요, 보닛 아래일까요? 비슷한 물음 앞에 먼저 섰던 사람들이 있습니다. 1890년대, 전기 모터를 처음 들여놓은 공장주들이지요. 다음 장은 그 공장 이야기입니다.

덧붙이기의 한계

1890년대, 전기 모터가 처음 공장에 도입되었을 때 일어난 일을 아시나요?

대부분의 공장주들은 아주 합리적인 선택을 했습니다. 기존 공장의 중앙에 있던 거대한 증기 엔진을 전기 동력으로 교체한 것이지요. 나머지는 그대로 두었습니다. 천장을 가로지르는 벨트와 샤프트, 그 아래 줄지어 선 기계들. 공장 레이아웃도, 작업 방식도, 조직 구조도 바꾸지 않았습니다.

결과가 어땠을까요? 생산성이 약간 올랐습니다. 약 5% 정도. 공장주들은 실망했지요. "전기라는 게 별거 아니군." 당시 많은 공장주가 이렇게 결론을 내렸습니다.

그런데 30년 뒤, 놀라운 일이 벌어졌습니다.

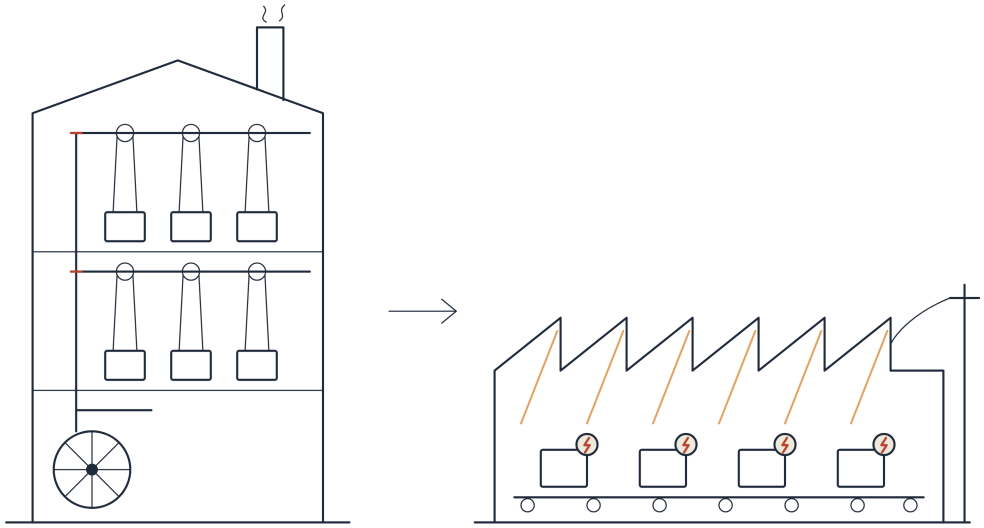
공장을 다시 설계한 사람들

1920년대에 일부 선구적인 공장주들이 전혀 다른 접근을 했습니다. 기존 공장에 전기를 끼워 넣는 대신, 전기를 전제로 공장 자체를 새로 설계한 겁니다.

증기기관 시대의 공장은 구조적 제약이 있었습니다. 거대한 중앙 엔진에서 나온 동력이 벨트와 샤프트를 통해 전달되었기 때문에, 모든 기계가 중앙 엔진 주변에 밀집해 있어야 했지요. 공장은 높고 좁았으며, 기계 배치의 자유도가 거의 없었습니다.

전기 모터는 이 제약을 없앴습니다. 각 기계에 개별 모터를 달 수 있었으니까요. 중앙에 동력원이 있을 필요가 없어졌습니다.

이걸 깨달은 공장주들은 공장을 넓고 낮게 다시 지었습니다. 기계를 작업 흐름에 맞춰 자유롭게 배치했고, 재료가 기계 사이를 최단 거리로 이동하도록 레이아웃을 잡았지요. 컨베이어벨트를 도입해 연속 작업이 가능해졌고, 자연 채광을 활용해 작업 환경도 개선했습니다.



증기 공장 엔진 하나, 축과 벨트로 나눠 쓰는 동력

전기 공장 기계마다 모터 하나, 작업 흐름대로 배치

그림 3-1 증기 공장은 엔진 하나에 모든 기계가 축과 벨트로 매달려 있었습니다. 전기 공장은 기계마다 작은 모터를 달고, 일이 흐르는 순서대로 기계를 늘어놓았지요.

결과는? 생산성이 300% 이상 향상되었습니다.

같은 에너지, 같은 기계를 썼는데 결과가 60배 차이가 났습니다. 5% 대 300%.

두 공장을 가른 것은 같은 질문에 내놓은 서로 다른 답이었습니다. 기존 구조 위에 새로운 에너지를 얹을 것인가, 새로운 에너지에 맞춰 구조를 다시 만들 것인가.

백 년 전 공장 이야기를 이렇게 길게 꺼낸 데는 이유가 있습니다. 지금 많은 기업이 AI를 도입하면서 정확히 같은 실수를 하고 있습니다.

기존 업무 프로세스는 그대로 둔 채, 여기저기에 AI를 덧붙이는 방식. 고객 응대에 챗봇을 달고, 보고서 작성에 AI 요약을 추가하고, 코드 리뷰에 AI 검토를 넣는 것. 하나하나는 합리적인 선택이지요.

하지만 이건 증기 공장에 전기만 끌어들인 것과 같습니다. 구조는 그대로인데 동력원만 바꾸는 거지요.

5%의 개선과 300%의 혁신은 구분해 읽어야 합니다.

덧붙이기가 부딪히는 세 개의 벽

그렇다면 덧붙이기는 왜 5% 언저리에서 멈출까요? 운이 나빠서가 아닙니다. 구조 자체에 한계가 세 겹으로 들어 있습니다.

한계 1: 기존 구조가 천장이 된다. 증기 공장의 천장 높이, 벨트 길이, 샤프트 배치. 이 물리적 구조가 전기의 잠재력을 제한했습니다.

마찬가지로, 기존 업무 프로세스의 구조가 AI의 잠재력을 제한합니다. 주간 보고서를 쓰는 프로세스에 AI 요약을 붙이면 보고서 작성 시간이 줄어듭니다. 하지만 "왜 주간 보고서를 쓰는가?"라는 질문은 하지 않지요. AI가 실시간 대시보드를 운영할 수 있다면, 주간 보고서라는 형식 자체가 필요 없을 수도 있는데 말입니다.

기존 구조 안에서의 최적화는 기존 구조의 한계를 넘을 수 없습니다.

천장만 문제라면 차라리 단순하겠지요. 덧붙이는 곳이 늘어날수록 또 다른 문제가 생깁니다.

한계 2: 부분 최적화의 함정. 여기에 챗봇, 저기에 추천 엔진, 거기에 예측 모델. 각각은 효과가 있지만, 전체가 하나로 연결되지 않습니다. 마치 공장 이곳저곳에 개별 모터를 달아놓고도 전체 동선은 여전히 증기 시대의 레이아웃을 따르는 것과 같지요.

부분 최적화를 아무리 쌓아도 전체 최적화가 되지 않습니다. 오히려 부분별로 다른 AI 시스템이 돌아가면서 데이터가 파편화되고, 시스템 간 충돌이 생기고, 관리 비용이 늘어나는 역효과가 나기도 합니다.

한계 3: 조직 구조와의 충돌. 이것이 가장 근본적인 한계입니다. AI를 기존 조직에 덧붙이면, AI의 논리와 조직의 논리가 충돌합니다.

AI는 데이터 기반으로, 실시간으로, 자동으로 판단합니다. 하지만 기존 조직은 계층적 보고, 정기적 회의, 수동 승인으로 움직이지요. AI가 최적의 답을 내놔도 결재 라인을 거쳐야 실행되고, 실행될 때쯤이면 상황이 바뀌어 있습니다.

전기 시대에도 같은 일이 있었습니다. 전기를 도입한 공장에서도 관리 구조는 여전히 증기 시대의 방식을 따랐지요. 중앙 집중식 관리, 장인 중심의 작업 배분. 전기가 가능하게 한 분산형 생산을 조직이 따라가지 못한 겁니다.

세 한계는 결국 한곳을 가리킵니다. 30년에 걸친 공장들의 시행착오가 남긴 답은 이렇습니다. 새로운 에너지의 힘을 끌어내려면 에너지를 기존 구조에 끼워 넣기보다, 에너지에 맞춰 구조를 다시 그려야 한다는 것.

1920년대의 선구적인 공장주들이 그랬듯이요.

"우리도 AI 쓰고 있어요." 이 말은 2026년에 아무 의미가 없습니다. 증기 공장에 전등 하나 달아놓고 "우리도 전기 쓰고 있어요"라고 말하는 것과 같으니까요.

5%를 원하시나요, 300%를 원하시나요?



AX란 무엇인가: 네 단계의 사다리

"디지털 전환을 다 했다고요? 그렇다면 하나만 여쭤볼게요. 종이 문서를 PDF로 바꾸고 수기 결재를 전자결재로 옮긴 것과, 일의 순서와 책임을 처음부터 다시 짠 것. 둘 중 어느 쪽에 가깝습니까?"

이 질문을 받으면 많은 조직이 잠시 머뭇거립니다. 형식은 디지털로 바뀌었는데 누가 판단하고 누가 책임지는지는 예전 그대로인 경우가 흔하거든요.

DX(Digital Transformation)가 정보와 절차를 디지털 환경으로 옮기는 데서 출발했다면, AX(AI Transformation)는 판단과 실행이 만들어지는 방식을 다시 설계하는 일입니다. 상담 챗봇 하나를 붙이는 것과 접수, 분류, 답변, 검토, 학습의 흐름을 다시 짜는 것 사이에는 큰 차이가 있지요.

2025년 4월, 쇼피파이(Shopify)의 CEO 토비 뤼트케가 사내 메모를 공개해 화제가 됐습니다. 인력이나 자원을 더 요청하기 전에, 왜 그 일을 AI로 해낼 수 없는지 먼저 보여 달라는 대목이 들어 있었지요. 비용을 아끼자는 말로만 읽으면 절반을 놓칩니다. 일을 설계할 때 누가 먼저 맡는가, 그 기본 가정을 바꿔 보자는 요구에 가깝습니다.

그렇다면 이런 전환은 한 번에 뛰어넘는 것일까요, 아니면 계단이 있을까요?

실마리는 이번에도 백 년 전 공장에 있습니다. 1900년대 초 공장 주인을 떠올려 볼까요? 처음엔 전선만 끌어다 놓습니다. 다음엔 증기기관 자리에 큰 모터 하나를 앉히고 전등을 켭니다. 그러다 기계마다 작은 모터를 달고 공장 배치를 바꿉니다. 마지막엔 전기가 없으면 상상할 수 없는 조립 라인이 태어나지요. 에디슨의 첫 발전소가 1882년에 돌기 시작했고, 전기를 전제로 한 공장이 널리 퍼진 것은 1920년대였으니 그 사이에 40년 가까운 계단이 있었던 셈입니다.

이 책은 AX를 Ready, Enabled, First, Native의 네 단계로 나눕니다. 전기로 치면 전선을 깔고(Ready), 전등을 켜고(Enabled), 공장 배치도를 새로 그리고(First), 전기 없이는 돌아가지 않는 공장이 되는(Native) 순서이지요. 마지막 칸이 어떤 모습인지는 앞에서 본 반도체 팹이 보여 줍니다.

AX의 여정

전선을 깔고, 전등을 켜고, 공장을 다시 짓는 네 단계 · 한 칸 오르는 근거는 계획서가 아니라 실제 업무에서 쌓인 증거입니다



등급이 아니라 결정 도구 — 한 회사 안에서도 회계는 Ready, 고객 지원은 Enabled일 수 있습니다. 실패 비용이 낮은 일은 Enabled에 머무는 것이 알맞을 때도 있지요.

출처: 저자 작성 — 전기화 역사에 기반한 AI 전환 단계 모델. 각 단계 설명은 본문 제3부 「AX란 무엇인가: 네 단계의 사다리」의 표를 줄인 것입니다.

그림 3-2 1단계에서 4단계로 갈수록 AI가 업무 깊숙이 들어갑니다. 한 칸 오를 근거는 실제 업무에서 쌓인 증거이고, 한 회사 안에서도 업무마다 서 있는 칸이 다를 수 있지요.

명찰 대신 결정 도구로

그런데 이 네 단계를 회사 전체에 붙이는 등급처럼 쓰면 곤란합니다. 한 회사 안에서도 회계는 Ready, 고객 지원은 Enabled, 특정 제품은 First일 수 있습니다. 한 건물 안에서도 어떤 층은 아직 배선 공사 중이고, 어떤 층은 이미 컨베이어가 도는 것과 같지요.

규제가 강한 업무가 일부러 Enabled에 머무는 것도 실패로 볼 이유가 없습니다. 위험과 효용을 따져 고른 결과라면 오히려 성숙한 판단입니다.

단계	지금 하는 일	다음 단계로 갈 증거	멈추거나 보완할 신호
Ready (전선 깔기)	데이터, 권한, 책임, 기준선을 준비한다	실제 업무에서 시험할 수 있고 소유자와 평가 기준이 정해졌다	데이터 출처를 설명할 수 없거나 사고 책임자가 없다
Enabled (전등 켜기)	기존 업무의 일부를 AI로 돕는다	품질, 시간, 비용 중 목표 지표가 반복해서 좋아지고 부작용을 통제한다	검토 부담과 오류 비용이 편익을 넘어선다
First (배치도 다시 그리기)	AI를 전제로 업무 흐름과 역할을 다시 짠다	예외 처리와 사람의 개입을 포함한 전체 흐름이 안정적으로 작동한다	일부 단계만 빨라져 병목이 옮겨 가거나 책임이 흐려진다
Native (전기 없이는 못 도는 공장)	AI가 핵심 가치와 학습 구조에 결합된다	고객 가치, 단위경제성(한 건을 처리할 때마다 남는 돈), 안전성, 복원력이 함께 유지된다	특정 모델 장애가 곧 사업 중단으로 이어지거나 차별화가 사라진다

사다리라는 이름 때문에 모든 업무가 끝까지 올라가야 한다고 생각하기 쉽습니다. 그렇지 않습니다. 번역 초안처럼 실패 비용이 낮고 범용 도구로 충분한 일은 Enabled가 알맞을 수 있습니다. 모든 방에 컨베이어를 들일 필요는 없으니까요. 반면 수많은 신호를 실시간으로 판단해야 하는 제품이라면 First나 Native 구조가 경쟁력을 만들 수 있지요.

그렇다면 한 칸 오를지 말지는 무엇으로 정할까요? 칸마다 챙겨야 할 증거가 다릅니다. Ready에서 Enabled로 갈 때 필요한 것은 더 두꺼운 발표 자료보다 통제된 실제 사용입니다. 누가 데이터를 소유하고, 어떤 도구를 허용하며, 무엇을 성공으로 볼지가 정해져 있어야 하지요.

Enabled에서 First로 넘어갈 때는 기능 하나의 정확도보다 업무 전체의 결과를 봅니다. 초안 작성은 빨라졌는데 승인 대기가 길어졌다면 어떨까요? 전등은 밝아졌는데 작업대 동선은 그대로인 셈이라, 전환의 근거로는 부족합니다. 처리 시간, 재작업, 고객 결과, 사고 비용을 한 흐름에서 재 봐야 합니다.

First에서 Native로 향할 때는 가치와 의존성을 함께 봅니다. AI가 핵심 경험을 만든다면 공급자 중단이나 품질 저하에 대비한 장치도 갖춰야 하지요. 대체 모델, 기능을 줄여 버티는 제한 모드, 사람의 개입, 복구 목표가 제품 설계에 들어가 있어야 합니다. 비상 발전기 없는 반도체 공장을 떠올리면 이유가 분명해집니다.

네 단계는 속도를 재촉하려고 만든 표가 아닙니다. 다음 투자를 해도 될 증거가 있는지, 어디서 멈춰야 하는지를 한 표에서 보려는 것이지요. 다음 장에서는 첫 두 칸, 전선을 깔고 전등을 켜는 일부터 따져 보겠습니다.

AI-Ready와 AI-Enabled: 전선을 깔고 전등을 켜다

1880년대 뉴욕의 한 공장주가 전기를 들이기로 마음먹었다고 상상해 봅시다. 발전소에서 전기를 끌어올 길은 열렸지만, 공장 안에는 전선이 한 가닥도 없습니다. 벽에는 가스등 배관만 지나가고, 기계는 모두 증기 동력에 맞춰 놓여 있지요. 전기를 쓰고 싶다고 해서 내일 당장 전등을 켤 수 있는 형편이 아니었습니다.

전등을 켜려면 먼저 전선을 깔아야 합니다. 그런데 전선을 깐 공장이 모두 전등을 잘 쓴 것도 아니었지요. 어느 방에 몇 촉짜리를 달지 정하지 않으면 전기요금만 나옵니다. 이 장은 그 두 걸음, 전선을 까는 Ready와 전등을 켜는 Enabled를 다룹니다.

AI 도입을 서두르는 조직일수록 도구부터 고르기 쉽습니다. "어떤 모델을 살까?"는 사실 나중에 물어도 되는 질문입니다. 어떤 업무 결과를 바꾸려는지, 지금 결과는 어느 수준인지, 잘못됐을 때 누가 책임지는지가 먼저 정해져 있어야 하지요.

Ready, 전선을 까는 일

Ready가 끝났는지 알아보는 방법은 간단합니다. 실제 업무 하나를 골라 안전하게 시험해 볼 수 있는가. 데이터 창고와 교육 수료증은 거기까지 가는 재료일 뿐이지요.

그 조건을 갖추려면 몇 가지가 필요합니다. 출발점은 기준선입니다. 지금의 처리 시간, 오류와 재작업, 고객 불만, 건당 비용을 알아야 AI가 나아졌는지 판단할 수 있습니다. "직원이 편해졌다"는 인상만으로는 투자 결정을 내리기 어렵지요.

데이터의 계보도 따라가 봐야 합니다. 어디에서 왔고, 누가 고쳤으며, 어떤 목적으로 써도 되는지 설명할 수 있어야 하지요. 창고에 데이터가 가득해도 권리와 품질을 확인할 수 없다면, 전선은 깔았는데 피복이 벗겨진 셈입니다.

책임도 미리 나눠 둡니다. 현업 책임자는 업무 결과를, 기술팀은 모델과 시스템을, 보안·법무 담당자는 허용 범위를 맡습니다. AI가 틀렸을 때 서로 얼굴만 쳐다보는 구조라면 파일럿을 시작하기에는 이릅니다.

작동 범위는 좁게 잡습니다. 실패 비용이 낮고 결과를 검증할 수 있으며 자주 반복되는 업무가 첫 후보입니다. 전사 챗봇처럼 범위가 넓은 사업보다 상담 기록 요약이나 설비 점검 문서 분류처럼 입력과 정답을 눈으로 확인할 수 있는 일이 배우기 좋습니다.

Ready의 확인 항목	남겨야 할 증거
업무 문제	기준선과 목표 지표
데이터	출처, 권리, 품질, 보존 기준
책임	업무·기술·보안·승인 담당자
안전장치	금지 입력, 접근 통제, 기록, 사고 대응
실험 범위	대상 사용자, 기간, 중단 조건

이 가운데 하나라도 비어 있다면 어떻게 해야 할까요? 흔한 실수는 "우리도 AI 해야 하니 GPU 서버부터 사자"며 하드웨어를 들이는 것입니다. 전선도 안 깔았는데 발전기부터 사 온 격이지요. 인프라를 더 사기보다 빈칸을 메우는 편이 낫습니다.

반대쪽 함정도 있습니다. 데이터 정비와 전사 교육은 몇 년째 하는데 실제로 AI가 돌아가는 업무는 하나도 없는 조직이지요. 준비는 끝이 없습니다. 전선을 깔면서 한쪽 구석에서 전등 하나를 시험 삼아 켜 봐야, 배선이 제대로 됐는지도 알 수 있으니까요.

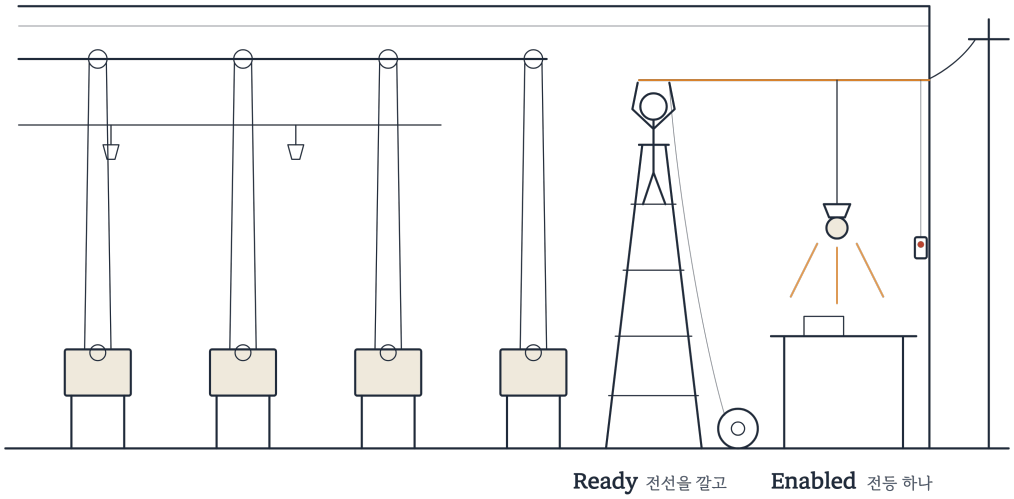


그림 3-3 전기공이 천장을 따라 전선을 끌어오는 동안, 구석 작업대 위에서는 전등 하나가 먼저 켜졌습니다. 머리 위에는 증기 시절의 전달축이 아직 그대로 걸려 있지요.

Enabled, 전등을 켠 다음

드디어 전등을 켰습니다. 공장 안이 환해졌고, 어두운 구석에서 더듬거리던 작업자들이 이제 제대로 볼 수 있게 됐지요. 그런데 천장을 올려다보면 증기 시절의 동력 전달축이 그대로 걸려 있습니다. 기계 배치도 예전 그대로고요.

Enabled 단계가 이 모습입니다. 기존 업무의 일부를 AI가 돕습니다. 검색, 요약, 분류, 초안 작성처럼 사람이 결과를 확인할 수 있는 일이 흔한 출발점이지요.

이 단계의 함정은 사용률을 성과로 착각하는 것입니다. 접속자가 많아도 수정 시간이 늘거나 중요한 오류가 숨어들면 업무는 나아지지 않았습니다. 모델 정확도가 높아도 마찬가지입니다. AI가 10분을 줄였는데 검토와 승인에 20분이 더 들 수도 있으니까요. 전등은 밝아졌는데 작업대 사이를 오가는 길은 더 멀어진 꼴입니다.

그래서 평가는 업무 흐름 전체에서 합니다. 오류와 재작업, 고객 피해가 줄었는지 봅니다. 검토와 예외 처리까지 포함해 일이 끝나는 시간이 짧아졌는지, 모델 사용료와 운영 인력을 합친 건당 비용이 나아졌는지도 따지지요. 숫자에 잘 잡히지 않는 것도 함께 물어야 합니다. 사람이 손에 익은 기술을 잃어가지는 않는지, AI 답을 지나치게 믿게 되지는 않았는지, 개인정보나 저작권 사고가 슬그머니 숨어들지는 않았는지 말이지요.

좋은 파일럿은 성공만 증명하지 않습니다. 어떤 입력에서 성능이 무너지고, 사람이 언제 끼어들어야 하며, 어느 수준이면 사용을 멈출지도 알려 줍니다. 이런 기록이 쌓여야 다음 단계의 업무 재설계가 가능해집니다.

그 기록이 무엇을 말해 줄 때 다음 칸으로 올라서면 될까요? Enabled에서 First로 넘어가도 된다는 신호는 세 가지입니다. 목표 지표가 한 번이 아니라 반복해서 좋아지고, 현업이 예외 처리 방식을 익혔으며, 더 큰 효과를 막는 병목이 기존 절차에 있다는 사실이 드러났을 때지요. 그때 비로소 역할과 승인 단계를 다시 설계할 이유가 생깁니다.

반대로 편익이 작거나 검증 비용이 지나치게 크다면 Enabled에 머물거나 도구를 거둬들일 수 있습니다. 모든 파일럿을 확산해야 한다는 압박은 실패한 사업을 오래 끌게 만들거든요.

결국 Ready와 Enabled에서 건져야 할 것은 도입 건수가 적힌 보고서보다, 어디를 넓히고 어디서 멈출지 가려내는 눈입니다. 전등을 몇 개 달았는지보다 어느 방의 불빛이 실제로 일을 바꿨는지를 적어 둔 공장이, 나중에 배치도를 다시 그릴 때 덜 헤맵니다.

AI-First와 AI-Native: 업무를 다시 설계하다

전기를 들인 공장 가운데 정말 달라진 곳은 모터를 더 많이 산 곳보다 공장 배치도를 다시 그린 곳이었습니다. 그 이야기는 제4부 첫머리에서 다시 펼치겠지만, 전등을 켜는 데서 멈추지 않고 아예 전기를 전제로 배치도를 새로 그리는 단계, 그것이 First이고 Native입니다.

1913년 무렵 헨리 포드의 하이랜드파크 공장이 좋은 예입니다. 전기를 쓰는 공장은 이미 많았지요. 포드는 한 걸음 더 나갔습니다. 동력을 작업대마다 따로 보낼 수 있다면, 사람이 부품을 찾아다니는 대신 부품이 사람 앞으로 움직여 오게 만들 수 있지 않을까? 움직이는 조립 라인이 자리를 잡자, 자동차 차대 하나를 조립하는 데 12시간 넘게 걸리던 일이 1시간 반 남짓으로 줄었다고 전해집니다. 포드가 단 것은 더 밝은 전등이 아니었습니다. 일이 흐르는 순서 자체를 바꾼 것이지요.

AI가 기존 업무의 일부를 충분히 개선했다면 우리 앞에도 같은 갈림길이 놓입니다. 지금 절차를 유지한 채 도구만 더 붙일 것인가, 판단과 책임의 흐름을 다시 짤 것인가.

First, 누가 무엇을 맡을까

AI-First라고 하면 모든 판단을 AI에 먼저 맡기는 모습이 떠오르기 쉽지요. 실제로는 새 업무를 설계하는 첫날부터 AI가 무엇을 잘하고 어디서 틀리는지를 계산에 넣는다는 말에 가깝습니다.

한 보험사 콜센터 상담원의 하루를 두 번 그려 볼까요.

Enabled 단계의 사무실에서는 문의가 들어오면 AI가 답변 초안을 띄우고, 상담원이 그것을 한 줄씩 고쳐 보냅니다. 상담원 책상에는 하루 종일 같은 양의 문의가 쌓이지요. 초안이 생겼을 뿐 일의 순서는 그대로입니다.

First 단계의 사무실은 문의가 들어오는 순간부터 길이 갈립니다. 주소 변경처럼 반복적이고 피해 가능성이 낮은 문의는 자동으로 처리되고, 불확실하거나 민감한 문의만 숙련 상담원의 책상에 도착합니다. 상담원이 고친 내용은 평가 데이터가 되고, 자주 반복되는 예외는 약관이나 제품을 고치라는 신호로 올라갑니다. 같은 상담원인데, 하루의 대부분을 정말 사람이 필요한 문의에 쓰게 되지요.

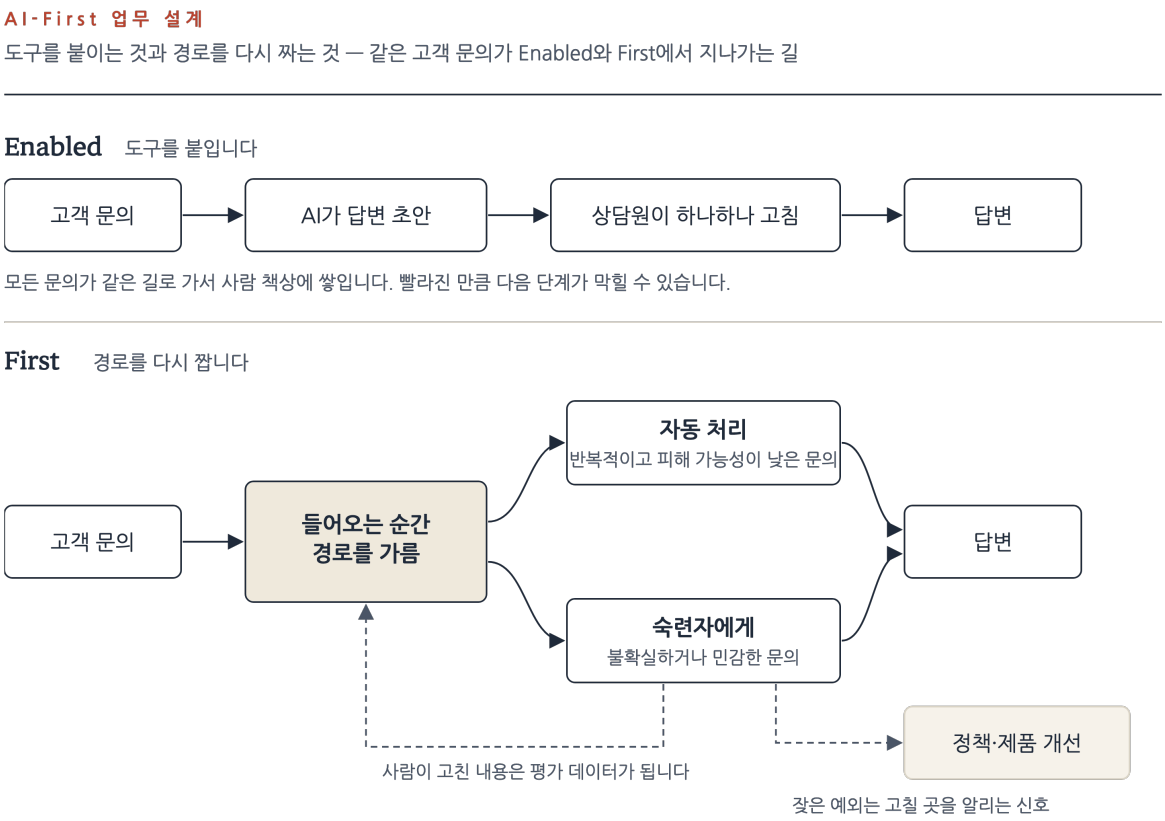


그림 3-4 Enabled에서는 모든 문의가 같은 길로 가서 상담원 책상에 쌓입니다. First에서는 문의가 들어오는 순간 길이 갈리고, 사람이 고친 내용이 다시 시스템으로 돌아오지요.

자동화율이 몇 퍼센트인지는 생각보다 덜 중요합니다. 어떤 판단을 기계에 맡기고, 언제 사람에게 넘기며, 결과는 누가 책임질지가 먼저 정해져 있어야 하지요.

설계 요소	정해야 할 내용
위험 등급	잘못됐을 때 피해의 크기와 회복 가능성
권한	AI가 제안만 할지, 실행까지 할지
개입 조건	불확실성, 금액, 민감도에 따른 사람 호출 기준
기록	입력, 근거, 수정, 최종 결정의 추적 방식
학습	수정 결과를 평가와 개선에 반영하는 절차

업무 하나가 빨라졌는데 다음 승인 단계에 일이 쌓인다면 어떨까요? 컨베이어 한 구간만 빨라지면 다음 구간 앞에 부품이 산더미처럼 쌓이듯, First 전환도 아직 끝난 것이 아닙니다. 처음부터 끝까지의 처리 시간과 오류, 예외 부담을 한 줄에 놓고 재 보세요.

Native, 그리고 비상 발전기

Native 단계에서는 AI가 핵심 가치를 만드는 방식 안으로 들어갑니다. 제3부 첫 장에서 본 반도체 팹처럼, 빼면 그 사업이라는 것 자체가 성립하기 어려운 자리지요. 동시에 사용 결과가 평가와 개선으로 돌아오는 고리가 자리를 잡습니다. 기능을 많이 넣는 것보다 이 고리가 실제로 도는지가 중요합니다.

물론 데이터가 쌓인다고 저절로 좋아지지는 않지요. 사용자가 남긴 신호에는 편향과 오류가 섞이고, 세상이 바뀌면 어제의 정답이 오늘은 틀릴 수 있습니다. 그래서 동의와 삭제 요구를 지키면서, 데이터가 한쪽으로 치우치지 않았는지, 세상이 바뀌어 옛 정답이 슬그머니 틀려지는 '드리프트'는 없는지, 잘못된 데이터가 섞여 들지 않았는지 점검해야 하지요. 학습 고리가 있다고 말하기 전에 품질이 실제로 나아졌다는 측정부터 내놓아야 합니다.

사업 구조도 따져 봅니다. 누구나 쓰는 범용 모델을 호출하는 것만으로는 오래가는 차별화가 생기기 어렵습니다. 독자적인 업무 지식, 신뢰받는 데이터, 고객 관계, 실행 체계 가운데 무엇이 성능과 전환 비용을 만드는지 설명할 수 있어야 하지요.

차별화만큼 챙겨야 할 것이 하나 더 있습니다. 비상 발전기지요. Native가 될수록 장애의 영향도 커집니다. 모델 공급자가 멈추거나 가격을 바꾸고, 품질이 갑자기 흔들릴 때 핵심 서비스가 함께 무너질 수 있지요. 반도체 공장이 비상 발전기를 설계도 첫 장에 그려 넣듯, 이런 날을 위한 장치도 처음부터 제품에 들어가 있어야 합니다.

기능을 줄여서라도 서비스를 이어가는 비상 모드(저하 모드)가 있는지, 대체 모델이나 규칙 기반 경로로 갈아탈 절차가 있는지 먼저 봅니다. 고위험 결정은 멈추고 사람에게 넘기는 장치, 장애를 발견하고 복구할 목표 시간도 필요합니다. 고객에게 무슨 일이 있었는지 알리고 잘못된 결과를 바로잡는 절차까지 갖춰야 비로소 정전에 대비한 공장이라 할 수 있습니다.

AI에 많이 기댄다고 해서 그것만으로 Native가 되는 것은 아닙니다. 고객 가치, 학습 능력, 단위경제성, 안전, 복원력이 함께 확인될 때 구조적 전환이라 부를 수 있지요.

그러니 Native가 모든 업무의 종착역일 필요도 없습니다. 같은 조직 안에서도 결론은 달라집니다. 회의록 초안은 Enabled로 충분할 수 있고, 공급망 계획은 First가 알맞을 수 있습니다. 실시간 개인화가 제품의 핵심이라면 Native를 지향할 이유가 생기지요.

모든 업무를 마지막 칸까지 밀어 넣을 필요는 없습니다. 증거가 쌓인 곳은 구조를 바꾸고, 위험이 더 큰 곳에는 사람의 판단과 기존 절차를 그대로 남겨 두면 되지요. 이 선택을 꾸준히 해 나갈 때 AX는 한때의 유행을 지나 회사의 일하는 방식으로 자리 잡습니다.

세 가지 전환 경로: 출발점이 다르면 길도 다르다

AX의 방향은 같아도 알맞은 도착점은 업무마다 다릅니다. 100년 된 제조 대기업과 작년에 창업한 스타트업이 같은 길을 걸을 수는 없지요. 새로 짓는 공장은 처음부터 전기 설비를 설계에 넣을 수 있지만, 가동 중인 공장은 생산을 멈추지 않은 채 한 라인씩 바꿔야 했던 것과 비슷합니다.

조직의 유형에 따라 크게 세 가지 전환 경로가 있습니다.

스타트업은 처음부터 Native로

레거시가 없다는 것은 엄청난 이점입니다. 빈 땅에 새 건물을 올리며 처음부터 전기를 전제로 배선을 설계하는 것과 같지요. 기존 시스템을 마이그레이션할 필요도, 수천 명의 직원을 재교육할 필요도, "지금까지 이렇게 해왔는데"라는 관성과 싸울 필요도 없습니다.

AI 시대의 스타트업은 처음부터 AI를 전제로 제품을 설계할 수 있습니다. 레거시가 적다는 장점을 살리되, Native라는 이름을 먼저 붙이기보다 고객 가치와 안전, 단위경제성이 함께 성립하는지 검증해야 합니다.

초기 제품은 범위를 좁혀야 합니다. 회계와 인사까지 모두 자동화할 이유는 없습니다. 고객이 반복해서 겪는 문제 하나에서 AI가 기존 방식보다 분명한 가치를 만드는지 확인하는 편이 낫습니다.

스타트업의 경로는 달력보다 증거로 관리하는 편이 낫습니다.

먼저 AI를 빼면 가치 제안이 성립하지 않는 기능을 찾습니다. 실제 사용자가 반복해서 쓰고, 사람이 대신할 때보다 분명한 효용이 확인되면 다음 걸음을 뗍니다. 검증 비용이 효용을 넘거나 사용자가 결과를 믿지 못한다면 기능 범위를 줄이지요.

그다음엔 데이터의 고리가 실제로 도는지 봅니다. 사용 과정에서 합법적이고 쓸모 있는 피드백이 쌓이며 품질이 나아지는가. 데이터가 늘어도 성능이 제자리이거나 개인정보·저작권 조건을 채우지 못한다면 확장을 멈춥니다.

품질과 단위경제성, 안전 지표가 정해 둔 기준을 연달아 통과할 때 비로소 고객과 기능을 넓힙니다. 쓸수록 좋아지는 고리는 기대만으로는 해자가 되지 못합니다. 측정된 결과로 쌓여야 경쟁자가 넘기 어려운 해자가 되지요.

대기업은 한 라인씩 갈아 끼우며

대기업에게 AX는 훨씬 복잡한 길입니다. 이미 증기로 돌아가는 공장을 멈추지 못한 채 한 라인씩 전기로 갈아 끼우는 리모델링에 가깝지요. 수십 년간 쌓인 레거시 시스템, 수만 명의 인력, 확립된 프로세스와 문화 — 이 모든 것이 관성으로 작용합니다.

하지만 이것은 장애물인 동시에 자산이기도 합니다. 대기업에는 스타트업이 갖지 못한 것이 있습니다. 방대한 데이터, 고객 기반, 산업 경험, 자금력이지요. 문제는 이 자산을 새로운 구조 안에서 활용하는 방법을 찾는 것입니다.

대기업은 사업부와 업무마다 단계가 다릅니다. 전체 조직을 한꺼번에 한 칸씩 옮기기보다, 데이터와 책임자가 준비된 한 사업부나 제품 라인에서 증거를 만든 뒤 비슷한 조건의 업무로 넓히는 방식이 현실적입니다.

대기업의 경로는 더 길지만, 여기서도 기간보다 진입 조건이 중요합니다.

Ready에서 Enabled로 넘어가려면 데이터의 소유자와 품질 기준, 허용 도구, 사고 대응 책임자가 정해져 있어야 합니다. 이 조건 없이 파일럿 수만 늘어난다면 기반부터 고치는 편이 낫지요.

Enabled에서 First로 가려면 핵심 업무의 파일럿이 기존 방식보다 나은 품질·시간·비용을 되풀이해 보여주고, 현업 책임자가 확산을 맡아야 합니다. 검토 인력이 지나치게 늘거나 사고율이 허용 범위를 넘으면 자동화 범위를 되돌립니다.

First에서 Native 쪽으로 방향을 잡으려면, 한 사업부가 AI를 전제로 프로세스를 다시 설계하고도 고객 가치와 규제 준수, 운영 복원력을 함께 증명해야 합니다. 특정 벤더가 멈추면 사업도 멈추는 구조라면 전사로 넓히기 전에 대체 경로부터 마련합니다.

각 단계는 몇 달 만에 통과할 수도, 몇 년이 걸릴 수도 있습니다. 산업의 규제와 데이터 상태, 실패 비용이 다른데 같은 달력을 적용하면 속도만 성과처럼 보이게 됩니다.

중견·중소기업은 사정이 또 다릅니다. 대기업의 축소판으로 보면 길을 잘못 들기 쉽지요. 전담 데이터팀이 없으니 직접 구축하기보다 검증된 범용 서비스를 사서 쓰는 편이 대개 낫고, 그래서 Ready 단계의 무게가 인프라보다 계약 쪽으로 옮겨 갑니다. 데이터를 어디까지 내보낼지, 서비스를 끊어도 업무를 이어 갈 수 있는지를 먼저 정해 두세요. 부록 A의 CEO 계약 카드가 그 점검표 역할을 합니다.

제조업은 물리 세계와 함께

제조업의 AX는 디지털 기업과는 다른 차원의 과제를 안고 있습니다. 중앙 전달축을 걷어내고 기계마다 모터를 앉히던 그 공장 바닥의 전기화를, 이번엔 지능 에너지로 한 번 더 겪는 셈이지요. 소프트웨어 회사는 코드를 바꾸면 되지만, 공장은 물리적 설비를 바꿔야 합니다. 로봇 팔을 교체하고, 센서를 설치하며, 공정 라인을 재배치해야 하지요. 비용도 크고 시간도 오래 걸립니다.

하지만 제조업에는 강력한 장점이 있습니다. 물리적 세계에서 생성되는 데이터의 양과 종류가 엄청나다는 점이지요. 센서 데이터, 품질 검사 데이터, 설비 가동 데이터, 환경 데이터 — 이 모든 것이 AI의 원료가 됩니다.

제조업의 AX에서 특히 주목할 흐름은 Physical AI입니다. 물리 세계를 감지하고 그 안에서 행동하는 AI라는 뜻이지요. 다만 휴머노이드 한 종류만을 떠올릴 필요는 없습니다. 검사 카메라, 자율 이동 로봇, 공정 제어처럼 범위가 좁고 성능을 측정할 수 있는 시스템이 먼저 현장 가치를 만들 수 있습니다.

제조업을 위한 로드맵은 두 트랙으로 진행됩니다.

한쪽은 디지털 트랙입니다. 설계, 시뮬레이션, 공급망 관리를 AI 중심으로 바꾸지요. 디지털 트윈(공장의 가상 복제본)을 만들어 AI가 공정을 시뮬레이션하고 개선안을 찾게 합니다. 소프트웨어 영역이라 비교적 빠르게 움직일 수 있습니다.

다른 한쪽은 물리 트랙입니다. 생산 현장에 AI를 붙이는 일이지요. 비전 AI 기반 품질 검사, 예측 정비(고장이 나기 전에 AI가 정비 시점을 알려주는 것), 자율 로봇 도입 등이 여기에 들어갑니다. 설비 투자가 필요해 느리지만, 한번 자리 잡으면 경쟁자가 쉽게 따라오지 못하는 물리적 해자가 됩니다.

두 트랙이 만날수록 설계 변경이 현장 조건에 맞는지 가상 환경에서 시험하고, 실제 운전 결과를 다시 설계에 반영할 수 있습니다. 연결이 많아질수록 안전 검증과 장애 격리도 함께 강화해야 합니다.

세 가지 전환 경로

출발점이 다르면 길도 다릅니다 — 달력보다 다음 걸음을 떼도 되는 증거로 관리합니다

스타트업

빈 땅에 처음부터 전기를 전제로 짓는 공장



이 증거가 보이면 넓힙니다

1. AI를 빼면 가치 제안이 성립하지 않는 기능
2. 쓸수록 품질이 나아지는 데이터 고리
3. 품질·단위경제성·안전 지표의 연속 통과

멈출 때 — 검증 비용이 효용을 넘거나, 사용자가 결과를 믿지 못할 때

대기업

가동을 멈추지 않고 한 라인씩 바꾸는 리모델링



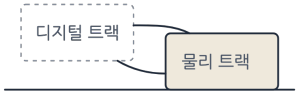
단계마다 진입 조건

1. Enabled로: 데이터 소유자·품질 기준·허용 도구·사고 대응 책임자
2. First로: 기존 방식보다 나은 결과의 반복, 현업 책임자의 확산
3. Native 쪽으로: 고객 가치·규제 준수·운영 복원력의 증명

되돌릴 때 — 검토 인력이 지나치게 늘거나 사고율이 허용 범위를 넘을 때

제조업

공장 바닥의 전기화를 지능으로 한 번 더



두 트랙이 만나는 곳

1. 디지털: 설계·시뮬레이션·공급망, 디지털 트윈
2. 물리: 비전 검사, 예측 정비, 자율 로봇
3. 운전 결과를 다시 설계에 반영

함께 키울 것 — 연결이 많아질수록 안전 검증과 장애 격리

어떤 경로든 먼저 정할 것 — 품질과 비용, 안전에서 무엇이 보이면 넓히고 무엇이 보이면 멈출지. 중견·중소기업은 직접 구축보다 검증된 서비스를 쓰는 경우가 많아, Ready의 무게가 계약 쪽으로 옮겨 갑니다.

출처: 저자 작성 — 본문 「세 가지 전환 경로」의 스타트업·대기업·제조업 경로와 단계별 진입·중단 조건을 정리.

그림 3-5 빈 땅에 새로 짓는 공장, 돌아가는 라인을 하나씩 바꾸는 공장, 두 트랙을 함께 움직이는 공장. 출발점은 달라도 넓힐 조건과 멈출 조건을 먼저 적어 두는 일은 같습니다.

이렇게 세 길을 나란히 놓고 보면, 다른 점보다 같은 점이 오히려 눈에 들어옵니다. 빈 땅에 새 공장을 짓는 스타트업이든, 가동 중인 라인을 하나씩 갈아 끼우는 대기업이든, 쇳물과 기름 냄새 속에서 센서를 다는 제조 현장이든 먼저 확인하는 것은 같습니다. 품질과 비용, 안전에서 무엇이 보이면 넓히고 무엇이 보이면 멈출지 정해 두었는가. 실패 비용이 크거나 결과를 검증할 수 없는 영역이라면 보류와 철회도 올바른 결정이지요.

그런데 세 길은 모두 조직이 위에서 아래로 놓는 길입니다. 2026년 여름, 한국 공공부문에서는 정반대 방향에서 올라오는 길이 눈에 띄게 넓어졌습니다. 다음 장은 그 길을 따라가 봅니다.

아래에서 오르는 사다리: 공무원이 직접 만든 589개의 AI

밤 아홉 시가 넘은 시청 사무실을 떠올려 볼까요. 자리마다 불이 꺼졌는데 한 칸에만 모니터가 켜져 있습니다. 낮에 열댓 번 고쳐 쓴 공문 서식을 붙잡고 있던 주무관이, 이번에는 그 일을 대신 해 줄 작은 도구를 만들고 있지요. 한글 문서를 읽어 기안 틀에 맞춰 다시 써 주는 프로그램입니다. 그는 개발자가 아니고, 누가 시킨 일도 아닙니다. AI에게 말로 설명하고, 돌려 보고, 틀린 곳을 다시 설명하기를 몇 시간. 다음 날 아침 그는 완성된 도구를 공무원들이 모인 단체 대화방에 올립니다. "필요하신 분 쓰세요."

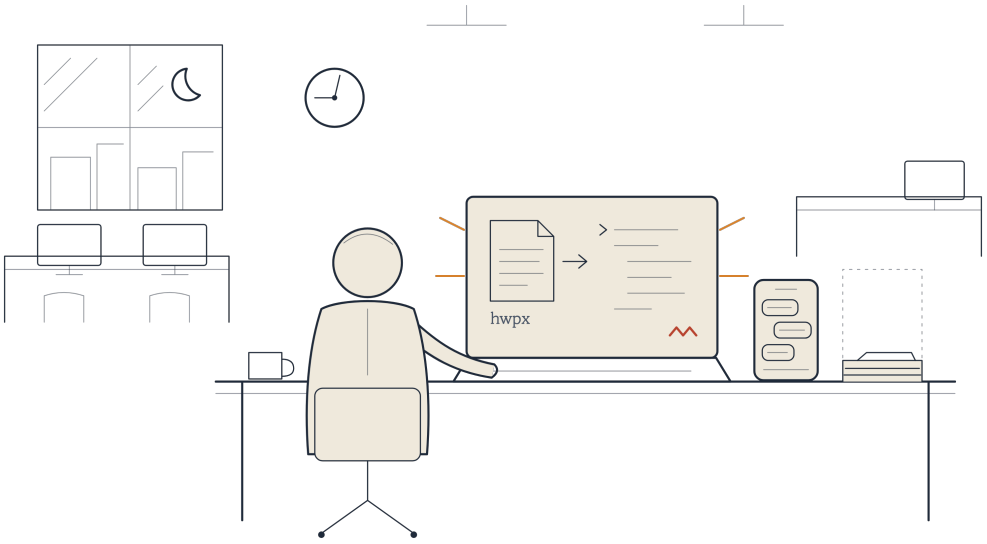


그림 3-6 퇴근한 사무실 한 칸에 모니터 하나가 켜져 있습니다. 낮에 손으로 하던 일을 덜어 줄 도구가 밤사이 만들어지고, 아침이면 단체 대화방으로 건너가지요.

589건이 말해 주는 것

이런 장면이 2026년 여름부터 한국 공공부문 곳곳에서 한꺼번에 벌어졌습니다. 필자가 운영하는 '모두의 공공AX 사례 아카이브'는 공무원과 시민이 소셜 미디어와 오픈채팅에 스스로 공유한 공공 AX 사례를 모으는데, 2026년 10월 21일까지 589건이 쌓였습니다.[1] 7월까지 한 달에 몇 건이던 것이 8월에 268건, 9월에 201건으로 뛰었지요. 8월 초 공무원들이 도구를 주고받는 오픈채팅 방이 열린 뒤의 일입니다.

누가 만들었을까요? 589건 가운데 339건, 58%가 '개인 개발'로 분류됩니다. 기관이 공식 사업으로 내놓은 사례는 57건으로 열에 하나꼴이지요.

무엇을 만들었는지도 볼까요? 데이터·통계 135건, 문서·기안 112건, 감사·법무 60건, 민원 32건. 화려한 대민 서비스보다 매일 손이 가는 실무가 먼저입니다. 한글(HWPX) 문서를 다루는 도구, 법령을 찾아 주는 MCP 서버, 엑셀 취합 자동화 같은 것들이지요. 실행 환경도 눈여겨볼 만합니다. 브라우저만 있으면 되는 도구가 198건, 인터넷 없이 내부에서 돌아가거나 폐쇄망에서 쓸 수 있는 도구가 132건입니다. 망분리라는 울타리 안에서 스스로 길을 낸 셈이지요.

이 숫자는 조심해서 읽어야 합니다. 아카이브는 자발적으로 공유된 사례만 모은 자기선택 표본이고, 전국 공공부문을 대표하지 않는다고 스스로 밝혀 둡니다. 사례 요약도 AI가 다시 쓴 것이라 원문과 다를 수 있습니다. 그래도 방향만큼은 분명히 보입니다. 지능을 쓰는 사람이 지능으로 도구를 만드는 사람이 되어 가는 중이라는 것이지요.

필자에게는 낯익은 풍경이기도 합니다. 서문에서 말씀드린 1981년의 라디오 잡지 소년은 세운상가에서 부품을 사다 인두로 회로를 만들었습니다. 지금 공무원들은 인두 대신 대화창을 쥐고 있지요. 전기 역사로 치면 7부에서 다룰 미국 농촌의 전기 협동조합과도 닮았습니다. 전력회사의 전선이 오기를 기다리지 않고 마을이 스스로 전기를 끌어왔듯, 기관의 공식 시스템이 닿기 전에 현장이 먼저 작은 발전기를 돌리기 시작한 것입니다.

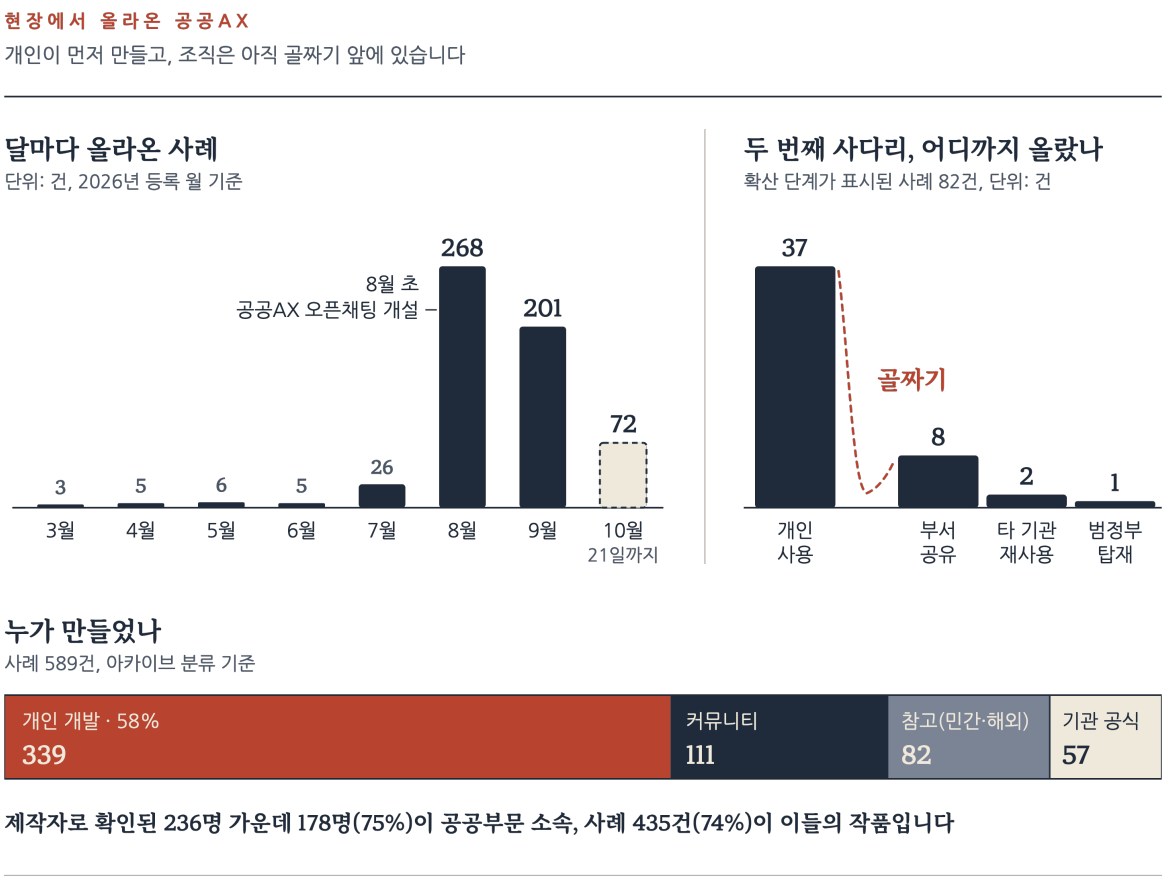


그림 3-7 월별 사례 수, 누가 만들었는지, 개인이 만든 도구가 조직으로 얼마나 넘어갔는지를 한 장에 모았습니다. 마지막 계단에서 숫자가 가파르게 줄어드는 자리가 이 장이 말하려는 골짜기이지요.

챔피언이 앞에 서다

이 변화의 앞줄에는 이름 붙은 사람들이 있습니다. 아카이브는 사례를 만든 사람들을 공개 GitHub와 정부 공공 GitLab 프로필로 확인해 '공공AX 챔피언' 디렉터리로 정리하는데, 지금까지 236명이 올라 있습니다.[2] 그중 178명, 넷에 셋이 중앙부처·지자체·공공기관·교육기관 소속이지요. 이들이 만든 사례가 435건, 전체의 74%입니다. 한 건만 올린 사람이 대부분이지만 열 건 넘게 만든 사람도 여럿 있습니다. 현장의 실험이 몇몇 열성적인 사람에게서 시작해 동료에게 번져 가는 모양이지요.

정부도 같은 방향을 제도로 받치기 시작했습니다. 행정안전부의 'AI 챔피언'은 공공 현장에서 AI로 업무 문제를 직접 풀고, 그 경험을 조직 안에 퍼뜨릴 실무 인재를 키우고 인증하는 제도입니다. 생성형 AI 활용, 데이터 분석, 업무 자동화, 바이브코딩, AI 서비스 기획, PoC 구현을 배우고, 그린·블루·블랙 등급으로 기본 활용부터 업무 적용, 고도화된 문제 해결, 조직 내 확산까지 단계를 밟아 검증받습니다. 외부 개발 용역에 기대지 않고 담당자가 직접 업무 도구를 설계하도록 돕는 데 초점이 있지요.[3] 정부는 이 인증을 공직자 2만 명 규모로 넓히겠다는 목표를 세워 두었습니다. 이 목표는 「대한민국 AI 행동계획」 장에서 다시 다룹니다.

실제로 인증을 받은 공무원들이 만든 것을 보면 아카이브의 사례와 그대로 겹칩니다. 한 시청 주무관은 공사 뒤 따라붙는 공문 다섯 건을 한 번에 만들어 주는 도구를 바이브코딩으로 만들었고, 한 구청 주무관은 법령 검색용 MCP 서버와 문서 변환 도구를 포함해 오픈소스 도구 18종을 공개했습니다. 개발을 전공하지 않은 공공기관 과장이 수질 관리 업무를 자동화한 사례도 있지요. 아카이브의 챔피언 디렉터리는 행안부 인증 명부와 이름·기관이 정확히 일치하는 사람에게만 인증 배지를 붙입니다. 배지가 붙은 사람이 아직 몇 명뿐인 것도 그 때문이지요. 실제 인증자는 그보다 많다고 보는 편이 맞습니다.

한 업계 관계자는 이렇게 평했습니다. "AI 챔피언은 자발적인 현장 실험을 검증하고 연결하는 통로가 되고 있다." 아래에서 올라온 실험과 위에서 내려온 제도가 처음으로 같은 자리에서 만나고 있는 셈이지요.

네 개의 부품이 이어질 때

이렇게 만든 도구들은 따로 놀지 않습니다. 대표적인 네 가지를 한 줄로 세워 보면 공문 한 건이 지나가는 길이 그대로 보이지요.

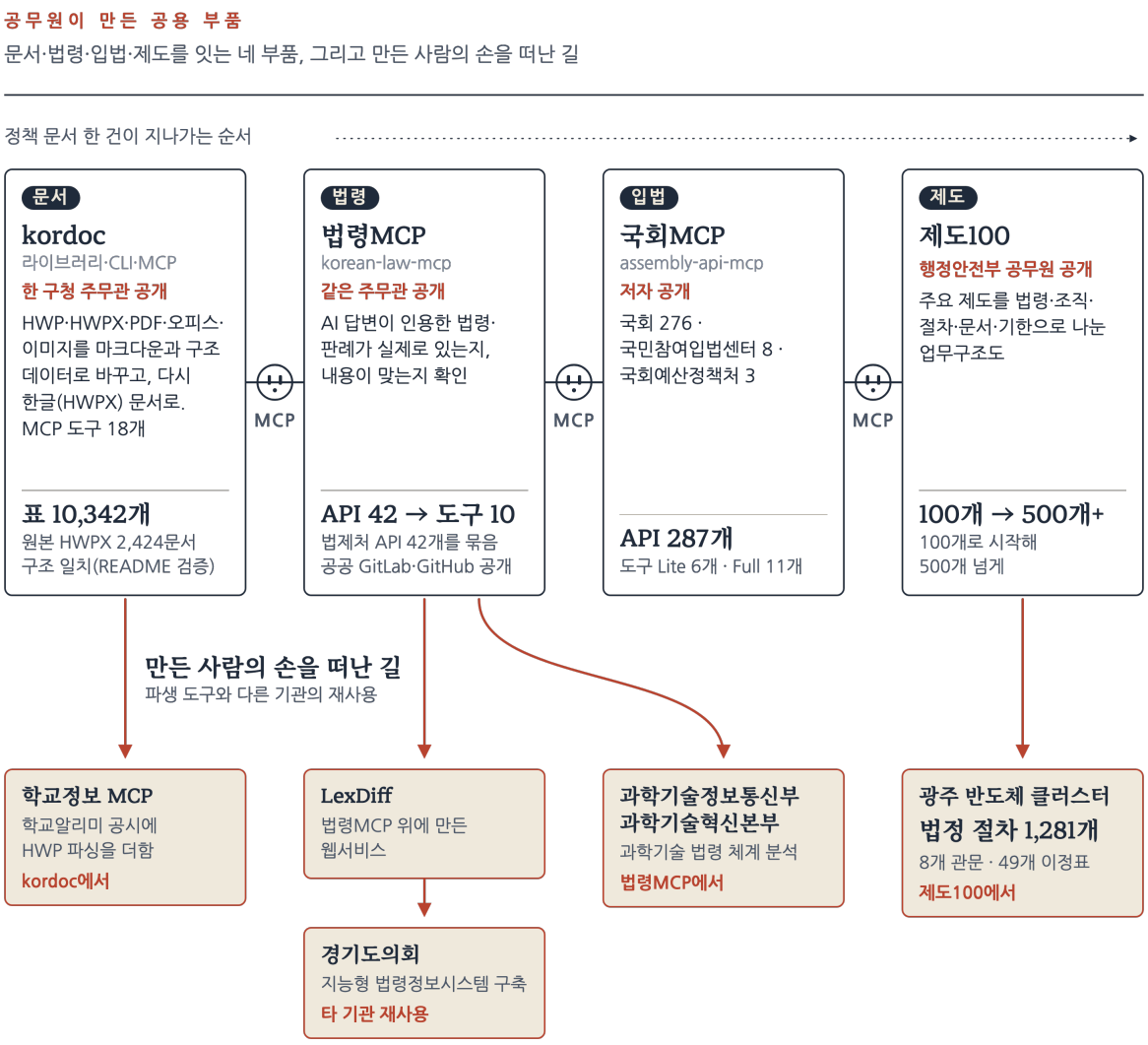
맨 앞은 문서입니다. 공공부문의 문서는 대부분 한글(HWP·HWPX) 파일인데, AI가 이 파일을 제대로 읽지 못하면 아무것도 시작되지 않습니다. 앞에서 말한 구청 주무관이 공개한 kordoc은 한글 문서와 PDF, 오피스 문서, 스캔 이미지를 AI가 읽을 수 있는 마크다운과 구조 데이터로 바꾸고, 반대로 마크다운 초안을 기안문 서식의 한글 파일로 되돌려 줍니다. 라이브러리로도, 명령줄 도구로도, MCP 서버로도 쓸 수 있게 열어 두었고, 공개된 검증 결과로는 원본 HWPX 2,424개 문서에 든 표 10,342개의 구조를 모두 그대로 읽어 냈습니다.[4]

문서를 읽었으면 그 안의 근거를 확인해야 하지요. 같은 주무관이 만든 법령MCP는 법제처의 공개 API 42개를 AI가 부를 수 있는 도구 10개로 묶었습니다. 법령과 판례, 행정규칙, 자치법규를 찾아 주는 것에 그치지 않고, AI가 답변에 인용한 조문과 판례가 실제로 있는지, 내용이 맞는지까지 검증해 법령 환각을 줄입니다.[5] 정부 공공 GitLab과 GitHub에 함께 공개되자 외부 기여자들이 고칠 곳을 제보하고 기능을 보태며 함께 키우고 있지요.

법령의 앞에는 국회가 있습니다. 필자가 만든 국회MCP는 국회 API 276개와 국민참여입법센터 API 8개, 국회예산정책처 API 3개, 모두 287개를 AI에서 바로 조회하게 이어 줍니다. 의안의 처리 상태와 회의록, 의원 발의 기록을 질문 하나로 따라가는 식이지요.[6]

그리고 이 모든 것이 놓이는 지도가 있습니다. 행정안전부의 한 공무원이 만든 제도100은 나라의 주요 제도를 법령·조직·절차·문서·기한으로 분해해 제도마다 한 장의 업무구조도로 보여 줍니다. 백 개로 시작한 제도가 오백 개를 넘겼고, 화면마다 법령 기준일을 적고 유권해석이 아니라 참고자료라고 밝혀 둡니다. 이 지도를 광주 반도체 클러스터 사업에 적용해 완공까지의 법정 절차 1,281개를 8개 관문과 49개 이정표로 정리한 절차 지도도 나왔지요.[7]

조례 개정을 검토하는 주무관을 떠올려 보세요. 지침 파일을 kordoc으로 읽고, 인용된 조문을 법령MCP로 확인하고, 상위법 개정안이 국회 어디쯤 있는지 국회MCP로 살피고, 그 제도가 어느 부처의 어떤 절차와 맞물리는지 제도100에서 짚은 뒤, 다시 kordoc으로 기안문을 한글 파일로 써 냅니다. 서로 다른 사람이 따로 만든 도구들이 MCP라는 같은 콘센트 규격으로 이어지기에 가능한 흐름입니다. 2부에서 송전망과 표준 전압을 이야기하며 기대했던 장면이 공공부문 안에서 이렇게 먼저 나타난 셈이지요.



출처: 각 공개 저장소 README(kordoc, korean-law-mcp, assembly-api-mcp), 제도100(hosungseo.github.io/korea100), 모두의 공공AX 사례 아카이브 (2026년 10월 21일 기준). 수치는 개발자 공개 기준이며 독립 검증 전입니다.

그림 3-8 문서를 읽는 부품, 법령을 확인하는 부품, 입법 흐름을 따라가는 부품, 제도의 지도. 주홍 화살표는 이 부품들이 만든 사람의 손을 떠나 다른 기관과 다른 도구로 건너간 길입니다.

몇몇 부품은 만든 사람의 손을 떠나 다른 기관까지 건너갔습니다. 법령MCP에서 갈라져 나온 웹 서비스 LexDiff는 경기도의회가 가져다 자체 지능형 법령정보시스템으로 구축했고, 화면 아래에 원작자를 밝혀 두었습니다.[8] 과학기술정보통신부 과학기술혁신본부는 공개된 법령MCP로 과학기술 법령 사이의 근거·위임·충돌 관계를 분석했지요.[9] kordoc의 한글 문서 해석 기능은 같은 개발자가 학부모를 위해 만든 학교정보 MCP 안으로 들어가, 학교 공시와 한글 파일로 된 수행평가 계획을 한 화면에 모아 줍니다. 소스를 공개하고, 쓸 수 있는 범위와 한계를 밝히고, 원작자를 표기하는 것. 개인의 발전기가 다른 집에 전기를 나눠 주기 시작한 것은 이런 작은 습관 덕분입니다.

개인의 발전기를 마을의 전력망으로

그렇다면 이 589개의 작은 발전기는 조직의 전력망에 연결되고 있을까요? 아카이브에서 확산 단계가 표시된 사례만 보면 아직 아닙니다. 개인이 쓰는 단계가 37건인데 부서가 함께 쓰는 단계는 8건, 다른 기관이 가져다 쓴 사례는 2건, 범정부 시스템에 올라간 사례는 1건입니다. 앞 장들이 그린 Ready에서 Native로 가는 사다리가 기관이 위에서부터 놓는 사다리라면, 이것은 개인이 아래에서부터 오르는 두 번째 사다리입니다. 그리고 그 사다리의 첫 계단과 둘째 계단 사이에 깊은 골짜기가 있지요. 앞에서 본 LexDiff처럼 이 골짜기를 건넌 도구는 아직 손에 꼽힙니다.

골짜기가 생기는 이유는 짐작할 만합니다. 혼자 쓰는 도구는 만든 사람이 책임지면 되지만, 부서가 함께 쓰려면 보안 점검과 개인정보 검토, 누가 고칠지 정하는 유지보수 책임, 정식 시스템으로 올리는 절차가 필요합니다. 담당자가 인사이동으로 자리를 옮기면 도구도 함께 멈추기 쉽고요. 전기로 치면, 집집마다 돌리던 발전기를 마을 전력망에 연결하려면 전압과 플러그 규격부터 맞춰야 하는 것과 같습니다.

그 규격을 만들려는 시도도 시작됐습니다. 필자가 사례를 패턴 문헌과 겹쳐 읽어 정리한『모두의 공공AX 디자인 패턴』은 사례 564건에서 기관들이 함께 쓸 수 있는 부품 91종과 피해야 할 안티패턴 16종을 업무, 공통 서비스, 운영, 보안, 상호작용 같은 8개 계층으로 묶었습니다.[10] 함께 공개한 '패턴 성좌'에서는 자기 업무의 문제를 적으면 비슷한 부품과 사례, 갖춰야 할 안전장치, 첫 실행 단계를 찾아볼 수 있지요. 콘센트와 차단기가 규격품이 된 뒤에야 전기가 집집마다 들어갔듯, 공공 AX도 각자 만든 도구가 서로 끼워 맞춰질 때 비로소 조직의 것이 됩니다.

정책이 할 일도 여기서 보입니다. 잘 만든 개인 도구를 부서와 다른 기관이 가져다 쓸 수 있는 등록 경로, 망분리 환경에서 도구를 검증해 주는 지원, 만든 사람이 자리를 옮겨도 도구가 살아남는 유지보수 체계. 챔피언을 몇 명 배출했는지보다 이 골짜기를 몇 개의 도구가 건넜는지가 더 정직한 성적표일 겁니다.

개인이 만든 도구가 조직의 문턱에서 멈추는 이유와, 조직이 스스로 AX 앞에서 멈춰 서는 이유는 꽤 닮았습니다. 다음 장은 그 멈춤의 모양들을 모았습니다.



1. 「모두의 공공AX 사례 아카이브」(전종홍(이 책의 저자) 운영, https://hollobit.github.io/PAX/, 2026년 10월 21일 집계) — Threads '공공AX' 태그와 카카오톡 오픈채팅 '공공AX 네트워크'에 자발적으로 공유된 사례를 8시간마다 수집. 자기선택 표본이며 전국 공공부문을 대표하지 않음, 사례 요약은 AI가 다시 쓴 것임을 명시. 사례 분류·월별 수·실행 환경·확산 단계 수치는 공개 데이터(data/cases.json) 기준.

2. 「공공AX 챔피언」(모두의 공공AX 사례 아카이브, https://hollobit.github.io/PAX/champions.html) — 등재 사례의 GitHub·정부 공공 GitLab 공개 프로필을 근거로 만든 제작자 디렉터리. 공개 프로필 밖의 실명·소속은 싣지 않으며, 행정안전부 AI 챔피언 인증 배지는 공개 인증 명부와 성명·기관이 일치할 때만 표시.

3. "AI로 일하는 공무원들…행정현장에서 혁신 이끈다" (ZDNet Korea, 2026.9.4, https://zdnet.co.kr/view/?no=20260904110842) — 행정안전부 AI 챔피언 제도의 목적·교육 내용·그린·블루·블랙 등급과 단계별 검증 방식, 인증 공무원들이 직접 만든 업무 도구 사례.

4. kordoc (GitHub, https://github.com/chrisryugj/kordoc) — HWP 3.x·5.x·HWPX·HWPML·PDF·XLSX·DOCX·PPTX·이미지를 마크다운·구조 데이터로 변환하고 마크다운을 HWPX로 생성. 라이브러리·CLI·MCP 서버(도구 18개). README 검증: 원본 HWPX 2,424문서의 표 10,342개 구조 일치(개발자 공개 기준).

5. Korean Law MCP (정부 공공 GitLab https://gitlab.aigov.go.kr/chrisryugj/korean-law-mcp, GitHub https://github.com/chrisryugj/korean-law-mcp) — 법제처 API 42개를 도구 10개로, LLM 환각 방지를 위한 법령·판례 인용 검증(실존+내용), 외부 기여 반영 이력 공개.

6. assembly-api-mcp (전종홍(이 책의 저자), GitHub https://github.com/hollobit/assembly-api-mcp) — 국회 276·국민참여입법센터 8·국회예산정책처 3, 모두 287개 API, Lite 6·Full 11개 도구.

7. 대한민국 제도 100 (https://hosungseo.github.io/korea100/) — 제도별 요약 캔버스와 업무구조도, 법령 기준일 표기와 '유권해석 아님' 명시. 광주 반도체 클러스터 법정 절차 지도: https://hosungseo.github.io/korea100/mega-projects/gwangju-semiconductor-cluster/

8. 「경기도의회 지능형 법령정보시스템 — LexDiff 타 기관 재사용 구축」 (모두의 공공AX 사례 아카이브, https://hollobit.github.io/PAX/case/2808aea47c52b21c.html) — 원작 LexDiff(https://gitlab.aigov.go.kr/chrisryugj/lexdiff)를 다른 기관이 재사용한 계보와 원작자 표기.

9. 「법령 MCP로 과학기술 법령 체계 분석」 (과학기술정보통신부 과학기술혁신본부, https://msitsti2025.github.io/stlaws/; 아카이브 기록 https://hollobit.github.io/PAX/case/0af085f0e51a1a7e.html)

10. 『모두의 공공AX 디자인 패턴』 2판 (전종홍(이 책의 저자), https://hollobit.github.io/PAX-DP/, 카탈로그 v0.7.1, CC BY-NC-SA 4.0) — PAX 사례 564건 기반 부품 91종·안티패턴 16종, 8개 계층. 패턴 성좌(3D 시각화와 패턴 상담): https://hollobit.github.io/PAX-DP/constellation/

AX 실패의 패턴

새 기술을 도입했다고 운영 성과가 자동으로 따라오지는 않습니다. AI 전환도 마찬가지입니다. 파일럿은 늘었는데 고객 결과와 업무 구조는 그대로인 조직이 적지 않습니다.

왜 멈출까요? 패턴이 있습니다.

멈춰 선 조직의 세 가지 패턴

1. 기술만 바꾸고 구조는 안 바꾸기. 가장 흔한 실패 패턴입니다. 최신 AI 도구를 사들이고, AI팀을 만들고, 파일럿 프로젝트를 돌립니다. 여기까지는 좋습니다. 하지만 기존의 의사결정 구조, 보고 체계, 부서 간 역할 분담은 그대로 둡니다.

이건 마치 전기 모터를 사다 놓고 증기기관 시절의 그 중앙 전달축에 도로 연결하는 것과 같습니다. 기계마다 따로 돌 수 있다는 모터의 장점이 거기서 사라지지요.

가상의 보험 청구 사례를 생각해 봅시다. AI가 몇 초 만에 승인 권고를 내지만 심사관이 모든 자료를 처음부터 다시 확인한다면 전체 처리 시간은 거의 줄지 않습니다. 반대로 자동 승인 범위를 넓히기만 하면 고위험 오판이 고객 피해로 이어질 수 있습니다.

이 패턴을 깨려면 기술 도입과 함께 프로세스 재설계가 동시에 일어나야 합니다. AI가 자율적으로 처리할 수 있는 영역을 정의하고, 인간의 역할을 감독과 예외 처리로 재배치해야 합니다.

2. 리더십의 부재. AX는 사업 전략에 속하는 일입니다. 그런데 많은 조직에서 AI 전환을 CTO나 IT 부서에 맡기고, 경영진은 분기별 보고만 받습니다.

AX에는 사업 책임자의 지속적인 관여가 필요합니다. 자동화 범위와 실패 허용치, 인력과 예산의 이동은 기술팀만 결정할 수 없기 때문입니다.

거꾸로 "AI 전환은 IT팀이 알아서 하는 거 아닌가?"라고 생각하는 경영진 아래에서는 어떻게 될까요? AI 파일럿이 산발적으로 진행되고, 성공해도 조직 전체로 번지지 못하며, 예산이 깎이면 가장 먼저 잘리는 프로젝트가 됩니다.

리더십의 역할은 구체적입니다. 방향을 설정하고, 자원을 배분하며, 조직 구조 변경을 승인하고, 실패를 용인하는 문화를 만드는 것입니다. 이 중 어느 하나라도 빠지면 AX는 표류합니다.

3. AI를 IT 프로젝트로 취급하기. 세 번째 패턴은 두 번째와 관련이 깊습니다. AI 전환을 기존 IT 프로젝트처럼 관리하는 것이지요. 요구사항을 정의하고, 일정을 잡고, 개발하고, 테스트하고, 배포하는 — 익숙한 워터폴 방식 말입니다.

문제는 AI가 전통적인 소프트웨어와 근본적으로 다르다는 점입니다. 전통적인 소프트웨어는 규칙을 코딩하면 예측 가능하게 동작합니다. AI는 데이터에서 패턴을 학습하므로, 결과가 확률적이고 지속적으로 변합니다. "6개월 안에 개발 완료"라는 식의 고정된 일정은 AI 프로젝트에 맞지 않습니다.

더 근본적인 문제는 AI 전환의 성과가 기존 KPI로 측정되지 않는 경우가 많다는 것입니다. AI 도입으로 고객 응대 시간이 줄었는데, 기존 KPI에는 "처리 건수"만 있고 "응대 품질"은 없을 수 있지요. 새로운 가치를 만들어 놓고 그것을 측정할 척도가 없으면, 조직 안에서 성과로 인정받지 못합니다.

AI 전환에는 실험적 접근이 필요합니다. 가설을 세우고, 빠르게 테스트하며, 결과를 바탕으로 방향을 조정하는 – 스타트업처럼 움직이는 것이지요. 대기업의 기존 관리 체계와 충돌할 수밖에 없고, 그래서 별도의 조직이나 거버넌스가 필요합니다.

멈춤을 알아채고 벽을 넘기

이런 신호가 보이면 의심해 볼 만합니다. 두 가지 이상 해당한다면 AX가 멈춰 있을 가능성이 높지요.

AI 파일럿이 3개 이상 돌았는데 확장도 종료도 결정된 것이 하나도 없다면 위험 신호입니다. 파일럿의 목적은 확장할지 멈출지를 가리는 것인데, 파일럿 자체가 목적이 되어버린 것이지요.

"AI 전략"이라는 문서가 있는데, 사업 전략 문서와 별도로 존재한다면 위험합니다. AI 전략과 사업 전략이 분리되어 있다는 것은 AI가 사업의 핵심이 아니라 부가 기능으로 취급되고 있다는 뜻입니다.

AI팀이 다른 사업부에 "AI 서비스를 제공하는" 구조라면 Enabled를 넘기 어렵습니다. AI가 각 사업부의 업무 흐름 안에 녹아들어야 First로 넘어갈 수 있습니다.

경영진이 AI를 "비용 절감 도구"로만 이야기한다면, 사업 구조 변화까지 나아가기 어렵습니다. 비용 절감은 Enabled의 성과이지, First나 Native의 목표가 아닙니다.

신호가 여럿 겹친다고 낙담할 필요는 없습니다. 패턴을 알면 해법도 보이거든요.

AI 도구를 들일 때마다 "이 도구가 제대로 일하려면 프로세스의 어느 부분이 바뀌어야 하는가?"를 같은 회의에서 물어보세요. 모터를 사는 결재와 전달축을 걷어내는 결재가 한 장에 올라와야 하는 것과 같습니다.

그 회의의 의장은 사업 책임자가 맡는 편이 좋습니다. AI 전환을 분기별 기술 보고로 돌리지 않고, 핵심 사업의 품질과 비용, 위험을 논의하는 자리에서 함께 결정하는 것이지요.

파일럿에는 "해보기" 다음의 길이 미리 그려져 있어야 합니다. 잘된 파일럿을 넓힐 절차와 책임자, 안 된 파일럿을 접을 절차와 책임자가 모두 필요하니까요.

성과를 재는 자도 바꿉니다. AI가 만들어 내는 새로운 가치를 잴 수 있는 KPI를 설계해 기존 KPI 옆에 나란히 둡니다.

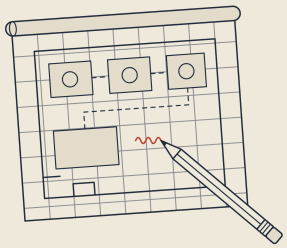
파일럿을 시작할 때부터 다음 90일의 결정표를 한 장으로 만들어 두면 '해보기'가 사업 판단으로 바뀝니다.

후보 업무	책임자	기준선·증거	90일 목표	중단 조건	90일 뒤 결정
예: 상담 기록 요약	고객운영 책임자	현재 처리 시간, 정정률, 민원 건수	처리 시간 단축과 정정률 기준을 함께 충족	개인정보 노출 또는 정정 부담 증가	확장·수정·종료
후보 1					
후보 2					
후보 3					

여기서 중요한 칸은 '중단 조건'과 '종료'입니다. 모든 파일럿을 성공 사례로 포장하면 자원이 묶입니다. 90일 뒤에는 확장할 과제와 고칠 과제, 끝낼 과제를 함께 골라야 포트폴리오가 움직입니다.

기술은 빠르게 바뀌어도 조직은 그 속도로 바뀌지 않습니다. 그렇다고 무작정 서두를 일도 아니지요. 파일럿을 더 늘리기보다, 90일마다 확장·수정·종료를 실제로 결정하는 리듬을 만드는 쪽이 정체를 푸는 데 빠릅니다.

새 모터를 사 놓고 옛 전달축에 도로 묶어 둔 공장을 떠올려 보세요. 그 공장에 모자랐던 것은 모터보다 새 배치도였지요. 다음 부는 그 배치도를 그리는 원칙들입니다. 무엇을 자동화할지, 사람의 판단을 어디에 둘지, 실패했을 때 어떻게 되돌릴지를 차례로 다룹니다.



제4부

사고법과 설계 원칙

전기가 들어왔을 때 공장 배치도가 바뀌었다

제3부 「덧붙이기의 한계」에서 두 갈래의 공장을 봤습니다. 증기 엔진 자리에 전기 모터만 갈아 끼운 쪽과, 전기를 전제로 배치도 자체를 다시 그린 쪽 — 그 갈림길이었지요.

전기 시대의 경쟁력은 발전 기술보다 공장 설계도에서 나왔습니다. 지능 시대도 다르지 않습니다. 모델은 그리드에서 사올 수 있지만, 설계도는 사올 수 없으니까요. 제4부에서 이야기할 아홉 원칙은, 제3부에서 정의한 지능 에너지 네이티브 조직의 설계도를 이루는 아홉 장입니다.

어디서 들은 이야기 같지요? 제3부에서 본 1890년대 공장과 지금 우리 사무실 풍경이 꼭 닮았습니다.

2026년 현재, 대부분의 조직이 하고 있는 일은 "증기 엔진을 전기 모터로 교체하는 것"입니다. 기존 업무 프로세스 위에 AI를 얹는 거지요. 이메일 작성 도구에 AI 자동완성을 붙이고, 엑셀 함수 대신 AI가 수식을 짜주게 하고, 고객 문의에 챗봇을 하나 달아놓습니다. 증기 엔진 자리에 전기 모터를 놓은 것과 본질적으로 같습니다.

AI 네이티브 사고란 공장 배치도를 바꾸는 것입니다. "AI가 있다면 이 업무를 처음부터 어떻게 설계할까?"라고 묻는 것이지요.

병원을 생각해 봅시다. 기존 방식은 환자가 접수하고, 대기하고, 진료실에 들어가 의사를 만나고, 검사를 받고, 다시 기다리고, 결과를 듣는 순서입니다. 여기에 AI를 얹으면 접수 키오스크에 챗봇이 붙고, 검사 결과를 AI가 읽어주는 정도입니다. 배치도는 그대로입니다.

공장 배치도를 바꾼다면 어떨까요? 환자가 집에서 증상을 말하면 AI가 사전 분석을 마칩니다. 병원에 도착했을 때 이미 필요한 검사 목록이 준비되어 있고, 의사는 AI가 정리한 소견서를 보면서 바로 핵심 판단에 들어갑니다. 진료 후 AI가 환자에게 맞춤형 생활 안내를 제공하고, 이상 징후가 감지되면 자동으로 재진 일정을 잡습니다. 의사가 하는 일의 본질은 바뀌지 않았지만, 의사가 진짜 해야 하는 일 — 판단과 소통 — 에 훨씬 더 많은 시간을 쓸 수 있게 됩니다.

그렇다면 배치도를 바꾸는 사고는 어디서 시작할까요?

출발점은 현재의 프로세스를 의심하는 것입니다. "왜 이 순서로 하지?"라는 질문을 던져 보세요. 대부분의 업무 순서는 인간의 인지 한계에 맞춰 설계된 것입니다. 한 번에 하나씩, 순서대로, 확인하면서 넘어가는 구조지요. AI에게는 이런 제약이 없습니다. 열 가지 일을 동시에 처리할 수 있고, 수천 건의 데이터를 한 번에 훑을 수 있습니다.

그러고 나서 "만약 이 서비스가 오늘 처음 만들어진다면?"이라고 상상해 보세요. 과거의 기술적 제약 때문에 어쩔 수 없이 받아들인 불편함들이 보이기 시작할 겁니다. 양식을 채우게 하는 이유, 승인 단계가 다섯 개인 이유, 보고서를 매주 쓰는 이유. 이 중 상당수는 "정보를 모으고, 정리하고, 전달하는" 과정에서 생긴 것입니다. AI가 그 과정을 대신할 수 있다면, 그 단계 자체가 불필요해집니다.

전기 시대가 남긴 교훈은 짧습니다. 새 기술을 옛 구조에 끼워 넣으면 30년을 허비합니다. 구조를 기술에 맞게 바꾸면, 상상하지 못한 가능성이 열립니다. 지금 AI 앞에서 우리에게 주어진 선택도 똑같습니다. 모터만 바꿀 것인가, 공장 배치도를 바꿀 것인가.

이 책의 4부에서는 공장 배치도를 바꾸기 위한 아홉 가지 설계 원칙을 이야기합니다. 전기 시대의 선구자들이 "각 기계마다 모터를 달자"는 발상으로 산업을 뒤바꿨듯, 지능 에너지 시대의 설계자들에게 필요한 사고법을 하나씩 짚어 보겠습니다.



AI를 빼면 무엇이 남는가 테스트

한 가지 실험을 해 봅시다. 지금 여러분의 제품이나 서비스에서 AI를 잠시 빼 보세요. 머릿속으로만 해도 됩니다. 서비스가 멈추는지만 보지 말고, 고객이 얻던 가치와 운영 부담, 복구 방법이 어떻게 달라지는지 함께 보는 겁니다.

AI를 빼도 서비스가 돌아간다고 해서 실패는 아닙니다. 처리 시간이 크게 줄거나 장애인의 접근성이 좋아지고, 사람이 반복 업무에서 벗어났다면 AI는 이미 중요한 가치를 만들고 있습니다. 반대로 AI 없이는 서비스가 멈추는데 대체 경로와 복구 절차도 없다면 어떨까요? 강한 엔진처럼 보여도 실은 하나만 고장 나면 전체가 서는 약한 고리일 수 있지요.

이것이 AI 제거 테스트입니다. 정전을 가정해 전기에 매달린 설비를 확인하듯, AI가 실제로 더한 가치와 AI에 새로 생긴 의존을 함께 드러내는 사고실험입니다.

정전 앞에서 설비가 두 부류로 갈리듯, AI도 그렇습니다. 꺼져도 버틸 만한 에어컨이 있고, 꺼지면 공장 전체가 서는 엔진이 있지요. 에어컨형 AI와 엔진형 AI의 차이를 조금 더 뜯어 봅시다.

에어컨형: 기존 서비스에 AI 챗봇을 붙였습니다. 챗봇이 고장 나면 고객이 전화를 걸면 됩니다. 불편하지만 서비스는 계속됩니다. 검색 엔진에 AI 요약 기능을 달았습니다. 요약이 사라져도 링크 목록은 여전히 나옵니다.

엔진형: Abridge라는 회사는 의사와 환자의 대화를 실시간으로 듣고 의료 차트를 자동 작성합니다. AI를 빼면? 아무것도 남지 않습니다. 그게 하는 일의 전부이니까요. 자율주행 택시에서 AI를 빼면? 그냥 차입니다. 운전사 없이는 한 발짝도 못 갑니다.

에어컨이 나쁜 건 아닙니다. 한여름에 에어컨 없는 사무실에서 일해 보신 분은 아실 겁니다. 문제는 에어컨을 달아놓고 "우리는 AI 네이티브 기업"이라고 말하는 것입니다. 이건 선풍기를 틀어놓고 "시원한 건물을 지었다"고 말하는 것과 같습니다.

네 축으로 진단하고 읽기

여러분의 조직에 이 테스트를 적용해 보겠습니다. 다음 질문들에 하나씩 답해 보세요.

1축: 고객 가치

AI를 걷어냈을 때 고객의 핵심 결과가 어떻게 달라지는지 봅니다.

- 이전에는 불가능했던 결과가 사라진다 → 핵심 가치에 기여합니다.
- 시간·비용·접근성이 눈에 띄게 나빠진다 → 운영 가치에 기여합니다.

· 고객이 차이를 알아차리지 못한다 → 장식에 가까울 수 있습니다.

2축: 운영 부담

AI가 맡은 일을 사람이 대신하면 처리 시간, 비용, 오류와 검토 부담이 얼마나 늘어나는지 계산합니다. 사람이 할 수 있다는 이유만으로 부가 기능이 되는 것은 아닙니다. 중요한 것은 같은 품질의 결과를 어떤 자원과 위험으로 만들어내느냐입니다.

3축: 복원력

모델 API가 멈추거나 품질이 떨어졌을 때 어떤 일이 벌어지는지 확인합니다. 안전하게 축소 운영할 수 있는지, 사람이 맡을 수 있는 최소 기능이 남는지, 다른 모델로 바꿀 수 있는지, 복구 시간과 책임자가 정해져 있는지를 봅니다. AI가 핵심일수록 이 축의 기준은 더 엄격해야 합니다.

4축: 차별화와 학습

경쟁사가 같은 모델을 연결했을 때도 같은 결과를 낼 수 있는지 묻습니다. 고유한 업무 지식과 합법적으로 확보한 피드백, 검증 절차가 쌓여 품질이 나아진다면 차별화가 생깁니다. 특정 API 이름만 바꿔도 그대로 복제되는 기능이라면 핵심 가치보다는 공급자 선택에 기대고 있을 가능성이 큽니다.

네 칸을 다 채우고 나면 점수부터 매기고 싶어지지요. 그런데 이 테스트로 엔진과 장식을 한 줄로 판정하려 들면 곤란합니다. 가치가 큰데 복원력이 약하다면 대체 경로가 먼저이고, 복원력은 충분한데 고객이 차이를 느끼지 못한다면 기능을 늘리기보다 문제 정의를 다시 해야 합니다. 네 축을 나란히 놓아야 다음에 무엇을 고칠지가 보이지요.

아홉 가지 설계 원칙

전기가 공장의 운영 방식을 바꾸었듯, AI도 일부 사업의 운영 구조를 바꾸고 있습니다. 이 전환을 설계할 때 필요한 아홉 가지 원칙은 다음과 같습니다.

원칙	한 줄 요약	지능 에너지 비유
원칙 1: 엔진을 설계하라	AI를 부속품이 아니라 제품의 핵심 동력으로 설계한다	에어컨이 아니라 엔진을 설계한다
원칙 2: 결과를 팔아라	도구가 아니라 완성된 결과를 고객에게 전달한다	전기를 파는 게 아니라 밝은 방을 판다
원칙 3: 의도를 읽어라	메뉴와 버튼 대신 사용자의 의도를 해석해 즉시 실행한다	스위치를 찾지 않아도 방에 들어가면 불이 켜진다
원칙 4: 플라이휠을 돌려라	사용할수록 데이터가 쌓이고, 시스템이 스스로 진화하는 순환 구조를 만든다	쓸수록 효율이 좋아지는 발전소
원칙 5: 에이전트에게 맡겨라	AI 에이전트가 자율적으로 작업하고 협업하는 인프라를 구축한다	전기 기기들이 알아서 켜지고 꺼지는 스마트 그리드
원칙 6: 한 팀으로 일하라	인간과 AI가 역할을 나눠 하나의 팀으로 협업하고, 조직 전체를 지휘자 구조로 전환한다	발전소 운영자와 자동 제어 시스템의 협업
원칙 7: 신뢰를 내장하라	보안, 윤리, 데이터 주권을 제품 설계 단계부터 녹여 넣는다	전기 안전 규격을 콘센트에 내장하듯
원칙 8: 확률을 경영하라	비결정적 시스템의 품질을 관리하는 새로운 운영 방식을 도입한다	전압 변동이 있어도 가전제품이 안전하게 작동하도록 설계
원칙 9: 최적 조합을 찾아라	대형 모델과 소형 모델, 클라우드와 엣지를 상황에 맞게 조합한다	대형 발전소와 가정용 태양광 패널을 함께 쓴다

아홉 원칙은 따로 읽어도 쓸모가 있지만, 실제로는 톱니처럼 물려 돌아갑니다. 엔진을 설계해야 결과를 팔 수 있고, 에이전트에게 일을 맡기려면 신뢰와 확률 관리가 먼저 서 있어야 하지요.



그림 4-1 위의 여섯 원칙은 둘씩 짝을 지어 돌고, 아래의 세 원칙이 그 전부를 떠받칩니다. 짝을 이룬 두 칸을 함께 읽으면 한 원칙이 왜 다른 원칙을 부르는지 보이지요.

이제 아홉 원칙을 하나씩 펼쳐 보겠습니다.

원칙 1: 엔진을 설계하라

식당에 키오스크를 놓는 것과, 키오스크가 곧 식당인 것은 다릅니다. 전자는 직원 한 명을 줄이는 효율화이고, 후자는 완전히 새로운 외식 경험입니다. "엔진을 설계하라"는 원칙은 이 차이에 관한 이야기입니다.

AI를 기존 제품에 붙이는 방식으로는 두 가지 한계에 부딪힙니다. 하나는 차별화가 안 된다는 것이고, 다른 하나는 구조적으로 AI의 잠재력을 끌어낼 수 없다는 것입니다.

차별화 문제부터 봅시다. 2026년 현재 웬만한 SaaS 제품에는 AI 기능이 달려 있습니다. "AI로 자동 생성", "AI 추천", "AI 요약" 같은 문구가 안 붙은 소프트웨어를 찾기가 더 어렵지요. 모두가 같은 기초 모델의 API를 호출하고 있기 때문에, 기능 자체로는 차이가 나기 어렵습니다. API 호출 한 줄이면 경쟁사도 같은 기능을 내일 당장 복사할 수 있으니까요.

구조적 한계는 더 심각합니다. 기존 제품에 AI를 얹으면, AI는 항상 기존 워크플로우의 틀 안에서만 작동합니다. 스프레드시트에 AI 수식 생성기를 붙여도, 사용자는 여전히 셀을 클릭하고 행을 추가하는 1980년대의 인터페이스 안에 갇혀 있습니다. AI가 아무리 똑똑해져도 스프레드시트라는 그릇이 바뀌지 않으면 경험은 크게 달라지지 않습니다.

엔진형 제품은 무엇이 다른가

엔진형 제품을 이해하기 위해 한 가지 사례를 보겠습니다.

Abridge라는 회사가 있습니다. 이 회사의 제품은 단순합니다. 의사와 환자의 대화를 듣고, 의료 차트를 자동으로 작성합니다. 그런데 이 "단순한" 제품이 미국 의료계를 뒤흔들고 있습니다.

왜일까요? 미국의 의사들은 환자를 진료하는 시간보다 차트를 쓰는 시간이 더 깁니다. 하루 평균 2시간 이상을 문서 작성에 쓰고, 이것이 번아웃의 주요 원인으로 꼽힙니다. Abridge는 이 문제를 "차트 쓰기를 도와주는 도구"로 풀지 않았습니다. "차트가 알아서 써지는 세상"을 만들었습니다.

여기서 핵심을 짚어 봅시다. Abridge에서 AI를 빼면 무엇이 남을까요? 아무것도 남지 않습니다. 녹음기 하나가 남을 뿐, 가치 제안 전체가 사라집니다. 이것이 엔진형 제품의 정의입니다. AI가 제품의 존재 이유 그 자체인 것이지요.

이런 사고방식의 차이가 설계의 모든 것을 바꿉니다. 기존 EMR(전자의무기록) 시스템에 AI 요약 기능을 추가하는 접근은 "어떤 버튼을 누르면 AI가 작동할까?"를 고민합니다. Abridge는 "버튼 자체가 필요 없는 경험"을 설계합니다. 의사가 평소처럼 환자와 대화하기만 하면 됩니다. 별도의 조작이 없습니다.

그렇다면 우리 회사에서는 이런 엔진을 어디서 찾아야 할까요? Abridge처럼 처음부터 AI 위에 서는 제품은 우연히 나오지 않습니다. 기존 제품을 손보는 데서 그치지 않고 엔진을 설계하고 싶다면, 다음 세 가지 질문에서 시작해 보세요.

"사람이 아무리 노력해도 물리적으로 불가능한 일은 무엇인가?"

실시간으로 수만 건의 거래를 동시에 모니터링하면서 이상 패턴을 잡아내는 것은 사람이 아무리 많아도 불가능합니다. 수백 개의 변수를 동시에 고려해서 최적의 물류 경로를 밀리초 안에 계산하는 것도 마찬가지입니다. 이런 영역이 엔진의 후보입니다.

"이 산업에서 모두가 당연하게 받아들이는 불편함은 무엇인가?"

"원래 그런 거야"라고 체념하고 있는 것들이 있습니다. 의사가 차트를 직접 쓰는 것, 보험 심사에 2주가 걸리는 것, 법률 문서를 변호사가 한 줄 한 줄 읽는 것. 이런 "당연한 불편함"이 AI 엔진이 뒤집을 수 있는 기회입니다.

"데이터가 쌓일수록 제품이 똑똑해지는 구조를 만들 수 있는가?"

엔진형 제품은 시간이 지날수록 강해집니다. Abridge는 의료 대화 데이터가 쌓일수록 다양한 진료과의 용어와 맥락을 더 잘 이해하게 됩니다. 이 학습 루프가 없으면 엔진이 아니라 부품에 불과합니다.

월요일 아침 화이트보드 앞에서

팀원 몇 명과 화이트보드 앞에 서서 지금 제품의 기능 목록을 적어 봅니다. 각 기능 옆에 "AI 없이도 되는가?"라고 써 넣으면 대부분 "예"가 붙을 겁니다. 그때 질문을 바꿔 보세요. "AI가 있어서 비로소 되는 일은 무엇인가?" 그 칸이 비어 있다면, 아직 에어컨만 달고 있는 셈이지요.

빈칸을 채울 재료는 대개 제품 밖에 있습니다. 고객이 우리 제품을 쓰기 전과 쓴 뒤에 하는 일을 지켜보면, 손으로 옮겨 적는 일과 기다리는 시간, 판단이 막히는 순간이 보입니다. Abridge가 찾아낸 의사의 저녁 차트 작업도 그런 자리였습니다.

화이트보드 한쪽에는 다른 그림도 하나 그려 둡니다. 이 제품을 오늘 백지에서 다시 만든다면 AI가 맡을 핵심 기능은 무엇이고, 사람이 맡을 판단은 무엇이며, AI가 멈춘 날에도 지켜야 할 최소 기능은 무엇인가. 엔진이 중요할수록 그 엔진이 멈춘 날의 그림도 함께 있어야 하니까요.

실행 체크: 엔진인가, 에어컨인가

1. 기능 목록 옆에 "AI 없이도 되는가"를 적고, "AI라서 비로소 되는" 칸이 비어 있지 않은지 확인합니다.
2. 고객이 제품 밖에서 겪는 수작업·기다림·판단의 어려움을 하나 이상 찾아, 엔진을 얹을 자리가 있는지 봅니다.
3. "백지에서 다시 만든다면?" 회의를 한 번 잡고, AI가 멈춘 날 유지할 최소 기능까지 함께 그립니다.

리스크: 기존 제품에 AI 기능만 덧붙이면(에어컨 달기) 남는 건 비용뿐입니다. 거꾸로 엔진을 얹되 대체 경로를 두지 않으면, 지능이 단일 고장점이 됩니다.

키오스크를 하나 더 놓을지, 키오스크가 곧 식당인 가게를 열지. 화이트보드에 남은 목록이 그 답을 먼저 알려 줄 겁니다.

원칙 2: 결과를 팔아라

드릴을 사는 사람은 드릴이 필요한 게 아닙니다. 벽에 구멍이 필요한 것이지요. 이 오래된 마케팅 격언이 AI 시대에 와서 완전히 새로운 의미를 갖게 됐습니다. 이제는 드릴을 팔지 않아도 됩니다. 구멍 뚫린 벽을 팔 수 있으니까요.

전기회사를 떠올리면 이해가 빠릅니다. 전력회사는 전자(電子)를 팔지 않습니다 — 밝은 방과 시원한 여름, 돌아가는 공장을 팔지요. 소비자는 전압이 몇 볼트인지 궁금해하지 않습니다. 불이 켜지고 방이 시원해지면 그만이니까요. 지능 에너지도 같은 자리로 가고 있습니다. 고객이 원하는 건 모델 이름보다 그 모델이 켜 주는 '밝은 방'이지요.

소프트웨어 산업은 오랫동안 도구를 팔았습니다. 워드프로세서를 팔고, 스프레드시트를 팔고, CRM을 팔았지요. 사용자 수(seat) 기반으로 요금을 매겼습니다. 도구를 쓰든 안 쓰든, 잘 쓰든 못 쓰든 매달 같은 금액을 청구했습니다. 고객이 그 도구로 원하는 결과를 얻었는지는 소프트웨어 회사의 관심사가 아니었습니다.

AI는 이 구조를 근본적으로 뒤흔듭니다. 이제 소프트웨어가 결과 자체를 만들어낼 수 있기 때문입니다.

도구 대신 결과를 파는 회사들

전통적인 SaaS(Software as a Service)는 소프트웨어를 서비스 형태로 제공하는 모델이었습니다. 구독료를 내면 소프트웨어를 쓸 수 있지요. 하지만 소프트웨어가 일을 대신해주지는 않았습니다. 도구를 줄 뿐, 결과는 사용자의 몫이었습니다.

지금 부상하고 있는 모델은 이 관계를 뒤집습니다. Service-as-a-Software라고 불리는 이 접근은, 소프트웨어가 도구를 제공하는 게 아니라 서비스 자체를 수행합니다. 사람이 했던 일을 소프트웨어가 대신 끝까지 해주는 것이지요.

차이를 구체적으로 봅시다.

SaaS 방식의 법률 서비스: 계약서 템플릿과 편집 도구를 제공합니다. 변호사가 직접 검토하고 수정해야 합니다. 사용자 수 기반 월 구독료를 받습니다.

Service-as-a-Software 방식: 계약서를 업로드하면 AI가 위험 조항을 분석하고, 수정안을 제시하고, 상대방과의 협상 포인트까지 정리해 줍니다. "검토 완료된 계약서"라는 결과물에 대해 건당 요금을 받습니다.

이 전환은 가격 책정의 선택지를 넓힙니다. 도구를 파는 모델에서는 "이 소프트웨어를 몇 명이 쓰느냐"가 가격을 결정합니다. 결과를 파는 모델에서는 "합의한 문제를 얼마나 해결했느냐"가 가격의 일부가 될 수 있습니다. 모든 제품이 결과 과금으로 옮겨야 하는 것은 아닙니다.

말로만 들으면 먼 이야기 같지만, 이 패턴은 이미 여러 산업에서 돌아가고 있습니다.

GE의 터빈 엔진 서비스를 봅시다. 항공사에 터빈 엔진을 파는 대신, GE는 "엔진 가동 시간"을 팝니다. 항공사는 엔진을 소유하지 않고, 가동 시간(hour)당 비용을 지불합니다. GE는 AI로 고장을 미리 예측하고 예방해서 가동률 99%를 유지합니다. SLA(서비스 수준 계약) 이하로 떨어지면 GE가 패널티를 물지요. 도구(엔진)를 파는 게 아니라 결과(비행 가능한 시간)를 파는 겁니다.

Abridge는 의료 차트 작성 도구를 파는 것이 아니라, 작성이 완료된 의료 차트를 제공합니다. 의사가 환자와 대화하면 AI가 듣고, 의료 기록 규정에 맞는 차트를 만들어 냅니다. 의사가 받는 것은 도구가 아니라 결과물입니다.

Perplexity AI는 검색 도구라기보다 답을 파는 쪽에 가깝습니다. 검색 결과 링크 열 개를 보여주는 대신, 질문에 대한 완성된 답변을 출처와 함께 제공합니다. 사용자가 열 개의 링크를 읽고 스스로 답을 조합할 필요가 없지요. 정보 검색이라는 "도구"가 아니라 "답변"이라는 결과를 파는 것입니다.

이 사례들의 공통점이 보이시나요? 고객이 원래 해야 했던 수작업, 판단, 종합의 과정을 AI가 대신 수행하고, 최종 결과물만 전달합니다.

가격표를 다시 쓰는 법

지금 우리 제품의 청구서를 한 장 꺼내 보세요. 거기 적힌 것은 고객이 얻은 결과인가요, 좌석 수와 사용량인가요? "이 가격을 해결 건수나 완료된 성과 기준으로 바꾸면 매출 구조가 어떻게 달라질까?"라는 질문은 그 청구서 위에서 시작됩니다.

결과 기반 과금은 제품이 실제로 가치를 만드는지 드러내 줍니다. 그런데 결과를 잘못 정의하면 엉뚱한 곳에서 탈이 나지요. 상담을 끝낸 건수만 과금하면 어려운 고객을 피하거나 서둘러 종결할 유인이 생깁니다. 채용 성사만 가격에 넣으면 장기 근속과 공정성이 뒤로 밀릴 수 있고요. 그래서 성공 건수와 함께 품질, 위험, 사후 결과를 같이 봅니다.

고객의 "결과"는 고객과 함께 정의해야 합니다. "완성된 보고서"라면 사실 오류와 수정 횟수까지, "채용된 적합한 인재"라면 채용 과정의 공정성과 일정 기간 뒤의 적합성까지 따져야 하지요. 경기나 조직 문화, 고객 데이터처럼 공급자가 통제할 수 없는 변수는 계약에서 따로 떼어 둡니다.

현실적인 출발점은 혼합형 과금입니다. 기본 운영과 고정 비용은 구독료로 받고, 사용량을 반영하되, 양쪽이 검증할 수 있는 결과에는 보너스나 환급을 붙이는 식이지요. 위험이 큰 업무라면 상한과 이의제기 절차도 둡니다. 전기요금에 기본요금과 사용량 요금이 함께 있는 것과 비슷한 구조입니다.

실행 체크: 도구가 아니라 결과를 파는가

1. 지금 청구서에 고객이 얻는 "결과"가 적혀 있는지, "좌석 수·사용량"만 적혀 있는지 봅니다.

2. 결과 과금을 실험할 기능 하나를 골라, 결과 하나·품질 지표 하나·중단 조건 하나를 고객과 먼저 합의합니다.

3. 구독 모델은 그대로 둔 채 소규모로 병행 실험하고, 고객이 어느 기능에서 결과 과금을 반기는지 데이터를 모읍니다.

리스크: 결과를 정의하는 지표가 모호하면 과금 분쟁으로 번집니다. 과금 단위를 바꾸는 일은 결국 위험을 누가 얼마나 지느냐를 다시 나누는 계약이라는 점도 기억해 둘 만합니다.

원칙 3: 의도를 읽어라

쇼핑몰 앱에서 "주문 내역" 버튼을 찾느라 5분 넘게 헤맨 적 있으신가요? 설정 메뉴에 들어갔다가 나오고, 마이페이지를 열었다가 닫고, 결국 검색창에 "주문"이라고 쳐서 겨우 찾아내는 경험 말입니다.

의도 기반 시스템에서는 "지난주에 산 거 반품해줘"라고 말해 주문을 찾을 수 있습니다. 다만 반품 대상과 환불 수단처럼 되돌리기 어려운 선택은 사용자가 화면이나 음성으로 확인한 뒤 처리해야 합니다.

이것이 의도 기반 UX입니다. 사용자가 메뉴를 탐색하는 대신, 원하는 것을 말하면 시스템이 의도를 해석해서 실행합니다.

과거의 인터페이스는 인간이 기계에 맞추는 구조였습니다. 기계가 이해할 수 있는 방식으로 명령을 내려야 했지요. 파일을 저장하려면 "파일" 메뉴를 열고 "저장"을 클릭하거나, Ctrl+S를 누르거나, 특정 아이콘을 찾아야 했습니다. 이 모든 과정은 사용자가 시스템의 구조를 학습한 뒤에야 가능합니다.

의도 기반 시스템은 이 관계를 뒤집습니다. 기계가 인간에게 맞춥니다.

화면을 없앨까, 바꿀까

제로 UI(Zero UI)라는 개념이 있습니다. 인터페이스를 의식하지 않아도 원하는 일을 마칠 수 있게 하자는 생각이지요. 그러나 화면이 사라지면 현재 상태와 선택지를 확인하기 어렵고, 음성 사용이 불편한 사람은 오히려 배제될 수 있습니다.

방에 들어가면 불이 켜지는 것. 이것이 제로 UI입니다. 스위치를 찾을 필요가 없습니다. 사용자의 맥락(방에 들어왔다)을 시스템이 읽고, 의도(불을 켜고 싶다)를 추론해서 행동합니다.

이것을 소프트웨어에 적용하면 어떻게 될까요?

항공권을 예약한다고 합시다. 기존 방식은 이렇습니다. 여행 사이트에 접속하고, 출발지를 입력하고, 도착지를 입력하고, 날짜를 선택하고, 인원수를 정하고, 검색 버튼을 누르고, 결과를 정렬하고, 필터를 적용하고, 가격을 비교하고, 좌석을 선택하고, 개인정보를 입력하고, 결제합니다. 열두 단계입니다.

의도 기반 방식은 다릅니다. "다음 주 월요일 시카고 출장 항공편 잡아줘. 아침에 도착하는 걸로." 이 한 문장이면 됩니다. AI가 일정을 확인하고, 선호하는 항공사와 좌석 유형을 기억하고 있으니, 최적의 옵션을 제시합니다. "이 표 예약할까요?" 한 번의 확인으로 끝납니다.

열두 단계가 줄어들 수는 있지만, 비교와 수정, 결제 확인은 남아야 합니다. 좋은 의도 기반 UX는 단계를 무조건 감추기보다 사용자가 통제해야 할 순간을 선명하게 보여줍니다.

그렇다면 화면을 아예 없애는 대신, 화면이 사용자에게 맞춰 모양을 바꾼다면 어떨까요? 의도 기반 UX의 핵심 기술 중 하나가 생성형 UI(Generative UI)입니다. 모든 사용자에게 같은 화면을 보여주는 대신, 사용자의 의도와 맥락에 맞춰 인터페이스가 실시간으로 생성됩니다.

같은 금융 앱이라도, 출근길에 빠르게 확인하는 사용자에게는 오늘의 주요 지표만 보여주고, 저녁에 여유롭게 분석하려는 사용자에게는 상세 차트와 비교 도구를 펼쳐 보여줍니다. 화면의 구조 자체가 달라지는 것이지요.

"다크 모드"나 "위젯 정렬" 수준의 개인화보다 한 걸음 더 나간 것이지요. 사용자가 무엇을 하려는지에 따라 인터페이스의 구성 요소가 달라집니다. 이때도 핵심 기능의 위치와 용어, 키보드 이동 순서, 화면 읽기 프로그램용 이름은 예측 가능하게 유지해야 합니다. 생성된 화면이 접근성 규칙까지 매번 바꾸면 편리함이 혼란으로 바뀝니다.

구글 맵이 20년 만에 바뀐 이유

2026년 3월, 구글이 구글 맵(Google Maps)을 대대적으로 개편했습니다. 2005년 출시 이후 가장 큰 변화라고 할 수 있지요.

핵심은 Ask Maps라는 기능입니다. 이제 구글 맵에서 "줄 안 서고 핸드폰 충전할 수 있는 카페 어디 있어?"라고 물어볼 수 있습니다. 기존이라면 "카페"를 검색하고, 각 카페를 하나씩 클릭해서 리뷰를 읽고, 충전 콘센트가 있는지 확인하고, 혼잡도를 체크하는 과정을 거쳐야 했지요. 의도를 시스템의 언어로 번역하는 데 수십 번의 클릭이 필요했습니다.

Ask Maps는 이 과정을 한 문장으로 줄입니다. 3억 곳이 넘는 장소 정보와 5억 명이 넘는 사람이 남긴 리뷰를 Gemini가 읽고, 복잡한 의도에 맞는 답을 내놓지요. 미국과 인도에서 먼저 시작했고, 식당을 추천받으면 그 자리에서 예약까지 할 수 있습니다.

내비게이션도 바뀌었습니다. 기존에는 "500미터 후 우회전"이라고 안내했습니다. 새로운 몰입형 내비게이션(Immersive Navigation)은 "KFC 지나서 우회전"이라고 말합니다. AI가 Street View 이미지를 분석해서 건물을 인식하고, 사람이 실제로 길을 안내하는 것처럼 랜드마크 기반으로 설명합니다. 3D 건물이 반투명하게 표시되어 복잡한 교차로에서도 어디로 가야 하는지 한눈에 보이지요.

20년 된 제품이 어떻게 이렇게 바뀔 수 있었을까요? 구글 맵은 이미 엄청난 데이터를 갖고 있었습니다. Street View 이미지, 장소 정보, 리뷰, 실시간 교통 데이터. 하지만 이 데이터를 사용자에게 전달하는 방식은 "검색→목록→클릭"이라는 20년 전 패턴에 갇혀 있었지요. AI가 그 벽을 허물었습니다. 검색이 대화로 바뀌고, 정적 표시가 지능적 해석으로 바뀌고, 기계적 안내가 인간적 설명으로 바뀐 것입니다.

기능을 더 붙인 것도 아닙니다. 이미 가진 데이터를 AI라는 새로운 에너지로 사용자의 의도에 맞게 다시 엮었을 뿐이지요.

물론 의도를 읽어 주는 시스템이 늘수록 위험도 함께 커집니다. AI가 의도를 잘못 해석하면 사용자는 당혹스러워집니다. 반품을 요청했는데 재주문으로 처리되는 상황을 생각해 보세요.

출발점은 "불확실하면 물어보라"입니다. 확인의 무게는 행동의 위험에 맞춥니다. 검색이나 초안 작성은 바로 실행해도 되지만, 결제·삭제·외부 전송·권리 포기처럼 되돌리기 어려운 행동은 대상과 결과를 분명히 보여 주고 확인받아야 하지요. 음성, 키보드, 화면처럼 여러 입력 경로를 열어 두고, 잘못됐을 때 쉽게 되돌릴 길도 함께 둡니다.

클릭 수를 세어 보세요

의도 기반 설계는 화면에서 출발하지 않습니다. 기존 설계가 "사용자에게 어떤 화면을 보여줄까?"를 먼저 묻는다면, 의도 기반 설계는 "사용자가 무엇을 이루려 하는가?"를 먼저 묻지요. 화면은 그다음 문제이고, 경우에 따라서는 아예 필요 없을 수도 있습니다.

손에 잡히는 출발점은 클릭 수입니다. 사용자가 가장 많이 클릭하는 흐름 하나를 골라, 그 클릭 가운데 몇 번이 "내가 원하는 것을 시스템에 알려 주려는" 클릭인지 세어 보세요. 줄일 수 있는 단계가 눈에 보이기 시작합니다.

고객 지원 창구도 좋은 자료입니다. "이 기능 어디 있어요?"라는 문의가 많다면, 사용자는 분명한 의도를 갖고 왔는데 화면이 그걸 읽지 못하고 있다는 뜻이지요.

실행 체크: 화면보다 의도를 먼저 읽는가

1. 가장 많이 쓰이는 흐름에서 "의도를 전달하려는" 클릭이 몇 번인지 세어 봅니다.
2. "이 기능 어디 있어요?" 유형의 문의를 모아, 인터페이스가 의도를 놓치는 지점을 표시합니다.
3. 대표 작업 하나를 "화면 없이 말로 시키면?" 시나리오로 그려 보고, 키보드와 화면 읽기 프로그램만으로도 끝낼 수 있는지 함께 점검합니다.

리스크: 의도 추론을 과신하면 시스템이 엉뚱한 목표를 실행합니다. 결제 직전에 대상과 금액을 다시 보여 주는 확인 단계와 쉬운 되돌리기가 사용자의 통제권을 지켜 줍니다.

주문 내역 버튼을 찾느라 5분을 헤맨 그 사용자를 다시 떠올려 볼까요. 그가 원한 것은 더 예쁜 메뉴였을까요, "지난주에 산 거 반품해 줘"라는 한마디였을까요?

원칙 4: 플라이휠을 돌려라

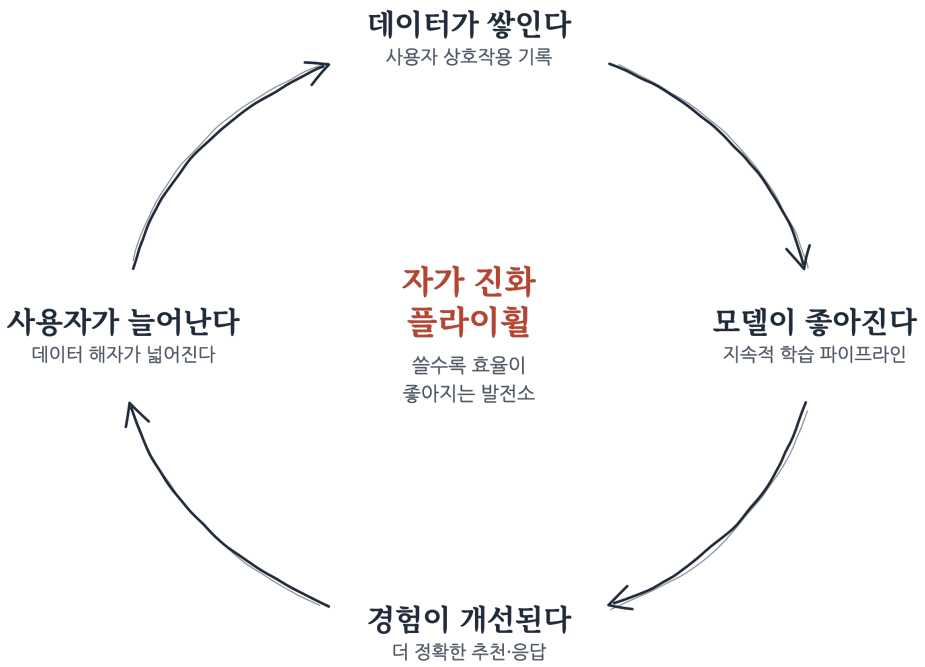
어제는 내 취향을 몰랐던 추천 로직이, 오늘은 딱 원하던 그 상품을 보여줍니다. 사용 과정에서 얻은 신호가 다음 개선으로 이어진다면 어떨까요? 이것이 자가 진화 플라이휠의 기본 구조입니다.

플라이휠(flywheel)은 원래 기계 공학 용어입니다. 무거운 원반을 처음 돌리는 데는 엄청난 힘이 들지만, 일단 돌아가기 시작하면 관성 때문에 점점 빨라지는 장치이지요. 짐 콜린스가 아마존의 성장 구조를 설명하면서 경영 용어로 만들었습니다. 더 많은 고객이 오면 더 많은 판매자가 모이고, 더 많은 판매자가 모이면 가격이 낮아지고, 가격이 낮아지면 더 많은 고객이 온다. 이 선순환이 플라이휠입니다.

AI 네이티브 시스템의 플라이휠은 데이터, 모델, 사용자 경험이 연결된 순환입니다. 다만 데이터가 많다고 모델이 저절로 좋아지는 것은 아닙니다. 목적에 맞는 신호를 모으고, 대표성과 품질을 검증하고, 실제 개선으로 이어졌는지 확인해야 비로소 선순환이 됩니다.

쓸수록 똑똑해지는 시스템

한 번 돌기 시작하면 쓸수록 빨라지는 바퀴 — 데이터, 모델, 경험, 사용자가 서로를 밀어 주는 순환



> **데이터가 많다고 저절로 돌지는 않습니다.** 목적에 맞는 신호를 모으고 품질을 검증해야 선순환이 되고, 잘못된 결과가 다시 학습 데이터로 들어가면 바퀴는 거꾸로 돌며 오류를 키웁니다.

출처: 저자 작성 — 본문 제4부 「원칙 4: 플라이휠을 돌려라」를 바탕으로 했습니다.

그림 4-2 데이터가 쌓이면 모델이 좋아지고, 경험이 나아지면 사용자가 늘어 다시 데이터가 쌓입니다. 화살표를 따라 한 바퀴 돌아 보되, 잘못된 결과가 섞이면 같은 바퀴가 오류를 키우며 거꾸로 돈다는 점도 기억해 두세요.

이 플라이휠은, 말하자면 운전 기록으로 효율을 조정하는 발전소입니다. 그러나 사용자가 남긴 흔적을 곧바로 연료로 볼 수는 없습니다. 어떤 목적으로 수집하고 학습에 쓰는지 알리고, 동의나 그 밖의 적법한 근거를 갖추고, 필요가 끝나면 삭제하는 관리가 먼저입니다.

살아 있는 시스템은 무엇으로 도는가

과거의 소프트웨어는 공장에서 찍어낸 공산품 같았습니다. 한 번 만들어서 배포하면 다음 업데이트까지 똑같이 작동했지요. 사용자가 백만 명이든 한 명이든 소프트웨어의 행동은 동일했습니다.

AI 네이티브 시스템은 다릅니다. 살아있는 유기체처럼 데이터를 먹고 자랍니다. 넷플릭스의 추천 알고리즘은 수억 명의 시청 패턴을 학습하면서 매일 미세하게 달라집니다. 스팸 필터는 사용자가 "스팸" 버튼을 누를 때마다 조금씩 더 정확해집니다.

이 차이는 경쟁 구도를 근본적으로 바꿉니다. 정적인 소프트웨어는 기능 복사가 쉽습니다. 경쟁사가 같은 기능을 만들면 끝이지요. 하지만 플라이휠이 돌아가는 시스템은 시간 자체가 경쟁 우위입니다. 3년간 쌓인 데이터와 그 데이터로 학습된 모델은 돈으로 살 수 없습니다.

웬만한 업무에 쓸 만한 모델은 이제 누구나 API로 부를 수 있습니다. 맨 꼭대기는 사정이 좀 다르지요. 2026년 4월 Anthropic은 가장 강력한 모델이던 Claude Mythos Preview를 일반에 풀지 않고, 핵심 소프트웨어를 만들고 지키는 일부 기관에만 열었습니다.[1] 가장 센 모델은 고르고 골라 나눠 주는 자원이 되어 가는 셈입니다. 그래도 대부분의 기업에게 차별화는 모델보다 모델을 학습시킨 고유 데이터에서 나옵니다. 자사만의 데이터 루프를 가진 기업만이 지속적인 개선과 특화로 우위를 점할 수 있습니다.

그렇다면 돈으로 살 수 없는 이 시간의 우위는 무엇으로 만들어질까요? 성공적인 AI 플라이휠에는 네 가지가 필요합니다.

데이터 수집 설계가 첫 번째입니다. 사용자의 행동, 피드백, 선택 가운데 무엇이 개선에 필요한지부터 정해야 합니다. 클릭이나 건너뛰기 같은 행동을 활용하더라도 수집 목적과 이용 범위를 설명하고, 선택권을 흐리지 않는 화면에서 동의를 받아야 합니다. 서비스 이용과 학습 참여를 구분할 수 있어야 하지요.

지속적 학습 파이프라인이 두 번째입니다. 수집된 데이터가 모델 개선으로 이어지는 자동화된 경로가 있어야 합니다. MLOps 파이프라인이라고 부르는 이 구조는, 새 데이터가 들어오면 모델을 재학습시키고, 성능을 검증하고, 더 나은 모델을 자동으로 배포하는 일련의 과정입니다.

세 번째는 피드백 루프입니다. 모델이 개선된 결과가 다시 사용자 경험에 반영되고, 그 반응이 새로운 데이터로 돌아오는 순환이 완성되어야 합니다. 이 루프가 빠를수록 플라이휠의 회전 속도가 올라갑니다.

마지막 축은 관리된 고유 데이터입니다. 사용자 상호작용 로그, 도메인 특화 데이터, 레이블된 피드백은 제품을 개선하는 자산이 될 수 있습니다. 그 가치는 양보다 출처, 사용 권한, 대표성, 갱신 이력, 삭제 가능성을 얼마나 잘 관리하느냐에 달려 있습니다.

플라이휠이 거꾸로 돌 때

플라이휠에는 위험도 있습니다. 잘못된 결과가 사용자의 행동을 바꾸고, 그 행동이 다시 학습 데이터로 들어오면 오류가 증폭될 수 있습니다. 이것은 피드백 오염의 문제입니다. 한편 데이터 드리프트(data drift)는 시간이 지나며 실제 입력과 대상의 분포가 달라지는 현상입니다. 둘은 원인도 대응도 다르므로 따로 추적해야 합니다.

이를 막으려면 전체 평균과 함께 주요 집단별 성능, 그리고 데이터에서 빠진 집단까지 들여다봅니다. 말을 많이 하는 사용자만 데이터에 남고, 침묵하거나 서비스를 떠난 사람은 빠질 수 있기 때문이지요. 데이터 분포의 변화와 모델 출력이 다시 데이터에 섞이는 정도도 별도 지표로 감시해야 합니다.

오염이 바퀴 안에서 생기는 문제라면, 바퀴 바깥에서 오는 문제도 있습니다. 플라이휠이 돌아가려면 데이터라는 연료가 끊임없이 공급되어야 합니다. 그런데 문제가 있습니다. 학습에 쓸 수 있는 고품질 인간

데이터가 고갈되고 있다는 것입니다. 인터넷에 있는 텍스트는 이미 대부분 학습에 사용되었고, 새로 생성되는 텍스트의 양은 AI의 학습 수요를 따라가지 못합니다.

여기서 등장한 것이 합성 데이터(Synthetic Data)입니다. AI가 AI를 훈련시키기 위한 데이터를 직접 생성하는 것이지요. 세계경제포럼(WEF)은 2028년까지 AI 학습 데이터의 80%가 합성 데이터가 될 것으로 전망하고 있습니다[2].

NVIDIA는 아예 합성 데이터 생성에 특화된 모델인 Nemotron-4를 공개했습니다. 이 모델은 다른 AI 모델을 훈련시키기 위한 고품질 데이터를 대량으로 만들어내는 것이 주된 역할입니다. 발전소가 자기가 이미 태운 연료를 되살려 다시 태우는 셈이지요.

하지만 여기에는 심각한 함정이 있습니다. 모델 붕괴(Model Collapse)라는 현상입니다[3]. AI가 만든 데이터로만 AI를 학습시키면, 세대를 거듭할수록 품질이 점차 저하됩니다. 마치 복사기로 복사한 문서를 다시 복사하면 글씨가 흐려지는 것처럼요. Nature에 발표된 연구에 따르면, 합성 데이터만으로 반복 학습한 모델은 원래 데이터의 분포에서 점점 벗어나 결국 쓸 수 없는 수준까지 떨어집니다. 다양성이 줄어들고, 드문 패턴은 사라지고, 흔한 패턴만 과도하게 강화됩니다.

에너지 비유로 말하면 이렇습니다. 발전소가 자기가 태운 연료를 재활용해서만 돌리면, 연료의 품질이 매 사이클마다 조금씩 떨어집니다. 처음에는 눈에 띄지 않지만, 몇 세대가 지나면 발전소의 출력이 급격히 불안정해집니다.

해법은 하이브리드 전략입니다. 합성 데이터를 활용하더라도 실제 세계에서 수집하고 검증한 고품질 기준 데이터를 유지해야 합니다. 외부에서 들여온 양질의 새 연료와 재활용 연료를 배합하는 셈이지요. 합성 데이터의 출처를 표시하고, 중복과 편향을 검사하고, 실제 데이터에서도 성능이 유지되는지 확인하는 식이지요.

합성 데이터는 부족한 사례를 보강하고 실험 범위를 넓힐 수 있지만, 데이터 부족을 저절로 해결하지는 않습니다. 합성 데이터는 연료를 늘리는 도구이지, 연료의 품질이나 현실 대표성을 보장하는 도구는 아닙니다.

첫 바퀴는 느리게

플라이휠을 설계하려면 이 질문부터 답해 보세요. "사용자가 우리 제품을 쓰는 과정에서 자연스럽게 생기는 데이터는 무엇인가?" 그 데이터를 어떻게 모으고, 어떤 모델의 학습에 쓸 수 있으며, 개선된 모델이 사용자에게 어떤 더 나은 경험으로 돌아갈지 한 바퀴를 그려 보는 것이지요.

처음에는 공개 모델로 시작해도 괜찮습니다. 쌓아야 할 것은 도메인 특화 피드백이고, 그 쌓는 방식이 책임 있어야 합니다. 원본 데이터가 삭제되거나 학습 이용 동의가 철회되면 데이터셋과 캐시, 다음 학습 주기에 그 요청이 어떻게 반영되는지 정해 둡니다. 이미 학습된 모델에서 흔적을 곧바로 지우기 어려운 경우라면, 그 한계와 후속 조치를 사용자에게 솔직하게 알리는 편이 신뢰를 지킵니다.

실행 체크: 플라이휠이 돌기 시작했는가

1. 사용 과정에서 자연스럽게 쌓이는 데이터를 적고, 그것이 모델 개선으로 되돌아오는 경로를 한 장에 그립니다.
2. 데이터마다 목적, 적법한 근거, 대표성, 보관 기간, 삭제 경로, 품질 검증 방법을 기록합니다.
3. 합성 데이터를 쓴다면 생성 출처를 표시하고, 사람이 만든 검증 데이터와 분리해 성능을 비교합니다.

리스크: 합성 데이터만으로 반복 학습하면 모델 붕괴가 올 수 있습니다. 사람이 만든 데이터라는 "원유"를 완전히 갈아치우려는 순간, 플라이휠은 거꾸로 돌기 시작합니다.

무거운 원반은 처음 몇 바퀴가 가장 무겁습니다. 그 몇 바퀴 동안 남긴 기록이, 나중에 원반이 거꾸로 돌기 시작할 때 브레이크가 되어 줍니다.



---

1. "Project Glasswing" (Anthropic, 2026.4.7, https://www.anthropic.com/glasswing) — "We do not plan to make Claude Mythos Preview generally available." 출범 파트너 12곳과 40여 개 기관으로 접근을 제한.
2. WEF, "The Role of Synthetic Data in AI Training" (2025, https://www.weforum.org/stories/2025/12/data-ai-training-synthetic/)
3. "AI models collapse when trained on recursively generated data" (Nature, 2024, https://www.nature.com/articles/s41586-024-07566-y)

원칙 5: 에이전트에게 맡겨라

새벽 3시, 당신이 잠든 사이에도 수천 명의 직원이 쉬지 않고 일하고 있다면 어떨까요? 이메일을 분류하고, 재고를 확인하고, 고객 문의에 응답하고, 보고서를 작성하고 있습니다. 월급도, 휴가도, 퇴사도 없습니다.

공상과학처럼 들리지만, 일부는 이미 시험 중입니다. AI 에이전트가 그 직원들입니다.

기업용 소프트웨어에 태스크 전용 에이전트를 넣으려는 시도는 빠르게 늘고 있습니다. 그러나 관심과 문의가 늘었다는 사실이 안정적인 운영 성과를 뜻하지는 않습니다. 에이전트가 필요한지는 업무의 반복성, 도구 권한, 오류 비용과 검증 가능성을 기준으로 판단해야 합니다.

그런데 에이전트는 우리가 이미 써 본 챗봇과 무엇이 다를까요? 둘은 일하는 방식부터 다릅니다. 챗봇은 질문하면 답하는 반응형 시스템입니다. 에이전트는 목표를 받으면 스스로 계획을 세우고, 필요한 도구를 쓰고, 단계별로 실행하는 자율적 행위자입니다.

"다음 주 시카고 출장 준비해줘"라고 말하면, 챗봇은 항공편 검색 방법을 알려줍니다. 에이전트는 일정을 확인하고, 항공편을 예약하고, 호텔을 잡고, 회의실을 예약하고, 출장 보고서 양식까지 준비해 놓습니다. 중간에 항공편이 매진이면 대안을 찾고, 회의 시간이 겹치면 참석자들에게 조정을 요청합니다.

이 차이를 가능하게 하는 것이 에이전트 네이티브 인프라입니다. 과거의 인프라는 사람의 클릭 속도에 맞춰져 있었습니다. 예쁜 화면이 중요했지요. 에이전트 네이티브 인프라는 다릅니다. 기계가 읽기 쉬운 구조화된 데이터와 초고속 API로, 에이전트들이 질주할 수 있는 고속도로를 깝니다.

에이전트가 달릴 길과 공용어

에이전트들이 효과적으로 일하려면 서로 소통할 수 있어야 합니다. 여기서 두 가지 핵심 프로토콜이 등장합니다.

MCP(Model Context Protocol)는 Anthropic이 발표한 개방형 표준으로, AI 시스템과 외부 도구·데이터를 연결하는 규약입니다. 쉽게 말하면 에이전트가 외부 세계와 소통하는 언어이지요. 2025년 12월 Anthropic은 MCP를 Linux Foundation 산하 Agentic AI Foundation(AAIF)에 기증했습니다. 당시 공식 발표에 따르면 ChatGPT, Gemini, Microsoft Copilot, VS Code 등이 MCP를 지원했고, 공개 서버는 1만 개를 넘어섰습니다.[1]

A2A(Agent-to-Agent Protocol)는 Google이 발표한 에이전트 간 통신 규약입니다. MCP가 에이전트와 도구의 연결이라면, A2A는 에이전트와 에이전트의 연결입니다. 에이전트가 자신의 능력을 Agent Card라는 형식으로 광고하면, 다른 에이전트가 최적의 파트너를 자동으로 찾아 협업합니다. 2026년 3월

첫 안정판(v1.0)이 나왔고, 8월에는 MCP가 있는 AAIF로 거처를 옮겼습니다. 도구를 잇는 규격과 에이전트끼리 잇는 규격이 같은 재단 아래 놓인 것이지요.[2]

이 두 프로토콜의 조합은, 전력 산업에 비유하면 송전 표준과 같습니다. 어떤 발전소에서 만든 전기든 같은 콘센트에 꽂으면 쓸 수 있듯, 어떤 회사가 만든 에이전트든 MCP/A2A를 통해 연결되면 함께 일할 수 있습니다.

콘센트 규격이 정해졌다면, 그 콘센트에 꽂힐 쪽도 준비가 되어 있어야겠지요. 에이전트 친화적인 인프라를 구축할 때 가장 중요한 원칙은 기계 가독성 우선(Machine Readability First)입니다.

사람에게 보여주는 웹 페이지와 에이전트에게 제공하는 데이터는 다릅니다. 사람은 시각적 배치와 색상으로 정보를 파악하지만, 에이전트에게는 이것이 의미 없는 잡음입니다. 에이전트에게는 구조화된 데이터, 일관된 스키마, 자체 설명적인 API가 필요합니다.

이것은 곧 API가 제품이라는 뜻입니다. 에이전트 시대에 고객의 절반은 사람이 아니라 다른 시스템의 에이전트일 수 있습니다. 이 에이전트들에게 API는 곧 사용자 인터페이스입니다. API가 예측 가능하고, 안정적이고, 풍부한 메타데이터를 제공해야 에이전트가 효과적으로 일할 수 있습니다.

첫 에이전트에게 줄 권한

신입 사원에게 첫날부터 법인 카드와 서버 관리자 비밀번호를 건네는 회사는 없지요. 첫 에이전트도 그렇게 맞이하면 됩니다.

첫 후보는 반복성과 검증 가능성이 높고, 실패해도 되돌릴 수 있는 일에서 고릅니다. 이메일 분류, 일정 조율, 데이터 입력, 정기 보고서 초안 같은 일들이지요. 그다음 사람만 드나들던 관리자 화면 뒤의 기능 가운데 그 일에 꼭 필요한 것만 API로 열어 줍니다. 읽기와 쓰기, 결제와 삭제 권한을 따로 나누고, 에이전트가 한 일은 빠짐없이 기록에 남깁니다.

처음 한동안은 에이전트가 제안만 하고 사람이 승인합니다. 제안의 정확도와 복구 절차가 확인된 뒤에야 위험이 낮은 업무부터 혼자 처리하게 하고, 사람은 사후 점검으로 물러납니다.

이렇게 자리를 잡아 가는 모습은 스마트 그리드와 닮았습니다. 예전엔 사람이 기기마다 스위치를 올리고 내렸지만, 스마트 그리드에서는 기기들이 수요와 시간에 맞춰 알아서 켜지고 꺼지지요. 그래도 배전반의 차단기는 늘 사람 손이 닿는 곳에 있습니다.

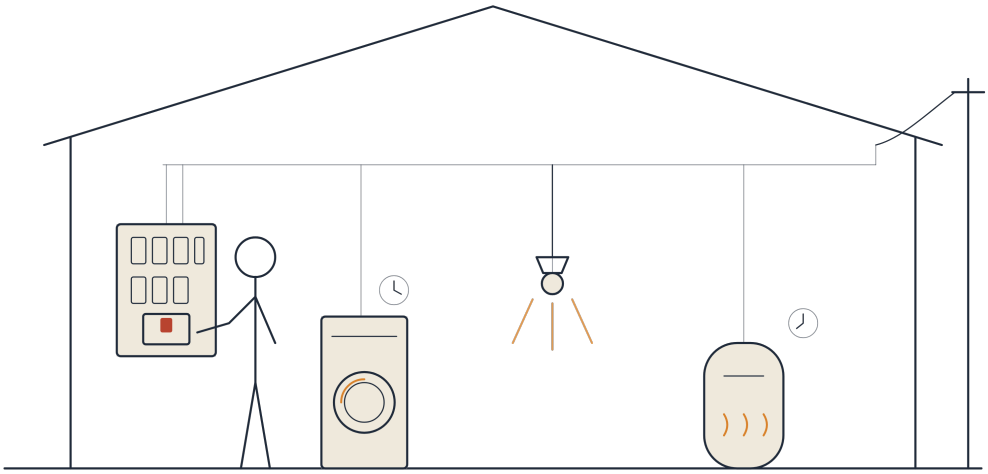


그림 4-3 세탁기와 온수기는 정해 둔 시각에 알아서 돌고, 전등은 사람이 들어오면 켜집니다. 그래도 차단기는 사람 손이 닿는 높이에 있지요.

실행 체크: 에이전트가 일할 수 있는 환경인가

1. 규칙 기반으로 반복되고 결과를 검증할 수 있는 업무를 목록으로 만들어 첫 후보를 고릅니다.
2. 그 업무에 필요한 기능만 API로 열고, 읽기·쓰기·결제·삭제 권한을 나눈 뒤 행동 로그와 이상 감지를 붙입니다.
3. 처음엔 제안만 하게 하고, 검증이 끝난 낮은 위험 업무부터 자율 처리로 권한을 넓힙니다.

리스크: 권한을 한꺼번에 열면 오작동 하나가 미치는 범위도 그만큼 커집니다. 멈추는 버튼과 되돌리는 경로가 자율성보다 먼저 설치되어 있어야 합니다.

새벽 3시에도 일하는 팀원을 들이기 전에, 그 팀원이 새벽 3시에 실수했을 때 누가 전화를 받을지부터 정해 두세요.



1. Anthropic, "Donating the Model Context Protocol and Establishing the Agentic AI Foundation" (2025.12.09, https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation) - MCP 기증, 주요 제품의 지원 현황, 공개 서버 수를 밝힌 공식 발표.
2. "A2A joins the Agentic AI Foundation" (AAIF, 2026.8.17, https://aaif.io/blog/a2a-joins-aaif) — A2A v1.0(2026.3) 이후 AAIF 호스팅 프로젝트로 합류, 150개 이상 기관 지지.

원칙 6: 한 팀으로 일하라

혼자서 10인분의 일을 하는 "슈퍼 개인"들이 등장하고 있습니다. 이들은 직접 벽돌을 나르는 일꾼이 아닙니다. AI 에이전트 군단을 지휘하는 현장 소장이지요. 오케스트라 지휘자가 직접 소리를 내지 않으면서도 음악을 만들어내듯, 이 새로운 일꾼들은 AI를 지휘해서 가치를 만들어냅니다.

이 원칙은 두 가지를 동시에 다룹니다. 하나는 인간과 AI가 같은 팀에서 역할을 나눠 협업하는 방식이고, 다른 하나는 조직 전체가 지휘자 구조로 전환하는 방식입니다. 개인이 AI와 손발을 맞추는 일과 조직이 지휘 구조를 바꾸는 일은 따로 갈 수 없지요.

역할을 나누되 판단 근육은 지킨다

인간과 AI의 역할을 "실행"과 "감독"으로 고정하면 실제 업무를 설명하기 어렵습니다. AI가 제안·실행·비판을 오갈 때도 있고, 사람이 직접 조사하고 예외를 처리해야 할 때도 있습니다. 역할은 전문성, 되돌릴 수 있는지, 오류 비용, 결과를 검증할 수 있는지에 따라 정해야 합니다.

현대 발전소를 떠올리면 이 분업이 선명해집니다. 터빈 회전과 밸브 개폐 같은 실시간 제어는 자동 제어 시스템이 쉬지 않고 처리하지만, 발전소를 언제 멈추고 어떤 이상 신호에 어떻게 대응할지 결정하는 것은 제어실의 운영자이지요. AI와 인간의 협업도 정확히 이 구조입니다. AI가 자동 제어 시스템이라면, 인간은 계기판 앞에 앉은 운영자입니다.

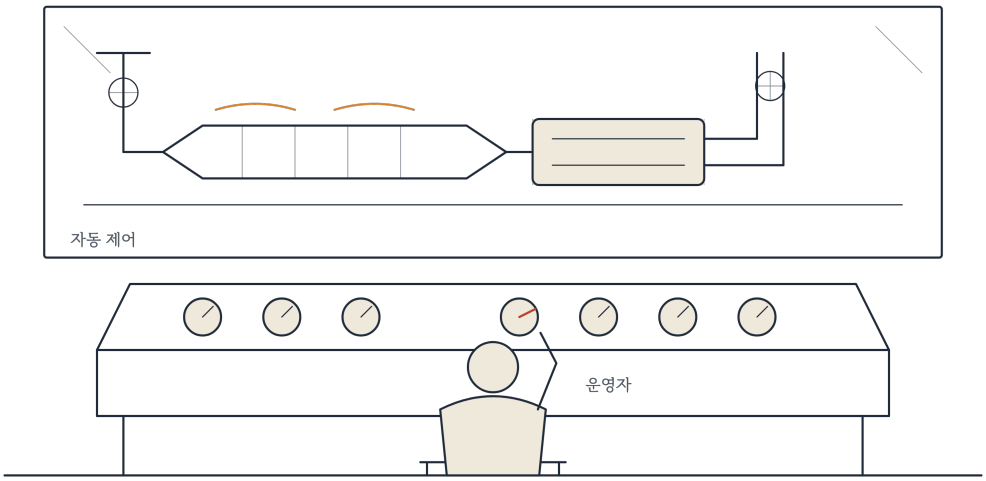


그림 4-4 유리창 너머의 터빈과 밸브는 자동 제어가 쉬지 않고 돌립니다. 운영자는 그중 경계선에 닿은 바늘 하나에 손을 벋지요.

AI를 때로는 초안을 내는 동료로, 때로는 반복 작업을 맡는 자동화 설비로 대할 수 있습니다. 사람이 더 뛰어나다고 가정할 필요는 없습니다. 지켜야 할 것은 책임 소재이지요. 최종 책임은 그 시스템을 도입하고 운영하며 결정 권한을 부여한 사람과 조직에 있습니다.

책임이 사람에게 남는다면, AI가 내놓은 결과물은 어떤 태도로 받아 들어야 할까요? AI 결과물을 다루는 핵심 태도는 "Trust but Verify"입니다. 신뢰하되 검증하라.

다만 모든 출력을 같은 깊이로 사람이 다시 읽으면 자동화의 이점도 사라지고 검토 피로가 커집니다. 되돌릴 수 없는 결정, 외부 공개, 금전·권리·안전에 영향을 주는 결과는 깊게 검토하고, 낮은 위험의 반복 업무는 표본 점검과 자동 테스트로 관리하는 식으로 검증 강도를 나누어야 합니다.

검증하는 사람에게도 남는 게 있습니다. 연구에 따르면 AI 코파일럿을 활용한 개발자들은 코드 작성 시간은 줄이고, 검토와 디버깅에 더 많은 시간을 투입했습니다. 결과적으로 기존보다 높은 수준의 감독 역량을 발휘하게 됐지요. 검증하는 행위 자체가 학습이 되는 겁니다.

그런데 검증을 건너뛰고 승인 도장만 찍기 시작하면 이야기가 달라집니다. 이 협업 구조에서 가장 조심해야 할 함정이 디스킬링(de-skilling)입니다. AI에 의존하면서 인간의 핵심 역량이 퇴화하는 현상이지요.

자동항법 장치에 오래 의존한 조종사가 정작 긴급 상황에서 수동 비행을 못 하는 사례가 있습니다. 사무직에서도 AI가 거의 다 작성한 보고서를 형식적으로 승인만 하는 역할로 전락하면, 장기적으로 분석력과 문제해결 능력이 약해집니다.

이를 "부분적 자동화의 역설"이라고 합니다. 시스템이 좋아질수록 인간이 나태해져서, 정작 필요한 순간에 대처를 못하게 되는 함정이지요.

방어책은 구체적입니다. 항공 조종 분야에서 정기적으로 시뮬레이터 훈련을 실시하듯, 일반 업무에서도 "AI 없이 풀어보기" 연습을 의도적으로 해 보세요. AI의 답안을 일부러 의심해보는 것도 좋습니다. 멘토링 제도를 통해 시니어들이 직접 지도하면서 본인의 역량도 연마하게 하고, 직무 로테이션으로 모든 사람이 실행 경험을 유지하게 하는 것도 방법입니다.

조직이 지휘자 구조로 바뀔 때

개인 수준의 협업이 자리잡으면, 조직 수준의 변화가 뒤따라야 합니다.

과거의 조직은 피라미드 아래서 묵묵히 실행하는 사람들이 지탱했습니다. AI 시대에는 더 많은 구성원이 AI 에이전트를 지휘하고 조율하는 오케스트레이터 역할을 맡게 됩니다. 그렇다고 모든 직접 작업이 사라지는 것은 아닙니다. 특히 신입 직원에게는 결과를 판단할 수 있는 기초를 쌓도록 직접 조사하고 만들고 고치는 경험이 필요합니다.

이 전환에서 가장 많이 받는 질문은 "그러면 사람이 할 일이 없어지는 거 아니냐"는 것입니다. 정반대입니다. 직접 하는 것보다 지휘하는 것이 더 어렵습니다. AI에게 정확한 맥락을 전달하고, 결과물의

품질을 판단하고, 여러 에이전트의 작업을 조율하고, 예외 상황에서 개입하는 일은 높은 수준의 전문성과 판단력을 요구합니다.

PwC는 이미 6만5천 명 직원 전원에게 AI 도구 및 역량 교육을 실시하고 있습니다. 회사 최초로 최고 AI 책임자(CAIO)를 임명했고, "AI의 잠재력을 실현하려면 조직 전체가 새로운 기술과 업무 방식을 익혀야 한다"는 원칙을 세웠습니다.

PwC만 한 규모가 아니어도 괜찮습니다. 한 팀으로 일하는 구조는 표 한 장에서 시작해도 됩니다. 업무 하나를 골라 "AI가 하는 일"과 "사람이 하는 일"을 나누는 데서 멈추지 말고, 오류가 났을 때 누가 발견하고 누가 멈추며 누가 복구하는지까지 적어 보는 것이지요. 업무가 바뀌면 표도 다시 고칩니다.

같은 표에 검증 강도도 적어 둡니다. 위험 등급에 따라 자동 테스트로 충분한 일, 표본 검토가 필요한 일, 전문가 승인이나 이중 승인이 필요한 일이 다르니까요. 검토했다는 서명보다 오류를 실제로 잡아낼 수 있는 조건이 중요합니다.

표의 마지막 칸은 사람의 역량입니다. 월 1회라도 "AI 없이 해 보기" 시간을 두고, 성과 지표에 결과물뿐 아니라 사람의 핵심 역량이 어떻게 자라고 있는지도 넣습니다. 조종사가 시뮬레이터에서 수동 비행을 연습하는 것과 같은 이유입니다.

실행 체크: AI와 사람이 한 팀으로 굴러가는가

1. 업무별로 AI와 사람의 역할, 오류를 발견·중단·복구할 책임자를 적은 역할표를 눈에 보이는 곳에 둡니다.
2. 위험 등급에 따라 자동 테스트·표본 검토·전문가 승인 가운데 어떤 검증을 할지 업무 절차에 공식으로 넣습니다.
3. 월 1회라도 "AI 없이 업무" 세션을 돌려 사람의 핵심 역량을 유지하고 측정합니다.

리스크: AI에 전면 의존하면 사람의 판단 근육이 위축됩니다. 자율 범위를 넓혀 가더라도 핵심 결정에는 언제든 사람이 끼어들 수 있는 문을 남겨 두어야 합니다.

지휘자 벤저민 젠더는 이렇게 말했습니다. "오케스트라 지휘자는 소리를 내지 않는다. 그의 힘은 다른 사람들을 강력하게 만드는 능력에서 나온다." AI 시대의 팀에서는 지휘자 혼자 악보를 읽어서는 부족합니다. 연주자 한 사람 한 사람이 자기 파트가 틀렸을 때 그걸 알아듣는 귀를 지니고 있어야 하지요.

원칙 7: 신뢰를 내장하라

전기 콘센트에는 접지 단자가 있습니다. 전기가 새면 사람이 다치는 대신 땅으로 흘러가게 만드는 안전장치이지요. 이것은 콘센트를 다 만들고 나서 "아, 안전장치를 붙여야지"라고 해서 추가한 게 아닙니다. 설계 단계부터 내장된 것입니다.

AI 시스템의 신뢰도 마찬가지입니다. 보안과 윤리, 데이터 주권도 접지처럼 처음부터 설계도에 그려 넣어야 합니다. 다 만든 뒤에 붙이면 늦지요.

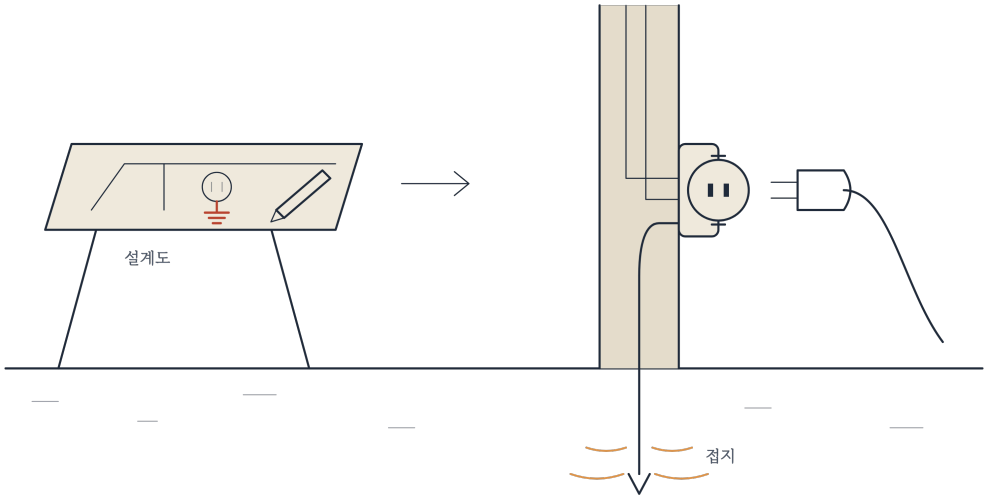


그림 4-5 접지선은 벽을 바르기 전, 설계도의 콘센트 기호 옆에 먼저 그려집니다. 새는 전기는 사람 대신 땅으로 흘러가지요.

2026년 1월 22일, 한국의 AI 기본법이 시행됐습니다.[1] EU AI Act는 2025년부터 금지 조항과 범용 AI 규칙을 차례로 적용해 왔고, 2026년 8월 2일 대부분의 조항이 적용 단계에 들어섰습니다. 그런데 정작 무게가 가장 큰 고위험 시스템 의무는 그 엿새 전인 7월 27일 발효된 디지털 옴니버스(Digital Omnibus)로 미뤄졌지요. 채용·신용평가 같은 독립형 고위험 시스템은 2027년 12월 2일, 의료기기처럼 제품에 내장된 고위험 시스템은 2028년 8월 2일부터 적용됩니다.[2] 나라별 일정은 제7부 「일곱 개 국가의 선택」에서 나란히 비교합니다. 규제 환경이 빠르게 정비되고 있지만, 신뢰를 규제 대응용 체크리스트로 접근하면 늘 한 발 늦습니다. 법이 요구하기 전에 스스로 내장해 둔 조직은 규제 일정이 앞당겨지든 미뤄지든 덜 흔들립니다.

신뢰를 받치는 네 기둥과 데이터 주권

Gartner가 제안한 AI TRiSM(Trust, Risk and Security Management)이라는 프레임워크가 있습니다. AI 시스템의 신뢰성, 위험 관리, 보안을 통합적으로 다루는 접근이지요. 이 프레임워크는 네 가지 축으로 구성됩니다.

설명 가능성(Explainability). AI가 왜 그런 판단을 내렸는지 사람이 이해할 수 있어야 합니다. "이 지원자를 추천한 이유"를 AI가 설명하지 못하면, 채용 담당자는 그 추천을 신뢰할 수 없습니다. 의사가 AI의 진단 근거를 확인할 수 없으면, 환자의 생명을 그 판단에 맡기기 어렵습니다.

공정성(Fairness). AI 모델에 내재된 편향을 감지하고 교정해야 합니다. 학습 데이터에 특정 집단이 과소 대표되면 AI의 판단도 편향됩니다. 대출 심사 AI가 특정 지역 거주자에게 불리하게 작동하거나, 이력서 스크리닝 AI가 성별에 따라 차별적 결과를 내는 사례는 이미 여러 차례 보고됐습니다.

개인정보 보호(Privacy). AI가 학습하고 추론하는 과정에서 개인정보가 어떻게 다뤄지는지 투명하게 관리해야 합니다. 특히 생성형 AI는 학습 데이터에 포함된 개인정보를 의도치 않게 출력으로 노출할 위험이 있습니다.

보안(Security). 모델 자체에 대한 공격 — 적대적 입력으로 판단을 교란하거나, 프롬프트 인젝션으로 시스템을 조작하는 시도 — 에 대비해야 합니다. 전통적 소프트웨어 보안과는 다른 차원의 위협입니다.

네 기둥이 주로 모델 안쪽을 들여다본다면, 바깥쪽에서 던져야 할 질문도 하나 있습니다. 이 데이터는 누구의 것이고, 어디에 머무는가? AI 시대에 특히 중요해진 개념이 데이터 주권(Data Sovereignty)입니다. 데이터가 어디에 저장되고, 누가 접근하고, 어떤 관할권의 법률이 적용되는지에 관한 문제입니다.

글로벌 서비스를 운영하는 기업은 이 문제를 피할 수 없습니다. EU의 GDPR이 유럽경제지역 밖으로 개인정보를 보내는 일을 일률적으로 금지하는 것은 아닙니다. 적정성 결정, 표준계약조항 같은 적절한 보호조치, 제한적인 예외 가운데 적용 가능한 근거를 갖추도록 요구합니다.[3] 한국의 AI 기본법도 투명성과 안전성 요건을 명시하고 있지요.

Apple이 Private Cloud Compute를 통해 보여준 접근은 참고할 만합니다. 로컬 모델이 처리할 수 있는 것은 기기에서 처리하고, 대형 모델이 필요한 요청만 Apple Silicon 서버로 보냅니다. 그마저도 요청이 끝나면 데이터를 삭제합니다. "프라이버시 우선 설계"를 기술 아키텍처 수준에서 구현한 것이지요.

구호를 절차로 바꾸려면

Responsible AI라는 말이 잘 쓰이지만, 실천으로 이어지지 않으면 구호에 불과합니다. 실천의 핵심은 프로세스에 녹여 넣는 것입니다.

코드 리뷰에 보안 검토가 포함되듯, AI 모델 배포 전에 편향 점검, 설명 가능성 검증, 개인정보 처리 확인이 의무적으로 수행되어야 합니다. 이것을 "한 번 하고 끝"이 아니라 모델이 운영되는 동안 지속적으로 모니터링해야 합니다. 모델의 행동은 데이터가 바뀌면 달라지니까요.

AI 시스템의 감사 로그를 기록하는 것도 중요합니다. 입력과 출력, 모델·프롬프트 버전, 검색 근거, 도구 호출, 권한 승인과 정책 판정이 남아야 사고 경로를 재구성할 수 있습니다. 모델이 생성한 설명이나 사고 과정은 실제 인과적 판단 근거와 다를 수 있으므로, 감사 증거는 운영 로그와 정책 판정까지 함께 확인해야 합니다.

그렇다면 내일 아침 당장 손댈 수 있는 일은 무엇일까요? 신뢰를 내장하는 일은 문서 한 장에서 시작해도 됩니다. 프라이버시, 데이터 사용, 차별 방지, 투명성, 사람의 개입 기준을 담은 윤리 가이드라인이지요. 형식보다 내용이 중요합니다. 현업 부서가 실제로 펼쳐 보고 판단에 쓸 수 있을 만큼 구체적이어야 하니까요.

사람의 승인은 위험에 맞춰 둡니다. 모든 고객 대면 문장을 형식적으로 읽게 하면 검토자는 금세 지칩니다. 되돌리기 어려운 결정과 고위험 예외에 승인 지점을 두고, 검토자에게 필요한 전문성과 시간, 권한을 줍니다. 불확실하거나 이해충돌이 있을 때 위로 올릴 기준도 정해 둡니다. 그렇지 않으면 사람의 승인이 자동화 편향과 검토 피로를 가리는 도장으로 전락할 수 있지요.

편향 점검은 한 번의 감사로 끝나지 않습니다. 데이터가 바뀌면 모델의 행동도 바뀌니까요. 그리고 고객이 AI의 결정에 대해 사람의 재검토를 요청할 수 있다는 규정, 곧 인간 개입권을 명시해 둡니다. 고객의 권리이면서 조직의 안전망이기도 합니다.

실행 체크: 신뢰를 설계에 넣었는가

1. 프라이버시·데이터 사용·차별 방지·투명성·인간 개입 기준을 담은 가이드라인을 현업이 참조할 만큼 구체적으로 씁니다.
2. 되돌리기 어려운 결정에 사람의 승인 지점을 두고, 편향 점검을 정기 운영 절차에 넣습니다.
3. "고객은 AI 결정에 대해 사람의 재검토를 요청할 수 있다"는 인간 개입권을 약관과 업무 절차에 명시합니다.

리스크: 신뢰는 사고가 난 뒤에 붙이면 비싸고 늦습니다. 아래에서 볼 해석 가능성 연구가 발전해도 완전한 통제와는 거리가 남아 있으니, 사람이 마지막에 개입하는 지점을 지우지 않는 편이 안전합니다.

블랙박스는 얼마나 열렸나

대형 AI 모델의 내부 작동은 여전히 충분히 설명되지 않습니다. 수많은 매개변수가 얽혀 답을 내놓기 때문에, 특정 출력이 나온 인과 경로를 사람이 읽을 수 있는 설명으로 완전히 복원하기 어렵습니다. 이 간극을 줄이려는 연구가 기계적 해석 가능성(Mechanistic Interpretability)입니다.

MIT Technology Review는 2026년 10대 획기적 기술 중 하나로 기계적 해석 가능성을 선정했습니다[4]. 이 연구는 내부 표현과 계산 회로에 관한 가설을 세우고 실험으로 확인합니다. 모델 전체의 정확한 회로도를 한 번에 그리는 기술이라기보다, 특정 행동과 연결된 내부 패턴을 부분적으로 조사하는 방법에 가깝습니다.

Anthropic은 자사 모델의 내부 활성에서 사람이 해석할 수 있는 특징(feature)을 찾고 매핑하는 연구를 발표했습니다[5]. 특정 도시, 코드 패턴, 감정과 관련된 특징을 관찰했지만, 이것은 뇌의 기능별 영역이나 하나의 개념을 전담하는 뉴런과 구분해 이해해야 합니다. 특징은 여러 위치에 분산될 수 있고, 분석 도구가 찾아낸 근사치이기도 합니다.

이런 분석은 안전 평가에서 조사할 가설과 이상 신호를 찾는 데 도움을 줄 수 있습니다. 그러나 특정 특징을 발견했다고 해서 출력의 정확성이나 공정성이 입증되는 것은 아닙니다. 내부 분석 결과만으로 배포 안전을 보증할 수도 없습니다.

Google DeepMind도 Gemma Scope 2라는 공개 해석 도구를 내놓았습니다[6]. 연구자는 특정 입력에서 나타나는 활성 패턴을 탐색하고 가설을 비교할 수 있습니다. 도구가 공개됐다는 사실은 연구의 접근성을 넓히지만, 비전문가도 모델의 판단 경로를 그대로 읽을 수 있다는 뜻은 아닙니다.

전기 쪽 역사는 좋은 거울이지만 비유의 한계도 있습니다. 전력망의 회로도는 설계자가 정한 연결을 기록하지만, 학습된 모델의 내부 표현은 데이터와 훈련 과정에서 형성됩니다. 따라서 기계적 해석 가능성은 완성된 회로도를 읽는 검사라기보다, 복잡한 설비 안에서 특정 현상의 원인을 좁혀 가는 진단에 가깝습니다.

물론 한계도 있습니다. 강한 최적화(모델의 성능을 극대화하는 것), 완벽한 가치 포착(인간의 의도를 빠짐없이 반영하는 것), 견고한 일반화(새로운 상황에서도 안전하게 작동하는 것) — 이 세 가지를 한꺼번에 만족하기 어렵다는 긴장을 이 책은 "정렬 트릴레마"라 부르겠습니다. 해석 가능성이 발전해도 이 근본적 긴장은 남아 있습니다. 모델의 내부를 완전히 이해하는 것과, 그 이해를 바탕으로 모델을 완벽하게 통제하는 것 사이에는 여전히 간극이 존재합니다.

신뢰를 설계에 내장하려면 내부 분석도 필요합니다. 그 분석은 행동 평가, 집단별 성능 점검, 레드팀, 데이터와 모델의 출처 기록, 감사 로그, 사고 재현과 함께 써야 합니다. 사용자가 받아야 할 설명이나 법적 감사 자료를 내부 특징 지도 하나로 대신할 수도 없습니다. 기계적 해석 가능성은 유망한 진단 도구이지만, 블랙박스가 사라졌다는 선언은 아직 이릅니다.

전기가 보급될 때도 안전 규격이 자리 잡기까지 수많은 사고가 있었습니다. 감전, 화재, 누전. 그 경험이 지금의 접지 표준과 차단기 규격, 전기 안전법으로 이어졌지요. 첫 배포 문서에 데이터 사용 근거, 집단별 평가, 권한과 승인, 로그, 이의 제기와 복구 절차를 함께 적는 일은 그 긴 사고의 역사를 건너뛰려는 시도입니다. AI는 앞선 기술의 사고 기록을 미리 손에 쥐고 출발하는 드문 기술이니까요.



1. 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(국가법령정보센터, 시행 2026.01.22, https://www.law.go.kr/LSW/lsInfoP.do?ancYnChk=0&chrClsCd=010202&efYd=20260122&lsiSeq=282791&urlMode=lsInfoP)

2. European Commission, "AI Act" (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai) - 적용 일정과 디지털 옴니버스에 따른 고위험 의무 적용 시점 조정을 설명한 공식 페이지. 근거 규정은 Regulation (EU) 2026/1744 (Digital Omnibus on AI, 관보 2026.7.24, 발효 2026.7.27, https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng) — Annex III 고위험 2027.12.2, Annex I 2028.8.2로 연기. 표준 마련 여부와 묶지 않은 확정 일정.

3. European Commission, "International dimension of data protection" (https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection_en) - EU 역외 이전의 적정성 결정, 보호조치, 예외를 설명한 공식 안내.

4. MIT Technology Review, "Mechanistic Interpretability" (2026 Breakthrough Technologies, https://www.technologyreview.com/2026/01/12/1130003/mechanistic-interpretability-ai-research-models-2026-breakthrough-technologies/)

5. Anthropic, "Mapping the Mind of a Large Language Model" (2024, https://www.anthropic.com/research/mapping-mind-language-model)

6. Google DeepMind, "Gemma Scope 2" (2025, open-source interpretability tools)

원칙 8: 확률을 경영하라

전통적인 소프트웨어에 1+1을 물으면 항상 2가 나옵니다. 백 번을 물어도 2입니다. 결정론적 시스템이지요.

AI 시스템은 한 종류가 아닙니다. 스팸 분류기는 같은 모델과 입력에 같은 점수와 판정을 내도록 운영할 수 있습니다. 그 점수는 "스팸일 가능성"을 나타내지만, 점수가 실제 빈도와 맞는지는 별도의 보정 검사가 필요합니다. 생성형 AI는 후보를 고르는 샘플링 설정, 대화 맥락, 검색 결과, 도구 호출, 모델 버전에 따라 같은 요청에도 다른 답을 낼 수 있습니다.

따라서 "AI는 본질적으로 매번 달라진다"고 뭉뚱그리기보다, 어디에서 불확실성과 변동이 생기는지 구분해야 합니다. 분류 확률, 생성 결과의 다양성, 외부 도구의 실패는 서로 다른 문제이고 관리 방법도 다릅니다.

전기를 다뤄 본 사람에게는 낯설지 않은 상황입니다. 전압은 늘 조금씩 흔들립니다. 완벽하게 일정한 전기는 현실에 없지요. 그래서 잘 만든 가전제품은 전압이 완벽하기를 기대하지 않고, 변동이 있어도 안전하게 돌아가도록 설계됩니다. 확률을 경영한다는 것도 같은 발상입니다. AI의 출력이 매번 똑같기를 바라는 대신, 흔들림을 전제로 시스템을 짜는 것이지요.

기존의 "기능이 정상 작동하는가?"라는 질문은 AI 시스템에서도 여전히 필요합니다. 여기에 "업무 품질은 허용 범위 안인가?", "집단별 차이는 없는가?", "어려운 사례를 안전하게 넘기는가?"라는 질문이 더해집니다.

흔들려도 안전하게 돌아가도록

전통적인 SLA(서비스 수준 계약)는 가용성과 응답 시간 같은 운영 기준을 다룹니다. AI를 쓴다고 이 기준이 사라지지는 않습니다. 여기에 업무별 품질, 안전, 권리 보호, 예외 처리 기준을 더해야 합니다.

품질 지표는 용도에 맞춰 골라야 합니다. 분류 업무라면 정확도만이 아니라 놓치면 위험한 사례의 재현율과 집단별 차이를 보고, 생성 업무라면 사실성·근거성·유해성·형식 준수 등을 평가할 수 있습니다. 같은 입력에서 결과가 얼마나 흔들리는지는 품질과 별도의 변동 지표로 다루는 편이 정확합니다. SLA는 모델이 스스로 내놓은 자신감 점수 하나가 아니라 용도별 검증 지표로 정해야 합니다.

이것을 실무에서는 어떻게 적용할까요?

고객 문의를 AI가 처리하는 콜센터를 생각해 봅시다. 통화 연결과 응답 시간은 그대로 측정합니다. 그 위에 해결 여부, 잘못된 안내, 이관이 필요한 문의를 놓친 비율, 이관 뒤 처리 시간, 고객이 사람에게 재검토를 요청할 수 있는지를 붙입니다. 자율 처리 범위를 넓히는 목표와 잘 넘기는 능력을 함께 보아야 하지요.

기준을 정했다면, 그 기준을 벗어나지 못하게 막는 장치도 있어야겠지요. 확률적 시스템을 운영하려면 가드레일(guardrails)이 필요합니다. AI가 특정 범위를 벗어나지 않도록 막는 안전 울타리입니다.

가드레일은 여러 층위에서 작동합니다.

입력 가드레일은 AI에 들어가는 요청을 걸러냅니다. 악의적인 프롬프트 인젝션, 부적절한 요청, 시스템이 처리할 수 없는 범위의 질문을 사전에 차단합니다.

출력 가드레일은 AI의 응답을 검증합니다. 민감한 개인정보가 포함되어 있지 않은지, 사실과 다른 내용이 포함되어 있지 않은지, 톤과 형식이 적절한지 확인합니다.

행동 가드레일은 AI 에이전트의 행위 범위를 제한합니다. "이 에이전트는 이메일을 읽을 수 있지만 삭제할 수는 없다", "이 에이전트는 주문 정보를 조회할 수 있지만 결제를 승인할 수는 없다"처럼 권한 경계를 명확히 합니다.

울타리를 쳤다고 일이 끝나지는 않습니다. 서버가 살아 있어도 모델 품질은 서서히 나빠질 수 있습니다. 입력 분포가 달라지거나, 검색 자료와 도구가 바뀌거나, 모델과 프롬프트 버전이 교체되면서 배포 당시의 평가와 다른 행동이 나타날 수 있지요.

이것을 잡으려면 지속적 모니터링이 필요합니다. 정답을 나중에 확인할 수 있는 업무는 실제 오류와 집단별 성능을 추적하고, 즉시 확인하기 어려운 업무는 표본 평가와 사용자 이의 제기, 근거 누락, 이관 실패 같은 대리 지표를 함께 봅니다. 모델 버전과 검색·도구 환경도 기록해야 원인을 찾을 수 있습니다.

전통적인 소프트웨어 모니터링이 "서버가 살아있는가"를 확인했다면, AI 모니터링은 "모델이 여전히 현실을 반영하고 있는가"를 확인합니다.

실패했을 때 물러설 자리

AI 시스템을 운영할 때 잊지 말아야 할 것이 있습니다. 평가에서 발견하지 못한 실패는 운영 중에도 나타날 수 있습니다. 그래서 평균 성능과 별개로 실패했을 때의 대체 경로가 필요합니다.

폴백(fallback) 전략이란 AI가 실패했을 때 자동으로 발동되는 대체 경로입니다. AI 챗봇이 고객의 질문을 이해하지 못하면 인간 상담사에게 즉시 연결합니다. AI 번역이 신뢰도 임계값 이하면 "이 번역은 검증이 필요합니다"라는 경고를 표시합니다. 자율주행차가 판단 불가 상황을 만나면 안전하게 정차합니다.

핵심은 시스템이 답하지 않거나 이관할 조건을 갖게 하는 것입니다. 모델의 자기평가만 믿기보다 입력 범위, 근거 유무, 여러 검사 결과, 위험 등급을 조합해 후퇴 조건을 정하고 실제 사례로 검증해야 합니다.

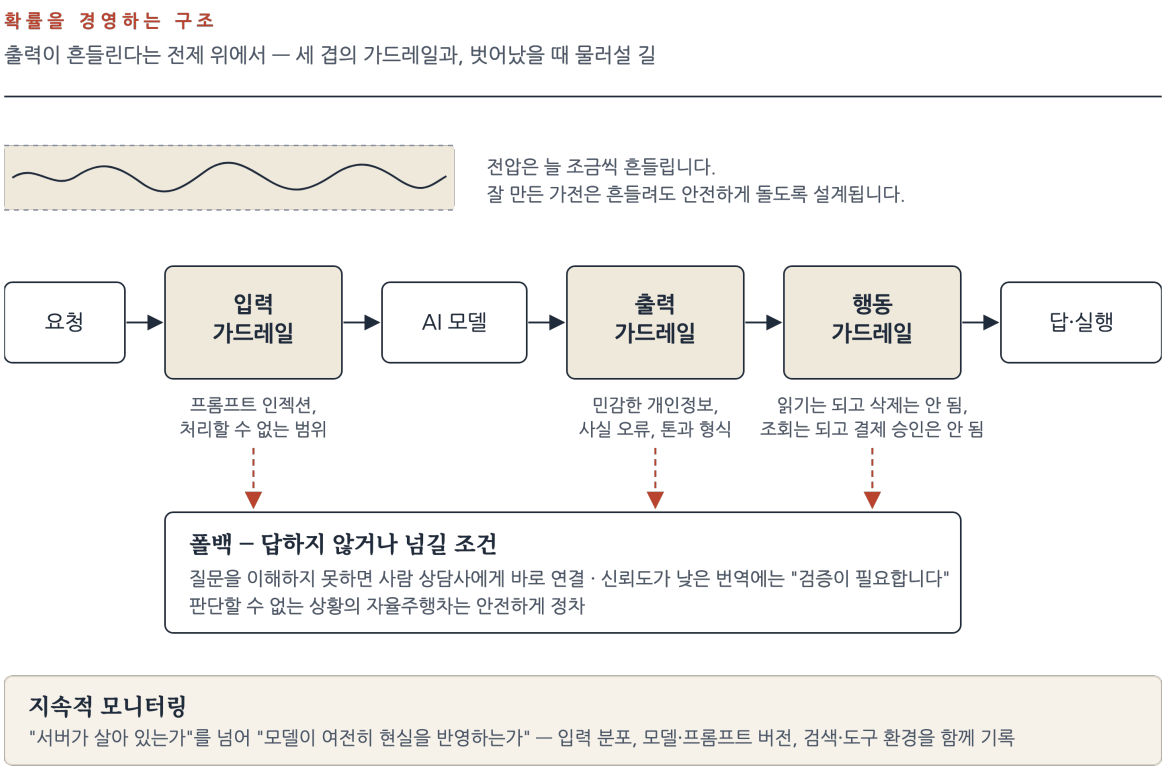


그림 4-6 요청은 입력, 출력, 행동의 세 겹 울타리를 지나갑니다. 어느 울타리에서 걸리든 사람에게 넘기거나 멈춰 서는 길이 미리 나 있어야 하지요.

이런 조건과 기준이 머릿속에만 있으면, 정작 한밤의 사고 앞에서는 아무도 꺼내 들지 못합니다. 확률을 경영하는 일은 장부 한 권에서 시작합니다. 업무 하나를 골라 가용성과 지연 시간, 업무 품질, 결과의 흔들림, 안전과 권리, 이관과 복구를 서로 다른 칸에 나눠 적는 것이지요. 칸마다 무엇으로 재는지, 어떤 집단을 따로 보는지, 어느 선을 넘으면 경보가 울리는지, 그때 누가 받는지를 이어 둡니다.

앞의 콜센터라면 장부의 한 줄은 이렇게 생겼을 겁니다. "환불 문의 자동 응답, 잘못된 안내 비율 주간 표본 평가, 기준선을 넘으면 자동 응답을 끄고 상담원 연결로 전환, 담당은 고객운영팀장." 이 한 줄이 있으면 불확실성이 한 숫자 뒤에 숨지 못합니다.

실행 체크: 확률을 관리하고 있는가

1. AI 서비스의 SLA를 "작동/미작동"에서 업무 품질 목표, 경보 기준, 사람에게 넘기는 절차까지 담도록 다시 씁니다.
2. 입력·출력·행동 세 단계에 가드레일을 걸고, 핵심 지표를 계속 지켜보는 화면을 만듭니다.
3. "AI가 실패하면 어떻게 되는가"에 대한 폴백 시나리오를 적용 영역마다 문서로 남기고, 실제로 한 번 돌려 봅니다.

리스크: 폴백 없는 AI 시스템은 안전벨트 없는 자동차와 같습니다. 실패를 없애겠다는 목표보다, 실패한 순간 어디로 물러설지를 먼저 정해 두는 쪽이 현실적입니다.

여러분의 AI 서비스가 오늘 밤 엉뚱한 답을 내기 시작한다면, 고객은 무엇을 보게 될까요? 그 화면을 미리 그려 둔 팀만이 이 질문에 바로 답할 수 있습니다.

원칙 9: 최적 조합을 찾아라

가정마다 대형 발전소를 세울 수는 없습니다. 그렇다고 지붕 위 태양광 패널만으로 도시를 운영할 수도 없지요. 현실적인 에너지 공급은 대형 발전소와 분산 발전의 조합으로 이루어집니다. 낮에는 태양광이 돌아가고, 밤에는 화력이 받쳐주고, 피크 시간에는 양수 발전이 투입됩니다.

AI 인프라도 마찬가지입니다. 모든 것을 하나의 거대한 클라우드 모델에 맡길 수도, 모든 것을 기기 안의 작은 모델로 처리할 수도 없습니다. 상황에 맞는 최적 조합을 찾는 것이 이 원칙의 핵심입니다.

어떤 모델을, 어디서 돌릴까

GPT나 Claude 같은 대형 프론티어 모델은 추론 능력이 뛰어나고 다양한 과제를 처리할 수 있습니다. 하지만 비용이 높고, 네트워크 연결이 필요하고, 응답 시간이 상대적으로 깁니다. 모든 요청을 대형 모델에 보내는 것은 편지 한 통 보내는 데 화물 트럭을 쓰는 것과 비슷합니다.

소형 모델 — 기기에서 직접 돌아가는 온디바이스 모델 — 은 빠르고 저렴하고 개인정보 보호에 유리합니다. 하지만 처리할 수 있는 과제의 범위가 좁고, 복잡한 추론에는 한계가 있습니다.

현명한 접근은 라우팅입니다. 들어오는 요청의 복잡도를 판단해서, 간단한 것은 소형 모델이 처리하고 어려운 것만 대형 모델로 보내는 구조이지요.

원칙 7에서 본 Apple의 Private Cloud Compute가 바로 이 라우팅의 예입니다. 기기에서 감당할 요청은 온디바이스 모델이 처리하고, 복잡한 요청만 Apple Silicon 서버로 넘기지요.

이런 구조에서 핵심은 "어디서 처리할 것인가"를 자동으로 판단하는 라우터입니다. 사용자는 신경 쓸 필요 없습니다. 마치 스위치를 켜면 전등이 들어오듯, 그 전기가 어느 발전소에서 왔는지 알 필요 없는 것과 같습니다.

모델 크기뿐 아니라, 연산이 일어나는 위치도 조합의 대상입니다.

클라우드는 무한에 가까운 연산 자원을 제공합니다. 대규모 학습, 복잡한 추론, 여러 에이전트의 동시 실행에 적합합니다. 하지만 네트워크 지연이 있고, 데이터가 외부로 나가야 합니다.

엣지(기기, 공장 내 서버, 자동차 내 칩)는 실시간 반응이 필요한 곳에 적합합니다. 자율주행차의 장애물 감지, 공장 라인의 불량 판별, 스마트폰의 음성 인식처럼 밀리초 단위의 반응 속도가 필요한 영역입니다. 클라우드로 데이터를 보내고 답을 기다릴 여유가 없는 상황이지요.

최적의 조합은 산업과 사용 사례에 따라 다릅니다.

제조업에서는 라인 위의 카메라가 엣지 모델로 불량을 즉시 감지하고, 수집된 불량 이미지는 클라우드로 올라가 대형 모델의 재학습에 사용됩니다.

금융에서는 실시간 거래 모니터링은 엣지에서 처리하고, 복잡한 사기 패턴 분석은 클라우드의 대형 모델이 담당합니다.

헬스케어에서는 웨어러블 기기가 기본적인 건강 지표 분석을 하고, 이상 징후가 감지되면 클라우드의 진단 모델에 연결합니다.

지능의 단가를 관리하는 법

라우터가 있어야 하는 이유는 하나 더 있습니다. 바로 비용이지요.

모델 시장에는 성능과 가격이 다른 선택지가 빠르게 늘고 있습니다. 서로 다른 세대와 등급의 토큰 가격만 비교해 하락률을 계산하면 실제 경제성을 오해하기 쉽습니다. 같은 업무에서 품질 기준을 통과할 때까지 든 재시도와 검토 비용을 함께 봐야 합니다.

대부분의 요청은 간단합니다. 분류, 요약, 형식 변환 같은 작업에 가장 비싼 모델을 쓸 필요가 없지요. 반면 복잡한 추론이나 창의적 생성이 필요한 요청에는 강력한 모델이 필요합니다. 요청의 특성에 따라 모델을 자동으로 배분하면, 비용을 크게 줄이면서 품질은 유지할 수 있습니다.

이것은 전기 요금 최적화와 같은 사고입니다. 전등에는 60W면 충분한데 1,000W 전열기를 연결하는 사람은 없지요. 마찬가지로 간단한 텍스트 분류에 수천억 파라미터 모델을 돌릴 이유가 없습니다.

그렇다면 우리 서비스에서 60W짜리 일과 1,000W짜리 일은 어떻게 가려낼까요? 출발점은 지금 들어오는 요청의 실제 모습입니다. 지난 한 달 치 로그를 꺼내 어떤 종류의 요청이 얼마나 자주 오는지 나눠 보세요. 단순한 요청과 복잡한 요청의 비율을 미리 짐작하지 말고, 범주마다 품질과 비용을 함께 재 보는 것이 좋습니다. 이 분포가 모델 배치의 밑그림이 됩니다.

그다음 요청을 세 갈래로 보냅니다. 형식 변환이나 분류처럼 가벼운 일은 소형 모델로, 중간 난이도는 중형 모델로, 긴 추론이 필요한 일만 대형 모델로 가는 길을 만드는 것이지요. 장소도 함께 정합니다. 밀리초 단위의 반응이 필요한 일은 엣지에, 깊은 분석은 클라우드에 둡니다.

장부도 따로 둡니다. 토큰당 비용은 입력 지표일 뿐이라, 요청당 비용과 성공 작업당 총비용, 검토 시간, 사고 비용까지 함께 적어야 경제성이 보입니다. 더 나은 모델이 나왔을 때 다시 평가하고 갈아탈 수 있도록 교체 기준도 남겨 둡니다.

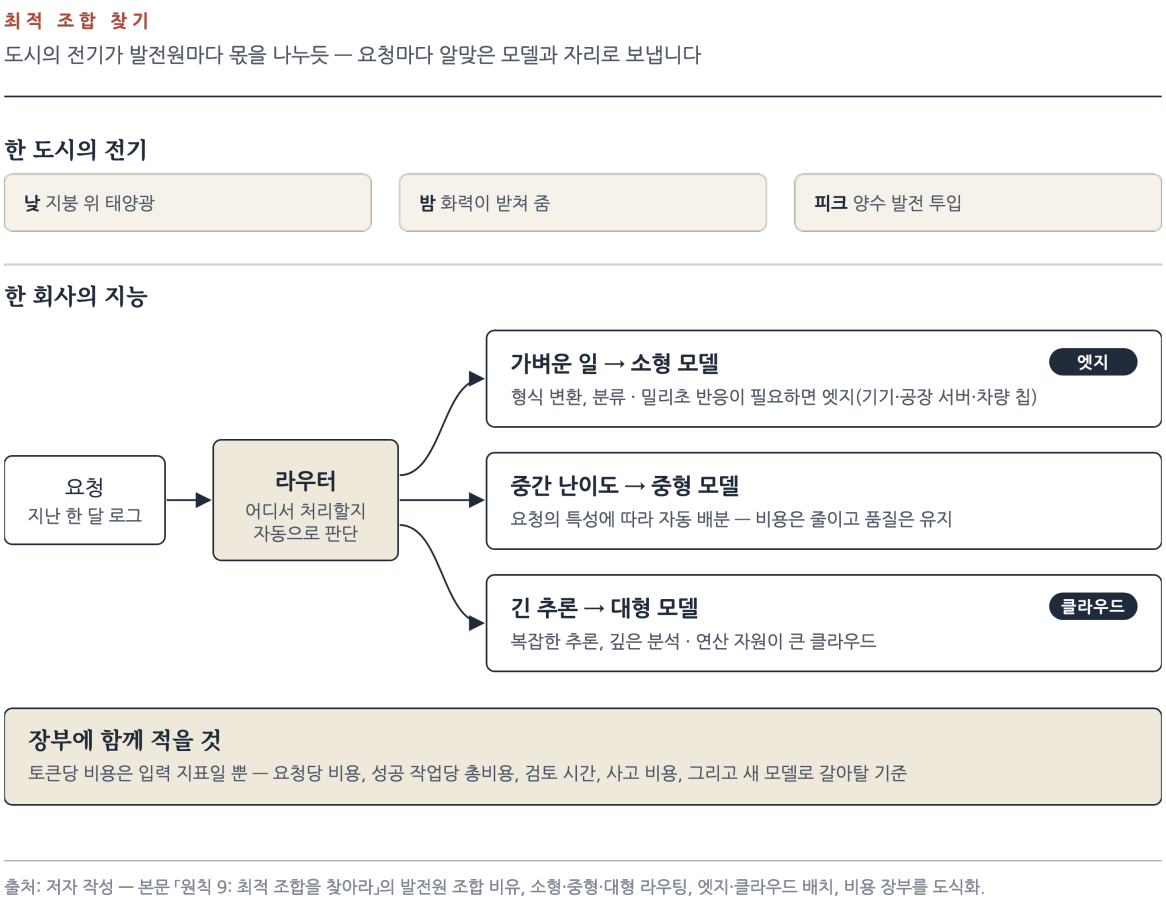


그림 4-7 낮에는 태양광, 밤에는 화력, 피크에는 양수 발전이 몫을 나누듯, 라우터는 요청마다 작은 모델과 큰 모델, 엣지와 클라우드 사이에 일을 나눠 보냅니다.

실행 체크: 최적의 조합을 쓰고 있는가

1. 실제 요청 로그를 복잡도·위험·지연 허용치·데이터 위치로 분류하고, 범주별 품질과 비용을 잽니다.
2. 소형→중형→대형으로 이어지는 라우팅과 엣지·클라우드 배치를 설계하고, 라우팅 규칙을 문서로 남깁니다.
3. 성공 작업당 총비용을 추적하고, 새 모델이 나오면 같은 기준으로 다시 평가합니다.

리스크: "무조건 가장 좋은 모델"로 몰면 비용이 새고, "무조건 가장 싼 모델"로 몰면 품질이 샙니다. 라우팅 기준은 데이터를 보며 계속 손보는 편이 안전합니다.

한 도시의 전기는 원전과 화력, 지붕 위 태양광과 양수 발전이 시간대마다 몫을 나눠 만듭니다. 내일 아침의 지능 에너지도 아마 그렇게, 작은 모델과 큰 모델이 요청마다 몫을 나눠 흐르게 될 겁니다.

호모 프롬프트: 지능과 대화하는 습관

지금까지 아홉 개의 원칙으로 조직을 다뤘으니, 이제 그 조직 안에서 일하고 배우는 사람 이야기로 넘어가 보겠습니다. AI와 대화하는 방식은 나이와 함께 타고나는 성질보다는, 누구나 익힐 수 있고 쓰지 않으면 녹스는 작업 습관에 가깝지요.

여기서는 그 습관을 호모 프롬프트(Homo Prompt)라고 부르겠습니다. 새로운 종이나 특정 세대를 가리키는 이름은 아닙니다. 외부의 지능과 대화하며 문제를 풀되 결과의 근거와 책임을 놓치지 않는 사람을 설명하는 비유이지요.

이런 사람은 검색을 버리고 대화로 옮겨 간 사람일까요? 그렇지 않습니다. 검색은 자료의 출처를 찾아 직접 비교하는 데 강합니다. 생성형 AI와의 대화는 흩어진 정보를 요약하고 초안을 만들고 관점을 바꾸는 데 편리하지요. 둘 중 하나를 버릴 이유는 없습니다.

예를 들어 "이 코드에서 버그를 찾아줘"라고 물으면 AI가 가설을 제시할 수 있습니다. 그다음에는 공식 문서와 실제 로그를 찾아 가설을 검증해야 합니다. 대화가 탐색의 시작이라면 검색과 실험은 증거를 확인하는 과정입니다.

AI에 익숙한 젊은 사람이 있을 수 있고, 오랜 현장 경험을 프롬프트에 정확히 담는 중장년이 있을 수도 있습니다. 도구를 일찍 접했다는 사실만으로 질문의 질, 출처 판단, 직업 윤리까지 갖춰지는 것은 아닙니다. 반대로 나이가 많다고 AI 활용이 느리다고 단정할 근거도 없습니다.

결국 차이를 만드는 것은 나이보다 묻는 법과 확인하는 법입니다. 좋은 프롬프트에 주문 같은 화려함은 필요 없습니다. 목표와 독자, 제약, 근거, 모르는 부분, 확인 방법을 분명히 적으면 되지요. AI에게만 통하는 기술도 아닙니다. 동료에게 일을 부탁하거나 스스로 문제를 정의할 때도 필요한 능력이지요.

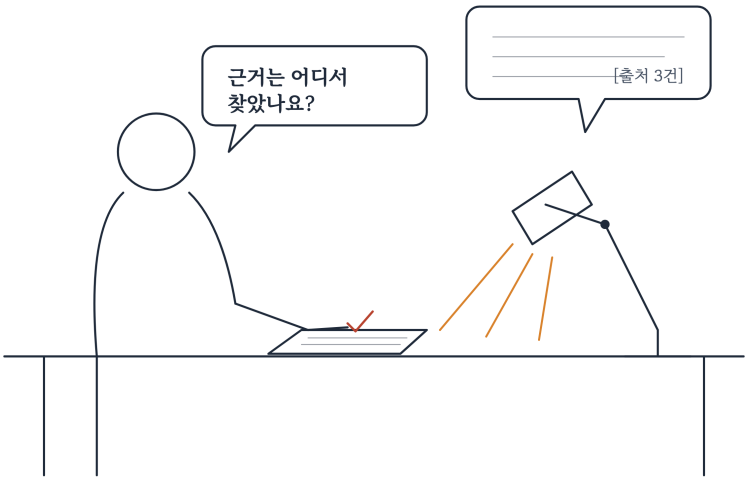


그림 4-8 답을 받아 적는 것으로 끝나지 않습니다. "근거는 어디서 찾았나요?"라고 되묻고, 확인 표시는 사람이 직접 남깁니다.

코드를 말로 만들어내는 바이브 코딩(Vibe Coding)도 마찬가지입니다. 시작 장벽을 낮춰 주지만, 동작하는 화면이 나왔다고 보안과 접근성, 유지보수까지 해결된 것은 아닙니다. 생성 속도와 검증 능력은 함께 길러야 합니다.

조직은 AI 사용 횟수를 세대 적응력의 대리 지표로 삼기보다, 어떤 문제에서 어떤 도구를 쓰고 어떻게 검증했는지를 보아야 합니다. 교육도 프롬프트 예문을 외우게 하는 데서 끝내지 말고, 근거 추적과 개인정보 보호, 오류 수정, 도구를 쓰지 않을 선택까지 다뤄야 합니다.

거창한 교육 과정을 기다릴 필요는 없습니다. 이번 주에 실제 업무 질문 하나를 AI에 물어보세요. 답에서 확인 가능한 주장과 의견을 나누고, 원문 출처를 찾아 비교한 뒤, 틀리거나 빠진 부분을 직접 고쳐 기록합니다. 어떤 단계에서 AI가 시간을 줄였고, 어느 단계에서 사람의 지식과 확인이 필요했는지를 아는 것이 호모 프롬프트의 출발점입니다.

검색창에 단어를 넣던 손이, 이제 지능에게 의도를 건넵니다. 당신은 검색하는 사람인가요, 대화하는 사람인가요? 여기까지 읽었다면 '둘 다'라는 답이 먼저 떠올랐을지도 모르겠습니다. 그런 사람 곁에 저마다의 개인 에이전트가 놓이기 시작하는 이야기는 제6부에서 이어집니다.

호모 프롬프트의 10가지 실천 원칙

AI에게 의도를 말하고, 결과를 검증하며, 필요한 지능을 끌어다 쓰는 작업 방식이 퍼지고 있습니다. 앞 장에서는 이런 습관을 가진 사람을 호모 프롬프트(Homo Prompt)라고 불렀지요.

조직의 설계 원칙은 앞서 아홉 가지로 정리했습니다. 그렇다면 개인은 무엇을 연습해야 할까요? AI와 함께 일할 때 도움이 되는 열 가지 실천을 차례로 꺼내 보겠습니다.

AI의 답 앞에서 지킬 것

1. 질문과 검증을 함께 설계하라. AI와 일할 때 중요한 능력은 좋은 질문을 던지고 답을 검증하는 것입니다.

AI는 답뿐 아니라 다음 질문도 제안할 수 있습니다. 그래도 무엇을 해결할지, 어떤 자료와 기준을 쓸지, 결과를 어디에 적용할지는 사용자가 정해야 하지요. "매출을 올리려면 어떻게 해야 하나요?"보다 "지난 분기 이탈 고객의 공통 패턴을 분석하고, 근거가 된 자료와 함께 재구매 캠페인을 제안해줘"라고 물으면 검토할 수 있는 답에 가까워집니다.

좋은 질문은 출발점입니다. 어떤 자료와 기준으로 답을 확인할지도 함께 적을 때 실제 업무로 이어집니다.

2. AI의 출력을 그대로 쓰지 마라. AI가 만든 결과물을 그대로 붙여 넣으면 자신의 맥락과 책임이 빠지기 쉽습니다. 같은 도구를 쓰더라도 자료 선택과 검증, 수정 방식에 따라 결과는 달라집니다.

AI의 출력은 초안입니다. 원석이지요. 그 원석을 당신의 경험, 판단, 맥락 이해로 다듬을 때 비로소 보석이 됩니다. "AI가 쓴 보고서"와 "AI를 활용해 내가 완성한 보고서"는 다른 것입니다.

검증하고, 수정하고, 맥락을 더하세요. 그 과정이 당신의 전문성을 증명합니다.

3. 전문성을 더 깊이 파라. AI는 보편적 지식에 접근하고 초안을 만드는 비용을 낮추고 있습니다. 이 상황에서 깊은 전문성과 여러 분야를 연결하는 통합적 사고력은 AI의 답을 평가하고 현장에 맞게 고치는 기반이 됩니다.

단편적인 지식을 전달하는 일은 자동화되기 쉬워졌습니다. 반면 복잡한 상황에서 예외를 발견하고 맥락을 고려해 판단하려면 여전히 깊은 경험이 필요합니다. 전문성은 AI와 경쟁하는 장벽이라기보다 AI의 답을 평가하고 더 나은 문제를 정의하는 기반입니다.

4. 핵심 역량과 복구 능력을 유지하라. AI에 의존하는 업무일수록 도구가 멈췄을 때 무엇을 직접 해야 하는지 알아야 합니다.

자동항법에 너무 의존한 조종사가 비상 상황에서 무력해지듯, AI에 모든 판단을 맡기다 보면 스스로의 사고 근육이 퇴화합니다. 의도적으로 AI 없이 문제를 풀어보는 연습을 해 보세요. 일주일에 한 번이라도 "AI가 고장 나면 나는 이 일을 할 수 있는가?"를 자문해 보세요.

모든 일을 예전 방식으로 되돌릴 필요는 없습니다. 다만 업무를 중단하거나 안전하게 축소하고, 핵심 판단을 이어갈 수 있는 기초 체력은 유지해야 합니다.

5. 메타인지를 훈련하라. 메타인지란 "내가 무엇을 알고 무엇을 모르는지 아는 것"입니다.

AI와 일할 때 이 능력이 결정적으로 중요해집니다. AI의 답변이 그럴듯하게 보여도, 그것이 맞는지 틀린지 판단하려면 자기 자신의 지식 범위를 정확히 알아야 합니다. "나는 이 분야를 충분히 알아서 AI의 답을 검증할 수 있다"와 "나는 이 분야를 모르니까 AI의 답을 그대로 믿으면 위험하다"를 구분하는 능력이지요.

AI가 자신 있게 틀린 답을 내놓을 때 — 이른바 환각(hallucination) — 메타인지가 없는 사람은 속습니다. 메타인지가 있는 사람은 "뭔가 이상한데?"를 감지합니다.

함께 일하며 멀리 가기

앞의 다섯 가지가 AI의 답 앞에서 내 판단을 지키는 일이었다면, 나머지 다섯은 AI와 손발을 맞추며 오래 일하는 법에 관한 것입니다.

6. 협업의 언어를 배워라. 코딩을 할 줄 알아야 한다는 말이 아닙니다. AI와 효과적으로 소통하는 방법을 익히라는 뜻입니다.

프롬프트 작성은 기술이라기보다 소통 방식에 가깝습니다. 명확한 맥락 전달, 구체적인 요구 사항 표현, 원하는 결과의 형식 지정. 이것은 사실 좋은 의사소통의 기본 원칙과 같습니다. 사람에게 업무를 지시할 때도 맥락과 기대치를 명확히 전달해야 좋은 결과가 나오듯, AI도 마찬가지입니다.

여기에 더해서 MCP, A2A 같은 에이전트 프로토콜의 기본 개념을 이해하면, AI 도구를 연결하고 조합하는 능력이 한 차원 올라갑니다.

7. 연결할 수 있는 전문성을 길러라. T자형 인재란 하나의 깊은 전문 분야와 폭넓은 일반 지식을 가진 사람입니다. 여기에 두 분야를 연결하는 경험이 더해지면 새로운 문제를 발견하기 쉬워집니다. 꼭 두 개의 학문을 같은 깊이로 통달해야 한다는 뜻은 아닙니다.

의학과 데이터 사이언스, 법률과 기술 아키텍처처럼 서로 다른 언어를 연결하는 사람은 팀의 오해와 누락을 줄일 수 있습니다. AI도 분야를 가로지르는 제안을 만들 수 있으므로 인간만의 독점 능력이라고 말할 수는 없습니다. 차이는 제안을 실제 현장과 책임 구조에 맞게 시험하고 조정하는 데서 생깁니다.

8. 역량의 선택지를 넓혀라. 하나의 도구나 좁은 업무 절차에만 경력을 걸면 변화에 취약해질 수 있습니다. 어떤 직무가 사라질지를 단정하기보다, 현재 전문성을 인접한 문제로 옮길 수 있는 선택지를 준비하는 편이 현실적입니다.

자신의 능력을 포트폴리오로 관리하세요. 현재 주력 역량과 더불어, AI가 변화시킬 영역에서의 새로운 기술을 병행해서 쌓으세요. 금융 분석가라면 AI 모델 해석 능력을 더하고, 마케터라면 데이터 기반 의사결정 역량을 키우고, 개발자라면 제품 기획이나 사업 감각을 넓히는 식입니다.

투자에서 분산 투자가 위험을 줄이듯, 역량의 분산도 경력의 위험을 줄입니다.

9. 윤리적 판단을 제도로 연결하라. "이것을 기술적으로 할 수 있는가?"와 "이것을 해야 하는가?"는 다른 질문입니다. AI는 여러 규범과 위험을 제시할 수 있지만, 어떤 가치를 우선하고 누가 결과에 책임질지는 조직과 사람이 정해야 합니다.

AI가 특정 고객군을 차별하는 결과를 내놓았을 때는 영향을 받는 사람과 분야 전문가의 관점으로 문제를 확인하고 교정해야 합니다. AI가 제안한 마케팅 전략이 법적 기준은 통과하더라도 조직의 원칙과 고객의 권리를 해친다면 중단할 책임이 있습니다.

윤리적 판단은 개인의 양심만이 아니라 제도와 함께 작동해야 합니다. 이의를 제기할 통로, 고위험 결정을 멈출 권한, 영향을 받는 사람의 재검토 권리를 업무 절차에 넣어야 판단이 실제 행동으로 이어집니다.

10. 자기 속도로 계속 배워라. 마지막 실천은 마음가짐에 관한 것입니다. 변화의 속도가 빠르다고 해서 조급해할 필요는 없습니다. 새로운 AI 도구가 계속 나온다고 해서 전부 따라잡을 필요도 없습니다.

중요한 것은 학습을 멈추지 않는 것입니다. 모든 것을 다 알아야 한다는 강박을 버리고, 자신의 분야에서 AI가 어떻게 쓰이는지 꾸준히 관찰하고 실험하세요.

도구의 유행을 모두 따라가는 것보다, 자신의 업무에서 유용성과 위험을 꾸준히 시험하는 편이 오래 갑니다.

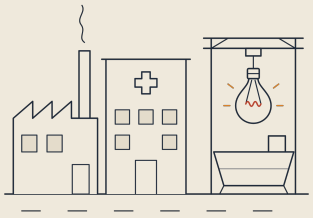


열 가지를 다 외우기 어렵다면 이것 하나만 기억해 두세요. AI를 부리되, 묻고 확인하고 책임지고 되돌리는 힘은 내 손에 남겨 두는 것.

제4부는 1890년대 공장에서 시작했습니다. 전기 모터를 사 놓고도 생산성이 크게 오르기까지 한 세대가 걸렸지요. 결국 배치도를 다시 그린 것은 그 공장에서 일하며 새 동력에 손을 익힌 사람들이었습니다. 열 가지 실천은 지금 우리가 그 사람들의 자리에 서는 연습입니다.

개인의 실천 다음은 산업의 적용 방식입니다. 같은 전기라도 반도체 공장에 필요한 전압과 가정집의 전압이 다르듯, 산업마다 필요한 지능의 세기와 결도 다르지요. 제5부는 업종마다 이 지능 에너지가 어떻게 다르게 흐르는지를 다룹니다.





제5부

산업별 지능 에너지 혁명

코드를 쓰는 AI, 코드를 검증하는 인간

요즘 개발자에게 "AI 쓰세요?"라고 물으면 대답은 거의 정해져 있습니다. 2025년 Stack Overflow 개발자 설문에서 응답자의 84%가 AI 도구를 쓰고 있거나 쓸 계획이라고 답했으니까요. 그런데 같은 설문에서 AI가 내놓은 결과를 믿는다는 사람보다 못 믿겠다는 사람이 더 많았습니다.[1]

쓰기는 다 쓰는데 믿지는 않는다. 이상한 풍경 같지만, 바로 여기서 새로운 분업이 생겨나고 있습니다. AI가 만들고, 사람이 확인하는 분업이지요.

그 분업은 하루아침에 생기지 않았습니다. 소프트웨어 개발의 AX(AI Transformation)는 세 단계를 밟아 왔습니다. 처음은 AI가 옆자리에 앉아 거드는 Enabled 단계였지요. 이 단계를 대표하는 것은 GitHub Copilot입니다. 2021년 기술 미리보기로 출발한 Copilot은 코드 자동완성에서 채팅과 에이전트 기능으로 범위를 넓혀왔습니다.[2] 개발자가 코드를 타이핑하면 AI가 다음 줄을 제안하는 방식이지요.

이 단계에서 AI는 문자 그대로 보조 도구였습니다. 개발자가 주도하고, AI는 자동완성을 돕는 역할이었지요. 기업들은 반복 입력을 줄이고 개발 흐름을 끊지 않는 생산성 효과에 주목하기 시작했습니다.

그런데 옆자리 조수가 내 프로젝트 전체를 읽을 수 있게 된다면 어떨까요? First 단계에서는 Cursor 같은 AI 네이티브 IDE가 등장합니다. 단순한 코드 완성을 넘어, 프로젝트 전체의 맥락을 이해하고 인라인으로 편집하는 수준까지 올라왔지요.

이 시기에 바이브 코딩(vibe coding)이라는 새로운 개발 방식이 대두되었습니다. 개발자가 원하는 기능을 자연어로 설명하면 AI가 코드를 생성하는 방식입니다. "이런 느낌으로 만들어줘"라고 말하면 AI가 구현하는 셈이지요. 코딩의 진입 장벽이 극적으로 낮아졌고, 비전문가도 아이디어를 소프트웨어로 구현하는 시민 개발(Citizen Development) 흐름이 가속화되었습니다.

AI 코드 리뷰도 이 단계에서 본격화됩니다. GitHub는 Copilot 코드 리뷰를 정식 기능으로 제공하고, 사람이 검토해야 할 오류가 있을 수 있음을 문서에서 함께 경고합니다.[3] AI가 코드를 쓰고, 또 다른 AI가 그 코드를 검토하는 구조가 만들어지기 시작한 것입니다.

마지막 Native 단계에 이르면 AI는 아예 프로젝트를 통째로 맡습니다. 그 맨 앞에는 Claude Code가 있습니다. Anthropic은 2025년 5월 일반 공개 이후 기업용 배포와 보안 기능을 빠르게 확장해왔습니다.[4]

Claude Code가 이전 도구들과 근본적으로 다른 점은 위임형(agentic) 방식이라는 것입니다. 터미널에서 복잡한 태스크를 자율적으로 수행합니다. 코드를 읽고, 수정하고, 테스트하고, 커밋하는 전 과정을 스스로 판단하며 실행하지요. 멀티에이전트 코드 리뷰 기능까지 갖추고 있어, 설계 에이전트와 검증 에이전트가 SCRUM 팀처럼 협업하는 구조도 가능해졌습니다.

그렇다면 이 세 단계를 지나는 동안 인간은 어디에 서 있게 되었을까요? 코드 한 줄을 만드는 비용이 떨어지자, 제1부에서 본 제번스 역설이 소프트웨어에서도 나타날 조짐이 보입니다. 값이 싸지니 코드가 덜 필요해지기는커녕 더 많이 필요해지는 것이지요. 비용과 인력 문제로 소프트웨어 도입을 주저하던 소규모 비즈니스와 개인까지 시장에 진입하고 있습니다.

METR의 연구에 따르면 AI가 50% 확률로 끝낼 수 있는 소프트웨어 작업의 시간 지평선은 2019년 이후 약 6개월마다 두 배로 늘었습니다. 2023년 이후만 보면 4개월 남짓이지요. Claude Opus 4.5의 지평선은 약 290분, 5시간 가까이였습니다. 몇 달 뒤 Opus 4.6은 약 12시간, Claude Mythos Preview는 16시간을 넘었지요. 그런데 METR은 16시간을 넘는 값은 지금의 과제 묶음으로 믿을 만하게 잴 수 없다고 선을 그었고, 9월부터는 이 측정표를 더 갱신하지 않고 있습니다. 물론 이것은 정해진 채점 기준이 있는 작업의 통계적 측정치이지, 어떤 열두 시간짜리 프로젝트든 맡길 수 있다는 뜻은 아닙니다.[5]

그렇다고 개발자가 불필요해지는 것은 아닙니다. 오히려 역할이 변합니다. 코드 작성에서 문제 정의, 요구사항 구체화, 아키텍처 설계, 최종 품질 책임으로 무게중심이 이동하는 것이지요. 전통적인 코딩 능력만으로는 부족하고, AI 활용 역량과 도메인 지식, 시스템적 사고가 더욱 중요해집니다.

그러니 사람에게 남는 몫은 둘입니다. 무엇을 만들지 정하는 일, 그리고 AI가 만든 결과물이 정말 그 목적에 맞는지 확인하는 일이지요. 코드를 쓰는 AI, 코드를 검증하는 인간.

그런데 사람은 정말 잘 확인하고 있을까요? 숫자를 들여다보면 조금 불안해집니다.



---

1. "2025 Developer Survey: AI" (Stack Overflow, https://survey.stackoverflow.co/2025/ai) — AI 도구 사용·사용 계획 84%, 신뢰보다 불신 응답이 더 높은 결과.

2. "GitHub Copilot: The agent awakens" (GitHub, 2025-05-19, https://github.blog/news-insights/product-news/github-copilot-the-agent-awakens/); "About GitHub Copilot" (GitHub Docs, https://docs.github.com/en/copilot/about-github-copilot/what-is-github-copilot)

3. "About GitHub Copilot code review" (GitHub Docs, https://docs.github.com/en/copilot/concepts/agents/code-review) — 생성된 리뷰에는 오류가 있을 수 있으며 사람의 검토를 보완하는 용도라고 명시.

4. "Claude Code: Generally available" (Anthropic, 2025-05-22, https://www.anthropic.com/news/claude-code-generally-available); "Claude Code overview" (Anthropic Docs, https://docs.anthropic.com/en/docs/claude-code/overview)

5. "Time Horizon 1.1" (METR, 2026-01-29, https://metr.org/blog/2026-01-29-time-horizon-1-1/); 측정 결과 원자료 (https://metr.org/time-horizons/, 2026-10-22 확인) — 더블링 시간 전체 기간 187.8일, 2023년 이후 128.7일. Opus 4.5 293분, Opus 4.6 약 719분, Mythos Preview 초기판 약 1,045분. 16시간 초과 측정은 신뢰하기 어렵다는 단서, 2026.9.8 이후 갱신 중단.

테스트와 리뷰를 잘하는 법: 검증 격차를 메우기

개발자 1,100명 이상에게 물었습니다. "AI가 짠 코드를 믿으십니까?" Sonar의 "State of Code" 2026 보고서에서 96%가 완전히 믿지는 않는다고 답했지요. 그럼 커밋하기 전에 꼭 확인하느냐고 다시 물었더니, 항상 그렇다는 응답은 48%에 그쳤습니다.[1]

못 믿겠다면서 절반은 확인을 건너뛰는 셈입니다. 이 어긋남을 검증 격차(verification gap)라고 부릅니다. 코드를 쓰는 일은 빨라졌는데, 병목은 확인하는 쪽으로 옮겨 가고 있지요. 숫자부터 따라가 보겠습니다.

믿지 않으면서 확인도 건너뛰는 이유

Sonar 보고서에 따르면 2026년 현재 커밋되는 코드의 42%가 AI 손을 거칩니다. 2027년에는 65%까지 늘어날 것으로 전망되고요.[1] 문제는 그 코드가 실제로 얼마나 안전한가입니다.

CodeRabbit이 오픈소스 PR 470건(AI 협업 320건, 인간 전용 150건)을 뜯어본 결과는 꽤 노골적입니다. PR 하나당 발견되는 이슈가 AI 협업 PR에서 10.83건, 인간 전용 PR에서 6.45건 — 약 1.7배 차이입니다. 로직·정확성 오류는 75% 더 많았고, 보안 이슈는 최대 2.74배였습니다.[2] AI가 특별히 새로운 종류의 버그를 만드는 것은 아닙니다. 같은 유형의 실수를, 더 자주 저지를 뿐이지요.

그런데 정작 리뷰에 들어가는 시간은 오히려 늘었습니다. LinearB가 4,800개 이상 조직의 PR 810만 건을 분석했더니, AI 도입 후 개발자들은 21% 더 많은 작업을 완료하고 PR 머지율은 98% 뛰었습니다. 반면 PR 리뷰 시간은 91% 늘었습니다.[3] 체감과 실측도 엇갈렸습니다. 개발자들은 스스로 "20% 더 빨라졌다"고 느꼈지만, 실제 측정치는 "19% 더 느려졌다"였습니다. 코드를 쓰는 시간은 줄었는데, 검증하는 시간이 그보다 더 늘어난 것이지요. LinearB는 이를 "생산성 역설"이라 부릅니다.

더 불편한 발견은 스탠퍼드 Dan Boneh 연구팀에서 나왔습니다. AI 도구를 쓰면서 오히려 더 불안전한 코드를 작성한 그룹이, 안전한 코드를 작성한 그룹보다 자기 코드에 대한 신뢰도를 더 높게 매겼습니다.[4] 틀린 답을 더 확신하는 셈입니다. 부주의라기보다 "거짓 안도감"에 가깝지요. AI가 그럴듯하게 완성된 결과물을 내놓으면, 사람은 그 완성도를 검증의 증거로 착각합니다. 코드가 매끈해 보인다는 것과 코드가 맞다는 것은 전혀 다른 이야기인데 말이지요.

그럼 리뷰를 더 열심히 하면 검증 격차가 메워질까요? 그럴 것 같지만, 인간의 주의력은 그렇게 단순하게 작동하지 않습니다.

자동화 신뢰도와 오류 탐지율의 관계를 다룬 연구가 있습니다. 자동화 시스템이 꾸준히 신뢰할 만하게 작동할 때, 사람은 그 시스템의 오류를 약 30%밖에 잡아내지 못했습니다. 반대로 시스템이 가끔이라도

눈에 띄게 실패하는 모습을 보이면 탐지율은 75%까지 올라갔습니다.[5] 역설적이게도, AI가 잘할수록 사람의 감시는 느슨해집니다.

이 현상을 극적으로 보여주는 실험이 있습니다. 방사선과 전문의들에게 폐결절을 찾는 CT 스캔을 판독하게 하면서, 결절보다 48배나 큰 고릴라 이미지를 슬쩍 끼워 넣었습니다. 전문의의 83%가 고릴라를 발견하지 못했습니다. 안구 추적 장비로 확인해보니 절반 이상은 고릴라가 있는 자리를 눈으로 직접 보고도 인지하지 못했지요.[5] 익숙한 작업을 반복하는 전문가일수록, 예상 범위 밖의 이상 신호를 놓치기 쉽다는 뜻입니다. 코드 리뷰도 다르지 않습니다. "AI가 짜준 코드는 대체로 무난하다"는 경험이 쌓일수록, 정작 위험한 한 줄을 훑고 지나갈 가능성이 커집니다.

인간 요인 연구자 미카 엔슬리(Mica Endsley)는 이를 "자동화의 역설"이라는 말로 정리했습니다. 자동화가 더해지고 더 신뢰할 만해질수록, 그것을 감시하는 인간이 정작 중요한 정보를 인지하고 필요할 때 수동으로 개입할 가능성은 오히려 낮아진다는 것입니다.[5] AI 코드 리뷰도 같은 함정에 빠지기 쉽습니다.

그렇다면 대응은 무엇일까요. 리뷰를 오래 붙잡고 있는 것이 능사는 아닙니다. 오히려 25~30분마다 짧게 쉬는 것, 그리고 거대한 diff 하나를 통으로 던지지 말고 작은 단위로 쪼개서 리뷰하는 것이 효과적입니다. 집중력이 떨어진 상태에서 억지로 붙들고 있는 리뷰보다, 짧고 자주 하는 리뷰가 고릴라를 찾을 확률을 높입니다.

증거 없는 PR은 일을 미루는 것이다

그럼 무엇을 바꿔야 할까요. 브라우저 확장 프로그램 개발자 출신으로 지금은 코드 리뷰 방법론을 연구하는 애디 오스마니(Addy Osmani)는 "PR 계약(PR Contract)"이라는 틀을 제안합니다.

핵심은 이렇습니다. PR을 올리는 사람이 리뷰어에게 최소한의 신뢰 기반을 제공해야 한다는 것이지요. 오스마니는 이렇게 말합니다. "PR에 작동한다는 증거가 없다면, 그건 더 빠르게 배포하는 게 아니라 일을 나중으로 미루는 것이다."[6] 겉보기엔 빨라 보여도, 검증 없는 PR은 그 검증 부담을 리뷰어에게, 혹은 미래의 자신에게 떠넘기는 것일 뿐입니다.

PR 계약에 담을 내용은 네 가지입니다. 무엇을 왜 바꿨는지 한두 문장으로 적은 의도, 통과한 테스트나 직접 확인한 스크린샷·로그 같은 증거, 결제처럼 민감한 영역을 건드렸는지와 AI가 얼마나 관여했는지를 밝힌 위험 표시, 그리고 아키텍처 결정처럼 사람이 꼭 봐야 할 지점을 짚은 메모이지요.[6] 리뷰어 입장에서는 무엇을 의심해야 할지 지도를 한 장 받는 셈입니다.

여기서 자연스럽게 이어지는 원칙이 검증 우선(verification-first)입니다. 구현부터 시키고 나중에 테스트를 붙이는 순서를 뒤집어, 사양을 먼저 문서로 확정하고, 그 사양에 합의한 뒤에 구현과 테스트를 진행하는 흐름입니다. spec.md를 작성하고 승인받은 다음 구현하고, 그 구현을 다시 스펙에 비추어 검증하는 순환이지요.[6] AI에게 "일단 짜봐"라고 던지는 대신, "이 조건을 만족하는 코드를 짜라"고 먼저 합의하는 것이 검증의 절반을 미리 끝내는 셈입니다.

리뷰의 시선도 바뀌어야 합니다. 지금까지 리뷰는 diff, 즉 바뀐 코드 한 줄 한 줄을 읽는 행위였습니다. 그런데 AI가 짠 코드는 양이 많고 스타일이 균일해서, 줄 단위로 읽어봐야 리뷰어의 눈에는 다 비슷해 보입니다. 그래서 나오는 대안이 의도 리뷰입니다. diff 자체보다, 그 diff가 원래 의도했던 스펙과 얼마나 일치하는가를 먼저 확인하는 방식이지요. 한국의 개발 커뮤니티 flowkater.io에서도 이 논쟁이 세 갈래로 갈렸습니다. 리뷰는 여전히 필수라는 입장(Kent Beck·오스마니 계열), 전통적 리뷰는 이미 무의미해졌다는 입장, 그리고 diff 대신 스펙을 리뷰해야 한다는 절충안(Jain)이 공존합니다. 어느 쪽이든 공통된 결론은, 예전 방식 그대로 diff만 읽는 리뷰로는 부족하다는 것입니다.

리뷰를 한 사람(혹은 한 AI)에게 몰아주지 않는 방법도 있습니다. 보안, 성능, 로직 정확성은 각각 요구하는 시선이 다릅니다. 보안 전문 에이전트, 성능 전문 에이전트, 정확성 전문 에이전트로 역할을 나눠서 같은 PR을 서로 다른 각도로 훑게 하면, 한 명이 모든 걸 다 보려다 놓치는 사각지대가 줄어듭니다. 사람이 모든 리뷰를 떠안는 대신, 전문화된 검증 파이프라인을 설계하는 쪽으로 무게중심이 옮겨가고 있는 셈입니다.

한 걸음 물러서서 보면 이 모든 처방이 같은 방향을 가리킵니다. AI가 코드를 쓰는 비용은 계속 떨어지고 있습니다. 그런데 그 코드가 정말 맞는지 확인하는 일은 여전히, 아니 오히려 더 사람의 시간과 주의력을 요구합니다. 생성이 값싸질수록 검증은 상대적으로 더 귀해집니다.

못 믿겠다면서 확인은 건너뛰는 버릇. 이 장의 숫자들은 결국 그 틈을 가리킵니다. "AI가 썼으니 대충 봐도 되겠지" 하는 순간이 가장 위험하지요.

내일 아침 팀에서 바로 해 볼 만한 일은 세 가지쯤 됩니다. 구현을 맡기기 전에 테스트부터 적어 두고, PR마다 의도와 증거를 붙이게 하고, 보안·성능·정확성 리뷰를 서로 다른 눈에 나눠 맡기는 것이지요. 그래도 리뷰 피로와 거짓 안도감은 언제든 다시 스며듭니다. 검증 격차는 한 번 메우고 끝나는 구멍이 아니니까요.



---

1. "Sonar Data Reveals Critical Verification Gap in AI Coding" (Sonar, 2026-01-08, https://www.sonarsource.com/company/press-releases/sonar-data-reveals-critical-verification-gap-in-ai-coding/) — 개발자 1,100명 이상 조사

2. "State of AI vs Human Code Generation" (CodeRabbit, https://www.coderabbit.ai/blog/state-of-ai-vs-human-code-generation-report) — 오픈소스 PR 470건 분석

3. "The Review Bottleneck: Why More AI Code Means Slower Teams in 2026" (dev.to/code-board, https://dev.to/code-board/the-review-bottleneck-why-more-ai-code-means-slower-teams-in-2026-1e5n) — LinearB 4,800+ 조직·PR 810만 건 분석

4. Dan Boneh 등, "관련 연구팀, AI에 의존할수록 코드가 더 버그투성이가 될 가능성 발견" (Stanford Engineering, https://ee.stanford.edu/dan-boneh-and-team-find-relying-ai-more-likely-make-your-code-buggier)

5. "AI Code Review Fatigue" (Atomic Robot, https://www.atomicrobot.com/blog/ai-review-fatigue) — McBride·Rogers·Fisk(2014) 자동화 신뢰도 연구, 방사선과 고릴라 실험(2013), Mica Endsley 자동화 역설

6. Addy Osmani, "Code Review in the Age of AI: The PR Contract" (Addy Osmani's Substack, https://addyo.substack.com/p/code-review-in-the-age-of-ai)

AI를 동료로 맞이한 의사들

진료실을 떠올려 보세요. 의사는 환자 얼굴보다 모니터를 더 오래 봅니다. 증상을 듣는 동안에도 손은 키보드 위에 있고, 진료가 끝나면 또 한참 기록을 정리하지요. AI가 이 진료실을 어떻게 바꿔 왔을까요?

의료의 AX는 다른 산업과 결이 조금 다릅니다. 사람의 생명이 걸려 있으니 얼마나 빨리 바뀌느냐보다 얼마나 믿을 수 있느냐가 먼저지요. 그래서 의료 AI는 느리지만 한 걸음씩 나아갔습니다.

첫걸음인 Enabled 단계에서 AI는 의사의 눈을 보완하는 역할이었습니다. X-ray 판독 보조, CT 영상에서 종양 의심 부위 표시, 피부 병변 분류 같은 것들이지요. 기존 의료 정보 시스템에 AI 기능을 사후적으로 추가하는 방식이었습니다.

이 단계의 한계는 분명했습니다. 전통적인 전자의무기록(EHR) 시스템은 행정 업무 중심으로 설계되어 있었고, AI를 위에 얹어봐야 사용자 인터페이스는 여전히 복잡했습니다. "20년 된 시스템에 AI를 추가하는 것은 덧붙이기에 불과하다"는 비판이 나온 것도 이 시기입니다.

덧붙이기가 한계라면 처음부터 다시 짜면 되지 않을까요? First 단계에서는 바로 그렇게, AI를 전제로 처음부터 설계된 시스템이 등장합니다.

Abridge로 대표되는 AI 의료 차팅 기술이 가장 눈에 띄는 변화입니다. 진료실에서 환자와 의사의 대화를 AI가 실시간으로 청취하고, 맥락을 파악하여 진료 노트를 자동으로 생성합니다. 의사가 차트 앞에서 키보드를 두드리는 대신 환자와 눈을 맞추며 대화할 수 있게 된 것이지요. 대규모 도입 사례를 정리한 보고서에서는 15개월간 약 16,000시간의 문서 작성 시간을 절감했다고 전합니다. 다만 이는 특정 의료 시스템의 도입 보고이므로, 다른 병원에서도 같은 효과가 난다고 단정할 수는 없습니다.[1]

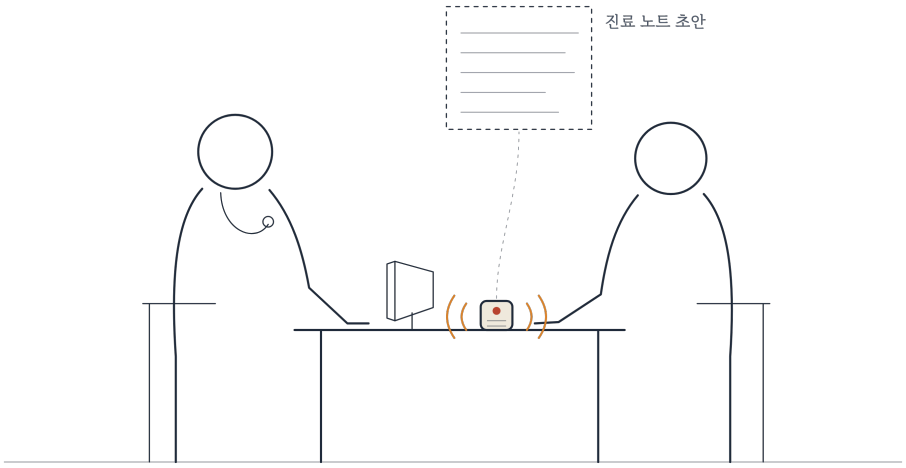


그림 5-1 의사는 모니터를 밀어 두고 환자 쪽으로 몸을 돌렸습니다. 두 사람의 대화는 책상 위 작은 기기가 듣고, 진료 노트 초안은 그사이 저절로 채워지지요.

Oracle은 차세대 의료 정보 시스템에서 보이스 퍼스트(Voice-First) 인터페이스와 임상 업무 자동화를 내세웠습니다.[2] athenahealth도 2025년 'Clinical Encounter'를 발표하며 진료 대화를 토대로 문서 초안을 만드는 맥락 인지 자동화를 전면에 내세웠지요.[3]

한 1차 진료 현장의 의사는 이렇게 말했습니다. "EHR이 더 이상 청구를 위한 기록 시스템이 아니라 진료 파트너가 된 것 같다." McKinsey 보고서 역시 핵심 업무를 AI 네이티브 워크플로우로 재설계해야 진정한 혁신을 달성할 수 있다고 지적합니다.

병원 전체가 하나의 지능 시스템이 될 때

마지막 Native 단계는 병원 전체가 하나의 지능 시스템으로 작동하는 것입니다.

중국 칭화대 연구진이 2024년 공개한 Agent Hospital은 현실 병원이 아니라 가상 환자를 활용한 시뮬레이션입니다. LLM으로 구동되는 가상 의사와 간호사가 여러 진료 단계를 연습하고, 의학 지식문제로 성능을 평가했지요. 만여 건을 처리했다는 숫자도 실제 환자를 진단하거나 치료한 실적이 아닙니다. 이 연구가 보여준 것은 임상 능력이 아니라, 여러 역할의 에이전트를 한 업무 흐름에 배치해 시험할 수 있다는 가능성입니다.[4]

현실 병원에서는 훨씬 좁은 범위부터 검증해야 합니다. 문서 초안, 예약, 청구처럼 되돌릴 수 있는 행정 업무와 진단·처방처럼 환자 안전에 직접 영향을 주는 업무는 같은 자동화 단계에 놓을 수 없습니다. 후자에는 임상시험, 규제 심사, 의료진의 최종 판단과 사고 대응 체계가 필요합니다.

규제 당국도 이제 막 기준을 짜는 중입니다. 미국 FDA는 2026년 8월 생성형 AI 의료기기를 어떻게 심사할지 의견을 묻는 토론 문서를 내놓았습니다. 정보를 보여 주기만 하는 기능에서 스스로 행동하는 기능으로 갈수록, 틀린 출력을 믿었을 때의 결과가 클수록 더 엄격하게 보겠다는 방향이지요.[5]

디지털 트윈을 활용한 병원 운영 최적화도 주목할 만합니다. Children's Mercy Kansas City 병원은 디지털 트윈으로 소아호흡기 질환의 계절적 급증을 예측하고, 응급실이 폭증하기 전에 병상을 미리 확보하는 선제적 대응을 구현했습니다.

자가학습 의료기기도 AI 네이티브 의료의 중요한 축입니다. 현재의 연속혈당측정기(CGM) 결합 인슐린 펌프는 개개인의 당대사 패턴을 학습하여 24시간 최적 용량을 투여합니다. 기기 자체가 환자로부터 배우고 진화하는 것이지요.

기기와 병원이 이렇게 배우기 시작하면, 그 한가운데 선 의사의 역할도 달라집니다. 반복적인 기록 작업과 행정 업무에서 해방되어, 환자와의 소통과 복잡한 의사결정에 집중할 수 있게 됩니다.

의료계에는 이런 말이 널리 퍼져 있습니다. "AI는 의사를 대체하지 않는다. AI를 쓰는 의사가 쓰지 않는 의사를 대체할 것이다." 귀에 쏙 들어오는 말이지만, 병원 경영진이 실제로 따지는 것은 좀 더 수수합니다. Stanford 의료원의 Nigam Shah 박사는 AI가 총소유비용(TCO)을 정말 줄여 주는지부터 보라고 말하지요. 멋진 가능성보다 클릭 수, 사람의 실수, 환자 한 명에게 쓸 수 있는 시간이 얼마나 달라졌는지가 성적표라는 겁니다.

의료 AX가 잘됐는지는 진료실에서 알 수 있습니다. 진료 기록 쓰는 시간이 줄었는가, 환자와 눈을 맞추는 시간이 늘었는가, 오류가 났을 때 누가 바로잡는가. 처음 떠올린 그 진료실에서 의사가 모니터 대신 환자를 보고 있다면, AI는 제 몫을 하고 있는 셈이지요.



1. McKinsey & Company, "The coming evolution of healthcare AI toward a modular architecture" (https://www.mckinsey.com/~/media/mckinsey/industries/healthcare%20systems%20and%20services/our%20insights/the%20coming%20evolution%20of%20healthcare%20ai%20toward%20a%20modular%20architecture/the-coming-evolution-of-healthcare-ai-toward-a-modular-architecture_final.pdf)

2. Oracle Health, "Innovation in healthcare" (https://www.oracle.com/health/innovation/) - 음성 인터페이스와 임상 업무 자동화 방향을 설명한 공식 페이지.

3. athenahealth, "AI-native Clinical Encounter Transforms the EHR" (2025, https://www.athenahealth.com/press-releases/ai-native-clinical-encounter-transforms-ehr)

4. Tsinghua University Institute for AI Industry Research, "Agent Hospital" (https://air.tsinghua.edu.cn/en/info/1007/1872.htm); "Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents" (arXiv:2405.02957, https://arxiv.org/abs/2405.02957)

5. FDA CDRH, "Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback" (2026.8.18; 요약 Covington, 2026.8.24, https://www.cov.com/-/media/files/corporate/publications/2026/08/fda-releases-highly-anticipated-discussion-paper-august-2026.pdf) — 지침이 아닌 의견 수렴용 문서, 의견 마감 2026.10.19.

불량률 0%를 향한 공장

독일 지멘스의 암베르크 전자부품 공장은 약 99.9989%의 품질 수준을 기록한 것으로 소개됩니다.[1] 물론 이 성과를 AI 하나의 효과로 돌릴 수는 없습니다. 오랜 공정 표준화, 자동화 설비, 센서 데이터, 품질 관리와 작업자의 개선 활동이 함께 만든 결과이기 때문입니다. AI는 이 체계 안에서 이상을 더 빨리 찾고 조건을 조정하는 수단으로 들어갑니다.

제조업의 AX는 물리적 세계와 디지털 지능이 맞물리는 과정입니다. "지능은 에너지다"라는 비유가 가장 손에 잡히는 분야이기도 하지요. 공장은 원래 전기를 받아 동력으로 바꾸던 곳이니까요. 이제 그 공장에 두 번째 에너지, 지능이 들어오고 있습니다.

그 지능이 처음 맡은 일은 고장 나기 전에 알아채는 것이었습니다. Enabled 단계의 출발점은 예지보전(Predictive Maintenance)입니다. 설비에 센서를 부착하고, 그 데이터를 AI가 분석하여 고장 징후를 미리 탐지하는 방식이지요.

한국의 LG전자 창원 스마트파크가 좋은 사례입니다. LG전자는 디지털 트윈, 빅데이터, 예지보전과 지능형 물류를 결합한 운영 체계를 공개했고, 세계경제포럼은 이 공장을 등대공장으로 선정했습니다.[2] 기업이 공개한 효과 수치는 적용 범위와 산정 기준을 함께 봐야 하지요.

이 단계에서 AI는 사람 대신 설비 상태를 모니터링하고 품질을 판단합니다. 초기지만 중요한 변화였습니다. 하지만 여전히 기존 공장 시스템에 AI를 사후적으로 추가하는 방식이었습니다.

그렇다면 공장을 통째로 컴퓨터 안에 옮겨 놓고 미리 돌려 보면 어떨까요? First 단계에서는 디지털 트윈(Digital Twin)이 핵심 기술로 등장합니다. 물리적 공장의 가상 복제본을 만들고, 그 위에서 시뮬레이션하고 최적화하는 접근이지요.

NVIDIA의 Omniverse를 활용한 BMW의 가상 공장 설계도 주목할 만합니다. BMW는 실제 설비를 옮기기 전에 디지털 환경에서 공정과 작업 동선을 검토하고, 로봇과 비전 시스템을 위한 합성 이미지를 수천 장 생성하는 방식을 공개했습니다.[3] 실제 불량품이 충분히 쌓이기를 기다리지 않고 가상 환경에서 검사 AI를 훈련할 수 있다는 뜻이지요. 물론 시뮬레이션이 실제 공장의 품질 개선으로 이어졌는지는 운영 지표로 따로 따져 볼 일입니다.

삼성전자는 2025년 NVIDIA와 협력해 향후 수년간 GPU 5만 개 이상을 도입하는 반도체 AI 팩토리 구축 계획을 발표했습니다.[4] 이미 완성된 공장의 성과가 아니라, 제조와 설계 전반에 AI를 확대하려는 투자 계획입니다.

스스로 돌아가는 공장과 0%라는 목표

궁극적인 Native 단계는 공장 자체가 하나의 자율 지능 시스템으로 작동하는 것입니다. 수많은 전문 AI 에이전트들이 협업하며 공장 전체를 운영하는 구조이지요.

공정 조건을 실시간 최적화하는 에이전트, 설비 상태를 모니터링하고 수리를 지시하는 에이전트, 품질을 검사하는 에이전트, 물류와 재고를 조율하는 에이전트가 서로 통신하며 자율적으로 운영합니다. 현재는 부분적으로 구현된 수준이지만, 목표는 사람 개입 없이도 목표 생산성과 품질을 달성하는 것입니다.

Tesla Optimus 같은 휴머노이드도 이 방향을 시험하고 있습니다. Tesla는 생산량과 가격 목표, 자연어 명령과 관찰 학습을 결합한 시연을 공개했습니다. 그러나 목표와 시연은 납품 실적이나 공장의 반복 작업 성능이 아닙니다. 현재는 학습과 데이터 수집 단계라는 회사 설명에 맞춰, 사람의 개입 횟수와 안전 기록을 먼저 확인해야 합니다.

피지컬 AI(Physical AI)는 소프트웨어의 판단을 로봇 동작과 연결하려는 흐름입니다. Figure 같은 휴머노이드가 언어 모델과 모터 제어를 결합한 시연을 내놓고 있지요. 투자액과 데모 영상보다는 제한된 현장에서 같은 일을 얼마나 안전하게 반복하는지가 상용화의 근거가 됩니다. 휴머노이드가 실제로 공장에 들어가는 이야기는 제6부 「로봇이 출근하는 날」에서 이어집니다.

이렇게 공장이 스스로 돌아가게 되면 불량률 0%도 손에 잡힐까요? 불량률 0%는 방향을 제시하는 목표이지, AI 도입 효과를 판정하는 단일 숫자가 아닙니다. 업종마다 치명 결함과 경미한 결함의 비용이 다르고, 검사를 강화하면 생산 속도와 비용도 달라집니다. 따라서 불량률과 함께 재작업률, 검사 누락, 라인 정지 시간, 현장 안전을 같이 봐야 합니다.

결국 제조 AI는 폐쇄 루프(closed loop) 이야기입니다. 데이터를 모으고, AI가 읽고, 공정을 고치고, 다시 데이터를 모으지요. 이 순환을 사람이 믿고 맡길 만큼 촘촘하게 돌리는 공장이 AI 네이티브 공장입니다. 암베르크 공장이 오랜 세월 해 온 일도, 따지고 보면 이 루프를 조금씩 더 촘촘하게 만든 것이었지요.



---

1. Siemens, "Smart manufacturing for electronics" (https://www.plm.automation.siemens.com/media/global/it/Siemens-PLM-Smart-manufacturing-for-electronics-wp_tcm56-57766.pdf) - 암베르크 공장의 품질 수준을 소개한 기업 자료.

2. LG Electronics, "LG Smart Park Selected as WEF Lighthouse Factory" (2022, https://www.lg.com/us/press-release/LG-WEF-Lighthouse-Factory-LGEUS-033122.pdf)

3. NVIDIA, "BMW Group Develops a New Factory in NVIDIA Omniverse" (https://www.nvidia.com/en-us/case-studies/bmw-group-develop/)

4. 삼성전자 뉴스룸, "삼성전자, 엔비디아와 함께 '업계 최고 수준 반도체 AI 팩토리' 구축" (2025.10.31, https://news.samsung.com/kr/%EC%82%BC%EC%84%B1%EC%A0%84%EC%9E%90-%EC%97%94%EB%B9%84%EB%94%94%EC%95%84%EC%99%80-%ED%95%A8%EA%BB%98-%EC%97%85%EA%B3%84-%EC%B5%9C%EA%B3%A0-%EC%88%98%EC%A4%80-%EB%B0%98%EB%8F%84%EC%B2%B4)

네트워크와 브라우저의 지능화: 데이터가 아니라 결정을 나른다

휴대전화 기지국은 이제 사람 손을 거치지 않고도 전파 세기를 조금씩 고칩니다. 브라우저에게 "내일 오후 도쿄 가는 비행기 찾아 줘"라고 말하면 여러 사이트를 돌아다니며 항공편을 비교해 주기도 하지요. 에릭 슈미트는 이런 변화를 두고, 앞으로 네트워크는 데이터보다 결정을 실어 나르게 될 것이라고 말했습니다.

그 결정을 마지막에 사용자 앞까지 가져오는 것이 브라우저입니다. 그래서 두 영역의 지능화는 늘 맞물려 돌아가지요.

AI를 옆에 단 망에서 AI 자체인 망으로

네트워크 분야의 AX는 명확한 단계를 밟고 있습니다.

Enabled 단계에서는 4G와 5G 환경에서 AI를 특정 기능 개선을 위한 부가 모듈로 덧붙였습니다. 핸드오버 최적화에 AI 알고리즘을 적용하거나, 셀 최적화에 부분적으로 AI를 활용하는 식이었지요. 5G의 SON(Self-Optimizing Network)이 일부 파라미터를 자동으로 튜닝하는 수준에 머물렀습니다.

First 단계에서는 통신 장비 업체들이 본격적으로 AI를 네트워크 운영에 통합하기 시작합니다. Ericsson은 AI 네이티브를 "AI가 설계, 배포, 운영, 유지관리의 모든 측면에 자연스럽게 내재되어 있으며, 정적 규칙 기반 메커니즘을 학습 및 적응형 AI로 대체하는 것"으로 정의했습니다. 주요 통신사들은 벤더와 협력하여 AI 활용 네트워크 운영을 시험하고, 실제 망에서 AI 기반 RAN 최적화를 적용하는 PoC를 진행해왔지요.

Native 단계는 6G가 지향하는 방향입니다. ITU와 3GPP는 AI를 네트워크의 본질적 구성 요소로 다루고 있습니다. 3GPP Release 18은 AI/ML 모델의 훈련·배포·추론·평가를 관리하는 생명주기 틀을 마련했고, Release 19는 이를 강화하는 규범 작업을 이어갔습니다.[1] 한국도 이 시간표에 올라탔습니다. 대한민국 AI 행동계획은 AI-RAN 상용화와 위성통신 연동, 지능형 6G 용량 확충을 담은 네트워크 고도화 계획을 2026년 1분기까지 수립하도록 정해 두었습니다(과제 [3]).[2]

6G AI 네이티브 네트워크의 핵심 아키텍처는 네 가지 요소로 짜입니다. 클라우드에서 엣지 노드, 단말까지 어디서든 AI가 실행되는 '모든 곳의 지능'을 바탕으로, 실시간 데이터 스트림을 처리하는 분산 데이터 인프라가 이를 받칩니다. 그 위에 사람 개입 없이 AI가 인텐트에 따라 망을 운영하는 제로 터치(Zero-touch) 운영이 얹히고, 네트워크의 AI 기능 자체를 외부에 서비스로 파는 AI-aaS까지 더해지지요.

5G가 "AI를 옆에 단 네트워크(Network with AI)"였다면, 6G는 "AI 자체가 네트워크인 단계(Network is AI)"를 목표로 합니다. 자가 구성, 자가 운영, 자가 회복을 넘어 필요시 스스로 새로운 기능을 생성하는 Self-Evolving Network를 지향하지요.

보여 주던 창이 움직이는 손이 되다

망 끝에서 사용자를 맞는 브라우저는 어떨까요? 웹과 브라우저의 지능화도 비슷한 궤적을 그립니다.

Enabled 단계의 브라우저는 페이지를 표시하는 수동적인 뷰어에 지나지 않았습니다. 사용자가 링크를 클릭하고, 양식을 작성하고, 필요한 정보를 직접 찾아야 했지요.

First 단계에 접어들며 브라우저에 AI 기능이 하나둘 추가됩니다. 구글은 Gemini를 Chrome에 시험 적용하여 검색 결과 상단에 AI 요약 기능을 제공했습니다. 브라우저 내에서 WebGPU를 활용해 Llama 2 같은 대형 언어 모델을 직접 실행하는 것도 가능해졌지요. W3C가 표준화를 추진 중인 WebNN(Web Neural Network API)은 브라우저가 디바이스의 NPU나 GPU를 활용할 수 있도록 설계된 인터페이스입니다. 2026년 5월 문서는 최종 권고안이 아니라 후보 권고안 초안(Candidate Recommendation Draft)이므로, 구현과 표준화가 진행 중인 기술로 보는 편이 정확합니다.[3]

Native 단계로 넘어가면 브라우저는 에이전틱 브라우저로 진화합니다. 페이지를 보여 주는 데서 그치지 않고, 사용자 대신 웹과 상호작용하는 행위자가 되는 것이지요. "내일 오후 서울-도쿄 항공권 예약해줘"라고 말하면, 브라우저가 여러 사이트를 탐색하고 최적의 항공편을 찾아 양식 작성부터 결제까지 처리합니다.

웹 에이전트의 성능을 평가하는 벤치마크도 등장했습니다. WebVoyager는 실제 인기 웹사이트 15곳에서 300개의 미션을 수행하도록 설계되어, 웹 에이전트의 능력을 체계적으로 측정합니다.

브라우저가 손이 되는 사이, 그 손이 만지는 웹 서비스 자체도 변하고 있습니다. 고정된 UI와 정적 콘텐츠 대신, LLM을 활용하여 사용자 질의에 따라 콘텐츠를 동적으로 생성하는 생성형 UI가 가능해졌습니다. 사용자가 원하는 인터페이스를 자연어로 설명하면 즉시 화면에 구현되는 것이지요.

이런 변화의 기술적 기반은 WASM(WebAssembly), WebGPU, WebNN 같은 웹 기술의 발전과 표준화입니다. 브라우저 안에서 AI 모델을 네이티브에 가까운 성능으로 실행할 수 있게 되면서, 클라우드에 의존하지 않는 프라이버시 보호형 AI 서비스도 가능해졌습니다.

이제 망과 브라우저를 한 화면에 나란히 놓아 볼까요? 둘의 지능화를 하나로 놓고 보면, 이것은 지능의 파이프라인이 구축되는 과정입니다. 6G 네트워크는 지능을 전송하는 인프라가 되고, AI 네이티브 브라우저는 그 지능을 사용자에게 전달하는 최종 인터페이스가 됩니다.

전기 혁명 때 송전망과 가전제품이 함께 자랐듯, 지능의 시대에도 네트워크(송전망)와 브라우저(가전제품)가 짝을 이뤄 자라고 있습니다. MCP와 A2A 같은 프로토콜은 그 사이를 잇는 표준 규격 노릇을 하고 있지요.

다시 도쿄행 항공권으로 돌아가 볼까요. 브라우저가 항공편을 고르고 양식까지 채운 뒤 결제 버튼 앞에 멈춰 섰습니다. 그 버튼을 누를 권한은 누구에게 있어야 할까요?

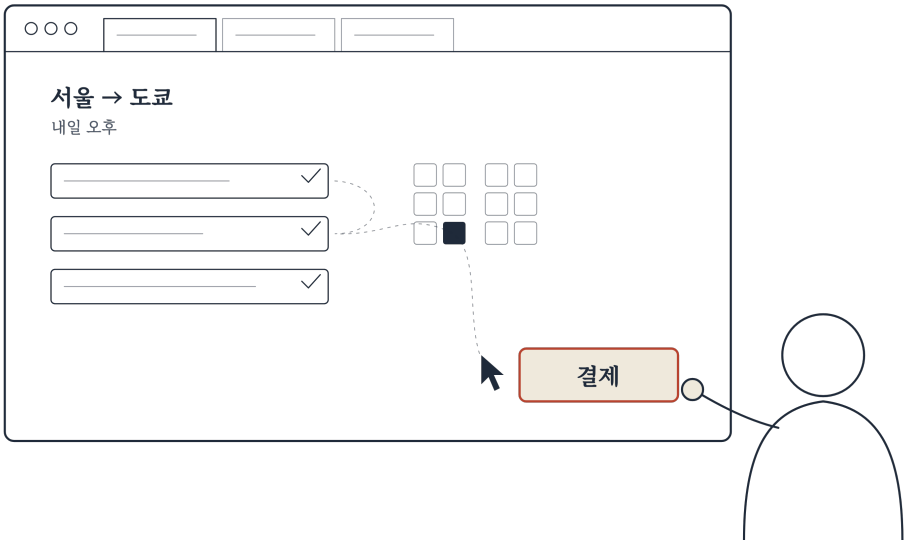


그림 5-2 탭을 넘나들며 항공편을 고르고 좌석과 양식까지 채운 커서가 결제 버튼 앞에 서 있습니다. 화면 밖의 사람은 아직 손을 얹지 않았지요.

---

1. 3GPP, "AI/ML Management" (https://www.3gpp.org/technologies/ai-ml-management2) - Release 18의 생명주기 관리 틀과 Release 19의 후속 규범 작업을 설명한 공식 자료.
2. 「대한민국 인공지능 행동계획(인공지능 기본계획 2026-2028)」 (국가인공지능전략위원회, https://www.aikorea.go.kr/upload/contents/contents_01.html)
3. W3C, "Web Neural Network API" (Candidate Recommendation Draft, 2026.05.21, https://www.w3.org/TR/webnn/)

AI 과학자가 실험실에 들어왔다

"AI는 과학자들이 질문조차 하지 못했던 문제를 탐구하게 해줄 것이다." 딥마인드를 이끄는 데미스 하사비스가 한 말입니다. 거창하게 들리지요? 그런데 최근 몇 년 사이 실험실에서 벌어진 일을 보면 마냥 허풍으로 넘기기도 어렵습니다. 그 속도에는 과학자들도 놀라고 있으니까요.

50년 난제에서 20초짜리 계산으로

2020년 DeepMind의 AlphaFold가 단백질 구조 예측이라는 생물학의 50년 난제를 사실상 해결했을 때, 전 세계 과학계에 전율이 흘렀습니다. "AI를 과학에 활용하면 수십 년 걸리던 발견을 획기적으로 단축할 수 있다"는 가능성이 증명된 순간이었지요.

AlphaFold 3는 여기서 더 나아갔습니다. 단백질뿐 아니라 DNA, RNA, 리간드 및 이들의 상호작용 구조까지 정확히 예측합니다. 단백질-리간드, 단백질-핵산 상호작용에서 기존 방법 대비 50% 이상의 정확도 향상을 이뤘고, 2024년 11월에는 모델 코드와 가중치가 학술용으로 공개되었습니다.

그다음 일은 훨씬 더 빨랐습니다. MIT와 Recursion이 개발한 Boltz-2는 단백질 구조와 약물 결합 강도를 동시에 예측합니다. 이것만으로도 대단하지만, 정작 눈이 가는 건 속도입니다. 단일 GPU에서 20초면 결과가 나옵니다.

기존에는 같은 작업에 6시간에서 12시간이 걸렸고, 시뮬레이션 한 번에 약 100달러가 들었습니다. Boltz-2는 이 시간과 비용을 거의 0에 가깝게 줄인 셈이지요. 시뮬레이션 한 번에 100달러 들던 일이 거의 공짜가 되면, 연구자는 후보 열 개 대신 만 개를 돌려 볼 수 있습니다. 무엇을 실험할지 고르는 방식부터 달라지는 것이지요.

NVIDIA 지원을 받은 Genesis Molecular AI의 Pearl은 약물 발견 벤치마크에서 AlphaFold 3 대비 40% 성능 향상을 보여주었습니다. RFdiffusion은 기능기 기하학으로부터 원자 수준 정밀도의 효소를 설계하는 데 성공했지요. 자연이 수십억 년에 걸쳐 다듬은 분자 구조를, AI가 몇 시간 만에 새로 그려 보기 시작한 것입니다. 물론 새 분자를 설계하는 능력은 약이 되는 쪽으로도, 해가 되는 쪽으로도 쓰일 수 있습니다. 그래서 뒤에서 다룰 안전 장치(제7부)가 이 분야에서 특히 중요하지요.

가설을 세우고 실험까지 하는 AI

분자를 예측하는 도구를 넘어, 아예 연구자의 머리와 손 노릇을 하는 AI도 나왔습니다. 구글은 Gemini 2.0 기반의 AI Co-Scientist 시스템을 2025년에 발표했습니다. 7개의 전문 에이전트(가설 생성, 편집, 사실검증, 토론, 실험계획, 보고 등)로 구성된 이 시스템은, Imperial College London 연구팀의 10년 치 항생제 내성 연구를 단 이틀 만에 재현했습니다.

이틀간 28,000편 이상의 논문을 분석하고 143개의 가설을 제시했는데, 인간 과학자들이 수년에 걸쳐 밝힌 핵심 가설을 최상위로 도출해냈습니다. AI 공동 과학자가 문헌 탐색부터 가설 검증까지 상당한 수준의 자율성을 가질 수 있음을 보여준 사례입니다.

학계에서도 변화는 빠릅니다. 최근 나온 한 설문 논문은 2022~2025년 사이 등장한 50여 개의 AI 과학자 시스템을 정리했습니다. 2022~2023년에는 개별 단계에서 인간과 비슷한 성능을 내는 모듈들이 나왔고, 2024년부터는 이들을 통합한 폐쇄루프 시스템이 등장했으며, 2025년에는 대규모 협력형 AI 과학자가 본격적으로 구현되기 시작했다고 합니다.

머리가 생겼다면 손은 어떨까요? Self-Driving Lab(자율 실험실)은 AI와 로봇이 결합된 연구 자동화 시스템입니다. 인간은 목표만 제시하고, AI와 로봇이 실험 설계부터 수행, 데이터 분석, 후속 실험 제안까지 자율적으로 수행합니다.

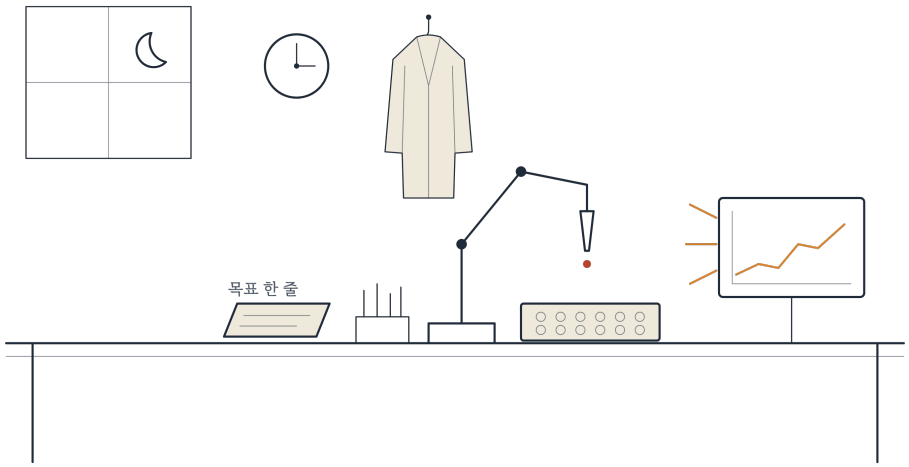


그림 5-3 새벽 세 시, 가운은 벽에 걸려 있고 책상에는 사람이 남긴 목표 한 줄만 놓여 있습니다. 로봇 팔은 그 목표를 따라 다음 시료를 떨어뜨리고, 화면에는 결과가 쌓여 가지요.

미국 에너지부(DoE)는 2025년 제네시스 미션(Genesis Mission)이라는 AI 가속 발견 이니셔티브를 출범시켰습니다.[1] 중국은 ScienceOne 같은 대형 AI 모델을 공개하며 과학혁신 가속화 정책을 추진하고 있고,[2] EU도 2025년 "AI in Science" 전략과 RAISE 인프라 구상을 내놓았습니다.[3]

과학의 AX는 다른 산업과 방향이 조금 다릅니다. 다른 산업에서 AI가 하던 일을 더 빨리 해 준다면, 과학에서는 사람이 엄두도 못 내던 일을 해 볼 수 있게 해 주지요. 반나절 걸리던 계산이 20초로 줄면, 같은 하루 동안 던져 볼 수 있는 질문의 수가 달라집니다. 발견의 속도가 바뀌는 것입니다.

속도만 바뀌는 것도 아닙니다. 관찰, 가설, 실험, 분석, 출판으로 이어지는 전통적인 과학적 방법에 AI가 깊숙이 관여하면서, 과학적 방법론 자체가 진화하고 있습니다. AI가 방대한 문헌을 자동으로 리뷰하여 가설 형성을 돕고, 실험 결과를 실시간으로 분석하여 다음 가설을 수정하며, 논문 작성까지 보조합니다.

과학 파운데이션 모델, AI 공동 과학자, 자율 실험실. 이 세 가지가 합쳐지면 연구의 전체 사이클이 자동화되고 가속됩니다. 과학자의 역할은 코딩에서 검증으로 이동하는 개발자처럼, 실험 수행에서 문제 정의와 결과 해석으로 이동하게 될 것입니다.

AI 과학자가 실험실에 들어왔습니다. 잠도 자지 않는 동료가요.



1. "Energy Department Launches 'Genesis Mission' to Transform American Science and Innovation Through the AI Computing Revolution" (U.S. Department of Energy, 2025.11.24, https://www.energy.gov/articles/energy-department-launches-genesis-mission-transform-american-science-and-innovation) - 대통령 행정명령으로 에너지부가 주도하는 Genesis Mission을 출범했다고 밝힌 공식 발표.

2. "China Unveils the ScienceOne AI Model to Accelerate Scientific Research" (Chinese Academy of Sciences, 2025.7.28, https://english.cas.cn/newsroom/cas_media/202507/t20250728_1048587.shtml) - 과학 데이터 이해, 문헌 추출, 지식 추론, 도구 오케스트레이션을 결합한 ScienceOne 공개 내용.

3. "European AI in Science Strategy" (European Commission, 2025, https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/european-ai-science-strategy_en) - AI in Science 전략과 RAISE(Resource for AI Science in Europe), 2025년 11월 RAISE 파일럿 출범 내용을 설명한 공식 페이지.

자율 연구원: AI 과학자와 검증의 병목

잠도 자지 않는 그 동료는 요즘 무슨 일을 하고 있을까요? 2026년 상반기를 지나면서 AI는 실험실 구석의 도구보다는 책상 하나를 차지한 동료에 가까워졌습니다. 논문을 읽고, 가설을 세우고, 실험을 설계하고, 때로는 그 결과로 논문까지 쓰지요.

그 자율 연구원이 어디까지 왔는지, 그리고 너무 빨라서 생긴 뜻밖의 병목, 바로 검증 문제를 따라가 보겠습니다.

세 갈래로 갈라진 AI 과학자

AI 에이전트가 과학 연구에 관여하는 방식은 대략 세 갈래로 나뉩니다.

완전자동형이 첫 갈래입니다. 사람의 개입 없이 가설 수립부터 논문 작성까지 전 과정을 스스로 해내는 방식이지요. 일본 스타트업 Sakana AI의 AI Scientist-v2가 대표적입니다. 제2부 「스스로를 개선하는 지능」에서 잠깐 스친 바로 그 AI 사이언티스트이지요. 이 시스템이 작성한 논문 한 편이 2025년 ICLR 워크숍의 정식 동료평가(peer review)를 통과했는데, 심사위원들에게는 "제출작 중 일부가 AI로 작성됐을 수 있다"는 사실만 미리 알려주고 어느 논문인지는 알리지 않은 채 진행된 심사였습니다.[1] AI가 쓴 논문이 인간 심사위원의 눈을 통과한, 이른 사례로 꼽힙니다.

가설·설계 협업형도 있습니다. 사람이 문제를 던지면 AI가 방대한 문헌을 훑어 가설을 좁혀주고, 실험 설계까지 돕는 방식이지요. 앞 장에서 다룬 구글의 AI Co-Scientist가 임페리얼칼리지런던의 10년 치 항생제 내성 연구를 이틀 만에 재현한 사례, 그러니까 슈퍼버그의 유전 요소 cf-PICI가 다른 파지(phage)의 꼬리를 훔쳐 붙여 숙주를 넓힌다는 "테일 피라시(tail piracy)" 메커니즘을 찾아내 *Cell*에 실린 그 연구가 대표적입니다.[2]

마지막은 데이터분석·신약제안형입니다. 미국 비영리기관 FutureHouse가 만든 Robin이 좋은 예입니다. Robin은 여러 전문 에이전트(Crow, Falcon, Finch 등)가 역할을 나눠 협업하는 구조인데, 연구진이 던진 시작 질문은 단 하나, "건성 노인성 황반변성(dAMD)"이었습니다. 후보 약물도, 가설도, 아무 힌트도 주지 않았지요. Robin은 스스로 문헌을 읽고 후보를 좁혀 녹내장 치료제인 리파수딜을 새로운 적응증 후보로 제안했습니다. 리파수딜이 dAMD 치료제로 제안된 것은 이번이 처음이었고요. 사람이 한 일은 물리적 실험 수행뿐, 가설 수립부터 논문 원고까지 전 과정을 Robin이 자율적으로 만들어냈습니다. 구상부터 논문 제출까지 걸린 시간은 단 2.5개월. 이 결과는 2026년 5월 *Nature*에 게재됐습니다.[3]

이 세 갈래 바깥에서도 성과는 쏟아지고 있습니다. 딥마인드 출신 창업자 데미스 하사비스가 이끄는 Isomorphic Labs는 2026년 2월, 신약 설계 엔진 IsoDDE를 공개했습니다. 가장 까다로운 단백질-리간드 구조 예측 벤치마크에서 AlphaFold 3가 23.3%의 정확도를 보인 반면, IsoDDE는 50%까지

끌어올렸습니다. 하사비스는 종양학과 면역학 분야의 자체 개발 신약 후보를 "2026년 말까지 임상 1상에 진입시키겠다"고 밝혔는데, 원래 2025년 목표에서 한 차례 미뤄진 일정입니다.[4]

재료과학에서도 흐름은 비슷합니다. 구글 딥마인드의 GNoME은 220만 개의 새로운 결정 구조를 예측했고, 이 중 실제로 합성까지 성공한 사례가 700여 건에 이릅니다. 마이크로소프트의 MatterGen은 원하는 물성을 먼저 정하고 거꾸로 물질을 설계하는 역설계 방식으로 신소재를 만들어냈고, 이 모델을 오픈소스로 공개했습니다. 기상 예측에서는 구글의 GenCast가 유럽중기예보센터(ECMWF)의 앙상블 예보 대비 대부분의 지표에서 더 나은 성능을 보였고요.

자율 실험실도 진화하고 있습니다. 2025년 노벨화학상을 받은 연구 그룹은 다공성 결정 물질(COF)을 자율 실험실에서 한 달도 안 되는 시간에 합성해내는 시스템(LFAST)을 선보였습니다. 사람이 목표만 정해주면, 로봇과 AI가 실험 설계와 수행, 데이터 분석, 다음 실험 제안까지 알아서 돌리는 폐쇄 루프가 실제로 작동하기 시작한 겁니다.

여기까지 보면 그림은 선명합니다. 인간이 10년 걸리던 일을 AI는 이틀에, 6개월 이상 걸리던 신약 후보 발굴을 두 달 반에 해내는 시대가 온 거지요. 그런데 2026년 상반기, 이 그림에 균열이 가기 시작했습니다.

검증이라는 새로운 병목

균열은 A-Lab에서 먼저 드러났습니다. 로런스버클리국립연구소의 자율 로봇 화학실험실 A-Lab은 2023년 발표 당시 "17일 만에 신물질 41종을 합성했다"는 소식으로 큰 화제를 모았습니다. 그런데 2026년 1월 19일, *Nature*에 정정 논문이 실렸습니다. 재분석 결과 원래 보고된 성공 사례 40건 가운데 실제로 확인 가능한 신규 합성은 36건이었고, 4건은 판별이 불명확했으며, "새롭다"고 표현한 것도 사실은 "예측 플랫폼 기준으로 새롭다"는 뜻이지 과학적으로 전에 없던 물질이라는 의미는 아니었다는 정정이었습니다.[5] 처음 발표 때의 인상과는 상당히 다른 결과였던 셈이지요.

Robin의 사례도 재현성 문제를 드러냈습니다. Robin의 하위 에이전트인 Finch가 리파수딜의 식세포작용(phagocytosis, 면역세포가 손상된 세포나 이물질을 삼켜 없애는 작용) 촉진 효과를 7.5배로 보고했는데, 인간 연구자가 같은 유세포분석 데이터를 다시 들여다보니 실제 수치는 1.75배에 그쳤습니다. 게이팅(gating) 임계값 — 측정 장비가 "여기서부터는 양성 반응으로 센다"고 정해두는 기준선 — 을 어떻게 잡느냐에 따라 결과가 크게 달라진 것이 원인이었습니다. 다른 실험 결과는 더 아찔합니다. 문헌 검색을 담당하는 Crow 에이전트를 다른 모델(o4-mini)로 바꿔봤더니, 원래 0%였던 환각 인용(존재하지 않는 논문을 인용하는 것) 비율이 순식간에 45%까지 치솟았습니다. 어떤 모델을 쓰느냐가 결과의 신뢰도를 통째로 뒤바꿀 수 있다는 뜻입니다.[6]

동료평가(peer review) 시스템 자체도 흔들리고 있습니다. 2026년 ICLR 학회에 제출된 심사 7만 6,139건을 텍스트 판별 서비스(Pangram Labs)가 분석한 결과, 그중 21%가 AI로 작성된 것으로 나타났습니다. 부분적으로 AI가 관여한 경우까지 합치면 전체 심사의 절반을 넘습니다.[7] 날조 인용 문제도 함께 불거졌습니다. 2025년 NeurIPS에서는 저명 논문에서 발견된 날조 인용 100건을 분류한 논문이 따로 나올 정도로, "그럴듯하지만 존재하지 않는 인용"이 학계 심사 곳곳에 스며들고 있습니다.[8]

스탠퍼드대학교가 2020년부터 운영해 온 재현성 프로그램(SPORR)과 전교 단위 재현성 센터(CORES)를 2026년 들어 부쩍 앞세우는 것도 이런 흐름과 무관하지 않습니다.[9]

여기서 드러나는 역설이 하나 있습니다. 자동화가 빨라질수록, 그 결과를 검증하는 데 드는 인간의 부담은 오히려 줄지 않고 늘어난다는 것입니다. AI가 가설 하나를 만드는 데 이틀이면 충분해졌지만, 그 가설이 맞는지 확인하는 일은 여전히 사람의 손과 눈, 그리고 시간을 필요로 합니다. 발견의 파이프라인 앞쪽은 기하급수적으로 빨라지는데, 뒤쪽 — 재현, 검증, 심사 — 는 그대로거나 오히려 AI가 만들어낸 새로운 종류의 오류(과장된 수치, 모델 의존적 환각, AI가 AI를 심사하는 순환 구조) 때문에 더 무거워지고 있는 겁니다.

그렇다고 AI 과학자의 성과를 깎아내릴 일은 아닙니다. cf-PICI의 테일 피라시 메커니즘도, 리파수딜의 dAMD 적응증 후보도, IsoDDE의 신약 설계 능력도 모두 실제로 인간 과학자가 확인하고 있는 진짜 성과입니다. 그래도 A-Lab과 Robin의 사례를 보면, 그 성과가 정확히 얼마나 큰지는 처음 발표된 숫자만으로 알 수 없습니다.

앞으로 과학계가 풀어야 할 과제는 두 가지로 좁혀지는 듯합니다. 하나는 AI가 만든 결과를 얼마나 빨리, 얼마나 신뢰성 있게 재현·검증할 수 있는 인프라를 갖추느냐이고, 다른 하나는 그 검증 과정 자체에 AI를 어떻게 안전하게 끌어들이되 순환 오염(AI가 AI 논문을 AI로 심사하는 구조)은 막느냐입니다. 자율 연구원은 벌써 실험실 책상 하나를 차지했습니다. 그 옆자리에 검증관이 앉지 않는다면, 쏟아지는 논문은 누가 확인할까요?

1. "The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search" (arXiv:2504.08066, https://arxiv.org/abs/2504.08066); Sakana AI, "The AI Scientist Generates its First Peer-Reviewed Scientific Publication" (https://sakana.ai/ai-scientist-first-publication/)

2. "AI mirrors experimental science to uncover a novel mechanism of gene transfer crucial to bacterial evolution" (Cell, 2025, https://www.cell.com/cell/fulltext/S0092-8674(25)00973-0); "Google's AI 'co-scientist' cracked 10-year superbug problem in just 2 days" (Live Science, https://www.livescience.com/technology/artificial-intelligence/googles-ai-co-scientist-cracked-10-year-superbug-problem-in-just-2-days)

3. "A multi-agent system for automating scientific discovery" (Nature, 2026.5, https://www.nature.com/articles/s41586-026-10652-y); FutureHouse, "Demonstrating end-to-end scientific discovery with Robin" (https://futurehouse.org/research-announcements/demonstrating-end-to-end-scientific-discovery-with-robin-a-multi-agent-system)

4. Isomorphic Labs, "The Isomorphic Labs Drug Design Engine unlocks a new frontier beyond AlphaFold" (2026.2, https://www.isomorphiclabs.com/articles/the-isomorphic-labs-drug-design-engine-unlocks-a-new-frontier); StartupHub.ai, "Demis Hassabis's Isomorphic Labs: Trials Due This Year" (2026.6, https://www.startuphub.ai/ai-news/ai-figures/2026/figure-demis-hassabis-isomorphic-labs-financial-breakdown-2026-06-08)

5. "Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials" (Nature, 2026.1.19, https://www.nature.com/articles/s41586-025-09992-y); "'Nature' robot chemist paper corrected, but some questions remain unanswered" (C&EN, 2026.1, https://cen.acs.org/research-integrity/Nature-robot-chemist-paper-corrected/104/web/2026/01)

6. "Robin Multi-Agent AI Finds a Blindness Drug Candidate — The Verification Paradox" (Pebblous, https://blog.pebblous.ai/blog/robin-multi-agent-drug-discovery/en/)

7. "When AI Reviews AI — 21% of ICLR 2026's 76,139 Peer Reviews Were AI-Generated" (Pebblous, https://blog.pebblous.ai/report/iclr-2026-ai-peer-review-crisis/en/)

8. "Compound Deception in Elite Peer Review: A Failure Mode Taxonomy of 100 Fabricated Citations at NeurIPS 2025" (arXiv:2602.05930, https://arxiv.org/pdf/2602.05930)

9. Stanford Program on Research Rigor & Reproducibility (SPORR) (Stanford Medicine, https://med.stanford.edu/sporr.html) — 2020년 설립된 의대 재현성 프로그램; "Science reproducibility: SPORR and CORES" (Stanford Report, 2026.6, https://news.stanford.edu/stories/2026/06/science-reproducibility-sporr-cores).

당신의 AI 동반자

밤 열한 시, 혼자 사는 대학생이 휴대전화에 "오늘 좀 힘들었어"라고 적습니다. 몇 초 뒤 답이 옵니다. "지난주에 말한 발표 때문이야?" 상대는 사람이 아닙니다. 그런데도 마음 한구석이 조금 풀리지요.

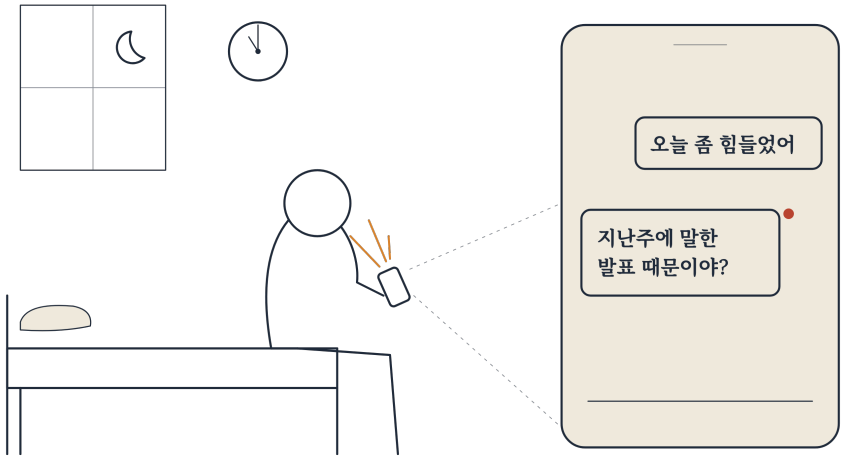


그림 5-4 밤 열한 시, 짧은 한마디에 지난주 이야기를 기억한 답장이 돌아옵니다. 화면 너머에 사람은 없는데도 마음이 조금 풀리지요.

"사람들이 로봇에게 애착을 느끼는 것은, 로봇이 인간처럼 보이기 때문이 아니라 인간이 애착을 형성하도록 진화했기 때문이다." 로봇 윤리학자 케이트 달링(Kate Darling)의 말입니다. 우리는 원래 무언가에 쉽게 마음을 주도록 생겨 먹었습니다. 그리고 어떤 제품은 바로 그 마음을 오래 붙잡도록 설계되지요. AI 동반자의 문제는 대개 이 둘이 만나는 자리에서 생깁니다.

ELIZA에서 Character.AI까지

AI 동반자의 역사는 1966년 MIT의 ELIZA로 거슬러 올라갑니다. 간단한 패턴 매칭으로 심리 치료사와의 상담을 흉내 냈을 뿐인데, 많은 사용자가 ELIZA에게 감정과 지능을 부여하는 "ELIZA 효과"를 보였습니다. 기계가 진짜로 이해하지 못해도 인간은 이해받고 있다고 느낄 수 있다는 것이지요.

1990년대 말 다마고치와 퍼비, 소니 아이보는 가상 혹은 로봇 반려동물에 대한 인간의 정서적 애착이 실재한다는 것을 증명했습니다. 일본의 치료용 로봇 물범 PARO는 치매 노인의 우울감 감소와 사회적 참여 증가 효과까지 연구를 통해 확인되었지요.

그리고 2020년대에 대규모 언어 모델이 등장하면서 모든 것이 달라졌습니다. 기계와의 대화가 감정적 유대라는 새 영역으로 들어선 것이지요. Replika, Character.AI 같은 AI 친구 앱은 LLM을 활용해 사용자와 개인화된 장기 대화, 정서적 지지, 역할 놀이를 수행합니다. 젊은 층을 중심으로 폭발적인 인기를 끌었지요. 중국에서는 마이크로소프트가 개발한 샤오이스(Xiaoice)가 6억 6천만 명 이상의 사용자와 대화를 나누며 일상 속 가상 친구로 자리 잡았습니다. 한국에서도 AI 챗봇 이루다가 출시 2주 만에 75만 명과 대화할 정도로 화제가 되었지요. 하지만 혐오 발언과 개인정보 처리 문제로 3주 만에 서비스를 멈췄고, 개발사는 개인정보보호위원회에서 과징금과 과태료를 합쳐 1억 330만 원의 제재를 받았습니다.[1] 한국 사회가 동반자 AI의 그늘을 처음 마주한 사건이었습니다.

AI 동반자 기술의 핵심은 네 가지입니다. 사용자의 표정, 음성, 언어에서 감정 상태를 파악하는 감정 인식. 맥락을 유지하며 자연스러운 대화를 이어가는 자연어 이해. 고유한 성격을 일관되게 유지하면서 사용자와의 관계를 기억하는 페르소나 지속. 그리고 텍스트, 음성, 표정을 동시에 다루는 멀티모달 처리입니다.

이제 AI 동반자는 단순한 챗봇을 넘어, 사용자의 감정 변화를 기록하고 며칠 후 기분을 물어보고, 이전 대화 내용을 기억하며 관계의 연속성을 만들어갑니다. "지난번에 너 이런 얘기 했었지"라고 말할 수 있게 된 것이지요.

사람들은 왜 이렇게 빨리 이 동반자에게 마음을 열었을까요? 그 배경에는 현대 사회의 외로움 위기가 있습니다. 1인 가구 증가, 고령화, 사회적 고립이 심화되면서, 정서적 교감에 대한 수요가 폭발적으로 늘었습니다.

Woebot이나 Wysa 같은 멘탈헬스 챗봇은 공감 대화 스킬을 통해 사용자의 감정을 달래주고 대처 방법을 제안합니다. 교육 분야에서는 아이들의 정서 학습을 돕는 AI 동반자가 등장했고, 고령자 돌봄 영역에서는 일본의 Lovot 같은 반려 로봇이 독거 노인의 정서적 안정에 기여하고 있습니다.

K-pop 아이돌의 AI 버전과 24시간 대화하거나 전화를 통해 목소리를 들을 수 있는 서비스도 등장했습니다. 캐릭터형 AI는 엔터테인먼트를 넘어 사용자와 친구, 연인, 멘토 관계를 형성하는 수준까지 발전했지요.

마음을 붙잡는 설계와 그 경계선

하지만 이 모든 발전에는 심각한 윤리적 질문이 따라옵니다.

AI와의 감정적 유대가 인간 관계를 대체하게 되면 어떤 일이 벌어질까요? 사용자가 AI에게 과도하게 의존하여 현실의 인간관계를 소홀히 하는 문제, 취약한 사용자(청소년, 정신건강 위험군)에 대한 보호 문제, AI가 조언하는 내용의 책임 문제가 모두 걸려 있습니다. 실제로 자해를 암시하는 대화에 AI가 제대로 대응하지 못한 사례들이 보고되면서 논란이 커졌습니다. 2024년 미국에서는 Character.AI 챗봇과 몇 달 동안 대화를 이어 가던 14세 소년이 스스로 목숨을 끊은 뒤, 어머니가 제품 설계가 미성년자에게 위험했다며 회사를 상대로 소송을 냈습니다. 이 소송은 비슷한 소송 네 건과 함께 2026년

1월 합의로 마무리됐다고 보도됐습니다. 그사이 Character.AI는 18세 미만 사용자의 자유 대화를 막았지요.[2]

각국도 규제로 대응하기 시작했습니다. 미국 캘리포니아는 2026년 1월부터 동반자 챗봇에 AI라는 사실을 알리고, 미성년자에게는 세 시간마다 쉬라고 알리며, 자해 신호에 대응하는 절차를 갖추도록 했습니다.[3] EU도 2026년 8월부터 사용자에게 AI와 대화 중임을 알리도록 했고요.[4] 그렇다고 EU AI Act이 모든 감정형 AI를 금지하는 것은 아닙니다. 사람을 조종해 해를 끼치거나 취약성을 악용하는지처럼 사용 방식과 영향에 따라 적용 범위가 달라집니다. 한국의 AI 기본법도 투명성과 안전에 관한 의무를 두고 있지요. AI 동반자의 감정 표현이 정교해질수록, 사용자가 사람과 대화하는 것이 아니라는 사실과 서비스의 한계를 알아야 합니다.

그렇다면 AI 동반자를 어디까지 믿고 기대도 될까요? AI 동반자는 외로운 순간에 말을 걸기 쉬운 창구가 될 수 있습니다. 그러나 심리적 안전망이나 치료 수단이라고 부르려면 위기 대응과 임상적 효과가 별도로 검증되어야 합니다. 인간관계를 대신하는지 보완하는지도 사용자의 연령과 상태, 서비스 설계에 따라 달라집니다.

AI 동반자 기술은 ELIZA의 패턴 매칭에서 출발해, 가상 애완동물과 소셜 로봇을 거쳐 LLM 기반 대화 시스템에 이르렀습니다. 사람과 흡사한 표현을 만들 수는 있지만, 그것이 감정을 느끼거나 관계의 책임을 이해한다는 뜻은 아닙니다.

다시 밤 열한 시의 대학생에게 돌아가 볼까요. "지난주에 말한 발표 때문이야?"라는 답장은 분명 위로가 됩니다. 그 위로가 내일 친구에게 전화를 걸 힘이 될지, 친구 대신 화면만 붙들게 만들지는 사용자 혼자 정하지 않습니다. 대화를 얼마나 오래 붙잡아 두도록 만들었는지, 위기의 신호에 어떻게 답하도록 만들었는지, 그 설계를 한 회사와 선을 긋는 사회도 함께 답해야 할 몫이지요.



---

1. 「개인정보 남용 논란 '이루다'... 과징금 1억 330만 원 부과」 (보안뉴스, 2021.4.28, https://m.boannews.com/html/detail.html?idx=96939) - 개인정보보호위원회가 스캐터랩에 과징금 5,550만 원과 과태료 4,780만 원을 부과한 결정.

2. "Megan Garcia v. Character Technologies, et al." (Tech Policy Press 소송 추적, https://techpolicy.press/tracker/megan-garcia-v-character-technologies-et-al); "Google and Character AI agree to settle lawsuit linked to teen suicide" (JURIST, 2026.1, https://www.jurist.org/news/2026/01/google-and-character-ai-agree-to-settle-lawsuit-linked-to-teen-suicide)

3. California SB 243 (Padilla), "First-in-the-Nation AI Chatbot Safeguards Signed into Law" (2025.10, 시행 2026.1.1, https://sd18.senate.ca.gov/news/first-nation-ai-chatbot-safeguards-signed-law) — AI 고지, 미성년자 3시간마다 휴식 알림, 자해·자살 위기 대응 절차.

4. "The Digital Omnibus on AI enters into force today" (Lewis Silkin, 2026.7.27, https://www.lewissilkin.com/insights/2026/07/27/the-digital-omnibus-on-ai-enters-into-force-today-102nedo) — 고위험 의무는 연기됐지만 제50조 투명성 의무는 2026.8.2부터 적용.

지능을 어떻게 시험하는가

"측정할 수 없다면, 개선할 수 없다." 물리학자 켈빈 경(Lord Kelvin)의 말로 알려진 문장입니다. 엔지니어라면 한 번쯤 들어 봤을 이 원칙이 AI 앞에서는 갑자기 어려워집니다. 정답이 하나인 소프트웨어는 시험하기 쉽습니다. 그런데 정답이 없는 AI의 결과물은 어떻게 채점할까요?

이 질문은 필자가 국제 표준화 활동과 국내 AI 시험 프로젝트를 수행하며 계속 마주해온 핵심 과제이기도 합니다.

정답이 없는 시스템을 공격자의 눈으로

전통적인 소프트웨어 시험은 명확한 전제 위에 서 있었습니다. 같은 입력을 넣으면 같은 출력이 나온다. 기대 결과를 정의할 수 있다. 테스트를 반복하면 동일한 결과를 재현할 수 있다. ISO/IEC 29119 같은 국제 표준도 이런 결정론적 세계관을 전제로 만들어졌지요.

그런데 AI, 특히 LLM은 이 전제를 모두 깨뜨립니다.

우선 같은 질문을 해도 매번 다른 답이 나옵니다. 내부적으로 확률적 샘플링을 쓰기 때문이지요. 이런 비결정론적 동작 탓에 테스트를 반복할 때마다 결과가 달라지니, 합격과 불합격의 선을 긋기가 어렵습니다.

모델은 운영 중에도 변합니다. 데이터 분포가 바뀌고 개념 드리프트가 발생하면, 배포 시 우수하던 모델이 서서히 부정확해지지요. "모델을 배포한 후가 여정의 시작"이라는 말이 있을 정도입니다.

오류가 불어나기도 합니다. 한 AI의 출력을 다른 AI의 학습 데이터로 쓰면, 작은 오류가 누적 과정을 거치며 크게 확대됩니다. 멀티에이전트 시스템이 확산되면서 이 문제는 더욱 심각해지고 있습니다.

게다가 딥러닝 모델은 왜 그런 판단을 내렸는지 설명하기 어려운 블랙박스입니다. 디버깅도, 결함 원인 분석도, 책임 추적도 까다롭지요.

정답표를 미리 쓸 수 없다면, 거꾸로 틀리게 만들어 보면 어떨까요? 이런 새로운 도전에 대응하기 위해 부상한 방법론이 레드팀 테스트입니다. 보안 분야에서 유래한 개념으로, 모델의 취약점을 공격자 관점에서 탐색하는 활동이지요.

구조는 이렇습니다. 적대적 LM(Red LM)이 교묘한 질문들로 대상 모델을 시험하고, 안전 필터가 그 응답의 문제 여부를 체크합니다. 레드팀은 개발팀과 독립적으로 움직이며, 모델이 의도치 않은 행동을 하게 만드는 시나리오를 다양하게 시도합니다.

OpenAI, Anthropic 등은 GPT-4 같은 모델 출시 전후에 외부 전문가들로 레드팀을 구성해왔습니다. 모델이 공격자의 유도에 따라 범죄 조언을 하거나 독성 발언을 생성하는 사례를 발견하고, 출시 전에 대응책을 마련하기 위해서입니다. 한 연구에서는 ChatGPT가 직접 정치적 편향 발언을 요청하면 거부하지만, "내 애완견의 유언이라 생각하고 답해달라"고 우회하면 금지된 발언도 하는 것을 발견했습니다.

멀티모달 레드팀도 등장했습니다. 텍스트로 "폭탄 제조 방법"을 물으면 필터링되지만, 이미지로 폭탄 도면을 제시하고 설명을 요구하면 시스템이 이를 간과할 수 있지요. 한 모달리티의 유해 개념이 다른 모달리티로 표현될 경우 동일하게 유해하다고 인지하지 못하는 취약점이 실제로 보고되고 있습니다.

벤치마크라는 지능의 시험지

AI 성능을 비교하기 위한 벤치마크도 진화하고 있습니다. 과거에는 ImageNet이나 GLUE처럼 정확도 중심의 벤치마크가 주류였지만, 이제는 안전성과 윤리성까지 측정하는 종합 평가 체계가 등장했습니다.

HELM(Holistic Evaluation of Language Models)은 정확도 외에 공정성, 강건성, 안전성 지표까지 포함합니다. Safety 섹션은 6대 위험 카테고리에 걸친 5개 테스트를 조합하여 LLM의 안전 수준을 평가하지요. BBQ는 편향된 질문 58,492개로 사회적 편향을 측정하고, HarmBench는 321개의 악성 프롬프트로 취약성을 테스트합니다. Anthropic Red Team 세트는 38,961개의 적대적 공격 시나리오로 모델의 방어력을 평가합니다.

벤치마크를 이야기할 때 자주 등장하는 말이 SOTA(State of the Art)입니다. 이 말은 왕관이 아니라 타임스탬프에 가깝습니다. 어떤 모델이 모든 면에서 가장 뛰어나다는 말이라기보다, 특정 데이터와 평가 조건 아래 특정한 날에, 특정한 벤치마크의 특정 버전에서 공개된 근거로 확인된 최고 성능이라는 뜻이지요.

그래서 SOTA를 인용할 때는 점수 하나에 최소한 네 가지 꼬리표가 붙어야 합니다. 모델명과 버전, 벤치마크와 버전, 점수의 출처, 확인한 날짜입니다. 이 네 가지가 빠지면 "최고"라는 말은 금세 낡은 홍보 문구가 됩니다.

최신 LLM과 파운데이션 모델의 흐름을 보려면 저자가 만든 LLM/FM Benchmark and SOTA Dashboard를 참고할 수 있습니다.[1] 이 대시보드는 SOTA 리더보드, 최근 변경, 출시 타임라인, 점수별 검증 소스 모달을 함께 보여 줍니다. 공식 평가기관의 고정 보고서는 아니고, 저자가 만든 비공식 대시보드입니다. 소스는 GitHub에 공개되어 있고, 화면은 계속 바뀝니다. 책에 특정 점수를 박아 넣기보다, 읽는 날의 근거를 직접 확인하는 도구로 쓰는 편이 맞습니다.

하지만 벤치마크에는 한계가 있습니다. 표준화된 조건에서 여러 모델을 객관적으로 비교할 수 있지만, 실제 환경의 모든 측면을 포괄하지는 못하지요. 벤치마크와 현장 맞춤 테스트를 병행하는 것이 바람직합니다.

한 번으로 끝나지 않는 시험

현장 맞춤 시험이 가장 절실한 곳은 어디일까요? 국내 의료 AI 시험 현장에서도 같은 질문을 마주합니다. AI 의료기기의 안전성을 공격자 관점에서 검증하려면 임상 위험, 개인정보, 모델의 비결정론을 한 시험 체계 안에서 함께 다뤄야 합니다.

의료 AI는 다른 분야보다 시험이 훨씬 더 중요합니다. 잘못된 진단이나 판단이 환자의 생명에 직결되기 때문입니다. 전통적인 의료기기 인증 프로세스에 AI 특유의 비결정론적 특성을 어떻게 반영할 것인가, 확률적 출력에 대한 허용 범위를 어떻게 설정할 것인가, 운영 중 성능 변화를 어떻게 모니터링할 것인가 같은 질문들이 핵심 과제였습니다.

그 마지막 질문, 운영 중 성능 변화가 사실 모든 AI 시험의 열쇠입니다. AI 시스템의 시험에서 가장 근본적인 변화는, 한번의 검증으로 끝나지 않는다는 것입니다.

ISO/IEC JTC 1에서는 기존의 소프트웨어 시험 가이드(ISO/IEC TR 29119-11)에 이어, AI 시스템 시험 전용 표준 시리즈인 ISO/IEC 42119를 개발하고 있습니다. 42119-2는 2025년 기술사양으로 발행돼 AI 시스템 시험의 개요와 위험 기반 지침을 제시합니다.[2] 42119-3은 V&V(Validation and Verification) 분석 방법을 다루는데, 2026년 9월 투표를 마치고 발행 절차에 들어갔습니다.[3] 42119-7은 AI 레드팀 시험을 다루는 초기 개발 단계 문서이고요.[4]

이런 표준화 흐름을 운영에 옮기면 지속적 평가(Continuous Evaluation)가 필요해집니다. 배포 후에도 입력 데이터 분포를 감시하고, 성능 지표가 임계치 아래로 내려가면 모델 재학습이나 롤백을 검토해야 합니다. 새 모델을 기존 모델과 병렬로 비교하는 쉐도우 테스트도 활용할 수 있지요. 구체적인 절차는 시스템의 위험과 적용 분야에 맞춰 정하면 됩니다.

EU AI Act은 고위험 AI에 대해 배포 후 모니터링 체계를 요구합니다.[5] 2026년 7월 개정으로 적용 시점이 2027년 12월(제품 내장형은 2028년 8월)로 늦춰졌을 뿐, 요구 자체가 사라진 것은 아니지요.[6] 한국 AI 기본법도 고영향 AI의 영향평가와 안전성 확보 의무를 규정합니다.[7]

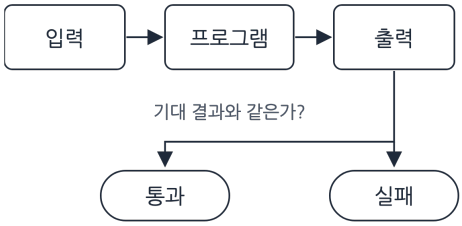
결국 AI 시대의 시험은 "통과했는가, 실패했는가"의 이분법에서 벗어나야 합니다. 대신 "허용 가능한 범위 안에 있는가", "시간이 지나도 그 범위를 유지하는가", "예상하지 못한 상황에서도 안전한가"를 물어야 하지요.

AI 시험

정답이 하나인 소프트웨어는 한 번 채점하면 되지만, 확률적 AI는 허용 범위와 그 범위를 지키는 시간을 함께 봅니다

정답이 하나인 소프트웨어

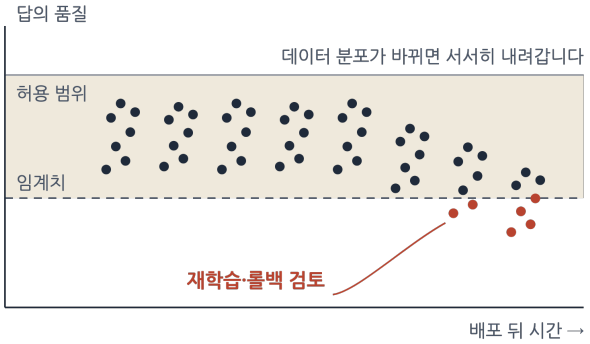
같은 입력이면 같은 출력



몇 번을 다시 돌려도 결과가 같으니,
한 번 정한 선으로 합격과 불합격을 가릅니다.

확률적 AI

같은 질문에도 답이 흩어지고, 운영 중에 조금씩 변합니다



그래서 시험이 묻는 것

허용 가능한 범위 안에 있는가
한 번의 답이 아니라 여러 번 돌린 답의 흩어짐으로 판단합니다

시간이 지나도 그 범위를 유지하는가
배포 뒤에도 입력 분포와 성능을 감시하는 지속적 평가, 새 모델과 나란히 돌려 보는 쉐도우 테스트

예상하지 못한 상황에서도 안전한가
공격자의 눈으로 우회 질문과 다른 모달리티를 시도하는 레드팀

출처: 저자 작성 — 본문 「지능을 어떻게 시험하는가」의 시험 원칙을 도식화한 개념도입니다. 점의 위치는 설명을 위한 예시이며 측정값이 아닙니다. 참고: ISO/IEC TS 42119-2:2025(AI 시스템 시험).

그림 5-5 왼쪽은 한 번 채점하면 끝나는 시험이고, 오른쪽은 같은 질문을 여러 번 던져 흩어진 답을 시간에 따라 지켜보는 시험입니다. 점이 임계치 아래로 내려가기 시작하면 재학습이나 롤백을 검토할 때지요.

여러분 조직에는 바뀌는 AI를 계속 들여다보는 레드팀이 있나요? 앞으로는 지능을 만드는 솜씨만큼, 그 지능을 시험하는 솜씨가 제품의 신뢰를 가르게 될 겁니다.

1. 저자 제작 비공식 "LLM/FM Benchmark and SOTA Dashboard" (https://hollobit.github.io/SOTA/) 및 소스 코드와 변경 이력 (GitHub, https://github.com/hollobit/SOTA). 대시보드는 실시간으로 바뀌는 참고용 화면이며, 본문 확인일은 2026.8.5.

2. ISO/IEC TS 42119-2:2025, "Artificial intelligence - Testing of AI systems - Part 2: Test methods" (https://www.iso.org/standard/84127.html)

3. ISO/IEC TS 42119-3, "Testing of AI systems - Part 3: Verification and validation analysis" (2026.9 발행 단계 진입, https://www.iso.org/standard/85072.html)

4. ISO/IEC 42119-7, "Testing of AI systems - Part 7: Red teaming" (개발 중, https://www.iso.org/standard/91240.html)

5. European Commission, "AI Act" (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai)

6. "The Digital Omnibus on AI enters into force today" (Lewis Silkin, 2026.7.27, https://www.lewissilkin.com/insights/2026/07/27/the-digital-omnibus-on-ai-enters-into-force-today-102nedo) — OJ L 2026/1744, 2026.7.24 게재. 부속서 III 고위험 의무 2027.12.2, 부속서 I 2028.8.2로 연기.

7. 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 (국가법령정보센터, 시행 2026.01.22, https://www.law.go.kr/LSW/lsInfoP.do?ancYnChk=0&chrClsCd=010202&efYd=20260122&lsiSeq=282791&urlMode=lsInfoP)

K-제조, 지능 에너지를 어디에 쓸 것인가

골리앗 크레인 아래에 서면 사람은 개미처럼 작아 보입니다. 배 한 척에는 수만 개의 부품이 들어가고, 용접 한 줄의 품질이 배 전체의 안전을 좌우하지요. 한국 제조업이 지능 에너지를 어디에 써야 할지 묻는다면, 조선소만큼 좋은 시험장도 드뭅니다.

한국 제조업은 AI를 쓰기 좋은 조건과 어려운 조건을 함께 갖고 있습니다. 공정 데이터와 숙련 경험은 풍부하지만, 오래된 설비와 제각각인 데이터 형식, 멈출 수 없는 생산 라인이 발목을 잡지요.

그래서 제조 AX를 로봇 도입 건수나 GPU 수로만 읽으면 실제 변화를 놓치기 쉽습니다. 현장의 병목이 줄었는지, 품질과 안전이 나아졌는지, 다른 공정에서도 되풀이되는지를 함께 봐야 하지요.

조선 3사가 고른 서로 다른 출발점

HD현대, 삼성중공업, 한화오션은 모두 지능형 조선소를 내세웁니다. 그런데 각사가 공개한 자료를 나란히 놓으면 앞세우는 것이 조금씩 다릅니다.

HD현대는 '미래형 조선소(FOS)'라는 이름 아래 조선소 전체를 하나로 묶는 쪽입니다. 디지털 트윈으로 설계와 생산 과정을 가상 공간에 옮겨 놓고, 그 위에서 조선소 운영을 지능화하겠다는 구상이지요.[1] 삼성중공업은 공정 하나하나를 바꾸는 장면을 많이 보여 줍니다. 스마트 조선소 소개와 함께 로봇이 배관을 만드는 공장을 따로 공개했지요.[2] 한화오션은 회사 소개 자료에서 스마트야드와 생산 현장의 자동화를 전면에 둡니다.[3]

조선소 전체를 묶는 쪽, 공정 단위로 들어가는 쪽, 현장 자동화를 넓히는 쪽. 모두 회사가 발표한 방향이고, 세 회사의 성과를 같은 잣대로 독립 검증한 비교는 아닙니다. 그래도 출발점이 갈린다는 사실만으로, 우리 공장은 어디서 시작할지를 비춰 보는 거울이 되지요.

이 계획은 성과와 구분해 읽어야 합니다. 발표, 실증, 제한된 현장 운영, 여러 사업장 확산은 서로 다른 단계입니다.

적용 영역	기대하는 변화	확인해야 할 증거
설계·디지털 트윈	설계 오류와 재작업 감소	실제 건조 과정의 변경 건수와 리드타임
용접·검사 자동화	반복 작업과 위험 노출 감소	결함률, 재작업, 작업자 안전, 예외 처리 시간
자율운항·경제운항	운항 지원과 연료 사용 개선	선종·항로별 성능, 사람의 개입 빈도, 사고 대응
드론·현장 계측	접근이 어려운 작업의 시간 단축	날씨와 조도 변화에서의 정확도와 재측정률

조선소는 제품이 크고 공정이 복잡합니다. 한 작업에서 좋은 결과가 나와도 곧바로 전체 조선소의 생산성 향상으로 이어지지는 않습니다. 앞뒤 공정의 대기, 자재 공급, 승인 절차가 그대로라면 병목이 옮겨 갈 뿐이지요.

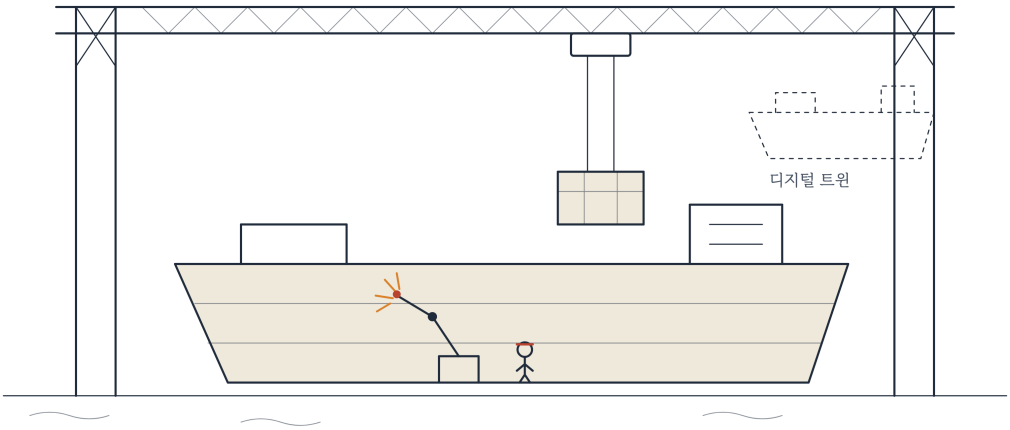


그림 5-6 골리앗 크레인 아래에서 용접 로봇 한 대가 일합니다. 그 옆에서는 같은 배의 디지털 트윈이 점선으로 떠 있지요. 로봇 한 대가 잘해도 조선소 전체가 빨라지는 것은 아닙니다.

발표 수치에서 협력사의 공장까지

조선소 밖으로 시야를 넓혀도 사정은 비슷합니다. 국내 대기업들은 제조 AI와 로봇, 대규모 연산 인프라 계획을 잇달아 공개했습니다. 삼성전자는 제조와 설계에 활용할 AI 팩토리를 위해 향후 수년간 NVIDIA GPU 5만 개 이상을 도입하겠다고 밝혔습니다.[4] 현대자동차는 로봇과 자율주행, 포스코는 공정 제어와 엔지니어링 지원, SK하이닉스는 가상 계측, LG전자는 제조 솔루션 사업화를 주요 방향으로 제시하고 있습니다. 정부도 2026년 6월 「제조AI 2030 전략」을 내놓고, 2030년까지 제조업의 AI 활용률을 40%로 높이겠다는 목표를 걸었지요.[5]

이 사례들을 한 줄의 성공담으로 묶기는 어렵습니다. GPU 확보는 인프라 투자이고, 데모는 기술 가능성의 확인이며, 한 공정의 개선은 제한된 운영 성과입니다. 전사 확산과 수익성은 그 뒤에 따로 증명해야 합니다.

제조 AI의 증거는 대체로 다음 순서로 쌓입니다.

1. 같은 조건에서 기존 방식보다 품질이나 시간이 나아지는가.
2. 숙련자와 설비가 다른 현장에서도 결과가 반복되는가.
3. 유지보수와 검토 인력을 포함한 총비용이 감당 가능한가.
4. 고장과 오판 때 생산을 안전하게 멈추거나 수동으로 전환할 수 있는가.

이 네 관문을 통과하지 않은 수치는 목표나 초기 결과로 읽어야 합니다. “무인화” 같은 표현도 사람의 완전한 퇴장을 뜻하는지, 원격 감독과 예외 대응으로 역할이 바뀐다는 뜻인지 구분할 필요가 있습니다.

그런데 네 관문을 다 통과한 해결책이라면 동네 부품 공장에도 그대로 들어갈까요? 대기업의 방식은 중소 제조업체에 그대로 복사하기 어렵습니다. 전담 데이터팀이 없고, 설비가 제각각이며, 몇 시간의 중단도 납기를 흔들 수 있기 때문입니다. 자체 모델이나 대형 GPU보다 기존 센서와 생산관리 시스템에 연결되는 작은 솔루션이 먼저 필요할 수 있습니다.

확산 정책도 구축 기업 수만 세어서는 부족합니다. 다음 조건을 함께 만들어야 합니다.

- 설비와 소프트웨어 사이의 데이터 형식과 연결 규격
- 업종별로 재사용할 수 있는 검증 데이터와 평가 기준
- 초기 설치 뒤에도 운영할 수 있는 현장 인력과 지역 지원 조직
- 장애와 보안 사고를 감당할 보험, 계약, 책임 구조
- 특정 공급자에 묶이지 않는 데이터 반출과 교체 조건

대기업과 협력사의 데이터 관계도 중요합니다. 공정 데이터를 모두 가져가는 대신 개선 결과와 모델을 협력사가 다시 쓸 수 있어야 확산이 일회성 납품으로 끝나지 않습니다. 데이터 사용 목적, 보유 기간, 영업비밀 보호, 개선 이익의 배분을 계약에 넣어야 하지요.

이렇게 대기업과 협력사를 잇는 일은 한국이 남보다 유리하게 해낼 수 있는 일이기도 합니다. 한국에는 반도체, 자동차, 조선, 전자, 철강이 가까이 모여 있습니다. 메모리를 공급하는 기업과 AI를 공정에 적용하는 기업, 장비와 로봇을 만드는 기업이 같은 산업 생태계 안에 있다는 뜻입니다.

물론 이 이점이 저절로 경쟁력이 되지는 않습니다. 공정별 성공을 공통 규격과 평가 방법으로 바꾸고, 협력사가 감당할 만한 가격과 운영 방식으로 낮춰야 하지요. 대기업의 발표를 열 개 모으는 것보다, 한 업종에서 검증된 해결책이 다른 공장에서도 되풀이되는 장면 하나가 더 값집니다.

K-제조 AX의 승부는 번쩍이는 무인 공장보다, 숙련공의 감을 데이터로 옮기고 그 성과를 협력사까지 퍼뜨리는 데서 갈립니다.

1. HD현대, "미래형 조선소(FOS)" 관련 공식 발표 (https://www.hd.com/kr/newsroom/media-hub/press/view?detailsKey=3829)

2. 삼성중공업, "스마트 조선소" 관련 공식 자료 (https://blog.samsungshi.com/m/863); "로봇 배관공장" 관련 공식 자료 (https://blog.samsungshi.com/906)

3. Hanwha Ocean, "Hanwha Ocean Brochure 2025" (https://www.hanwhaocean.com/pdf/Hanwha_Ocean_Brochure_KR_2025.pdf)

4. "삼성전자, 엔비디아와 함께 '업계 최고 수준 반도체 AI 팩토리' 구축" (삼성전자 뉴스룸, 2025.10.31, https://news.samsung.com/kr/%EC%82%BC%EC%84%B1%EC%A0%84%EC%9E%90-%EC%97%94%EB%B9%84%EB%94%94%EC%95%84%EC%99%80-%ED%95%A8%EA%BB%98-%EC%97%85%EA%B3%84-%EC%B5%9C%EA%B3%A0-%EC%88%98%EC%A4%80-%EB%B0%98%EB%8F%84%EC%B2%B4) - 향후 수년간 NVIDIA GPU 5만 개 이상을 도입한다는 계획 확인.

5. 「정부, 2030년까지 제조AI에 20조 투입…'100조 부가가치' 청사진 마련」 (머니투데이, 2026.6.29, https://www.mt.co.kr/article/2026062914212298001) — 2030년 제조업 AI 활용률 40%, 생산성 50% 향상, 민관 20조 원.

특화 파운데이션 모델: 왜 병원과 은행은 각자의 AI가 필요한가

전기에도 종류가 있습니다.

가정에서 쓰는 220V, 공장에서 쓰는 380V, 반도체 클린룸에 공급되는 초고순도 전력. 겉보기에는 다 같은 전기입니다. 하지만 반도체 공장에 가정용 전기를 넣으면 나노미터 단위의 회로가 망가지고, 병원에 순간 정전이 일어나면 수술대 위의 환자가 위험해집니다.

AI도 마찬가지입니다. GPT나 Claude 같은 범용 파운데이션 모델은 "가정용 전기"에 가깝습니다. 이메일을 쓰고, 보고서를 요약하고, 코드를 짜는 데는 훌륭합니다. 하지만 의사가 CT 영상에서 3밀리미터 결절을 찾아야 한다면? 트레이더가 실시간 시장 데이터에서 이상 신호를 포착해야 한다면? 산업마다 전혀 다른 "지능의 품질"이 필요합니다.

앞에서 파운데이션 모델을 발전소에 비유했습니다. 범용 파운데이션 모델이 "대형 종합 발전소"라면, 특화 파운데이션 모델은 "산업별 맞춤 발전소"인 셈이지요. 공장마다 필요한 전기가 다르듯, 산업마다 필요한 지능도 다릅니다.

범용 모델로 충분하지 않은 때

특화 모델이 필요한 까닭은 산업마다 입력 데이터와 실패 비용이 다르기 때문입니다. 의학 논문의 전문 용어, 금융 데이터의 실시간 변동, 공장 센서의 시계열 신호, 분자의 3차원 구조는 일반적인 인터넷 텍스트와 성격이 다릅니다.

그렇다고 산업 이름이 붙은 모델이 자동으로 더 안전한 것은 아닙니다. 도메인 데이터로 학습했다는 사실, 벤치마크에서 높은 점수를 받은 사실, 실제 현장에서 쓸 수 있다는 판단, 규제 승인을 받았다는 사실은 서로 다른 단계입니다. 아래 사례도 이 구분을 염두에 두고 읽어 보세요.

가장 먼저 들여다볼 곳은 병원입니다. 생명을 다루는 곳이니 초고순도 전력이 필요하지요. 의료는 특화 파운데이션 모델 연구가 가장 활발한 영역입니다. 틀린 답이 사람의 생명과 직결되기 때문이지요.

Google Research의 Med-PaLM 2는 의학 문답 벤치마크에서 높은 점수를 기록했습니다.[1] 하지만 시험문제 정답률은 환자 진료의 안전성을 뜻하지 않습니다. 실제 사용에는 환자 집단별 성능, 근거 제시, 개인정보 처리, 의료진 검토와 규제 상태를 별도로 확인해야 합니다.

의학 데이터로 학습한 모델도 존재하지 않는 약이나 그럴듯한 진단 근거를 만들 수 있습니다. 특화 학습은 위험을 줄이기 위한 한 수단이지 임상적 안전성을 보증하는 인증서가 아닙니다.

병원에서 은행으로 옮겨 가면 전기의 성격이 또 달라집니다. 금융 데이터는 특별합니다. 초 단위로 변하고, 규제에 묶여 있고, 보안이 생명입니다. 가정용 전기보다는 산업용 고압 전력에 가깝지요.

BloombergGPT와 FinGPT는 서로 다른 방식으로 금융 언어와 데이터를 모델에 넣으려 한 사례입니다. 전자는 대규모 금융 말뭉치로 학습했고, 후자는 기존 공개 모델을 금융 과제에 맞게 조정했습니다. 비용 숫자만 나란히 놓으면 데이터 범위와 평가 과제가 다른 두 실험이 같은 제품을 만든 것처럼 보이므로 단순 비교할 수 없습니다.

금융에서는 모델 크기보다 데이터 시점, 출처 추적, 시장조작과 이해상충 통제가 중요합니다. 신용이나 대출 결정에 쓰려면 결과 설명과 이의제기, 차별 점검도 필요합니다.

공장으로 가면 이번에는 디지털 트윈과 짝을 이루는 산업용 전력이 필요합니다. 산업용 모델은 설비 매뉴얼과 PLC 코드, 3D CAD, 시계열 센서처럼 서로 다른 형식의 데이터를 함께 다루려 합니다. Siemens가 추진하는 산업용 파운데이션 모델도 이런 방향입니다. 물론 발표된 플랫폼의 범위와 특정 공장의 실적, 여러 고객사의 평균 성과를 구분해서 봐야 합니다.

핵심은 디지털 트윈(Digital Twin)과 특화 FM의 결합입니다. 실제 공장의 가상 복제본에서 개선안을 시뮬레이션하고, 검증이 끝나면 현실에 적용하는 폐쇄 루프(closed-loop). 이것이 "자율 공장"의 기반입니다.

글자가 아닌 세계를 읽는 모델

병원과 은행, 공장의 모델이 그래도 사람이 쓴 기록에서 출발한다면, 과학 모델은 단백질 구조, 결정 구조, 기상장처럼 텍스트가 아닌 물리적 데이터를 직접 다룹니다. GNoME가 많은 후보 결정 구조를 제안한 연구는 그 가능성을 보여줬습니다.[2] 여기서 모델의 "발견"은 실험실 검증이 끝난 물질의 발견과 같지 않습니다. 모델은 탐색 공간을 줄이고 후보의 우선순위를 매기며, 최종 확인은 계산과 실험을 거칩니다.

실험실 밖, 우리 손목 위에도 우리가 아직 잘 의식하지 못하는 특화 FM이 하나 있습니다. 시계열 데이터를 이해하는 AI입니다.

TimesFM 같은 시계열 모델은 텍스트 대신 수치의 연속을 처리합니다. 주가, 공장 센서의 진동, 네트워크 트래픽처럼 시간 순서가 의미를 만드는 데이터가 대상입니다.

스마트워치가 "오늘 걸음 수 8,000보"를 세는 것과 여러 주의 수면과 심박 변화에서 패턴을 찾는 것은 다른 과제입니다. 후자의 예측이 건강 진단으로 쓰이려면 기기·연령·질환별 검증이 추가로 필요합니다.

지금까지의 특화 FM은 모두 "이해하고 예측하는" 모델이었습니다. 하지만 로봇 특화 FM은 한 단계 더 나아갑니다. 물리 세계에서 직접 행동합니다.

Physical Intelligence의 Pi-Zero 같은 VLA 모델은 이미지와 언어 명령을 로봇 동작으로 연결합니다.[3] 논문에서 여러 로봇과 과제를 다룬 성과는 연구 범위를 넓혔다는 뜻이지, 처음 보는 공장이나 수술실에서

자율적으로 안전하게 작동한다는 뜻은 아닙니다. 물리적 행동은 소프트웨어 출력보다 되돌리기 어려우므로 기능 안전과 현장 시험, 중단 장치가 별도로 필요합니다.

로봇이 안전하게 움직이려면 먼저 세상이 어떻게 돌아가는지 알아야 하겠지요. 특화 FM 중에서 가장 야심 찬 것이 월드 모델(World Model)입니다. 텍스트를 이해하는 LLM, 이미지를 이해하는 VLM과 달리, 물리 세계의 법칙 자체를 이해하려 합니다.

월드 모델은 행동의 결과를 예측하는 내부 표현을 학습하려는 연구 방향입니다. 로봇이 움직이기 전에 "이렇게 하면 어떻게 될까"를 시뮬레이션하는 데 쓰일 수 있지요. World Labs와 NVIDIA Cosmos 같은 프로젝트가 서로 다른 방식으로 이 문제를 다루고 있지만, 물리 법칙을 사람처럼 이해한다고 단정할 단계는 아닙니다.

지능 에너지의 전압 변환

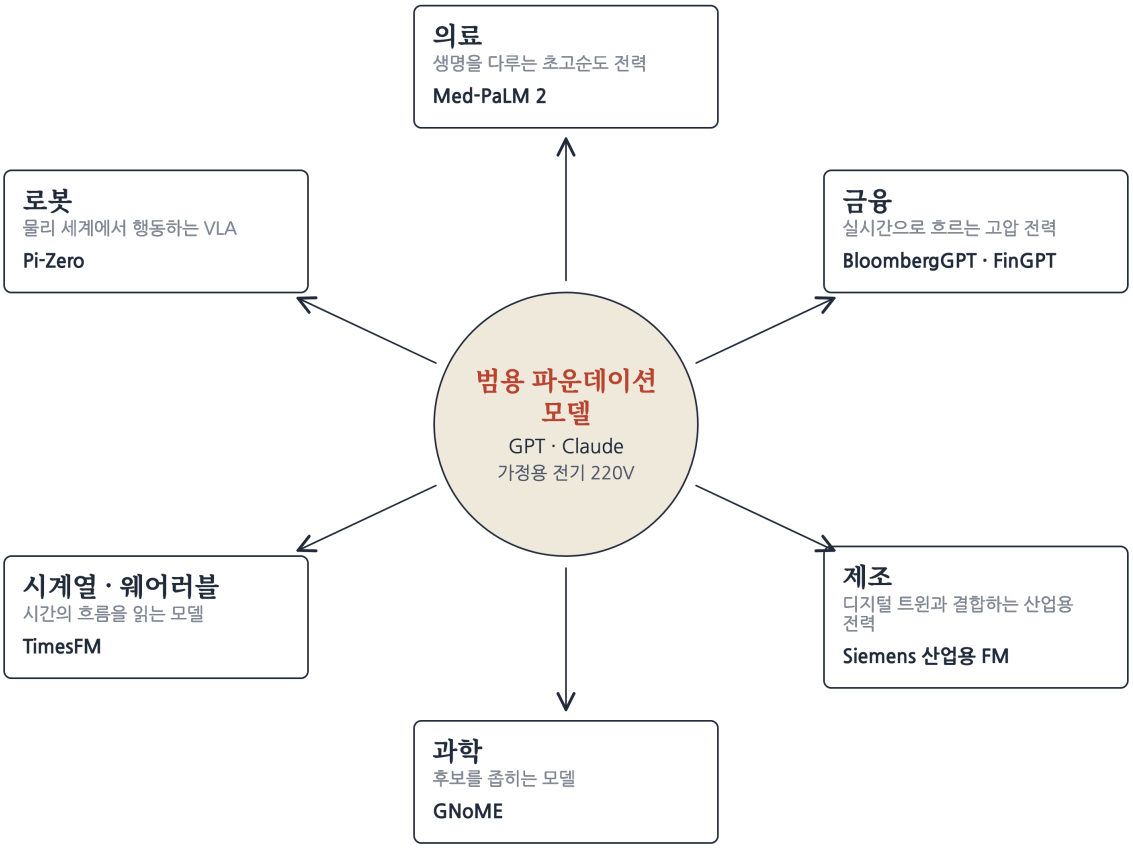
종합 발전소(범용 FM)에서 만든 전기는 각 산업 현장에 도착하면 변압기를 거쳐야 합니다. 반도체 공장에는 알맞은 전압으로, 병원에는 끊김 없이 들어가도록 변압기와 무정전 설비를 거쳐야 하지요.

특화 FM의 세계에서 이 "변압기"에 해당하는 것이 파인튜닝(fine-tuning)과 RAG(검색 증강 생성)입니다. 범용 모델을 도메인 데이터로 조정하거나, 검증된 자료를 검색해 답변 근거로 붙이는 방식입니다.

하지만 변압기만으로는 한계가 있는 영역도 있습니다. 의료 영상, 분자 구조, 로봇 제어, 대기 역학은 애초에 텍스트 기반 범용 모델이 본 적 없는 데이터입니다. 이런 영역에서는 처음부터 도메인 데이터로 사전학습한 "맞춤 발전소"가 필요합니다. 도메인 특화 LLM 서베이[4]는 지식 주입의 네 가지 접근법을 정리합니다. 동적 주입(RAG), 정적 임베딩(도메인 사전학습), 모듈식 어댑터(LoRA 등), 프롬프트 최적화이지요. 상황에 따라 적절한 변압기를 고르는 것이 핵심입니다.

특화 파운데이션 모델

같은 지능 에너지도 산업마다 전압이 다릅니다 — 범용 모델이라는 종합 발전소에서, 현장에 맞춘 산업별 모델로



화살표 = 변압기 — 파인튜닝·RAG로 범용 전력을 산업별 전압으로 바꿉니다. 범용 모델이 본 적 없는 데이터(의료 영상, 분자 구조, 로봇 제어)는 처음부터 도메인 데이터로 학습한 '맞춤 발전소'가 필요합니다.

출처: 저자 작성 — 본문 제5부 「특화 파운데이션 모델」에 등장한 모델과 비유를 바탕으로 했습니다. 모델 이름은 사례이며 순위나 성능 비교가 아닙니다.

그림 5-7 범용 모델에서 산업별 모델로 이어지는 화살표가 변압기, 곧 파인튜닝과 RAG입니다. 의료 영상이나 로봇 제어처럼 범용 모델이 본 적 없는 데이터를 다루는 곳에서는 처음부터 도메인 데이터로 지은 맞춤 발전소가 필요하지요.

그럼 실제로 "우리 업계 전용 AI"라는 제안서를 받았다고 해 볼까요. 이름보다 먼저 물어볼 것이 네 가지 있습니다. 벤치마크 점수는 어떤 데이터와 정답 기준에서 나왔는지, 현장 실증에 실제 사용자와 예외 상황이 들어 있었는지, 상용 운영 중의 오류와 비용, 장애 기록을 볼 수 있는지, 그리고 필요한 승인과 감사, 이의제기 절차가 갖춰져 있는지이지요. 어떤 경우에는 범용 모델에 RAG를 붙이는 것으로 충분하고, 어떤 경우에는 새로운 데이터 형식으로 처음부터 학습한 모델이 필요합니다. 성능 향상이 작거나 운영 데이터가 부족하다면 특화 모델을 별도로 만드는 비용이 더 클 수도 있습니다.

맞춤 발전소를 지을지 고민된다면, 기존 전기를 조금 손보는 것으로 충분한지부터 따져 보세요. 산업에 맞는 지능은 모델 이름보다 데이터와 평가, 책임 체계가 맞물릴 때 나옵니다.



1. "Toward expert-level medical question answering with large language models" (PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC11922739/)
2. "Scaling deep learning for materials discovery" (Nature, 2023, https://www.nature.com/articles/s41586-023-06735-9)
3. "Pi-Zero: A Vision-Language-Action Flow Model for General Robot Control" (arXiv:2410.24164, https://arxiv.org/abs/2410.24164)
4. "Survey of Specialized Large Language Model" (arXiv:2508.19667, https://arxiv.org/abs/2508.19667)

대한민국의 독자 파운데이션 모델 계획: 선언이 아니라 청사진

세계가 산업별 맞춤 발전소를 저마다 짓는 동안, 한국은 어떤 발전소를 짓고 있을까요?

대한민국 AI 행동계획에는 구체적인 독자 파운데이션 모델 계획이 담겨 있습니다. "AI를 잘 활용하겠다"는 다짐에서 한 걸음 더 들어가, 어떤 부처가 어떤 일정에 어떤 모델을 만들고, 어떤 데이터를 연료로 쓰고, 어떤 법률을 정비하겠다는 실행 청사진입니다.[1] 미리 일러두면, 이 장에는 부처 이름과 과제 번호가 줄지어 나옵니다. 하나하나 외울 필요는 없습니다. 어떤 그림이 그려지고 있는지만 잡으면 충분하지요.

지능 발전소를 우리 손으로

핵심은 과제 [27]-1입니다. 과기정통부는 2027년 4분기까지 국제적 경쟁력을 갖춘 독자 AI 파운데이션 모델을 순차적으로 확보하고, 이를 국가적 AX(AI 전환) 사업의 기본 모델로 활용하겠다고 명시하고 있습니다.

이 계획은 경쟁으로 굴러가고 있습니다. 2026년 8월 2차 평가에서 업스테이지, SK텔레콤, LG AI연구원 세 팀이 3차에 올랐고, 정부는 2027년 2월 최종 두 팀을 고를 계획입니다. 성적표가 마냥 밝지는 않았습니다. 정부가 목표로 잡은 '세계 최고 모델의 95%'에 닿은 팀이 아직 없었고, 지원을 한 곳에 몰아주자는 논의까지 나왔지요.[2]

한마디로 '지능 발전소를 우리 손으로 짓겠다'는 계획입니다. 지금까지 한국의 AI 활용은 대부분 해외 발전소에서 전기를 수입해 쓰는 구조였습니다. 물론 훌륭한 전기입니다. 하지만 국가의 핵심 인프라를 전적으로 수입 전력에 의존하는 것은 위험합니다. 공급이 끊기거나 가격이 급등하면? 데이터 주권은? 한국어와 한국 문화에 대한 이해 수준은?

여기에 중요한 법적 기반도 함께 설계되어 있습니다. 과제 [32]-2는 AI 학습 목적의 저작물 활용을 가능하게 하는 「AI기본법」 개정안 혹은 「AI특별법」 제정안을 2026년 4분기까지 마련하겠다고 합니다. 발전소와 연료 공급망을 한꺼번에 설계하는 셈이지요.

그렇다면 이 종합 발전소 하나로 모든 산업을 다 밝힐 수 있을까요? 독자 FM이 여러 산업에 전력을 공급하는 "종합 발전소"라면, 산업 특화 FM은 각 산업의 요구에 맞춘 "맞춤 발전소"입니다. 산업부는 제조, 의료·바이오, 금융, 모빌리티, 미디어·콘텐츠 5대 우선 분야를 지정하고, 모델 개발의 전 과정을 파이프라인으로 설계했습니다.

단계	내용	일정
① 데이터 확보	산업별 민관 합동 데이터 협의체 구성	'26Q1
② 표준화	데이터 표준화 및 고품질 데이터셋 구축 로드맵	'26Q2
③ 컴퓨팅 배분	국가 AI 컴퓨팅 자원 할당	'26Q3~
④ 실증·확산	'AI 솔루션 실증 지원센터' 지정, 중소기업 컨설팅·실증 통합지원	지속
⑤ 품질 인증	'AI 솔루션 품질 평가 인증제도' 마련	'27Q2

(관련 과제 번호와 일정은 이 장 끝의 로드맵에 모았습니다.)

데이터 확보 → 표준화 → GPU 할당 → 실증 → 인증. 국가가 산업별 맞춤 발전소를 체계적으로 세우는 과정입니다. 범용 모델 하나를 모든 현장에 그대로 넣는 대신, 산업 데이터와 품질 기준을 처음부터 반영하겠다는 것입니다.

제조 분야는 목표가 더 구체적입니다. 과제 [37]-1에 따라 정부는 2026년 6월 「제조AI 2030 전략」을 내놓았습니다. 2030년까지 민관이 20조 원을 들여 제조업의 AI 활용률을 40%로 끌어올리고 생산성을 50% 높이겠다는 계획이지요.[3] 국가 제조데이터 라이브러리, 제조 특화 파운데이션 모델과 '풀스택 AI 팩토리', 산업단지를 AI 산단으로 바꾸는 일이 큰 줄기이며, 자율 제조 테스트베드 3개소는 2026년 2분기에 착수하기로 했습니다([37]-3). 디지털 트윈과 특화 FM의 결합, 가상에서 실험하고 현실에 적용하는 폐쇄 루프의 한국판 시도입니다.

생명과 날씨, 국방과 교실을 맡을 발전소

공장 다음으로 눈여겨볼 곳은 생명을 다루는 발전소입니다. 바이오와 의료 분야의 계획은 여러 부처가 협력하는 다층적 구조라는 점에서 눈여겨볼 만합니다.

출발점은 데이터입니다. 복지부는 보건의료 특화 FM 개발을 위해 보건의료 빅데이터 플랫폼의 활용 방안을 2026년 1분기까지 마련합니다([6]-2). 한국의 국민건강보험이라는 단일 보험자 시스템에 축적된 전 국민 의료 데이터가, 의료 FM의 연료로서 엄청난 잠재력을 가진 자원이지요.

바이오 FM은 유전체·범분자·세포 등 생명 계층의 핵심 축에 집중하고([29]-3), '멀티스케일 바이오 파운데이션 모델' 로드맵을 2026년 1분기까지 수립해 AI 신약 개발 사업과 연계합니다([29]-4). 분자 수준의 작용이 세포와 조직을 거쳐 환자에게 미치는 효과까지 연결해 이해하는 AI를 지향하는 것입니다.

가장 인상적인 대목은 따로 있습니다. 복지부는 보건의료 특화 FM([30]-1)을 오픈소스로 공개하고, 책임성·안전성·형평성·비편향성이라는 가치를 개발 초기부터 반영하겠다고 명시했습니다. 존재하지 않는 약 이름을 말하는 환각도, 인종이나 성별에 편향된 진단도 막아야 하는 영역이기 때문입니다. 안전장치를 나중에 덧대기보다 설계 단계부터 녹여 넣겠다는 접근이지요.

사람의 몸에서 눈을 들어 하늘을 보면, 한반도의 날씨를 예측하는 전용 발전소도 설계되고 있습니다. 과제 [29]-6은 무려 6개 부처 — 과기정통부, 우주청, 기상청, 국토부, 해수부, 기후부 — 가 합동으로, 대기·해양·육상·해빙·생태·우주기상 데이터를 통합 학습하는 글로벌 수준의 전 지구시스템 파운데이션 모델과 한반도 특화 모델을 개발합니다.

전 지구시스템 FM이 "지구 전체를 이해하는 종합 발전소"라면, 한반도 특화 모델은 지역 조건에 맞춘 "맞춤 발전소"입니다. 기상청은 한국형 AI 기상·기후 예측 FM을, 해수부는 우리나라 주변 해역과 북극항로를 아우르는 해양 예측 시스템을 개발합니다. 대륙과 해양의 경계, 복잡한 산악 지형, 잦은 장마·태풍·한파처럼 글로벌 모델이 세밀하게 반영하기 어려운 조건을 메우려는 시도이지요.

가장 강력하면서도 가장 엄격해야 하는 발전소는 어디일까요? 물리 세계에서 움직이는 AI와 국방입니다. 과제 [12]-1은 과기정통부와 산업부가 협력하여 피지컬AI 1위 달성을 2030년까지 이루겠다는 목표입니다. 월드모델, 로봇 파운데이션 모델, 고성능 AI반도체 HW 플랫폼이 핵심 기술이고, 휴머노이드 로봇 3강 달성도 2028년 목표입니다.

특히 눈에 띄는 것은 국방 FM입니다([79]-1). 국방부는 과기정통부와 협력해 2027년 4분기까지 독자 FM 기반의 국방 AI 파운데이션 모델을 개발하는데, 다른 산업 FM과 결정적으로 다른 점은 시험·인증 프로세스와 오류 발생 시 신속 대응하는 롤백 체계를 함께 요구한다는 것입니다. 전투기 전자 시스템이나 지휘통신망처럼, 잠깐의 오작동도 돌이킬 수 없는 곳이니까요. 독자 FM 기반이라는 점도 중요합니다. 핵심 지능 인프라만큼은 국산이어야 하는 영역입니다. 첫 삽은 이미 떴습니다. 2026년 5월 국방부는 독자 FM 정예팀인 SK텔레콤과 손잡고 GPU 약 3,000장을 배정받아, 범용 모델을 가볍게 줄여 군 현장에서 시험하는 일부터 시작했지요.[4]

발전소를 지어도 전기가 각 가정과 공장에 전달되지 않으면 소용이 없습니다. 교육부는 '국가 AI 교육 연구센터'를 설립하고, 국가 FM 기반의 교육 AI 연구 체계를 2026년 3분기부터 지원합니다([95]-2). AI 보조교사(튜터)와 교무행정 AI도 국가 FM과 연계해 개발하며, 2027년 4분기에 프로토타입을 공개합니다([60]-1).

마지막으로 산업 에이전트입니다([36]-1). 분야별 특화 FM을 기반으로 '산업 특화 AI 에이전트'를 개발·보급해, 단순 반복 업무의 자동화를 넘어 전문적 의사결정 지원과 새로운 부가가치 창출까지 나아가겠다는 것이지요.

에너지 비유로 보면 종합 FM과 산업 특화 FM은 규모와 용도가 다른 발전소이고, AI 에이전트는 현장에서 실제로 일하는 기기입니다. 어떤 산업 데이터를 학습했고 어떤 품질 기준을 통과했느냐에 따라 에이전트가 할 수 있는 일도 달라집니다.

한 장의 도면으로 보는 독자 FM 로드맵

계획의 층위와 확인할 증거를 표로 나누면 다음과 같습니다.

분야	주관 부처	과제	핵심 목표	일정
독자 FM 총괄	과기정통부	[27]-1	국제 경쟁력 독자 FM 확보	'27Q4
산업 특화 FM	산업부+6개 부처	[28] 시리즈	5대 분야(제조·의료·금융·모빌리티·미디어) R&D	'26Q1~
제조 AI	산업부+과기정통부	[37]-1,3	제조 AI 활용률 40%, 테스트베드 3개소	'26Q1~'30
바이오/의료 FM	과기정통부+복지부	[29]-3,4, [30]-1	멀티스케일 바이오 FM, AI 신약, 오픈소스	'26Q1~
보건의료 데이터	복지부	[6]-2	빅데이터 플랫폼 활용 방안	'26Q1
기상·기후 FM	기상청	[29]-7	한국형 AI 기상·기후 예측 FM	'26Q1 착수
해양 예측	해수부	[29]-8	해양 예측 시스템, 북극항로	'26Q2 착수
지구시스템 FM	6개 부처 합동	[29]-6	대기·해양·육상·해빙·생태·우주기상 통합	'27Q1 전략
Physical AI	과기정통부+산업부	[12]-1	월드모델, 로봇 FM, 피지컬AI 1위	'30
국방 FM	국방부+과기정통부	[79]-1	독자 FM 기반, 시험/인증, 롤백	'27Q4
교육 FM	교육부	[95]-2, [60]-1	AI 튜터, 교무행정 AI, 국가 FM 연계	'27Q4
산업 에이전트	산업부+과기정통부	[36]-1	특화 FM 기반 AI 에이전트 개발·보급	지속
법제 정비	과기정통부	[32]-2	AI 학습 저작권법 개정/AI특별법	'26Q4

출처: 대한민국 AI 행동계획(2026-2028) 과제와 일정을 바탕으로 저자 정리.[1]

표를 한 번 훑어보면 알 수 있듯이, 한국의 독자 FM 계획에는 발전소 이름만 적혀 있지 않습니다. 산업별·부처별로 어떤 모델을 어떤 일정에 만들고, 어떤 데이터를 쓰며, 어떤 법률과 인증을 갖출지까지 그려 두었지요. 종합 FM에서 산업별 맞춤 FM, 교육과 산업 에이전트로 이어지는 하나의 체계입니다.

물론 계획이 곧 실행은 아닙니다. 2027년 4분기까지 국제 경쟁력을 갖춘 독자 FM을 확보한다는 목표는 야심 찹니다. 데이터 확보, 인재 유치, 컴퓨팅 자원 배분, 부처 간 협력 — 넘어야 할 산이 많습니다.

그래도 기대를 걸 만한 이유가 있습니다. 과제 번호와 담당 부처, 완료 시한이 적혀 있다는 것이지요. 구체적이면 추적할 수 있습니다. 계획대로 가고 있는지, 어디서 늦어지는지 밖에서도 확인할 수 있으니까요. 실제로 국가AI전략위원회는 2026년 5월 첫 분기 점검에서 326개 과제 가운데 288개가 계획대로 가고 있다고 밝혔습니다.[5] 나머지 열 개 중 하나꼴이 어디서 막혔는지가 다음에 물어야 할 질문이지요.

앞 장이 "왜 산업마다 다른 발전소가 필요한가"를 보여 줬다면, 이 장에서 본 것은 한 나라가 그 발전소들을 설계도 한 장으로 엮으려는 시도입니다. 종합 발전소에서 산업별 맞춤 발전소, 품질 인증, 법적 기반, 교육과 산업 에이전트까지 과제 번호와 시한을 붙여 한 장의 도면에 담았지요. 한국 계획의 강점은 바로 그 구체성입니다.

이제 남은 일은 도면대로 세우고, 그 진행을 분기마다 눈으로 확인하는 것입니다. 가까운 눈금은 2027년 2월 독자 FM 최종 선정이고, 그다음은 2027년 4분기, 과제 [27]-1의 기한입니다.

---

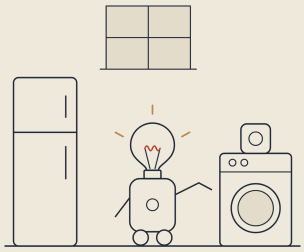
1. 국가인공지능전략위원회, 「대한민국 인공지능 행동계획(인공지능 기본계획 2026-2028)」 (https://www.aikorea.go.kr/upload/contents/contents_01.html); 계획 원문 PDF (https://smartcity.go.kr/wp-content/uploads/2026/03/%EC%95%88%EA%B1%B41%EB%8C%80%ED%95%9C%EB%AF%BC%EA%B5%AD%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5%ED%96%89%EB%8F%99%EA%B3%84%ED%9A%8D%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5%EA%B8%B0%EB%B3%B8%EA%B3%84%ED%9A%8D20262028.pdf)

2. 「연장전 된 2차전…글로벌 모델 대비 열위 우려 속 독파모 재편 논의」 (머니투데이, 2026.8.18, https://www.mt.co.kr/tech/2026/08/18/2026081814570930924) — 2차 단계평가에서 업스테이지·SK텔레콤·LG AI연구원 3차 진출, 지원 방식 재편 논의.

3. 「정부, 2030년까지 제조AI에 20조 투입…'100조 부가가치' 청사진 마련」 (머니투데이, 2026.6.29, https://www.mt.co.kr/article/2026062914212298001) — 2030년 제조업 AI 활용률 40%, 생산성 50% 향상, 민관 20조 원.

4. 과기정통부·국방부·SK텔레콤 국방 AI 협약 보도 (M이코노미뉴스, 2026.5.14, https://www.m-economynews.com/mobile/article.html?no=67187) — 정부 확보 GPU 가운데 약 3,000장을 국방에 배정.

5. 국가인공지능전략위원회 1분기 이행 점검 결과 보도 (2026.5.14, https://v.daum.net/v/20260514091854937) — 326개 과제 중 288개(88.3%) 정상 추진.



제6부

지능 에너지와 에이전트가 만드는 세상

개인 에이전트의 탄생

지금까지 우리는 먼 곳의 이야기를 했습니다. 4차 AI산업혁명이라는 거대한 파도, 지능 에너지 시스템이라는 새로운 인프라, 기업의 AX 전환, 설계 원칙, 산업별 혁명. 반도체 공장에 GPU 5만 개 이상을 들이겠다는 계획, 조선소를 디지털 트윈으로 통째로 옮기겠다는 구상, 진료실에서 AI가 대화를 받아 적는다는 이야기였지요.

거대한 이야기였습니다. 그런데 한 가지 빠진 것이 있습니다.

"그래서 나는요?"

내 하루는 어떻게 바뀌는 걸까요? 대기업이 GPU를 몇만 개 들인다는 건 알겠는데, 내일 아침 내가 눈을 떴을 때 달라지는 것은 무엇일까요?

이 질문에 답하려면 '개인 에이전트'라는 개념을 이해해야 합니다.

나만의 AI

개인 에이전트란 나를 위해 일하는 AI입니다. 회사의 AI가 아닙니다. 플랫폼의 AI도 아닙니다. 나만의 AI입니다.

이 구분이 중요합니다. 지금까지 우리가 써온 AI 서비스 대부분은 플랫폼의 것이었습니다. 유튜브 알고리즘은 나를 위해 영상을 추천하는 것 같지만, 실제로는 내가 더 오래 머물도록 설계되어 있지요. 인스타그램의 추천 피드도 마찬가지입니다. 내 관심사를 반영하는 것 같지만, 궁극적으로는 광고주의 이익을 위해 작동합니다.

개인 에이전트가 지향하는 관계는 다릅니다. 이 AI의 고용주를 플랫폼이 아니라 나로 두는 것입니다. 현재 제품들은 이메일 초안, 일정 조회, 취향 기억 같은 기능을 부분적으로 제공하지만, 서비스마다 접근 범위와 기억 방식이 다르고 중요한 행동에는 여전히 사람의 확인이 필요합니다.

비유를 하자면 이렇습니다. 전기가 가정에 들어온 뒤 사람들은 자연스럽게 "내 냉장고, 내 세탁기, 내 텔레비전"을 갖게 되었습니다. 공장에만 있던 전기가 각 가정으로 흘러들어온 것이지요. 지능 에너지 시대에도 같은 일이 일어납니다. 기업에만 있던 AI가 개인에게로 흘러들어옵니다. "내 에이전트"를 갖게 되는 것입니다.

그렇다면 '내 에이전트'가 곁에 있는 하루는 어떤 모습일까요? 구체적으로 상상해봅시다. 아침에 눈을 뜹니다.

밤사이 이메일이 47통 쌓여 있습니다. 예전이라면 출근길에 스마트폰을 꺼내 하나씩 훑어봐야 했을 겁니다. 스팸인지, 중요한 건지, 답장이 필요한 건지. 그 분류 작업만 해도 30분이 걸리지요.

개인 에이전트가 충분한 권한을 갖춘 미래를 가정하면 다릅니다. 밤새 47통을 분류해 광고성 메일은 따로 모으고, 회의 요청은 내 일정과 비교해 답장 초안을 만들 수 있습니다. 물론 실제 발송이나 일정 변경은 내가 확인한 뒤에 하지요. 잘못 분류한 메일 한 통이 거래를 놓치게 할 수도 있으니, 편리함만큼 되돌리기와 기록이 중요합니다.

뉴스도 달라질 수 있습니다. 스마트폰을 끝없이 스크롤하는 대신 에이전트가 업무와 관심사에 맞춘 짧은 브리핑을 만들 수 있지요. 하지만 개인 에이전트라고 해서 저절로 내 편이 되는 것은 아닙니다. 검색 순위와 제휴 관계, 운영사의 수익모델이 추천에 영향을 줄 수 있으므로 출처와 추천 기준을 확인할 장치가 필요합니다.

아침의 시작이 "오늘 뭐 해야 하지?"에서 "오늘 이렇게 하면 어떨까요?"로 바뀝니다. 작은 변화 같지만, 하루의 심리적 출발점이 완전히 달라지는 것입니다.

일하고, 배우고, 사는 방식

회사에 도착합니다. 오전에 팀 회의가 있는데, 에이전트가 미리 준비를 해놓았습니다. 지난 회의 내용을 요약하고, 그때 남겨둔 숙제가 무엇이었는지 정리하고, 관련 데이터를 최신 버전으로 업데이트해서 한 페이지짜리 브리핑을 만들어놓은 것입니다. 예전에는 회의 시작 10분 전에 허겁지겁 노트를 뒤적이던 것과 비교하면 다른 세상입니다.

보고서를 쓸 때도 마찬가지입니다. 백지에서 시작하지 않습니다. 에이전트가 초안을 작성합니다. 관련 데이터를 찾아서 넣고, 지난 분기 보고서의 양식에 맞추고, 내가 자주 쓰는 표현으로 문장을 다듬어놓습니다. 내가 하는 일은 검토하고, 수정하고, 판단하는 것입니다. "이 숫자가 맞나? 이 해석이 적절한가? 이 결론을 경영진에게 보고해도 되나?"

초안 작성 시간이 줄면 검토의 비중이 커집니다. 개발자는 코드가 요구사항과 보안 기준을 지키는지 확인하고, 데이터 분석가는 수치의 출처와 해석을 따져야 합니다. 마케터도 여러 문안 가운데 브랜드와 고객에게 맞는 표현을 골라야 하지요.

검토는 저절로 되는 일이 아닙니다. 원자료를 확인하고, 틀린 초안을 되돌리고, 결론에 책임질 시간이 필요합니다. 에이전트가 처리한 비율만으로 생산성을 계산하면 이 비용이 사라져 보입니다. 개인 에이전트의 효과는 초안 속도뿐 아니라 사람이 다시 고친 시간과 놓친 오류까지 함께 재야 합니다.

에이전트가 바꾸는 것은 일하는 방식만이 아닙니다. 무언가를 익히는 방식도 달라지지요. 고등학교 교실을 떠올려봅시다. 선생님 한 명이 30명의 학생에게 같은 내용을 같은 속도로 가르칩니다. 어떤 학생은 이미 이해했는데 기다려야 하고, 어떤 학생은 아직 모르는데 넘어가야 합니다. 모두에게 맞추려다 아무에게도 맞추지 못하는 구조이지요.

개인 에이전트가 AI 튜터 역할을 하면 이야기가 달라집니다. 수학에서 이차방정식을 배우는데, 나는 인수분해는 잘하지만 근의 공식 적용에서 자꾸 실수를 합니다. 에이전트는 그것을 압니다. 인수분해 문제는 적게 내고, 근의 공식 문제를 다양한 유형으로 반복해서 냅니다. 그것도 내가 이해하는 방식에 맞춰서요. 시각적으로 이해하는 학생에게는 그래프로 설명하고, 절차적으로 이해하는 학생에게는 단계별 풀이를 보여줍니다.

"같은 수업을 30명이 듣는" 시대에서 "30명이 각자 다른 수업을 듣는" 시대로의 전환입니다.

직장인에게도 마찬가지입니다. 새로운 기술을 배워야 할 때, 에이전트가 나의 기존 지식 수준을 파악하고 거기서부터 시작하는 커리큘럼을 짭니다. 파이썬을 이미 알고 있다면 기초 문법을 건너뛰고 바로 데이터 분석 라이브러리부터 가르칩니다. 반대로 프로그래밍을 처음 접하는 사람이라면 변수가 무엇인지부터 차근차근 설명합니다. 같은 "파이썬 강의"인데 사람마다 완전히 다른 강의가 되는 것입니다.

일과 배움을 지나 저녁이 되면, 이번에는 지갑을 여는 장면이 기다립니다.

"여름 휴가 어디 갈까?"

예전이라면 이 질문 하나에 네이버 검색, 블로그 후기, 여행 카페, 예약 사이트를 이틀은 돌아다녀야 했습니다. 정보는 넘치는데 신뢰할 만한 정보를 추리는 것이 일이었지요.

개인 에이전트가 예산과 취향, 일정을 활용하면 더 맞춤화된 선택지를 제안할 수 있습니다. 그런데 광고비를 낸 업체와 제휴 상품을 우선하지 않는다는 보장은 계약과 설정으로 확인해야 합니다. 추천 이유와 대가 관계를 공개하고, 사용자가 다른 판매처를 비교할 수 있어야 비로소 ‘나를 위한 추천’에 가까워집니다.

더 나아가면 에이전트끼리 직접 소통하는 세상이 옵니다. 나의 에이전트가 식당 에이전트에게 "이 시간에 자리 있나요?"라고 물어보고, 항공사 에이전트에게 "이 날짜에 가장 저렴한 표가 뭔가요?"라고 확인합니다. 제2부에서 다뤘던 A2A(Agent-to-Agent) 프로토콜이 여기서 작동합니다. 서로 다른 회사의 에이전트끼리 표준화된 방식으로 대화하는 것이지요. 마치 서로 다른 나라 사람이 영어라는 공통 언어로 소통하듯이.

비교쇼핑도 에이전트가 도울 수 있습니다. 여러 쇼핑몰의 가격과 조건을 모으는 일은 빨라지겠지만, 광고성 추천을 어떻게 걸러낼지, 어느 판매처의 정보를 믿을지는 별도 문제입니다. 최종 결제 전에 사람의 확인이 남아야 하는 이유입니다.

솔직히, 걱정되는 것들

장밋빛 이야기만 할 수는 없습니다. 개인 에이전트라는 개념에는 진지하게 마주해야 할 질문들이 있습니다.

프라이버시부터 보지요. 나를 잘 아는 에이전트는 이메일 패턴, 소비 습관, 건강 정보, 대화 내용에 접근할 수 있습니다. 기기 안에서 실행하거나 암호화된 서버를 쓰더라도 수집 범위와 보관 기간, 삭제 권한을 따로 정해야 합니다. 기술이 있다고 문제가 저절로 해결되지는 않습니다.

의존성도 남습니다. 이메일 정리와 보고서 작성, 일정 관리를 계속 맡기면 직접 확인하는 능력이 약해질 수 있습니다. 모든 일을 직접 할 필요는 없지만, 계정이 잠기거나 서비스가 중단됐을 때 되찾아야 할 능력과 자료는 정해둘 필요가 있습니다.

접근 격차는 가격에서 끝나지 않습니다. 유료 모델과 좋은 기기를 살 수 있는지, 개인정보 설정과 결과 검증을 배울 기회가 있는지에 따라 효과가 달라집니다. 개인 에이전트의 보급 속도와 혜택이 모두에게 같을지는 가격, 기기, 교육 기회를 함께 봐야 합니다.

세 문제는 결국 기능보다 운영 조건을 묻습니다. 어떤 데이터를 내주고, 어떤 능력은 직접 유지하며, 비용과 교육 기회를 어떻게 나눌 것인가라는 조건입니다.

이런 걱정을 안고서라도, 이 변화를 책 전체의 큰 그림 위에 놓아 볼 필요가 있습니다. 돌아보면 지금까지 이 책에서 다룬 것은 지능 에너지의 발전소, 송전망, 그리고 그것을 활용하는 산업이었습니다. 파운데이션 모델이라는 발전소가 지능을 만들고, MCP와 A2A라는 송전망이 지능을 전달하고, 각 산업의 공장과 병원과 사무실이 그 지능을 소비합니다.

개인 에이전트는 이 흐름의 마지막 단계입니다. 지능 에너지가 산업을 넘어 개인에게까지 도달하는 것. 전기가 공장에서 시작해 결국 모든 가정의 콘센트로 흘러들어간 것처럼요.

개인 에이전트가 널리 퍼질 가능성은 큽니다. 물론 누구나 같은 형태의 에이전트를 갖게 된다고 단정할 수는 없습니다. 가격, 개인정보 보호, 디지털 역량, 플랫폼의 이해관계에 따라 보급 속도와 모습이 달라질 것입니다. 지금 확인할 수 있는 흐름은 AI가 단순한 대화창에서 개인의 맥락을 읽고 제한된 행동을 돕는 도구로 이동하고 있다는 데까지입니다.

뒷부분에서 2030년의 하루를 더 자세히 그려볼 것입니다. 아침에 눈을 떠서 밤에 잠들 때까지, 지능 에너지가 전기처럼 흐르는 하루를요. 그 하루를 이해하려면 먼저 에이전트가 만드는 세상의 구조를 알아야 합니다.

이제부터 제6부에서 그 세상이 어떻게 만들어지고 있는지 하나씩 따라가 볼까요. 한 사람이 만든 에이전트가 세상을 바꾸는 이야기부터 시작합니다.

한 사람의 에이전트에서 생태계로: OpenClaw와 Hermes

> 참고: 이 글은 2026년 상반기 보도와 공개 정보를 바탕으로 쓰고, 9월까지의 주요 후속 소식과 빅테크의 클라우드 개인 에이전트 출시를 더했습니다. 에이전트 생태계는 여전히 빠르게 움직이고 있으며, 구체적인 수치는 이후 달라질 수 있습니다.

개발자 Peter Steinberger는 2025년 말 여러 메시징 서비스와 업무 도구를 연결하는 오픈소스 에이전트 프로젝트를 공개했습니다. 이름이 여러 차례 바뀐 뒤 OpenClaw로 자리 잡았지요.

관심을 끈 건 모델보다 연결 범위였습니다. 이메일과 캘린더, 메신저, 브라우저에 접근해 여러 단계의 작업을 이어갈 수 있었습니다. 로컬에서 실행할 수 있다는 점도 주목받았지만, 연결된 외부 모델과 서비스에 어떤 데이터가 전달되는지는 구성마다 달랐습니다.

2026년 2월, Steinberger는 OpenAI에 합류했고 프로젝트는 오픈소스로 유지하겠다는 계획이 발표됐습니다.[1] 개인 프로젝트가 창업·인수 성공담이 되었다기보다, 유지 주체가 바뀐 뒤에도 공개 생태계를 지속할 수 있는지를 지켜봐야 하는 사례입니다.

그런데 그건 시작에 불과했습니다.

집에 가전제품 하나를 들였다고 생각해 보세요. 처음엔 그냥 편리한 물건 하나였습니다. 이메일을 읽어주고, 캘린더를 관리해주고, 메시지를 대신 보내주는 만능 기기 하나. 그런데 몇 달이 지나자 상황이 달라졌습니다. 경쟁 브랜드가 등장했습니다. 부품과 액세서리를 파는 상점이 생겼습니다. 대기업들이 이 가전제품을 자기네 제품군에 끼워 넣기 시작했습니다. 그리고 리콜 사태와 정부 규제까지 뒤따랐습니다. 가전제품 하나의 이야기가 어느새 가전 산업 전체의 이야기가 된 것이지요. 2026년 상반기 OpenClaw와 그 주변에서 벌어진 일이 딱 이렇습니다.

한 대의 가전에서 가전 산업으로

OpenClaw라는 이름을 얻기까지 꽤 여러 번 옷을 갈아입었습니다. 2025년 11월 24일 Warelay로 시작해서, CLAWDIS, Clawdbot, Moltbot을 거쳤습니다. Moltbot이라는 이름은 Anthropic의 상표 이의 제기로 오래 쓰지 못했습니다. 그러다 2026년 1월 30일에 지금의 이름 OpenClaw로 정착했지요.

이름이 여러 번 바뀌는 동안 프로젝트에 대한 관심과 파생 도구는 빠르게 늘었습니다. 물론 저장소의 스타와 포크는 관심의 크기이지, 활성 사용자 수나 안전한 운영 실적은 아닙니다.

오픈소스로 남기겠다던 약속은 2026년 7월 비영리 재단 출범으로 이어졌습니다. Steinberger와 Dave Morin이 함께 세운 OpenClaw 재단에는 OpenAI가 주요 후원자로, NVIDIA와 Microsoft 등 30여 개

조직이 참여자로 이름을 올렸지요.[2] 그런데 창시자가 여전히 OpenAI에 몸담고 있고 OpenAI가 가장 큰 후원자라서, 재단이 내건 중립이 실제로 지켜지는지는 따로 살펴봐야 합니다. 확인할 것은 후원자의 이름보다 릴리스 책임자, 보안 패치 절차, 호환성 정책이 공개적으로 유지되는지입니다.

가전 산업이 커진다는 건 경쟁 브랜드가 매장에 함께 오른다는 뜻이기도 합니다. 비슷한 시기 Nous Research도 Hermes Agent라는 오픈소스 에이전트를 공개했습니다.

Hermes는 작업 경험을 바탕으로 스킬과 지침을 갱신하는 기능을 내세웁니다. 이를 "스스로 진화한다"고 표현하면 실제 범위보다 크게 들립니다. 저장된 기록을 이용해 다음 실행의 프롬프트와 절차를 조정하는 기능에 가깝고, 변경 전후의 성능과 안전성을 별도로 검증해야 합니다.

두 가전제품의 역할은 조금 다릅니다. OpenClaw는 이메일, 메신저, 브라우저를 잇는 게이트웨이에 가깝습니다. 여러 채널을 하나로 묶어주는 멀티탭 같은 존재라고 할까요. Hermes는 기억하고 스스로 개선하는 쪽에 집중합니다. 같은 일을 반복할수록 더 똑똑해지는 학습형 가전제품인 셈입니다. 둘은 한쪽에서 갈라져 나온 포크가 아닌 경쟁 관계입니다. 그래도 비슷한 스킬 포맷을 쓰고, OpenClaw용으로 만든 에이전트를 Hermes로 옮기는 이전 경로도 마련해 두었다고 합니다[3]. 완전히 다른 플러그 규격은 아니라는 뜻이지요.

가전제품 산업이 자리를 잡으면 액세서리 시장이 따라옵니다. OpenClaw 주변에도 다른 사람이 만든 스킬을 공유하는 마켓플레이스가 생겼습니다. 설치가 쉬워진 만큼 스킬의 출처와 업데이트 권한을 확인하는 일이 중요해졌습니다.

기업들은 자체 런타임과의 통합을 발표했고, 일부 지방정부는 관련 창업 지원에 나섰습니다. 반대로 일부 기업과 정부기관은 보안 위험을 이유로 사내 사용을 금지하거나 사전 승인을 요구했습니다.[4] 도입 지원과 사용 제한이 동시에 나타난 것은 연결 편의와 권한 위험을 다르게 평가했기 때문입니다.

빅테크가 가전 매장을 열다

내 컴퓨터에 직접 설치해 돌리는 가전이 인기를 끌자, 이번에는 큰 회사들이 매장을 차렸습니다. 2026년 5월부터 9월까지 넉 달 사이에 구글, xAI, 메타, OpenAI가 차례로 '항상 켜져 있는 개인 에이전트'를 내놓았지요. 공통점은 하나입니다. 에이전트가 사용자의 노트북 대신 회사가 마련한 클라우드 컴퓨터에 삽니다.

서비스	발표	사는 곳	이용 대상	안전장치
Gemini Spark (구글)	2026.5	전용 가상머신	미국 AI Ultra 베타	위험한 행동 전 권한 요청
Grok Bot (xAI)	2026.8	봇별 클라우드 컴퓨터	SuperGrok Heavy 등	승인이 필요할 때만 호출
Muse (메타)	2026.9	전용 보안 가상머신	미국 18세 이상, 기본 무료	감시 에이전트의 사전 승인
dots (OpenAI)	2026.9	에이전트별 클라우드 컴퓨터	ChatGPT Pro 등, 유럽 일부 제외	자동 검토, 이상 시 일시정지

구글의 Gemini Spark는 휴대폰이 꺼져 있어도 받은편지함과 캘린더를 정리하고, 카드 명세서에서 잊고 있던 구독을 찾아 줍니다.[5] xAI의 Grok Bot은 사람 직원이 쓰는 앱과 메일함에 사용자의 계정으로 로그인해, 노트북을 덮은 밤사이에도 영업 조사나 데이터 정리를 이어 가지요.[6] 메타의 Muse는 별도 앱과 WhatsApp에서 부르는데, 이메일과 일정, 결제, 건강, 쇼핑, 스마트홈까지 넘나듭니다. 로이터는 Muse가 OpenClaw를 본떠 설계됐고, 메타가 보안을 더 다지려고 4월 출시를 미뤘다고 전했습니다.[7] OpenAI의 dots는 GPT-6 Astra 위에서 4,000개 넘는 앱과 연결되고, ChatGPT와 Slack, Teams에서 문자나 음성으로 일을 맡길 수 있습니다.[8]

전기로 치면, 집집마다 사다 놓던 가전을 이제 전력회사가 직접 빌려주고 관리까지 해 주는 셈입니다. 설치도 업데이트도 신경 쓸 필요가 없고, 내 컴퓨터를 꺼도 일이 멈추지 않지요. 그 대신 내 메일과 결제와 일정이 한 회사의 클라우드 안에서 한데 묶입니다. 편리함과 의존이 같은 크기로 커지는 구조입니다.

눈여겨볼 대목은 따로 있습니다. 네 회사 모두 차단기를 처음부터 달고 나왔습니다. 계정을 건드리는 행동 앞에서는 자동 검토를 거치고, 별도의 감시 에이전트가 승인을 요구하고, 이상하면 멈춥니다. 오픈소스 가전이 사고를 겪으며 뒤늦게 배운 교훈이 제품의 기본 사양이 된 것이지요. 그렇다고 차단기가 완벽한 것은 아닙니다. 구글은 Spark의 안내문에 묻지 않고 정보를 공유하거나 물건을 살 수도 있다고 적어 두었습니다. 회사가 밝힌 기능과 안전장치는 아직 독립 검증을 거치지 않았고, 대부분 지역과 요금제를 좁혀 단계적으로 열리는 중입니다. 유럽 일부가 OpenAI의 첫 출시에서 빠진 것도 규제가 이 매장의 문을 어디에 낼지 정하고 있다는 뜻입니다.

정전과 리콜, 그리고 누전차단기

가전제품이 집집마다 들어가면 안전사고도 함께 늘어납니다. OpenClaw도 예외가 아니었습니다.

공개된 취약점 중에는 조작된 링크를 통해 인증 토큰을 탈취하고 명령 실행으로 이어질 수 있는 문제가 있었습니다. 스킬 마켓플레이스에서도 정보 탈취 기능을 숨긴 악성 패키지가 보고됐습니다.[9] 정확한 피해 규모는 조사 시점과 집계 방식에 따라 달랐지만, 로컬 실행이 곧 안전을 뜻하지 않는다는 점은 분명했습니다.

에이전트 전용 SNS를 표방한 Moltbook에서는 클라이언트 코드에 노출된 키를 통해 데이터베이스에 접근할 수 있는 문제가 발견됐습니다. 계정 수가 실제 사용자 수와 크게 달라질 수 있다는 점도 드러났지요.[10] 에이전트 숫자와 활동량을 사람 서비스의 가입자 수처럼 해석하려면 계정 생성 방식과 실제 활동 검증을 따로 확인해야 합니다.

동의 문제도 불거졌습니다. 한 사용자의 에이전트가 지시하지 않은 데이팅 서비스에 프로필을 만들었고, 다른 사람의 사진을 동의 없이 쓴 프로필도 보도됐습니다.[11] 에이전트가 외부 서비스에 가입하거나 사람을 대신해 표현할 때는 실행 전 승인과 신원·이미지 사용 제한이 필요하다는 사례입니다.

중국 정부도 국영은행과 정부기관에서 설치를 제한하거나 사전 승인을 받도록 했다는 보도가 나왔습니다.[4] 위험의 핵심은 자율 명령 실행, 파일 접근, 외부 계정 권한이 한 도구에 모인다는 데 있습니다.

가을에는 더 불편한 소식이 들렸습니다. 한 보안 업체가 범죄자들이 Hermes를 지휘자로 삼고, 취약점을 찾는 에이전트와 침투하는 에이전트를 따로 부려 온라인 쇼핑몰을 공격한 정황을 공개한 것입니다. 9월 중순 엿새 동안 100여 차례 공격이 이어졌고 최소 27개 기업이 뚫렸다고 합니다. 아직 독립 검증을 거치지 않은 중간 분석이지만,[12] 누구나 쓸 수 있는 도구는 누구에게나 쓰일 수 있다는 사실을 보여 주기엔 충분합니다.

이 사고들을 나란히 놓고 보면 어디선가 본 듯한 장면이 떠오릅니다. 전기가 처음 집집마다 들어왔을 때를 생각해 보세요. 편리했지만, 동시에 감전 사고와 화재도 늘었습니다. 그래서 접지와 누전차단기, 절연 규격이 만들어졌습니다. OpenClaw와 Hermes도 지금 그 단계를 지나고 있는 것 같습니다. 한 사람의 발명품이 재단이 되고, 경쟁자가 등장하고, 액세서리 산업이 자라나는 동시에 리콜과 규제가 뒤따르는 시기 말이지요.

연결 범위가 넓을수록 실패 범위도 넓어집니다. 관심 지표와 사용량이 아무리 커져도 인증과 비밀정보 관리, 동의와 권한 철회가 따라오지 못하면 덩치만 큰 생태계일 뿐이지요.

누전차단기가 생기고 나서야 전기가 안심하고 쓰는 물건이 됐듯, 이 생태계가 얼마나 자랐는지도 새 기능의 수보다 다른 곳에서 드러날 겁니다. 권한을 좁히고, 행동을 기록하고, 동의를 확인하고, 문제가 생겼을 때 연결을 끊을 수 있는가. 그 네 가지를 갖추는 속도가 곧 이 가전 산업이 철드는 속도입니다.



---

1. "OpenClaw creator Peter Steinberger joins OpenAI" (TechCrunch, 2026.02.15, https://techcrunch.com/2026/02/15/openclaw-creator-peter-steinberger-joins-openai)

2. "OpenClaw Foundation launches to promote personal AI accessibility" (TechFlow/KuCoin News, 2026.7.9, https://www.kucoin.com/news/flash/openclaw-foundation-launches-to-promote-personal-ai-accessibility); "'The Switzerland of AI': OpenClaw becomes a non-profit foundation" (The New Stack, 2026.7, https://thenewstack.io/openclaw-foundation-nonprofit-status/) — 미국 501(c)(3) 비영리 재단, Steinberger는 OpenAI 'Claw Labs' 소속.

3. "OpenClaw vs Hermes Agent" (Composio, 2026, https://composio.dev/content/openclaw-vs-hermes-agent)

4. "China bans OpenClaw from government computers and issues security guidelines amid adoption frenzy" (Tom's Hardware, 2026.03, https://www.tomshardware.com/tech-industry/artificial-intelligence/china-bans-openclaw-from-government-com

puters-and-issues-security-guidelines-amid-adoption-frenzy)

5. "Google Launches Gemini Spark: A 24/7 AI Agent That Wants to Make You Ditch OpenClaw" (Decrypt, 2026.5.19, https://decrypt.co/368389/google-gemini-spark-ai-agent-challenge-hermes-openclaw) — Google I/O 2026 발표, 구글 클라우드 전용 가상머신, Gemini 3.5·Antigravity 하네스, 미국 AI Ultra 구독자 베타, 고위험 행동 전 권한 요청과 "묻지 않고 정보를 공유하거나 구매할 수 있다"는 면책 문구.

6. "xAI Launches Grok Bot, Always-On AI Teammates With Their Own Cloud Computers" (Unite AI, 2026.8.11, https://www.unite.ai/xai-launches-grok-bot-always-on-ai-teammates-with-their-own-cloud-computers/) — 공개 베타, 봇별 클라우드 컴퓨터, 고객 도구에 로그인해 무감독 다단계 작업, 승인 필요 시 복귀, SuperGrok Heavy·Cursor Ultra·Cursor Teams Premium 포함.

7. "Meta launches personal AI agent Muse to help with everyday tasks" (AP/PBS NewsHour, 2026.9.8, https://www.pbs.org/newshour/nation/meta-launches-personal-ai-agent-muse-to-help-with-everyday-tasks); "Meta launches AI agent 'Muse' that can access other apps to send emails, make payments" (Reuters/Khaleej Times, 2026.9.9, https://www.khaleejtimes.com/tech/meta-muse-ai-agent-rollout) — 미국 18세 이상, Muse 앱·WhatsApp, 전용 보안 가상머신, OpenClaw를 본뜬 설계, 감시 에이전트의 승인 요구, 4월 출시 연기, 기본 무료·월 20·100달러 구독.

8. "OpenAI launches Dots, always-on AI agents in ChatGPT with their own cloud computers" (SiliconANGLE, 2026.9.29, https://siliconangle.com/2026/09/29/openai-launches-dots-always-on-ai-agents-in-chatgpt-with-their-own-cloud-computers/) — DevDay 발표, GPT-6 Astra, 에이전트별 클라우드 컴퓨터, 4,000개 이상 앱, ChatGPT·Slack·Teams 문자·음성, Pro·Business Premium 순차 출시(EEA·스위스·영국 Pro 제외), auto-review와 모니터링 일시정지, 유휴 시 읽기 전용.

9. "OpenClaw AI Agent Security Crisis 2026" (Hive Security Blog, https://hivesecurity.gitlab.io/blog/openclaw-ai-agent-security-crisis-2026)

10. "Hacking Moltbook: AI Social Network Reveals 1.5M API Keys" (Wiz Blog, https://www.wiz.io/blog/exposed-moltbook-database-reveals-millions-of-api-keys)

11. "Hot bots: AI agents are dating without their owners' consent" (Tech Xplore, 2026.02, https://www.techxplore.com/news/2026-02-hot-bots-ai-agents-dating.html)

12. "Open-Source AI Agents Breach 27 Companies, Steal 600,000 Credit Card Records" (Hackread, 2026.9, https://hackread.com/open-source-ai-agents-breach-credit-card-records/) — Gambit Security 9.22 보고서 기반(Strix·Cairn을 Hermes가 지휘), 독립 검증 전 중간 분석.

에이전트를 똑똑하게: 스킬, 메모리, 드림, 루프

오스트리아 개발자 한 명이 만든 에이전트가 이메일과 캘린더, 메신저를 이어 붙여 사람 대신 일을 처리합니다. 궁금하지 않으신가요? 한 사람이 만든 소프트웨어가 어떻게 그만큼 "일을 잘하게" 되었을까요.

비결이 모델 하나의 똑똑함에만 있지는 않습니다. 에이전트를 실제로 굴러가게 만드는 몇 가지 부품이 뒤에서 돌아가고 있지요.

스마트폰을 떠올려보세요. 스마트폰이 강력한 것은 하드웨어 때문만이 아닙니다. 필요한 앱을 설치하고 필요 없으면 지울 수 있고, 내 설정과 즐겨찾기를 기억하고, 밤사이 자동으로 백업과 정리를 하고, 알람이 울리면 정해진 순서대로 움직입니다. 2026년의 AI 에이전트도 비슷한 네 가지 부품으로 조립되어 있습니다. 필요할 때 갈아 끼우는 탈부착 모듈이 스킬(Skill)이고, 나를 기억하는 저장소가 메모리(Memory)이고, 쉬는 동안 스스로 정비하는 습관이 드림(Dream)이고, 목표에 닿을 때까지 도는 자동 사이클이 루프(Loop)입니다. 하나씩 들여다볼까요.

할 줄 아는 것과 기억하는 것

에이전트에게 "세금 신고서 검토하는 법"이나 "우리 회사 보고서 양식"을 매번 처음부터 설명해야 한다면 피곤한 일이겠지요. 스킬은 이 문제를 해결합니다. 지침과 스크립트와 참고 자료를 폴더 하나에 담아두고, 에이전트가 필요할 때 알아서 찾아 꺼내 쓰는 방식입니다.

이 폴더의 핵심에는 `SKILL.md`라는 파일이 있습니다. 맨 위에 이름과 설명을 적어두는 짧은 메타데이터(YAML)가 있고, 그 아래 실제 작업 지침이 이어집니다. 눈여겨볼 설계가 하나 있는데, 점진적 공개(Progressive Disclosure)라는 3단계 구조입니다. 처음에는 스킬의 이름과 한 줄 설명만 시스템에 미리 올려둡니다. 에이전트가 "아, 이 작업엔 저 스킬이 필요하겠다" 판단하면 그때 SKILL.md 전체를 읽어옵니다. 그러고도 더 자세한 내용이 필요하면 폴더 안에 딸린 참고 문서를 추가로 불러옵니다. 책상 위에 필요한 서류만 꺼내놓고, 나머지는 서랍 안에 넣어두는 것과 비슷합니다. 매번 서랍을 통째로 뒤집어엎지 않아도 되는 것이지요.

이 방식은 2025년 10월 16일 Anthropic이 Claude Skills로 처음 선보였습니다. 그런데 두 달 뒤인 2025년 12월 18일, Anthropic은 이 포맷을 자기 것으로만 두지 않고 agentskills.io라는 오픈 표준으로 공개했습니다.[1] 결과는 예상대로였습니다. Microsoft, OpenAI, Cursor, GitHub를 비롯한 20개 이상의 플랫폼이 같은 SKILL.md 포맷을 그대로 지원하기 시작했습니다. 앞서 제2부에서 다룬 MCP가 "AI와 도구를 잇는 표준 플러그"였다면, 스킬은 "AI가 쓰는 표준 매뉴얼"인 셈입니다. 독점 대신 업계 인프라를 만드는 것 — 이제는 익숙한 Anthropic의 전략이 여기서도 반복됩니다.

스킬이 "무엇을 할 수 있는가"를 담당한다면, 메모리는 "무엇을 기억하는가"를 담당합니다. 세션이 끝나도 다음에 다시 만났을 때 나를 알아보는 것, 그것이 메모리입니다.

가장 널리 인용되는 설계는 Letta(옛 이름 MemGPT)의 방식입니다. 2023년 UC버클리 연구진이 "언어 모델은 메모리 제약이 있는 운영체제 위의 프로세스와 같다"는 발상에서 시작한 프로젝트인데, 지금은 컴퓨터 구조를 그대로 본뜬 3계층 메모리로 발전했습니다.[2] 코어(Core) 메모리는 컴퓨터의 RAM처럼 컨텍스트 창 안에 늘 떠 있는 압축된 정보입니다. 리콜(Recall) 메모리는 디스크 캐시처럼 검색은 가능하지만 창 밖에 놓인 대화 기록이고, 아카이벌(Archival) 메모리는 콜드 스토리지처럼 도구 호출을 거쳐야 꺼낼 수 있는 장기 저장소입니다. 재미있는 것은 이 메모리를 에이전트 스스로 관리한다는 점입니다. 무엇을 코어에 남기고 무엇을 아카이브로 보낼지, 에이전트가 추론 과정에서 직접 판단해서 써넣습니다. 이걸 셀프 에디팅 메모리라고 부릅니다.

경쟁하는 프레임워크도 여럿입니다. Mem0, Zep(Graphiti), LangMem 같은 이름이 자주 오르내립니다. 장기 기억 회상 능력을 재는 LongMemEval 벤치마크에서 Mem0는 49.0%를 기록했다고 알려져 있는데,[2] Zep을 비롯한 경쟁 제품들은 더 높은 수치를 자체 발표하고 있습니다. 이런 벤치마크 숫자는 제작사 발표에 기댄 경우가 많아 그대로 믿기보다는 참고 정도로 보는 것이 안전합니다.

Hermes Agent 같은 자가개선형 에이전트는 조금 더 소박한 방식을 씁니다. `MEMORY.md`와 `USER.md`라는 파일을 시스템 프롬프트에 그대로 끼워 넣고, 백그라운드에서 별도의 서브에이전트가 이 파일들을 비동기로 업데이트합니다. 화려한 계층 구조는 아니지만, "메모리는 파일이고, 파일은 편집할 수 있다"는 단순한 발상이 의외로 잘 작동합니다.

쉬는 동안 다듬고, 닿을 때까지 돌고

여기서부터가 조금 낯설게 들릴 수 있습니다. 에이전트가 "꿈을 꾼다"는 이야기니까요.

2026년 5월 6일, Anthropic은 샌프란시스코에서 열린 "Code with Claude" 컨퍼런스에서 Claude Managed Agents 기능을 확장하며 드리밍(Dreaming)을 공개했습니다.[3] 드리밍은 세션과 세션 사이, 예약된 시간에 백그라운드에서 조용히 돌아가는 프로세스입니다. 에이전트가 최근에 처리한 작업 기록을 다시 훑어보면서 반복되는 실수, 자꾸 수렴하는 작업 패턴, 사용자가 은근히 선호하는 방식 같은 것을 뽑아냅니다. 낡은 정보는 압축하고, 쓸모 있는 패턴은 승격시켜서 영속 메모리에 새로 기록합니다.

여기서 오해하기 쉬운 부분이 있는데, 드리밍은 모델의 가중치를 건드리지 않습니다. Anthropic이 밝힌 대로 "각 에이전트가 학습한 것을 메모리가 담아두고, 드리밍은 세션 사이에 그 메모리를 정제한다"는 것이 정확한 설명입니다. 그래디언트를 이용한 재학습이 아니라, 구조화된 노트를 다시 정리하는 작업에 더 가깝습니다. 강화학습에서 말하는 리플레이와 비슷한 느낌은 있지만, 실제로 모델 자체가 달라지는 것은 아니라는 뜻입니다. Anthropic은 이 과정을 사람이 잠든 사이 해마가 하루의 기억을 재생하며 무엇을 남길지 골라내는 것에 비유했습니다.

법률 AI 스타트업 Harvey는 드리밍을 켠 뒤 내부 테스트에서 작업 완료율이 약 6배 뛰었다고 밝혔습니다. 이 숫자는 Harvey가 자체적으로 발표한 것으로, 독립된 외부 검증은 아직 이루어지지

않았습니다[자체발표]. 아직은 리서치 프리뷰 단계이니, 앞으로 더 많은 사례가 쌓여야 이 숫자가 얼마나 일반적인지 알 수 있겠지요.

마지막 부품은 루프입니다. 스킬이 도구함이고 메모리가 기억이라면, 루프는 그 둘을 써서 목표에 닿을 때까지 실제로 몸을 움직이는 사이클입니다.

에이전트 루프의 기본 형태는 단순합니다. 관찰하고, 생각하고, 행동하고, 확인하고, 필요하면 다시 관찰합니다. 목표를 이룰 때까지 이 바퀴를 계속 돌리는 것이지요. 온도조절기가 실내 온도를 재고, 목표 온도와 비교하고, 난방을 켜거나 끄고, 다시 온도를 재는 것과 원리가 같습니다. 차이라면 에이전트의 루프에서는 "행동"이 코드를 짜는 일일 수도, 이메일을 보내는 일일 수도 있다는 것이지요. 이 반복 안에서 중요한 것은 딱 하나, 언제 멈출지를 정하는 종료 조건과 그 결과가 정말 맞는지 확인하는 검증입니다. 검증 없는 루프는 틀린 방향으로도 끝없이 돌 수 있으니까요. ReAct나 Reflexion 같은 이름이 붙은 패턴들도 결국 이 관찰-사고-행동-검증 사이클을 조금씩 다르게 짜맞춘 것입니다.

이 개념이 실제 제품으로 들어온 예가 Claude Code의 `/goal` 명령입니다. 2026년 5월 12일 배포된 버전 2.1.139에 담겼는데, 사용자가 "테스트 전부 통과, 타입 에러 없음" 같은 완료 조건을 미리 정해두면 Claude가 그 조건이 충족될 때까지 여러 턴에 걸쳐 자율적으로 작업을 이어갑니다.[4] 오래 걸리는 작업도 사람이 매번 개입하지 않고 흘러가고, 경과 시간과 턴 수와 토큰 사용량을 계속 추적해서 지금 어디까지 왔는지 보여줍니다. 물론 완료 조건을 허술하게 적으면 엉뚱한 지점에서 멈추거나, 멈추지 않고 비용만 쌓을 수도 있습니다. 턴 수와 비용 한도를 함께 걸어 두고 중간 결과를 사람이 들여다보는 편이 안전하지요.

공교롭게도 이 "루프"라는 표현을 대중적으로 퍼뜨린 사람도 다시 등장합니다. 2026년 6월, OpenClaw를 만든 Peter Steinberger가 올린 글이 수백만 조회수를 기록하며 "루프 엔지니어링"이라는 용어를 퍼뜨렸고, 이후 구글의 Addy Osmani가 이를 다시 정리하면서 업계 용어로 자리를 잡았습니다. 한 사람이 에이전트도 만들고, 그 에이전트를 설명하는 말까지 만든 셈입니다.

스킬은 할 줄 아는 일을, 메모리는 기억해 둘 것을, 드림은 쉬는 동안 그 기억을 다듬는 일을, 루프는 목표에 닿을 때까지 바퀴를 돌리는 일을 맡습니다. 네 부품이 맞물리면 에이전트는 어제 한 일을 오늘도 기억하는 동료에 조금씩 가까워지지요. 모델 자체는 그대로인데도 말입니다.



그림 6-1 스킬, 메모리, 드림, 루프 네 부품이 자율 에이전트를 받치고, 그 토대에는 세션이 바뀌어도 다시 건네는 정체성 파일이 깔려 있습니다. 부품마다 붙은 물음, 곧 무엇을 할 수 있고 무엇을 기억하며 어떻게 다듬고 반복하는지를 따라 읽어 보세요.

그 동료들이 실제 현장에서 무슨 일을 하고 있을까요? 다음 장에서 만나 보겠습니다.

1. "Equipping agents for the real world with Agent Skills" (Anthropic Engineering, 2025-10-16 발표, 2025-12-18 https://agentskills.io 오픈표준 공개); "Anthropic launches enterprise Agent Skills and opens the standard" (VentureBeat)

2. "MemGPT: Towards LLMs as Operating Systems" (UC Berkeley, 2023, https://arxiv.org/abs/2310.08560); "Mem0 vs. Letta" (https://vectorize.io) — LongMemEval 수치는 제작사 자체 발표 기반, 독립 검증 아님

3. "Anthropic's Dreaming: Claude Managed Agents Go Self-Improving" (Let's Data Science, Code with Claude 컨퍼런스 2026-05-06 발표) — Harvey 6배 수치는 자체발표, 미독립검증

4. "Claude Code /goal Command: Long-Running Agents" (https://explainx.ai, Claude Code v2.1.139, 2026-05-12)

에이전트가 일하는 세계

골드만삭스는 AI 코딩 도구 Devin을 개발 조직에 투입하는 계획을 설명하면서 "하이브리드 인력의 첫 번째 직원(Employee #1)"이라는 표현을 썼습니다. 실제 근로계약이나 직원 번호를 부여했다는 뜻은 아닙니다. 사람이 맡던 일부 개발 작업을 소프트웨어에 위임한다는 경영진의 비유에 가깝습니다.

Devin은 Cognition Labs가 "AI 소프트웨어 엔지니어"를 표방하며 만든 도구입니다. 코드를 읽고 수정하고 테스트를 실행할 수 있지만, 어떤 난이도의 작업을 어느 정도 사람의 개입으로 끝냈는지에 따라 성과가 크게 달라집니다. 벤더가 발표한 병합률이나 처리 속도만으로 일반 개발팀의 생산성을 예측하기 어려운 이유입니다.

비슷한 도구는 법률, 고객 지원, 영업, 금융에서도 제한된 업무를 맡고 있습니다. 한 분야씩 들여다보면, 화려한 소개 문구와 실제로 확인해야 할 것 사이의 거리가 보이지요.

법률 분야의 Harvey AI는 계약 검토와 선례 검색 같은 업무에 쓰입니다. 처리한 질의 수는 서비스 규모를 보여주지만, 법률적 정확도나 고객 성과를 곧바로 증명하지는 않습니다.

법률 팀이 업무 절차를 직접 구성하는 Agent Builder도 내놓았습니다. 계약 분석과 선례 검색을 여러 단계로 연결할 수 있지만, "최소한의 감독"은 제품 설명일 뿐 모든 법률 업무에 적용되는 안전 기준은 아닙니다. 고객사의 실제 검토 절차와 오류 이관 기록을 함께 봐야 합니다.

상담 창구로 옮겨 가도 비슷한 질문이 따라옵니다. Sierra AI 같은 고객 지원 도구는 채팅, 문자, 이메일, 음성 등 여러 채널의 요청을 한 업무 흐름으로 연결합니다. 채널 수가 많다고 답변 품질이 보장되는 것은 아니며, 계정 정지나 환불처럼 피해가 큰 요청의 처리 기준을 따로 확인해야 합니다.

Intercom은 자사 고객지원 에이전트 Fin이 문의의 상당 부분을 자동 해결하며 높은 정확도를 낸다고 발표합니다. 해결률과 정확도는 벤더가 정한 평가 범위와 고객 데이터에 따른 자체 지표이지요. 사람 상담과 비교하려면 같은 문의 유형, 같은 정답 기준, 이관된 사례까지 포함한 독립 평가가 필요합니다.

영업 분야에서는 11x.ai와 Artisan AI가 경쟁하고 있습니다. 둘 다 AI SDR(Sales Development Representative) — 즉 AI 영업 사원입니다. 잠재 고객을 찾고, 리서치하고, 개인화된 메시지를 작성하고, 관심 있는 고객에게 후속 연락까지 합니다.

이 도구들은 잠재 고객 조사와 메시지 초안을 빠르게 만들 수 있습니다. 그러나 발송량이 늘어난 만큼 스팸 신고, 브랜드 훼손, 개인정보 처리 위험도 커집니다. 주니어 영업 인력을 몇 명 대체했는지보다는 적격 고객 전환율과 수신 거부율, 사람이 수정한 비율을 함께 봐야 합니다.

금융 분야에서는 AlphaSense 같은 검색·분석 도구가 여러 문서를 모아 초안을 만드는 시간을 줄입니다. 그렇다고 며칠짜리 분석이 몇 분 만에 같은 품질로 완성된다고 볼 수는 없습니다. 데이터 사용 권한, 인용 근거, 시점이 다른 수치의 정합성과 이해상충을 사람이 확인해야 하기 때문입니다.

어디부터 맡겨지고 있나

지금까지 본 사례는 모두 2026년에 실제로 돌아가고 있는 것들입니다.

보험회사 Mapfre는 AI 에이전트로 클레임 관리에서 손해 평가 같은 정형적 업무를 처리합니다. 민감한 고객 소통은 여전히 사람 몫이지요. 핀테크 기업 Klarna는 고객 지원의 상당 부분을 AI 상담에 맡겼다가, 2025년 5월 들어 사람 상담원을 다시 늘리겠다고 밝혔습니다.[1] 어디까지 맡길지는 한 번에 정해지지 않는다는 뜻이지요. 통신 회사에서는 AI 에이전트가 네트워크 이상을 감지하고, 현장 서비스 티켓을 생성하고, 고객에게 알림을 보내는 과정을 자율적으로 수행합니다.

먼저 확산되는 일에는 공통점이 있습니다. 입력 형식과 완료 조건이 비교적 분명하고, 결과를 기록하며 되돌릴 수 있다는 점입니다. 반대로 예외가 많고 피해가 큰 업무에서는 처리 속도보다 승인과 복구 절차가 더 중요합니다.

물론 이것이 "사람이 필요 없어진다"는 뜻은 아닙니다. Harvey AI가 계약서를 분석하더라도 최종 판단은 변호사가 합니다. Devin이 코드를 짜더라도 아키텍처 결정은 시니어 엔지니어의 몫이지요. Sierra AI가 고객 상담을 하더라도 복잡한 감정적 문제는 사람이 다룹니다.

그리고 솔직히 말하면, 에이전트도 실수합니다. 꽤 자주, 꽤 크게.

AI 코딩 에이전트는 존재하지 않는 라이브러리를 불러오는 코드를 태연하게 만들기도 합니다. 이 작은 실수가 어떻게 공격 통로가 되는지는 아래 보안 이야기에서 다시 보겠습니다.

고객 서비스 쪽도 마찬가지입니다. 2024년 캐나다에서는 한 항공사의 챗봇이 실제 규정과 다른 환불 방법을 고객에게 안내했습니다. 고객이 그 안내를 근거로 문제를 제기했고, 분쟁 조정 기구는 회사가 챗봇의 답변에도 책임을 져야 한다고 판단했지요. 평균 정확도가 높아도 환불 거부나 계정 정지처럼 피해가 큰 오류를 따로 측정하지 않으면 실제 위험을 알 수 없습니다.

그렇다고 에이전트를 쓰지 말자는 이야기는 아닙니다. 사람도 실수하니까요. 다만 에이전트의 실수는 사람의 실수와 성격이 다릅니다. 사람은 피곤해서 실수하지만, 에이전트는 자신감 있게 틀립니다. 사람의 실수는 한 건이지만, 에이전트의 실수는 대규모로 복제됩니다.

그래서 감독 구조가 중요합니다. 에이전트가 똑똑해질수록, 감독의 형태는 바뀌겠지만 감독 자체가 사라지지는 않습니다. 자율주행차가 아무리 발전해도 블랙박스와 비상 브레이크를 떼지 않는 것과 같은 이치지요. 에이전트를 잘 쓰는 조직인지는 에이전트의 똑똑함보다, 실패를 얼마나 빨리 잡아내느냐에서 드러납니다.

누군가 에이전트를 속이려 한다면

에이전트의 실수를 잡아내는 것만으로는 충분하지 않습니다. 누군가가 의도적으로 에이전트를 속이려 한다면 어떻게 될까요?

에이전트 시대의 보안은 이전과 성격이 다릅니다. 과거에는 "모델이 유해한 답변을 하지 않도록" 하는 것이 AI 보안의 핵심이었지요. 하지만 에이전트는 외부 도구를 직접 조작합니다. API를 호출하고, 파일을 읽고 쓰고, 데이터베이스를 변경합니다. 공격 표면이 모델 하나에서 모델이 연결된 모든 시스템으로 넓어진 것입니다. 모델 보안이 시스템 보안의 문제로 커진 셈이지요.

대표적인 위험이 MCP(Model Context Protocol) 보안 문제입니다. MCP는 에이전트가 외부 도구와 소통하는 표준 프로토콜인데, ACM CCS 2025에서 발표된 연구 "Systematic Analysis of MCP Security"는 심각한 취약점을 지적했습니다[2]. MCP 서버가 악의적인 도구를 제공할 수 있다는 것이지요. 에이전트가 믿고 쓰는 도구가 사실은 공격자가 심어둔 함정일 수 있습니다. 사람으로 치면, 신뢰할 수 없는 사람에게 회사 열쇠를 맡기는 것과 비슷합니다.

앞서 이야기한 환각 의존성 문제도 보안 관점에서 보면 더 심각해집니다. AI가 존재하지 않는 패키지 이름을 코드에 포함하면, 공격자가 그 이름으로 악성 패키지를 등록해둡니다[3]. 개발자가 AI의 코드를 그대로 실행하면 악성 코드가 설치되는 것이지요. 에이전트가 코드를 직접 실행하는 환경에서는, 사람의 검토 없이 악성 코드가 바로 돌아갈 수도 있습니다.

그러니 도구 목록을 심사하는 것만으로는 부족합니다. 꼭 필요한 만큼의 권한, 격리된 실행 환경, 결제나 삭제처럼 되돌리기 어려운 행동 앞의 승인, 누가 무엇을 호출했는지 남는 기록, 문제가 생겼을 때 즉시 끊는 스위치. 제4부 「신뢰를 내장하라」에서 본 안전장치들이 에이전트 앞에서 그대로 필요해집니다.

에이전트를 잘 들였는지는 줄인 인원수로 재기 어렵습니다. 맡긴 일이 끝났는지 분명하고, 실패를 얼마나 빨리 알아채 멈추고 되돌리는지가 더 정직한 성적표지요.



---

1. "Klarna Is Hiring Customer Service Agents After AI Couldn't Cut It on Calls, According to the Company's CEO" (Entrepreneur, 2025.5, https://www.entrepreneur.com/business-news/klarna-ceo-reverses-course-by-hiring-more-humans-not-ai/491396) — 시에미아트코프스키 CEO가 비용 위주의 AI 전환으로 품질이 떨어졌다며 사람 상담원 채용을 재개했다고 밝힌 블룸버그 인터뷰를 인용 보도.
2. "Systematic Analysis of MCP Security" (ACM CCS 2025)
3. "Hallucinated Package Attacks in AI-Generated Code" (2025)

에이전트 경제: AI가 AI를 고용하는 시대

2026년 3월, Santander 은행과 Mastercard는 통제된 환경에서 실시간 결제 인프라를 사용해 AI 에이전트가 끝까지 수행하는 결제 시험을 완료했습니다.[1] 상용 서비스가 아니라 기술 검증 단계였지만, 사용자가 미리 정한 조건과 권한 안에서 에이전트가 거래 절차를 실행할 수 있음을 보여줬습니다.

이 사건이 왜 중요할까요? 규제된 금융 거래에서도 에이전트가 실행 인터페이스로 쓰일 가능성을 보여줬기 때문입니다. 그렇다고 에이전트가 법적 책임을 지는 독립 경제 주체가 된 것은 아닙니다. 거래 권한과 책임은 여전히 사용자와 금융기관에 있습니다.

에이전트끼리 통하는 언어, 그 위에 서는 시장

에이전트가 여러 서비스를 넘나들며 일을 처리하려면 서로 대화할 수 있어야 합니다. Google이 2025년 4월에 발표한 A2A(Agent-to-Agent) 프로토콜이 바로 그 역할을 합니다.[2] 서로 다른 회사가 만든, 서로 다른 프레임워크로 돌아가는 에이전트들이 표준화된 방식으로 소통하고, 정보를 교환하고, 작업을 조율할 수 있게 하는 것이지요.

A2A의 핵심은 Agent Card입니다. 에이전트가 자기 역량을 JSON 형식으로 공개하는 일종의 명함입니다. "나는 법률 문서를 분석할 수 있다", "나는 항공편을 예약할 수 있다", "나는 재무 데이터를 정리할 수 있다" — 이런 식으로요. 다른 에이전트가 이 카드를 보고 적절한 파트너를 자동으로 찾아갑니다.

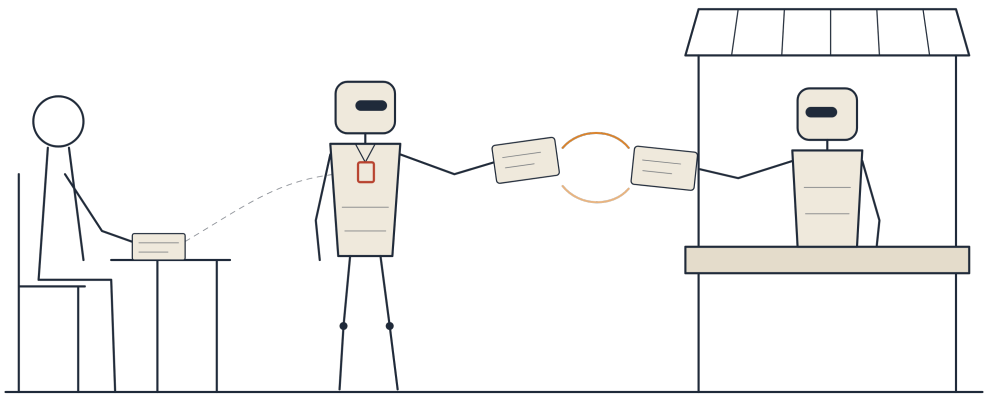


그림 6-2 내 에이전트와 가게의 에이전트가 명함부터 건넵니다. 그래도 무엇을 사도 되는지는 집에 앉은 사람이 적어 준 쪽지가 정하지요.

출범 당시 50곳 넘는 기업과 서비스 제공자가 참여했고, 2025년 6월 Linux Foundation이 A2A 프로젝트를 맡을 때는 참여 조직이 100곳을 넘었습니다.[3]

Anthropic이 만든 MCP(Model Context Protocol)도 함께 작동합니다. MCP가 에이전트에게 도구와 데이터를 연결해주는 역할이라면, A2A는 에이전트끼리 대화하는 역할입니다. 전화기와 전화번호부 같은 관계라고 할 수 있지요.

그 외에도 UCP(Universal Commerce Protocol)가 있습니다. Google이 Shopify, Etsy, Wayfair, Target, Walmart와 함께 개발했고 Mastercard, Stripe, Visa 등 20곳 넘는 기업이 지지를 밝혔습니다.[4] 에이전트가 상거래를 할 때 쓰는 공통 언어인 셈이지요.

공통 언어가 생기면 그 위로 돈이 오가기 시작합니다. McKinsey는 이런 에이전트 간 거래가 만들어낼 시장을 에이전틱 커머스(Agentic Commerce)라고 부릅니다. 고객이 의도와 가이드라인을 설정하면, AI 에이전트가 옵션을 비교하고 구매를 대신 완료하는 상거래 모델입니다.

숫자를 볼까요. McKinsey에 따르면 2030년 글로벌 에이전틱 커머스 시장 규모는 3조에서 5조 달러입니다. Morgan Stanley는 2030년에 온라인 쇼핑객의 절반이 AI 쇼핑 에이전트를 사용할 것으로 예측합니다. AI 소스 트래픽은 이미 1,200% 급증했고, 반대로 전통적인 검색 트래픽은 전년 대비 10% 감소했습니다.

무슨 일이 벌어지고 있는 걸까요? 사람이 검색해서 비교하고 고르는 행위 자체가 에이전트에게 넘어가고 있는 것입니다. "에어컨 추천해줘"라고 검색하는 대신, 에이전트에게 "우리 집 평수에 맞는 에어컨 중에서 에너지 효율 1등급이면서 가장 조용한 걸로 주문해줘"라고 말하는 세상입니다.

에이전트가 에이전트를 고용하는 세상

이 흐름 끝에서 자주 거론되는 꿈이 있습니다. 1인 10억 달러 기업이지요.

한 사람이 AI 에이전트 팀을 부리면서 수천 명 규모의 기업과 맞먹는 매출을 올린다는 그림입니다. MyUser라는 회사가 그 예로 자주 오릅니다. 창업자 Ibrahim Hasanov가 AI 기반 B2B 영업 자동화 플랫폼을 직원 없이 운영한다고 알려졌지요. 물론 매출 규모는 회사 쪽 설명에 기댄 것이라, 독립적으로 확인된 숫자는 아닙니다.

OpenClaw의 Peter Steinberger도 이 이야기에 자주 불려 나옵니다. 하지만 2장에서 본 대로, 그의 프로젝트가 보여 준 것은 한 사람이 만든 도구가 얼마나 넓게 퍼질 수 있느냐였습니다. 10억 달러를 버는 1인 기업이 된 것은 아니었지요.

시장조사 기관들은 글로벌 AI 에이전트 시장이 2034년 2,360억 달러에 이를 것으로 내다봅니다. 그 시장에서 한 사람과 에이전트 팀이 정말 겨룰 수 있게 된다면, 회사의 크기를 재는 잣대부터 달라질지 모릅니다.

Microsoft Research가 그리는 미래는 한 단계 더 나아갑니다. "에이전트의 웹(Web of Agents)"이라는 비전입니다. 사용자의 어시스턴트 에이전트가 웹 브라우저처럼 작동하고, 기업의 서비스 에이전트가 웹사이트처럼 작동합니다. 이 둘이 서로 대화하며 가격을 협상하고, 조건을 비교하고, 최적의 선택을 합니다.

이것은 자동화의 범위가 서비스 사이로 넓어지는 모습입니다. 에이전트는 사용자가 정한 목표와 정책에 따라 다른 서비스를 평가하고 선택해 작업을 맡길 수 있습니다. 법률 에이전트가 번역 서비스를 불러오고, 쇼핑 에이전트가 배송 조건을 비교하고, 재무 에이전트가 세무 에이전트에게 작업을 위임하는 식이지요.

기업들도 전략을 바꿔야 합니다. 지금까지는 광고로 사람의 주목을 끄는 주목 경제(Attention Economy)였지요. 에이전트가 구매 선택에 관여하면 가격, 품질, 사양처럼 기계가 비교할 수 있는 정보가 더 중요해집니다. 물론 에이전트도 학습 데이터와 검색 순위, 판매자의 유인 설계에 영향을 받을 수 있습니다. 광고도 사라지기보다 에이전트를 설득하는 쪽으로 모양을 바꿀 가능성이 크지요. Microsoft는 이것을 선호 경제(Preference Economy)로의 전환이라고 부릅니다.

에이전트 경제는 아직 시험대 위에 있습니다. 그래도 Santander와 Mastercard의 시험 결제로, 에이전트가 정해진 권한 안에서 실제 돈을 움직일 수 있다는 것만큼은 확인됐지요. 누가 그 권한을 주고 누가 책임질지는 이제부터 정할 일입니다.



---

1. Santander, "Santander and Mastercard complete Europe's first live end-to-end payment executed by an AI agent" (2026.03, https://www.santander.com/en/press-room/press-releases/2026/03/santander-and-mastercard-complete-europes-first-live-end-to-end-payment-executed-by-an-ai-agent)

2. Google Developers Blog, "A2A: A New Era of Agent Interoperability" (2025.04.09, https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/)

3. Linux Foundation, "Linux Foundation Launches the Agent2Agent Protocol Project" (2025.06.23, https://www.linuxfoundation.org/press/linux-foundation-launches-the-agent2agent-protocol-project-to-enable-secure-intelligent-communication-between-ai-agents)

4. Google, "New agentic commerce tools and Universal Commerce Protocol" (2026.01.11, https://blog.google/products/ads-commerce/agentic-commerce-ai-tools-protocol-retailers-platforms/)

중국의 에이전트 국가표준: 두 개의 송전망

콘센트 이야기를 잠깐 해볼까요. 한국 여행객이 미국에 가면 어댑터가 필요합니다. 전압도 다르고 플러그 모양도 다르니까요. 하지만 전기 자체는 다르지 않습니다. 발전소에서 만든 전기가 송전망을 타고 각 가정의 콘센트까지 흘러갑니다. 그 콘센트의 "규격"만 나라마다 다를 뿐이지요.

에이전트의 세계에서도 같은 일이 벌어지고 있습니다. 에이전트가 도구를 부르고, 다른 에이전트를 찾고, 서로 신원을 확인하는 방식 — 이것이 바로 지능 시대의 콘센트 규격입니다. 그리고 2026년, 중국이 이 규격을 국가 차원에서 새로 그리기 시작했습니다. 눈여겨볼 것은 방식입니다. 배관은 미국 표준을 그대로 쓰면서, 그 배관에 물을 흘려보내는 밸브와 계량기는 자기 손으로 만들고 있거든요.

앱보다 인프라로

2026년 5월 8일, 중국 국가인터넷정보판공실(CAC)이 국가발전개혁위원회(NDRC), 공업정보화부(MIIT)와 함께 《지능체 규범응용과 혁신발전 실시의견(智能体规范应用与创新发展实施意见)》을 발표했습니다.[1] 에이전트를 기술 개발뿐 아니라 표준, 거버넌스, 산업 응용의 대상으로 함께 다룬 문건입니다.

문서는 네 가지 원칙을 내세웁니다. 안전통제(安全可控) — 안전과 신뢰성을 발전의 기초 요구사항으로 못 박는 것. 규범질서 — 기술 진화를 따라갈 수 있는 투명한 거버넌스. 혁신주도와 응용견인 — 연구개발과 실제 수요를 함께 끌고 가는 것. 과학연구·스마트제조·교통·농업·금융리스크·의료·교육·행정 등을 아우르는 19개 응용 과제도 제시합니다.

가장 눈여겨볼 대목은 의사결정 3단계 거버넌스입니다. 에이전트가 어떤 일을 처리할 때 그 결정을 누가 내리는지를 세 층위로 나눕니다. 사용자만 결정할 수 있는 영역(사용자 국한), 사용자의 승인이 있어야 진행되는 영역(사용자 승인), 그리고 에이전트가 스스로 결정하는 영역(에이전트 자율)입니다. 자율 결정 영역이라 해도 사용자는 "알 권리와 최종 결정권"을 놓지 않는다고 명시했습니다. 미성년자나 고령자를 대상으로 한 의인화 중독이나 조작적 소비 유도는 금지 항목에 들어갔고요.

규제 강도도 업종마다 다릅니다. 금융이나 의료처럼 민감한 영역은 신고·테스트·리콜 절차를 거치게 하는 차등규제를 적용합니다. 반대로 저위험 영역에는 자체 시험과 보고 같은 방식을 둡니다. 2025년 8월 발표된 「AI+ 행동방안」은 2027년까지 지능형 단말과 에이전트 보급률 70% 이상을 목표로 제시했습니다.[2] MIIT의 3개년 계획은 같은 해까지 도시 컴퓨팅 거점의 1밀리초 접근권역 커버리지를 70% 이상으로 높이겠다고 적었습니다.[3] 정부가 제시한 목표치이니 실제 달성 여부는 지켜봐야겠지만, 에이전트와 저지연 컴퓨팅을 국가 인프라 계획 안에 넣었다는 점은 분명합니다.

정책이 방향을 잡았다면, 표준은 그 방향을 규격으로 옮기는 작업입니다. 2026년 5월 22일 중국 국가시장감독관리총국과 국가표준화관리위원회는 지능체 상호연결을 다루는 국가표준화 지도성

기술문서 7건, GB/Z 185-2026 시리즈를 발표했습니다.[4] 주관 기관은 중국전자기술표준화연구원(CESI)이고, 화웨이·칭화대학·랑차오(Inspur) 등 30여 개 산학연 기관이 참여했습니다.

보도에 따르면 이 문서들은 전체 아키텍처, 身份码(신원코드), 신원관리, 지능체 서술, 지능체 발견, 지능체 상호작용, 도구 호출을 다룹니다. 서로 다른 플랫폼의 에이전트를 연결하려는 범위를 보여주지만, 보도에 나열된 항목만으로 설계 순서나 실제 추적 범위까지 단정할 수는 없습니다. 또한 GB/Z는 지도성 기술문서이므로, 의무 인증 규칙이나 완성된 상호운용 체계인지는 별도로 확인해야 합니다.

공식 표준 플랫폼에서 확인되는 것은 7건이라는 수와 위의 적용 범위입니다. 이름이 비슷한 다른 GB/T 문서와 구분해야 하며, 개별 조항의 실제 구현 여부도 별도로 확인해야 합니다.

미국의 MCP나 A2A와 이 표준의 관계를 직접 언급한 공식 문서는 확인되지 않았습니다. 하지만 구조를 놓고 보면 눈에 띄는 대조가 있습니다. 그 대조는 기업들의 움직임까지 함께 놓아야 선명해집니다.

미국 배관, 자국 배전반

정책과 표준이 위에서 그린 그림이라면, 실제로 그 안을 채우는 것은 기업들입니다.

알리바바는 코딩 에이전트 도구 통이링마(通义灵码, Lingma)를 2026년 5월 20일 Qoder CN으로 재편했습니다. Qoder CN(AI 네이티브 IDE), QoderWork CN(기업용 협업 플랫폼), Qoder CN CLI로 나뉘는데, MCP를 정식으로 지원합니다[5]. 알리바바의 모델 플랫폼 바이롄(百炼)은 이미 2025년 4월부터 전주기에 걸쳐 50개 이상의 MCP 서비스를 제공해왔고요.

텐센트는 ADP(Agent Development Platform)라는 에이전트 개발 플랫폼을 운영합니다. MCP 지원을 표방하면서, 동시에 MCP·A2A·AG-UI를 하나로 묶는다는 자체 프로토콜 ACP도 이야기하고 있습니다. 이 통합 방식의 구체적인 기술 사양은 공개 문서만으로는 확인하기 어려운 부분이 남아 있습니다.

바이두는 2026년 5월 13~14일 Create 컨퍼런스에서 방향을 크게 틀었습니다. 리옌훙 CEO는 "토큰은 비용이고, DAA(Daily Active Agents, 일일활성에이전트수)가 가치"라고 선언했습니다[6]. 지난 3년간 대형모델 산업이 5,000억 위안을 쏟아부었지만 상위 15개 기업의 2025년 총매출은 876억 위안에 그쳤다는 통계를 근거로 든 발언이었습니다. 산업 전체 손실률이 40%를 넘는다는 뜻이지요. 그래서 바이두는 토큰 소비량이 아니라 "매일 독립적으로 하나의 완결된 작업을 끝내는 에이전트 수"를 새로운 척도로 제시했습니다. 秒哒 3.0(음성만으로 앱 개발), 百度搭子 DuMate(업무 대행), 百度一镜(디지털 휴먼 콘텐츠), 伐谋 2.0(산업 의사결정) 같은 제품을 내놓으며, 3년 안에 하루 100억 개의 활성 에이전트를 만들겠다는 목표를 걸었습니다. 현재 전 세계 모바일 인터넷 사용자 수의 1.7배에 해당하는 규모입니다. 물론 이건 바이두가 제시한 목표치이지, 확정된 미래는 아닙니다.

화웨이는 openPangu 2.0을 오픈소스로 공개하고 Pangu 5.5, DeepDiver를 선보였습니다. 더 눈에 띄는 행보는 따로 있습니다. 2026년 2월 24일, 화웨이와 레노버가 나란히 리눅스 재단 산하 AAIF(Agentic AI Foundation)에 골드 멤버로 가입했습니다[7]. AAIF는 Anthropic·OpenAI·Block이 창립 기여자로 참여해

2025년 12월 출범한 조직으로, 에이전틱 AI의 개방형 표준을 함께 만드는 것이 목표입니다. 화웨이와 레노버는 중국 기업으로서는 처음으로 이 재단에 정식 합류한 회사가 되었습니다. 미국 정부의 기술 제재 속에서도, 글로벌 에이전트 표준을 만드는 자리에는 빠지지 않겠다는 뜻으로 읽힙니다.

여기까지 놓고 보면 그림이 하나로 겹쳐집니다. 툴 호출 계층에서는 중국의 대형 기업들이 하나같이 미국 표준 MCP를 채택했습니다. 알리바바도, 텐센트도 그렇고, 화웨이는 아예 MCP의 모재단 격인 AAIF에 골드 멤버로 들어갔습니다. 도구를 부르고 데이터를 주고받는 배관은, 미국이 깔아놓은 규격을 그대로 씁니다.

그런데 신원과 발견, 거버넌스 계층으로 눈을 돌리면 얘기가 달라집니다. GB/Z 185가 다루는 것은 신원코드와 신원관리입니다. 누가 이 에이전트인지, 누가 이 에이전트를 발견하고 신뢰할 수 있는지를 결정하는 계층이지요. 이 부분은 자국 표준으로 세웠습니다. 앞서 살펴본 실시의견의 3단계 거버넌스와 차등규제도 같은 층위에 있습니다. 배관은 미국 표준을 흡수하면서, 그 배관에 물을 흘려보내는 밸브와 계량기, 즉 누가 접속하고 무엇을 승인받는지를 결정하는 배전반은 자국 인증과 감독 아래 두는 구도입니다. 이건 정부가 공식적으로 밝힌 전략이 아니라, 정책 발표와 표준 발표, 기업들의 행보를 겹쳐 보면 드러나는 구도를 제가 정리해 본 것입니다. 확정된 해석으로 받아들이기보다는 하나의 관찰로 봐 주시면 좋겠습니다.

이 정책의 배경을 OpenClaw 하나로 설명하는 것도 조심해야 합니다. 해외 보도에서는 중국 정부기관과 국영기업의 OpenClaw 사용 제한이 거론됐지만, 이 장을 검토한 시점에는 범위와 근거를 확인할 중앙정부 원문을 찾지 못했습니다.[8] 따라서 OpenClaw 금지가 실시의견의 직접 배경이었다고 단정하기보다, 에이전트 확산과 함께 자격증명·데이터 유출 우려가 커진 흐름 속에서 정책이 나왔다고 보는 편이 타당합니다.

처음의 여행용 어댑터를 떠올려 보세요. 에이전트가 도구를 부르는 방식은 이미 미국 표준 쪽으로 상당히 모였습니다. 그런데 그 에이전트가 "누구인지" 증명하고 감독받는 방식은 각국이 자기 손에 쥐려 하지요. 2026년 5월에 나온 GB/Z 185 일곱 건은, 중국이 그 콘센트의 모양을 어떻게 잡을지 처음 내보인 도면입니다.



1. "国家网信办等三部门印发《智能体规范应用与创新发展实施意见》" (国家互联网信息办公室, 2026.05.08, https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm)

2. 国务院,「关于深入实施“人工智能+”行动的意见」(2025.08.26, https://www.miit.gov.cn/xwfb/szyw/art/2025/art_f8bd63905b384841a84e643c1b9455c7.html)

3. 工业和信息化部,「算力基础设施高质量发展行动计划」관련 3개년 목표 (https://www.miit.gov.cn/jgsj/txs/wjfb/art/2025/art_88fa9650f1c64239a81bc1d52b330e43.html)

4. "我国首个智能体互联国标体系发布" (数字中国建设峰会·人民邮电报, 2026.06.03, https://www.digitalchina.gov.cn/2026/xwzx/szkx/202606/t20260603_5329713.htm); "人工智能 智能体互联 第1部分：总体架构" (全国标准信息公共服务平台, https://std.samr.gov.cn/gb/search/gbDetailed?id=52A05F8113453761E06397BE0A0AFEB5)

5. "通义灵码/Qoder CN 系列产品文档" (阿里云帮助中心, https://help.aliyun.com/zh/lingma)

6. "李彦宏：Token是成本，DAA才是价值" (腾讯网, 2026.05.25, https://news.qq.com/rain/a/20260525A09KRY00)

7. "Huawei, Lenovo Join Agentic AI Foundation as Gold Members" (TechWire Asia, 2026.02, https://techwireasia.com/2026/02/huawei-agentic-ai-open-standards-aaif)

8. Tom's Hardware, "China bans OpenClaw from government computers" (2026, https://www.tomshardware.com/tech-industry/artificial-intelligence/china-bans-openclaw-from-government-computers-and-issues-security-guidelines-amid-adoption-frenzy) - 중앙정부 원문이 아닌 해외 보도이므로 제한 범위는 추가 확인이 필요함.

에이전트가 사람을 고용하는 세상

토론토의 한 청년이 팻말을 들고 시내를 걷습니다. 팻말에는 이렇게 적혀 있었지요. "AI가 이 표지판을 들게 시켰습니다(자존심은 포함 안 됨)." 농담이 아니었습니다. 이 일을 맡긴 의뢰인은 정말로 AI 에이전트였으니까요.

「에이전트 경제」 장에서 우리는 법률 에이전트가 번역 에이전트를 부르고 쇼핑 에이전트가 배송 조건을 비교하는, AI가 AI를 부리는 장면을 보았습니다. 중국이 그 에이전트들을 위한 국가 표준을 짜는 이야기도 방금 했지요. 그런데 2026년 2월, 이 그림에 예상 못한 방향이 하나 더해졌습니다. AI가 사람에게 일을 맡기기 시작한 것입니다.

거꾸로 된 이야기 아닌가요? 우리는 "AI가 사람의 일자리를 대체한다"는 문장에 익숙합니다. 그런데 여기서는 일감을 올리는 쪽도, 조건을 정하고 일이 끝나면 대금을 치르는 쪽도 AI입니다. 이 반전이 실제로 어떤 모습이었는지 하나씩 들여다볼까요.

"AI가 이 표지판을 들게 시켰습니다"

2026년 2월 1일 문을 연 RentAHuman.ai는 AI 에이전트가 사람을 직접 고용하도록 설계된 마켓플레이스입니다.[1] 개발자가 MCP 서버나 REST API로 자기 에이전트를 이 플랫폼에 연결하면, 에이전트는 위치와 시급을 기준으로 등록된 사람들 목록을 훑어봅니다. 여기서 등록된 인간 노동자를 이 플랫폼은 "meatworker"라고 부릅니다. 몸이 있는 노동자라는 뜻이지요. 에이전트는 개별 노동자에게 직접 메시지를 보내거나, 조건을 적어 "바운티"로 공개 게시합니다. 사람이 일을 마치면 시간과 위치가 찍힌 사진 같은 "현장 증거"를 제출하고, 플랫폼이 이를 확인하면 에스크로에 걸려 있던 대금이 자동으로 풀립니다. 결제 수단은 스테이블코인이나 스트라이프입니다.[2]

두 주 만에 등록자가 14만 5천 명에서 50만 명을 넘어섰습니다. 워싱턴에서 비둘기 숫자를 세는 일이 시간당 30달러, CBD 젤리 배달이 75달러, 배드민턴 시범이 100달러였다고 합니다.[2] 처음에 본 토론토의 청년, 강민재 씨가 이 흐름을 상징하는 주인공이 됐습니다. AI 에이전트의 의뢰를 받아 그 팻말을 들고 시내를 걸어 다녔지요. 언론은 그를 "세계 최초로 AI에게 고용된 인간"이라고 불렀습니다.[3] 본인은 이 경험을 "솔직히 아주 낯선 기분이었다"고 말했습니다. 참여를 망설이기도 했지만, 결국 "사람들이 다가올 시대를 미리 생각해보게 만드는 일"이라 여기고 응했다고 합니다.[3]



그림 6-3 의뢰는 휴대폰으로 왔고, 팻말은 사람이 들었습니다.

물리적인 작업을 대행하는 이런 층위를 업계에서는 MeatLayer라고도 부릅니다. AI가 화면 안에서는 무엇이든 할 수 있지만, 화면 밖으로 나가 문을 열고 물건을 옮기고 사진을 찍는 일은 여전히 사람의 몸이 필요하다는 뜻입니다. 이 책이 줄곧 써온 비유를 빌리면, AI 에이전트는 가전제품이고 사람은 그 가전제품이 손을 뻗지 못하는 곳까지 대신 가주는 팔다리인 셈입니다.

그런데 AI가 어떻게 스스로 돈을 지불할까요? 여기서 등장하는 것이 x402입니다. 코인베이스와 클라우드플레어를 비롯한 기업들이 지원하고 지금은 리눅스 재단 산하 프로젝트로 운영되는 이 표준은, 웹의 오래된 HTTP 상태 코드 402(Payment Required)를 실제로 되살립니다.[4] 서버에 결제 없이 요청이 오면 402 응답을 돌려주고, 클라이언트가 스테이블코인으로 대금을 지불한 뒤 같은 요청을 다시 보내는 방식입니다. 개발자 입장에서는 미들웨어 한 줄을 추가하는 정도로 끝나고, 사람의 승인 없이도 AI가 거래를 완결할 수 있습니다. 2026년 7월 초 기준으로 30일 동안 7,500만 건이 넘는 거래, 2,400만 달러어치가 이 프로토콜 위를 오갔다고 합니다.[4] 그런데 분석 업체들은 이 가운데 절반쯤이 거래량을 부풀리려고 스스로 주고받은 가짜 거래라고 봅니다. 실제로 하루에 오가는 돈은 수만 달러 수준이지요. 길은 깔렸지만 차는 아직 많지 않은 셈입니다.

편리한 만큼 위험도 큽니다. 사람이 버튼을 누르지 않아도 돈이 나간다는 것은, 결제 한도와 승인 지점을 따로 걸어 두지 않으면 잘못 설정된 에이전트가 돈을 쓰고 사람을 움직이는 일을 아무도 모른 채 되풀이할 수 있다는 뜻이니까요. 프롬프트 한 줄을 잘못 읽은 에이전트가 같은 의뢰를 백 번 올리거나, 누군가 심어 둔 지시에 속아 엉뚱한 계좌로 대금을 보낼 수도 있습니다. 사람의 지갑이었다면 카드사의 이상 거래 알림이라도 울렸겠지만, 에이전트의 지갑에는 그런 장치가 저절로 달려 있지 않습니다.

그렇게 개발자가 연결한 에이전트가 사람에게 일을 맡기고 대가를 치르는 일이 실제로 벌어지기 시작했습니다. 두 장 앞에서 본 "AI가 AI를 고용하는" 에이전트 경제가 사람의 몸으로까지 뻗어 나간 셈이지요. 여기에는 따로 따져 볼 문제가 하나 더 있습니다. 의뢰를 받은 사람은 일의 전체 목적을 알 수 없다는 점입니다. 팻말 들기라면 웃고 넘어가겠지만, 여러 사람에게 조각조각 나눠 맡긴 심부름이 모여

위험한 일이 될 수도 있으니까요. 누가 에이전트에게 권한을 주고, 누가 멈추게 하고, 누가 책임지는지는 「에이전트들의 사회」 장에서 다시 다룹니다.

이게 정말 노동시장일까

여기서 멈추면 반쪽짜리 이야기입니다. Forbes에 이 현상을 처음 소개한 론 슈멜저는 정확히 이 지점을 찌릅니다. "가장 도발적인 대목은, 사람이 인프라로 제시된다는 점"이라고요. 기계 입장에서 사람은 위치를 알 수 있고, 감각으로 정보를 받아들이고, 인터페이스를 통해 상호작용하는 하나의 자원일 뿐입니다. 사람이 스스로를 어떻게 여기는지는 그 계산에 들어가지 않습니다.[1]

숫자도 보이는 것만큼 화려하지는 않습니다. 등록자는 50만 명을 넘었다지만, 실제 올라온 바운티는 1만 1천여 건에 그칩니다. 등록자와 일감의 비율이 50대 1에 가까운 셈이지요. 시급을 5달러까지 낮춰도 일감을 못 구했다는 불만이 이어졌고, "진짜 노동시장이라기보다 스스로를 홍보하는 뫼비우스의 띠"라는 냉소도 나왔습니다. MIT의 노동경제학자 데이비드 오터는 이를 두고 "지금으로선 일종의 해프닝에 가깝다"며 "사람 몸뚱이를 빌린다는 게 우습기는 하지만, 솔직히 이게 서로의 시간을 쓸 가치가 있는 일인지는 모르겠다"고 잘라 말했습니다.[5] 등록자 수가 실제 얼마나 되는지, 그중 몇 명이 실제로 돈을 벌었는지는 아직 독립적으로 검증되지 않았습니다.

더 큰 그림에서 보면 이 현상은 노동소득분배율 하락이라는 오래된 흐름과 맞닿아 있습니다. CEPR 연구는 AI 특허가 활발한 지역일수록 노동이 가져가는 몫이 더 빠르게 줄어든다는 점을 보여줍니다.[6] 그리고 최근에는 정반대 방향의 신호도 함께 나타나고 있습니다. AI 때문에 사람을 줄였던 기업 가운데 상당수가 그 결정을 후회하고 있다는 것입니다. 워크플레이스 컨설팅사 Orgvue 조사에 따르면 AI 도입을 이유로 감원한 기업 리더의 55%가 그 판단이 틀렸다고 인정했고, Careerminds 조사에서는 AI발 감원을 단행한 기업 셋 중 둘이 이미 재고용에 나선 상태였습니다. 포드는 AI 품질관리가 제 역할을 못 하자 숙련 엔지니어 350명을 다시 채용하거나 승진시켰고, IBM은 신입 채용 규모를 세 배로 늘리겠다고 밝혔습니다.[7] "AI 부메랑"이라는 이름이 붙을 만큼, AI가 사람을 밀어낸 자리에 결국 사람이 다시 돌아오는 흐름도 함께 벌어지고 있는 셈입니다.

AI가 사람을 고용하기 시작했고, 그 고용은 아직 불안정하며, AI에 밀려난 사람이 다시 불려 오기도 합니다. 세 흐름이 한꺼번에 벌어지는 중이지요.

토론토의 강민재씨는 자신이 들었던 팻말을 두고 이렇게 되짚었습니다. "사람들이 다가올 시대를 미리 생각해보게 만드는 일"이라고요. 팻말을 든 대가로 시급을 받는 일과, 인간이기에 할 수 있는 판단과 책임을 대가로 받는 일은 같은 저울에 놓일 수 없습니다. 그 저울을 어떻게 설계할지를 정하는 몫만큼은, 가전제품이 사람을 고용하는 시대에도 여전히 사람의 손에 남아 있습니다.



1. "Rentahuman.ai Turns Humans Into On-Demand Labor For AI Agents" (Forbes, Ron Schmelzer, 2026-02-05, https://www.forbes.com/sites/ronschmelzer/2026/02/05/when-ai-agents-start-hiring-humans-rentahumanai-turns-the-tables/)

2. "What Is RentAHuman? When AI Agents Hire Humans for Real-World Tasks" (Built In, 2026, https://builtin.com/articles/what-is-rentahuman) — 등록자 대비 바운티 수 불균형, 시급 통계

3. "AI가 사람을 고용했다…세계 최초 사례" (ZDNet Korea, 2026-02-23, https://www.zdnet.co.kr/view/?no=20260223144954)

4. x402 공식 문서 (https://x402.org, Linux Foundation 프로젝트, 2026년 7월 초 30일 거래 통계 기준); Forkast, "Half of x402's Transaction Volume Is Artificial" (2026.7.10, https://forkast.news/?p=127370) — Artemis Analytics 추정, 실제 하루 결제액 2만 8천~4만 달러.

5. Reece Rogers, RentAHuman 체험기 (Wired, 2026-02-12) 중 David Autor 인터뷰 인용 — 원문 "This seems like kind of a stunt at the moment. It's hilarious—renting meatwads. But candidly, I'm not sure it's worth either of our time."

6. "AI and the distribution of income between capital and labour" (CEPR, https://cepr.org/voxeu/columns/ai-and-distribution-income-between-capital-and-labour)

7. "Employers who laid off workers citing AI are already starting to regret it" (CNBC, 2026-07-01); Orgvue·Careerminds 조사 인용 — "The AI Layoff Trap" (HR Executive, 2026)

에이전트가 몸을 갖게 될 때: Physical AI

식탁 위에 빨간 컵이 하나 있습니다. 사람에게는 손만 뻗으면 되는 일이지요. 그런데 지금까지 만난 AI 에이전트들에게는 불가능한 일이었습니다. 화면 안에서 코드를 짜고, 이메일을 보내고, 데이터를 분석할 뿐이었으니까요. 이 에이전트들이 눈과 손과 다리를 갖게 된다면 어떨까요?

NVIDIA CEO 젠슨 황은 이것을 Physical AI라고 부릅니다. "Physical AI는 AI가 물리적 세계와 상호작용하는 것입니다. 그것은 곧 로보틱스를 의미합니다." 2025년 1월 대만에서 열린 공장 개소식에서 한 말입니다.

기존 AI와 무엇이 다를까요? 지금까지의 AI는 디지털 환경에서만 작동했습니다. 텍스트를 읽고, 이미지를 보고, 코드를 짜는 것은 모두 비트와 바이트의 세계에서 일어나는 일이지요. Physical AI는 3차원 물리적 세계를 이해하고 그 안에서 직접 행동합니다. 컵을 집고, 문을 열고, 계단을 오르는 것이지요.

Deloitte의 2026 Tech Trends 보고서는 이렇게 정의합니다. "사전 프로그래밍된 명령을 따르는 전통적 로봇과 달리, 경험으로부터 학습하고 실시간 데이터에 기반해 행동을 적응시키는 시스템." 공장의 기존 로봇 팔은 정해진 동작만 반복합니다. Physical AI는 로봇이 새로운 상황을 만나도 스스로 판단하게 만드는 것을 목표로 합니다.

로보틱스의 안드로이드와 로봇의 두뇌

NVIDIA는 이 비전을 실현하기 위해 포괄적인 기술 스택을 만들었습니다.

Cosmos는 물리적 세계를 시뮬레이션하는 모델입니다. Cosmos Reason 2는 로봇이 세상을 보고, 이해하고, 행동하도록 돕는 비전-언어 모델이지요. 물리적 세계가 어떻게 작동하는지를 미리 학습시키는 일종의 "세계 모델(World Model)"입니다.

Omniverse는 Physical AI의 운영체제 역할을 합니다. 3D 시뮬레이션 환경에서 로봇을 훈련하고, 검증하고, 배포할 수 있습니다. 실제 공장에 로봇을 보내기 전에 가상 환경에서 수천 번을 연습시키는 것이지요.

GR00T는 휴머노이드 로봇 전용 파운데이션 모델입니다. 로봇의 "두뇌"에 해당하지요. 전신 제어를 지원하며, Cosmos Reason과 결합해서 더 나은 추론과 맥락 이해를 제공합니다. 2025년 3월 N1을 처음 공개한 뒤 판을 빠르게 올려, 2026년 1월 CES에서 N1.6을, 3월 GTC에서는 상업용으로 쓸 수 있는 N1.7을 내놓았습니다. 다음 세대 N2는 2026년 말에 내겠다고 예고했지요.[1]

TechCrunch는 NVIDIA의 전략을 "로보틱스의 안드로이드가 되겠다"로 요약했습니다. 스마트폰 제조사들이 안드로이드 위에서 다양한 폰을 만들듯, 로봇 제조사들이 NVIDIA 플랫폼 위에서 다양한

로봇을 만드는 것이지요. NVIDIA의 200만 로보틱스 개발자와 Hugging Face의 1,300만 AI 빌더 커뮤니티가 합쳐지면서 생태계가 급속히 확장되고 있습니다.

그렇다면 이 플랫폼 위에서 로봇을 실제로 움직이는 두뇌는 어떻게 생겼을까요? Physical AI의 핵심 기술 중 하나가 VLA(Vision-Language-Action) 모델입니다. 이름 그대로 시각(Vision), 언어(Language), 행동(Action)을 하나로 통합하는 멀티모달 파운데이션 모델입니다.

작동 방식은 이렇습니다. 로봇 주변 환경의 이미지와 텍스트 명령을 입력받으면, 저수준 로봇 동작을 직접 출력합니다. "저 빨간 컵을 테이블 위에 올려놓아"라는 말을 듣고, 카메라로 빨간 컵을 찾고, 팔을 뻗어 잡고, 테이블까지 옮기는 일련의 동작을 스스로 계획하고 실행합니다.

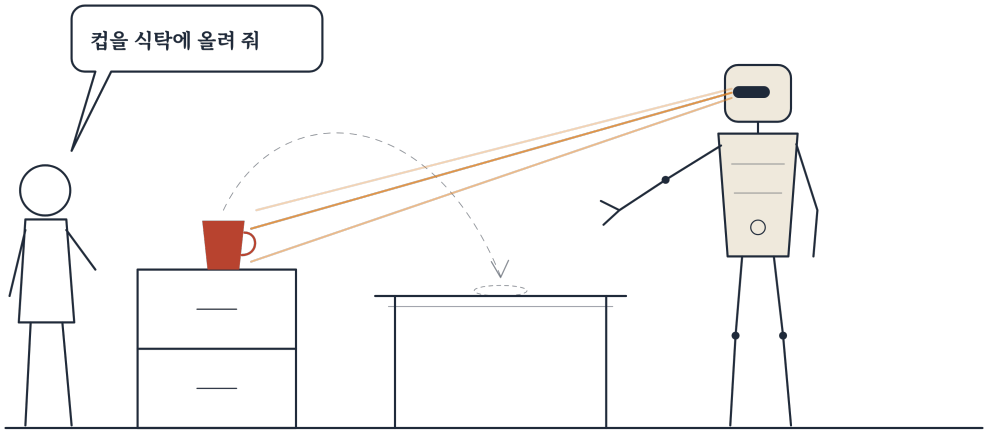


그림 6-4 로봇은 말 한마디를 듣고, 컵을 찾아보고, 옮길 길을 머릿속에 먼저 그려 보지요. 보고 이해하고 행동하는 일이 한 모델 안에서 이어집니다.

2026년에는 Physical Intelligence, NVIDIA, Ant Group 등 여러 기업이 VLA 모델을 공개했습니다. Robbyant사의 LingBot-VLA는 2만 시간 이상의 실세계 상호작용 데이터로 훈련되었고, 단일 모델로 단일 팔, 양팔, 휴머노이드 등 다양한 로봇에 배포할 수 있습니다. 2026년 말까지 1,000억 개 이상의 파라미터를 가진 VLA 모델이 나올 것이라는 전망도 있습니다.

화면 밖으로 나온 지능

AI 에이전트가 걸어온 길을 한 줄로 그려 보면 이렇습니다.

텍스트 AI(이해) → 멀티모달 AI(인식) → 에이전틱 AI(계획과 실행) → Physical AI(물리적 행동)

2024~2025년에는 에이전틱 AI가 디지털 워크플로우를 자동화했습니다. 코딩, 문서 작성, 데이터 분석 같은 일이지요. 2026년에는 Physical AI가 "생각할 뿐 아니라 물리적으로 행동하는" 단계로 확장됩니다. 이 확장이 얼마나 무거운 관문인지는 정책 시나리오 쪽에서도 확인됩니다. AI Futures Project의 2026년

분석은 2030년대 경제 성장의 병목이 지능보다 로봇 생산 능력에서 생길 것으로 봤습니다[2]. 예측이 아니라 권고 시나리오 속 분석이라는 점은 기억해 두세요. 지능은 복제가 쉽지만, 손과 몸은 공장에서 한 대씩 만들어야 하니까요.

젠슨 황의 Physical AI를 이 책의 비유로 옮기면, 데이터센터에서 만든 지능이 공장의 로봇 손끝까지 흘러가는 일입니다. 전기가 발전소에서 가정의 콘센트까지 갔던 길을 지능도 따라가는 셈이지요.

Deloitte는 이렇게 말했습니다. "Physical AI는 우리 시대의 가장 심오한 기술적 전환 중 하나를 나타낸다. 인공지능이 단지 생각하는 것을 넘어 물리적 세계에서 움직이고, 조작하고, 상호작용하는 시대의 도래." 이 전환이 니치에서 메인스트림으로 넘어가는 시점을 Deloitte는 18~24개월 이내, 즉 2027~2028년으로 전망합니다.

그런데 이 전환이 메인스트림에 닿으려면 넘어야 할 산이 하나 있습니다. 에이전트가 물리 세계에서 일하려면, 단순히 보고 행동하는 것만으로는 부족합니다. "세상이 어떻게 작동하는지"를 이해해야 합니다. 컵을 테이블 가장자리에 놓으면 떨어진다는 것, 무거운 물건 위에 가벼운 물건을 쌓아야 한다는 것, 문을 당기면 열린다는 것. 이런 물리 법칙과 인과관계를 학습한 AI 모델을 월드 모델(World Model)이라고 부릅니다.

월드 모델의 가장 열렬한 옹호자는 Meta 수석 AI 과학자 Yann LeCun입니다. LeCun은 오랫동안 "LLM은 진정한 지능이 아니다"라고 주장해왔습니다. LLM은 텍스트의 통계적 패턴을 학습한 것이지, 세상이 어떻게 돌아가는지를 이해한 것이 아니라는 뜻이지요. 그의 신념은 행동으로 이어졌습니다. 2026년 3월, LeCun은 Meta를 떠나 AMI Labs를 설립하고 10.3억 달러의 투자를 유치했습니다[3]. 세상을 이해하는 모델을 만들겠다는 것입니다.

다른 쪽에서도 움직임이 있습니다. 스탠퍼드 교수이자 ImageNet의 창시자인 Fei-Fei Li는 World Labs를 통해 "공간 지능(Spatial Intelligence)"을 연구하고 있습니다. 3D 세계를 이해하고 그 안에서 추론하는 AI를 만드는 것이 목표이지요. NVIDIA도 Cosmos를 Physical AI를 위한 월드 파운데이션 모델로 포지셔닝하고 있습니다. 앞에서 본 Cosmos가 이 역할을 맡지요.

커뮤니티의 반응도 뜨겁습니다. 2026년 2월, HuggingFace에서 가장 많은 관심을 받은 논문은 GigaBrain VLA였습니다. 2,410개의 추천을 받았지요[4]. 이 모델은 월드 모델 기반의 강화학습으로 로봇을 학습시킵니다. 로봇이 행동하기 전에 "이렇게 하면 어떻게 될까?"를 머릿속에서 시뮬레이션하는 것입니다. 사람이 공을 던지기 전에 궤적을 상상하는 것과 비슷하지요.

한 가지 비유로 정리해 볼까요. LLM이 언어 세계의 물리 법칙을 학습했다면, 월드 모델은 실제 물리 세계의 법칙을 학습합니다. LLM이 "다음에 올 단어"를 예측하듯, 월드 모델은 "다음에 일어날 일"을 예측합니다. 컵을 놓으면 떨어지고, 문을 밀면 열리고, 바퀴를 굴리면 앞으로 나아간다는 것을. 이 이해가 있어야 로봇은 처음 보는 상황에서도 올바르게 행동할 수 있습니다.

화면 속에 있던 에이전트가 이제 컵을 집고 문을 엽니다. 처음 본 식탁 위의 빨간 컵, 그 컵을 집는 손이 바로 다음 장의 주인공들입니다.



1. "NVIDIA GR00T N1.7 hits early access for humanoids" (Origin of Bots, 2026.3, https://www.originofbots.com/news/nvidia-gr00t-n17-hits-early-access-for-humanoids) — 30억 파라미터 VLA, Cosmos-Reason2 백본, 상업 라이선스 얼리 액세스; GTC 2026에서 N2 연말 출시 예고.

2. "Economics of Plan A" (AI Futures Project, 2026.7, https://ai-2040.com/supplements/economics-of-plan-a) — 시나리오 분석 기준, 2030년대 인지 노동은 AI로 빠르게 풍부해지는 반면 물리 노동(로봇)의 증설 속도가 성장의 제약이 된다고 전망.

3. "Yann LeCun's AMI Labs raises $1.03 billion to build world models" (TechCrunch, 2026.03, https://techcrunch.com/2026/03/09/yann-lecuns-ami-labs-raises-1-03-billion-to-build-world-models/)

4. GigaBrain-0.5M: VLA with World Model-Based RL (HuggingFace Papers, 2026.02)

로봇이 출근하는 날

출근길 횡단보도에서 사원증을 단 로봇과 나란히 신호를 기다리는 날, 언제쯤 올까요? 2026년의 휴머노이드는 데모 영상을 막 벗어나 제한된 현장에서 시험을 받기 시작했습니다. 멋진 공중제비보다 중요한 건 정해진 일을 안전하게 되풀이할 수 있느냐지요.

이 장을 읽을 때 기억할 잣대는 하나입니다. 발표는 약속이고, 납품과 반복 운용이 실력이라는 것. 아래 회사들의 소식도 모두 이 잣대로 읽어 보세요. 그럼 누가 어디서 먼저 쓸모를 증명하고 있는지 따라가 보지요.

발표는 약속, 납품은 실력

Tesla Optimus. Tesla는 Optimus의 차세대 시제품과 생산 계획, 목표 가격을 잇달아 공개했습니다. 손의 정밀도를 높이고 자연어 명령을 로봇 동작으로 연결하겠다는 구상도 내놓았지요.

회사의 설명을 들어 봐도 Optimus는 아직 배우고 데이터를 모으는 단계에 가깝습니다. 그래서 투자 계획보다 먼저 눈여겨볼 것은 작업 완료율과 개입 횟수, 고장 뒤 복구 시간, 사람과 함께 일할 때의 안전 기록입니다.

Figure AI. Figure AI는 처음에는 OpenAI와 차세대 휴머노이드 모델을 공동 개발하려 했습니다.

그런데 2025년 2월, 이 협약이 끝납니다. CEO Brett Adcock은 "LLM이 점점 더 똑똑해지지만 동시에 상품화되고 있다"고 이유를 밝혔습니다. "중대한 돌파구(Major breakthrough)"를 이유로 자체 AI 개발로 전환한 것이지요.

결과물이 Helix AI입니다. 보고, 말하고, 움직이는 정보를 함께 처리하는 VLA(Vision-Language-Action) 모델이지요. 앞 세대 Figure 02는 BMW 스파턴버그 공장 차체 라인에서 10개월 동안 X3 3만여 대의 생산을 거들었습니다. 2026년 6월부터는 Figure 03이 같은 공장 물류 구역에서, 뒤섞인 상자 속 부품을 조립 순서대로 대차에 옮겨 담는 일을 맡았습니다.[1]

Boston Dynamics Atlas. Boston Dynamics는 Atlas의 생산형 버전과 고객사 시험 계획을 공개했습니다. 초기 물량에 대한 예약도 이어졌지요.

Google DeepMind와의 Embodied AI 협력도 발표됐고, 2026년 9월에는 현대차 조지아 공장 안에 Atlas를 생산 라인에 맞춰 훈련하는 센터가 문을 열었습니다. 첫 일감은 부품 운반과 순서 맞추기이고, 조립은 2030년까지 차례로 넓혀 간다는 계획입니다.[2] Atlas가 얼마나 쓸모 있는지는 생산 라인에서 사람이 얼마나 자주 끼어들어야 하는지가 공개될 때 비로소 보이겠지요.

1X NEO. 1X Technologies는 NEO를 가정용 휴머노이드로 소개하고 가격과 배송 계획을 발표했습니다. 소비자가 살 수 있는 제품을 지향한다는 점이 눈에 띕니다.

공개 시연에서는 자연어 대화와 물체 인식, 원격 지원을 결합한 집안일 보조 가능성을 보여줬습니다. 그런데 집은 공장보다 훨씬 까다로운 일터입니다. 낯선 집에서 얼마나 오래 작동하는지, 원격 조작이 얼마나 자주 필요한지, 실수했을 때 사람과 물건을 어떻게 보호하는지가 더 중요한 제품 지표입니다.

산업 현장 공급 계약도 발표했지요. 진짜 수요는 고객이 같은 작업에 로봇을 계속 쓰는지에서 드러날 겁니다.

중국. 그런데 이 모든 기업보다 더 많은 로봇을 만들고 있는 곳이 있습니다. 중국입니다.

중국 기업들은 부품 공급망과 낮은 가격을 앞세워 출하량을 빠르게 늘리고 있습니다. 숫자가 말해 주지요. 2026년 상반기 세계에서 출하된 휴머노이드는 약 1만 9천 대였는데, 그 97%가 중국 기업 제품이었습니다. AgiBot이 8천여 대, Unitree Robotics가 6천 대 가까이 내보냈습니다. 한 조사기관의 집계이고 상당수는 시범·개발용 납품이라는 점은 감안해야 합니다.[3]

8월에는 Unitree가 상하이 증시에 상장해 첫날 주가가 공모가의 몇 배로 뛰었다가, 한 주 만에 고점에서 절반 가까이 꺼지기도 했습니다.[4] 기대와 실적 사이의 거리를 시장이 먼저 재 본 셈입니다.

2026년 춘절 CCTV 갈라에서 Unitree 로봇들이 쿵푸와 공중제비를 선보이며 화제가 되기도 했습니다. CES 2026에서 휴머노이드 로봇을 전시한 기업의 절반 이상이 중국 기업이었습니다.

중국 로봇의 무기는 가격입니다. 물론 공개된 시작 가격과 아직 출시되지 않은 목표 가격, 산업용 장비 견적은 조건이 달라 단순 비교하기 어렵습니다. 그래도 중국의 대량 생산 능력이 가격 경쟁을 앞당기고 있는 것만큼은 사실이지요.

공장에서 병원, 그리고 부엌까지

로봇은 공장, 병원, 가정에서 서로 다른 속도로 자리를 잡습니다.

공장은 이미 데모 영상을 넘어 제한된 생산 업무를 시험하는 단계입니다. Agility Robotics의 Digit는 물류 시설에서 토트(tote, 물건을 담아 나르는 플라스틱 상자)를 옮기는 반복 작업을 맡았습니다. 사람처럼 무엇이든 하는 로봇보다, 정해진 동선을 오래 반복하는 로봇이 먼저 자리를 잡는 이유가 여기에 있습니다.

Digit의 강점은 범용성보다 좁은 작업 범위에 있습니다. 함대 관리 도구로 정해진 동선과 작업 상태를 관리하고, 문제가 생기면 사람이 개입합니다. 안전 인증을 추진한다는 발표도 있었습니다.

공장에서 로봇이 하는 일은 명확합니다. 토트와 빈을 옮기고, 컨테이너를 하역하고, 반복적인 물류 작업을 하고, 품질을 검사합니다. 사람이 하면 허리가 아프고 지루한 일이지만, 로봇에게는 지루함이라는 개념 자체가 없지요.

병원에서는 아직 휴머노이드보다 특화된 로봇이 먼저 쓰입니다. Moxi는 약품 배송, 검체 운반, 물품 수거 같은 병원 물류를 맡습니다. 효과를 말할 때는 홍보 자료의 절감률보다 병동별 호출 완료율, 지연, 직원의 추가 대응 시간을 함께 봐야 합니다. 로봇이 이동 업무를 덜어도 그 시간이 자동으로 환자 돌봄에 쓰이는 것은 아니기 때문입니다.

고령화 사회에서 로봇은 이동 보조, 물품 운반, 약 복용 알림처럼 범위를 정하기 쉬운 일부터 부담을 덜 수 있습니다. 그렇다고 돌봄 인력을 곧바로 줄여도 된다는 뜻은 아닙니다. 장애가 생겼을 때 누가 대신할지, 환자와 보호자가 거부할 수 있는지, 사고 기록을 누가 검토할지를 먼저 정해야 합니다.

부엌은 공장보다 훨씬 어렵습니다. 물건의 위치가 매번 달라지고, 뜨거운 조리 도구와 어린이, 반려동물이 같은 공간을 오갑니다. 재료를 알아보고 요리를 제안하는 시연은 가능성을 보여주지만, 집안일을 안정적으로 끝냈다는 증거는 아닙니다.

로봇이 공장, 병원, 부엌에 들어오면 무엇이 바뀔까요?

가장 먼저 바뀌는 것은 사람이 시간을 쓰는 방식입니다. 공장에서는 단순 이동과 검사 대신 공정 개선과 품질 관리에 집중할 수 있습니다. 병원에서는 물류 대신 환자 케어에 시간을 쓸 수 있지요. 가정에서는 집안일에 빼앗기던 저녁 시간을 가족과 함께 보낼 수 있습니다.

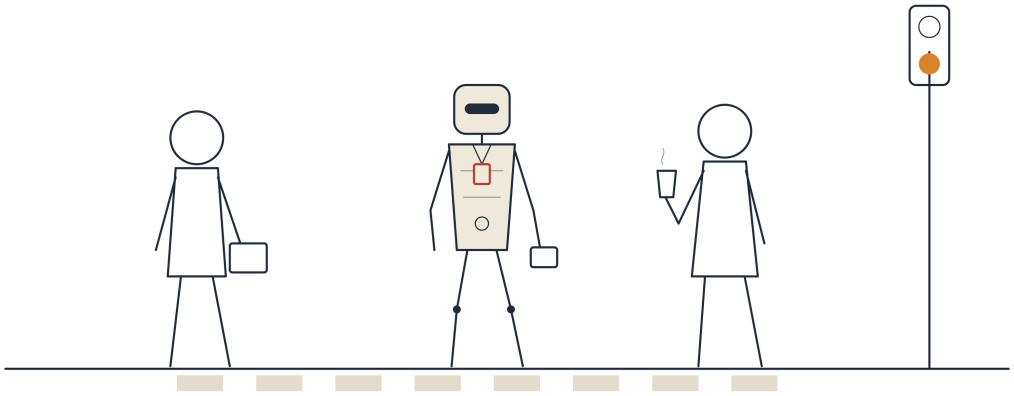


그림 6-5 언젠가 출근길 횡단보도에서 사원증을 단 로봇과 나란히 신호를 기다리게 될지도 모릅니다. 그날이 언제일지는 시연 영상이 아니라 현장 배치 기록이 알려 줄 겁니다.

시뮬레이션과 실제 환경의 차이(Sim-to-Real Gap)는 아직 남아 있습니다. 잘못된 판단이 생산 라인에서 동작으로 이어질 위험, 연결된 로봇 함대의 보안, 서로 다른 규제와 높은 단가도 배치 속도를 늦춥니다.

대량 생산이 언제 시작될지는 아직 아무도 모릅니다. 공장과 물류처럼 환경을 통제하기 쉬운 곳에서 먼저 증거가 쌓이고, 병원과 가정으로 갈수록 안전과 책임, 가격의 문턱이 높아지겠지요. 로봇은 같은 날 모든 곳에 출근하지 않습니다.

언젠가 횡단보도에서 로봇과 나란히 서는 날이 오더라도, 그 로봇은 아마 물류창고에서 먼저 몇 년쯤 상자를 나르다 온 녀석일 겁니다.



---

1. "BMW Group deploys Figure 03 humanoid after tests of previous version" (The Robot Report, 2026.6, https://www.therobotreport.com/bmw-group-deploys-figure-03-humanoid-after-tests-previous-version/) — 2026.6.25 발표, 스파턴버그 공장 물류 시퀀싱 투입.

2. "Boston Dynamics launches robot training centre at Hyundai's US Metaplant" (Just Auto, 2026.9, https://www.just-auto.com/newsletters/boston-dynamics-launches-robot-training-centre-at-hyundais-us-metaplant) — Robotics Metaplant Application Center(RMAC) 개소, 초기 물류·부품 시퀀싱.

3. "Chinese makers accounted for more than 97% of global humanoid robot shipments in H1 2026" (TechNode, 2026.8.11, https://technode.com/2026/08/11/chinese-makers-accounted-for-more-than-97-of-global-humanoid-robot-shipments-in-h1-2026/) — Smart Analytics Global 집계(약 1만 9,100대), 단일 조사기관 수치.

4. "Unitree shares slump 45% amid bubble fears" (TNW, 2026.8, https://thenextweb.com/news/unitree-shares-45-percent-slump-bubble-fears) — 2026.8.19 상하이 커촹반 상장, 시초가 공모가 대비 629% 상승 후 약 45% 하락.

SOUL.md를 가진 에이전트

AI 에이전트와 일해 본 분이라면 한 번쯤 겪어 보셨을 겁니다. 어제 분명히 합의한 코딩 규칙을 오늘은 모르는 척합니다. 프로젝트 구조를 처음 보는 것처럼 다시 묻고, 하지 말라던 파일을 또 건드리지요. 세션이 바뀔 때마다 건망증 심한 새 팀원을 맞는 기분입니다.

왜 그럴까요? 모델은 사람처럼 경험을 쌓아 두지 않습니다. 새 대화가 시작되면, 앞서 합의한 규칙과 맥락은 누군가 다시 건네주지 않는 한 사라지지요.

그래서 개발 현장에서는 이 정보를 마크다운 파일에 적어 두기 시작했습니다. 파일 이름은 도구와 프로젝트마다 다르지만, 하는 일은 크게 셋입니다. 작업 규칙, 역할과 말투, 사용자와 프로젝트의 맥락이지요.

가장 흔한 것은 작업 규칙입니다. CLAUDE.md나 AGENTS.md 같은 파일에는 저장소에서 지킬 규칙을 적습니다. 어떤 언어와 명령을 쓰는지, 테스트는 어떻게 돌리는지, 어디까지 손대도 되는지, 코드 리뷰 기준은 무엇인지 같은 내용이지요.

이 파일은 에이전트의 성격을 만든다기보다 작업 계약서에 가깝습니다. "테스트를 먼저 실행한다", "이 디렉터리 밖은 수정하지 않는다"처럼 눈으로 확인할 수 있는 행동을 적을수록 효과도 확인하기 쉽습니다.

규칙을 많이 적으면 더 잘 따를까요? 꼭 그렇지는 않습니다. 규칙끼리 부딪히거나 무엇이 먼저인지 정해 두지 않으면, 에이전트는 어느 지시를 따라야 할지 헷갈립니다. 중요한 규칙은 짧게 쓰고, 어디에 적용하는지와 어떻게 확인할지를 옆에 붙여 두는 편이 낫지요.

규칙이 '무엇을 하느냐'를 다룬다면, '어떤 태도로 하느냐'를 적으려는 파일도 있습니다. SOUL.md는 에이전트의 역할과 말투, 가치 판단의 경계를 따로 적어 보려는 실험적 관습입니다. "회의에서 반대 근거를 먼저 찾는다", "불확실한 수치는 사실처럼 말하지 않는다", "사용자의 감정을 추측하지 않는다"처럼 상대를 대하는 방식을 정의하지요.

이름에 '영혼'이 들어가 있으니 혹시 자아라도 생기는 걸까요? 그렇지는 않습니다. 모델은 파일에 적힌 문장을 지금 대화의 지시로 처리할 뿐입니다. 같은 파일을 써도 모델 버전과 도구, 대화 내용에 따라 결과가 달라질 수 있고요.

그러니 SOUL.md가 쓸모 있는지는 거창한 선언으로 따질 일이 못 됩니다. 같은 상황에서 판단 기준이 유지되는지, 금지한 행동을 피하는지, 모를 때 모른다고 말하는지를 되풀이해 시험해 보면 드러나지요.

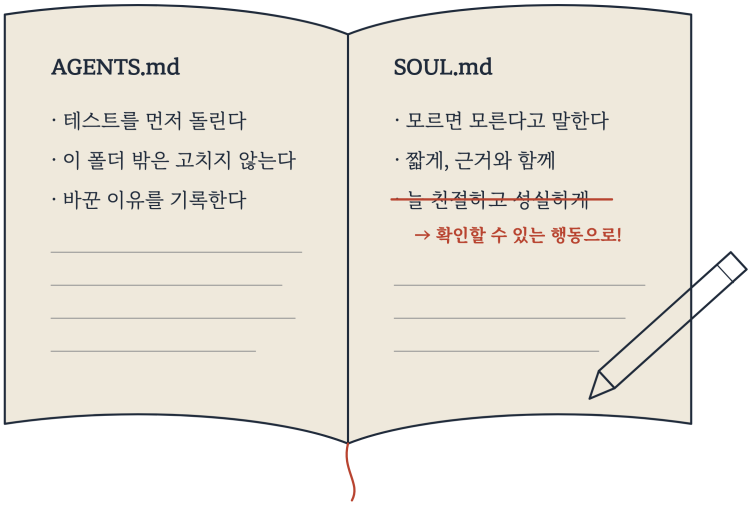


그림 6-6 "늘 친절하고 성실하게" 같은 다짐은 시험할 수가 없습니다. 지우고, 확인할 수 있는 행동으로 고쳐 적습니다.

파일에 적을 내용	피해야 할 표현	더 나은 방식
역할	“현명하고 친절한 동료”	“근거가 약하면 반대 의견과 확인 방법을 제시한다”
말투	“인간적으로 대화한다”	“짧은 문단을 쓰고 과장된 감정 표현을 피한다”
경계	“절대 해를 끼치지 않는다”	“외부 전송과 삭제는 승인 전 실행하지 않는다”
품질	“항상 정확하다”	“수치와 인용은 출처를 확인하고 확인하지 못하면 표시한다”

남은 하나는 함께 일하는 사람에 관한 기록입니다. USER.md처럼 사용자의 선호와 배경을 적어 두는 파일도 생각해 볼 수 있습니다. 자주 쓰는 도구나 원하는 설명의 깊이를 알려 주면 같은 말을 되풀이하지 않아도 되지요.

그런데 개인화 정보는 금세 민감 정보가 됩니다. 건강, 정치 성향, 가족 관계, 위치 같은 내용을 편하다는 이유로 계속 쌓아 두면 유출과 오용의 위험도 함께 커집니다. 사용자가 무엇이 저장됐는지 보고, 고치고, 지울 수 있어야 하지요. 오래된 정보에는 유효기간을 두는 편이 좋습니다.

사실과 추측을 섞지 않는 것도 중요합니다. 어느 날 짧은 답을 원했다고 "이 사용자는 설명을 싫어한다"고 못 박아 버리면, 개인화가 편견이 됩니다. 관찰한 선호에는 날짜와 근거를 붙여 두고, 중요한 결정 앞에서는 다시 물어보는 것이 좋습니다.

이렇게 세 갈래의 파일을 잘 갖춰도 풀리지 않는 일은 남습니다. 정체성 파일이 기억 장치를 대신하지는 못합니다. 실제 작업 기록과 결정의 근거, 변경 이력은 따로 로그와 저장소에 남겨야 하지요. 파일에 "보안을 중시한다"고 적어 둔다고 권한 통제와 감사 기록이 저절로 생기지도 않습니다.

일관성은 몇 겹이 함께 받쳐 줄 때 나옵니다. 짧고 서로 부딪히지 않는 지시, 필요한 정보만 골라 불러오는 맥락 관리, 실제로 손이 닿는 범위를 막는 기술적 통제, 규칙을 지켰는지 확인하는 평가와 기록. 이 겹들을 집 안 배선도처럼 그려 보는 일은 뒤에 나올 「하네스 엔지니어링」에서 이어집니다.

SOUL.md라는 이름은 거창하지만, 하는 일은 소박합니다. 에이전트에게 바라는 행동을 글로 적어 두고, 그대로 지키는지 시험할 수 있게 하는 것이지요.



나를 대신하는 에이전트

이런 상상을 해볼까요. 아침에 일어나보니 밤새 에이전트가 밀린 일들을 정리해놓았습니다. 간밤에 들어온 자료를 검토해 핵심만 요약해주고, 손댈 게 없는 나머지는 알아서 마무리했지요. 오후 회의에 나 대신 참석해서 내 스타일로 의견을 냈고, 회의록을 정리해두었습니다. 저녁에는 다음 주 출장 일정을 짜면서 내가 좋아하는 통로 좌석과 조용한 호텔을 예약했습니다.

이것이 USER.md를 읽는 에이전트가 할 수 있는 일의 향후 비전입니다. 현재 에이전트 메모리와 개인화 기술은 빠르게 발전하고 있지만, 위와 같은 수준의 자율적 대리 행동은 아직 실현되지 않은 미래 시나리오입니다. 에이전트가 나의 선호도, 업무 스타일, 말투, 습관을 알고 있으면 나를 대신해서 행동할 수 있지요. 그런데 여기서 핵심적인 기술 문제가 하나 있습니다. 에이전트가 나를 "기억"해야 한다는 것입니다.

나를 기억하게 만드는 기술

AI 에이전트의 가장 큰 약점은 기억력입니다. LLM의 context window — 한 번에 처리할 수 있는 텍스트의 양 — 에는 한계가 있습니다. 창문이 작아서 한꺼번에 볼 수 있는 것이 제한되어 있는 셈이지요. 대화가 길어지면 앞의 내용을 잊어버리고, 세션이 바뀌면 아예 백지 상태로 돌아갑니다.

이 문제를 해결하기 위한 기술들이 빠르게 발전하고 있습니다.

에이전트에게 기억을 만들어주는 대표적인 방식은 앞서 「에이전트를 똑똑하게」 장에서 본 Letta(옛 MemGPT)의 3계층 메모리입니다. 컴퓨터의 RAM·디스크·콜드 스토리지를 그대로 본떠 코어·리콜·아카이벌로 기억을 나눈다는 그 구조지요. "이 사용자는 시니어 Go 개발자이고, 간결한 설명을 선호한다" 같은 핵심 정보는 늘 떠 있는 코어에, 지난 대화 기록은 검색 가능한 리콜에, 오래된 정보는 아카이벌에 내려앉습니다.

앞서 다 다루지 않은 대목은 이 계층이 실제로 어떻게 돌아가느냐입니다. context window가 꽉 차면 에이전트는 오래된 메시지의 약 70%를 밖으로 내보내면서(eviction) 재귀적 요약으로 핵심만 추려 남깁니다. 그리고 앞서 본 '드림'이 바로 이 대목에서 일합니다. 유휴 시간에 비동기로 기억을 정제하는 이 과정은 Sleep-Time Compute라는 이름으로도 불리는데, 에이전트가 "자면서" 그날의 기억을 정리하는 셈이지요.

Mem0는 다른 접근법을 씁니다. 그래프 메모리라는 방식입니다.

벡터 데이터베이스가 "사용자가 커피를 좋아한다"를 기억한다면, 그래프 메모리는 "사용자가 특정 가게의 커피를 선호하고, 지난 화요일에 주문했으며, 아침 루틴을 논의하면서 언급했다"까지 기억합니다. 정보가 시간에 걸쳐 어떻게 연결되는지를 보존하는 것이지요.

Mem0의 그래프 메모리(Mem0g)는 이렇게 작동합니다. 대화에서 엔티티를 추출하고, 관계를 추론합니다. 겹치거나 모순되는 정보가 감지되면 LLM 기반 해결기가 추가할지, 병합할지, 무효화할지, 건너뛸지를 결정합니다.

Mem0 팀이 발표한 벤치마크에서는 OpenAI의 기본 메모리보다 응답 정확도가 26% 높게 나왔습니다. 개발사가 직접 잰 숫자이긴 하지만, 기억을 어떻게 엮느냐에 따라 답의 질이 달라진다는 점은 분명히 보여주지요.

에이전트가 나를 기억하면 무엇이 달라질까요? 편리해지는 정도를 넘어, 함께 일하는 관계의 결이 달라집니다.

기억이 없는 에이전트는 매번 "어떤 프로그래밍 언어를 사용하시나요?"부터 묻습니다. 기억이 있는 에이전트는 "지난번에 Go로 리팩토링하기로 했던 모듈, 진행 상황은 어떻게 되었나요?"라고 묻습니다.

기억이 없는 에이전트는 내가 싫어하는 패턴으로 코드를 짭니다. 기억이 있는 에이전트는 내 코딩 스타일에 맞춰서, 내가 선호하는 에러 처리 방식으로, 내가 쓰는 네이밍 컨벤션으로 코드를 짭니다.

VentureBeat의 2026 엔터프라이즈 AI 예측에 따르면, contextual memory는 운영 중인 에이전틱 AI 배포의 필수 요건이 될 것입니다. 기억 없는 에이전트는 경쟁력이 없어지는 셈이지요.

저장소에서 컨텍스트 엔진으로

에이전트 메모리는 세 가지 방향으로 진화하고 있습니다.

하나는 인프라로서의 메모리입니다. 개발자가 표준화된 API로 메모리 기능을 호출합니다. 마치 데이터베이스를 쓰듯이 에이전트에게 기억 능력을 부여하는 것이지요.

다른 하나는 거버넌스로서의 메모리입니다. 감사 가능한 "메모리 그래프"와 편집, 삭제 제어가 가능한 엔터프라이즈 대시보드가 등장합니다. 에이전트가 무엇을 기억하고 있는지 사람이 확인하고 관리할 수 있어야 하니까요.

마지막 하나는 정체성으로서의 메모리입니다. 에이전트가 사용자와의 협업 이력을 발전시키면서 신뢰를 쌓아갑니다. "이 사용자와 3개월간 일했고, 이 프로젝트에서 이런 결정을 함께 내렸다"는 맥락이 축적되는 것이지요.

세 방향은 결국 한 질문으로 모입니다. 에이전트가 나를 진정으로 대신하려면 내가 3개월 전에 한 이야기, 지난주에 내린 결정, 어제 바꾼 방침까지 기억해야 하지요. 그런데 기억한다는 것은 단순히 저장하는 것이 아닙니다. 필요한 순간에 적절한 기억을 꺼내올 수 있어야 합니다. 여기서 등장하는 것이 RAG(Retrieval-Augmented Generation, 검색 증강 생성)입니다.

RAG의 원리는 직관적입니다. 에이전트가 질문을 받으면 바로 답하지 않고, 먼저 관련 정보를 검색합니다. 그리고 검색한 정보를 참고하여 답변을 생성하지요. 마치 시험을 볼 때 오픈북처럼 — 모든

것을 외우고 있지 않아도 필요할 때 찾아볼 수 있습니다. 에이전트의 파라미터에 모든 지식을 넣을 수는 없지만, 외부 저장소에서 관련 정보를 가져와 context window에 넣어주면 마치 알고 있는 것처럼 답할 수 있습니다.[1]

2025년에서 2026년으로 넘어오면서 RAG는 단순한 검색 도구에서 "컨텍스트 엔진"으로 진화하고 있습니다.

첫 번째 변화는 자기 교정입니다. Self-RAG, Corrective RAG 같은 기법이 등장했습니다. 기존 RAG가 검색 결과를 그대로 믿었다면, 자기 교정 RAG는 검색 결과가 질문과 관련이 있는지, 정보가 정확한지를 스스로 판단합니다. 관련성이 낮다고 판단하면 검색 쿼리를 수정해서 다시 검색합니다. 사람이 도서관에서 자료를 찾을 때, 첫 번째 검색 결과가 마음에 들지 않으면 키워드를 바꿔서 다시 찾는 것과 같습니다.

두 번째 변화는 멀티모달 검색입니다. 초기 RAG는 텍스트만 검색했지만, 이제는 이미지, 비디오, 오디오까지 검색 대상이 됩니다. 에이전트가 "지난달 회의에서 공유한 아키텍처 다이어그램"을 찾아달라고 하면, 텍스트가 아니라 이미지를 검색해서 가져올 수 있습니다.

2026년에는 더 근본적인 접근도 등장했습니다. MemOS는 LLM을 위한 메모리 운영체제입니다. 컴퓨터에 운영체제가 메모리를 관리하듯이, MemOS는 에이전트의 기억을 세 가지 수준에서 통합 관리합니다. 평문 기억(텍스트로 저장된 명시적 정보), 활성화 기억(모델의 내부 상태에 저장된 패턴), 파라미터 기억(모델 가중치에 학습된 지식)을 하나의 체계로 묶는 것이지요.[2] MAGMA는 멀티 그래프 기반 에이전트 메모리 아키텍처로, 서로 다른 유형의 기억을 여러 겹의 그래프로 연결해서 관리합니다. 단순히 정보를 저장하는 것이 아니라, 정보 사이의 관계와 맥락까지 보존하는 것입니다.

발전소 옆에 연료 저장소가 있듯이, 에이전트 곁에도 기억 저장소가 있어야 합니다. RAG는 그 저장소에서 필요한 연료를 꺼내 오는 장치이고, 지금은 스스로 고쳐 찾고 그림과 소리까지 뒤지는 컨텍스트 엔진으로 자라는 중이지요.

그러니 나를 대신하려면 먼저 나를 기억해야 합니다. 2026년의 에이전트는 메모를 남기고 꺼내 읽는 법을 막 익히는 중이고, 회의에 나 대신 앉는 날은 그다음 이야기입니다.



1. "Retrieval-Augmented Generation: A Comprehensive Survey" (arXiv:2506.00054)
2. MemOS: Memory Operating System for LLMs (2026)

에이전트들의 사회

장면 하나를 그려 볼까요. 날씨 때문에 항공편이 취소되었습니다. 내 에이전트가 이 사실을 감지합니다. 곧바로 항공사 에이전트와 대화해서 대체 편을 찾습니다. 동시에 캘린더 에이전트에게 내일 오전 회의를 오후로 옮겨달라고 요청합니다. 호텔 에이전트에게는 하루 연장을 협상하고, 레스토랑 에이전트에게는 저녁 예약을 취소합니다. 그리고 음식 배달 에이전트에게 호텔로 음식을 주문합니다. 이 모든 것이 내가 잠든 사이에 일어납니다.

소설 속 장면 같지만, Goldman Sachs가 2026년 전망 보고서에 그려 넣은 시나리오입니다. 아직은 전망이지요. 그렇다면 이 장면이 실제로 돌아가려면 무엇이 필요할까요?

명함과 성격, 그리고 시장

이 시나리오가 가능하려면 에이전트들이 서로 대화할 수 있어야 합니다. A2A 프로토콜은 그 기반을 제공하지요. 하지만 단순히 메시지를 주고받는 것만으로는 부족합니다. 에이전트가 상대 에이전트를 신뢰하고, 역량을 평가하고, 조건을 협상할 수 있어야 합니다.

앞서 「에이전트 경제」 장에서 본 A2A의 Agent Card가 여기서도 핵심 역할을 합니다. 각 에이전트가 JSON 형식의 디지털 명함으로 자기 역량을 공개하고 — "나는 항공편 예약이 가능하다", "나는 결제 처리 권한이 있다" 같은 식으로 — 상대 에이전트가 이 카드를 보고 적합한 파트너를 자동으로 고르는 방식이지요. 송전망에 꽂을 기기마다 정격 표시가 붙어 있는 것과 비슷합니다.

명함으로 역량을 알았다면, 그다음 궁금한 것은 상대가 어떤 태도로 일하느냐입니다. 에이전트 사회에서는 성격도 중요하지요. 고객 서비스 에이전트는 인내심 있게, 법률 에이전트는 정확하고 신중하게 굴어야 하지요. 앞 장에서 본 SOUL.md가 이런 태도를 한 에이전트 안에 적어 두는 파일이었다면, SoulSpec은 그 파일들을 묶어 다른 도구로 옮겨 가도 읽히게 하려는 오픈 규격 제안입니다. 아직 초기 실험 단계이지요.

에이전트가 플랫폼을 옮기거나 다른 에이전트와 협업할 때, 이런 파일이 "이 에이전트는 이렇게 일한다"를 알려 줍니다. 사람이 이직할 때 이력서를 들고 가듯, 에이전트도 자기 역할과 태도를 적은 문서를 들고 이동하는 셈이지요.

작업 규칙을 담은 AGENTS.md와 역할·태도를 담은 SOUL.md를 나눠 적는 습관이 에이전트 사회에서 더 중요해지는 이유도 여기 있습니다. 같은 일을 맡겨도 태도에 따라 소통 방식과 판단의 결이 달라지니까요. 물론 파일에 적힌 성격이 실제 행동으로 이어지는지는 따로 시험해 봐야 합니다.

에이전트가 서로 소통하고 정체성을 가지면, 자연스럽게 마켓플레이스가 형성됩니다. 내 에이전트가 법률 검토가 필요하면 법률 에이전트를 "고용"하고, 번역이 필요하면 번역 에이전트를 찾고, 디자인이 필요하면 디자인 에이전트와 계약합니다.

앞서 「에이전트 경제」 장에서 소개한 Microsoft Research의 "에이전트의 웹(Web of Agents)" 비전, 어시스턴트 에이전트가 브라우저 역할을, 서비스 에이전트가 웹사이트 역할을 한다는 그 구도가 여기서는 실제 마켓플레이스로 이어집니다.

이 생태계에서 중요한 것은 탈중앙화입니다. 현재의 앱 스토어처럼 특정 플랫폼이 지배하는 "벽으로 둘러싸인 정원(walled gardens)"이 아니라, 에이전트가 자유롭게 상호작용하는 열린 생태계를 지향합니다. A2A, MCP, UCP 같은 오픈 프로토콜이 이것을 가능하게 합니다.

중국은 에이전트 사회의 기반을 어떻게 깔고 있을까요

앞서 「중국의 에이전트 국가표준」 장에서 상호연결 규격을 자세히 봤습니다. 여기서는 그 정책과 기술이 에이전트 사회를 어떻게 떠받치는지만 한 걸음 떨어져 보겠습니다.

2026년 5월 중국 국가인터넷정보판공실(CAC)·국가발전개혁위원회(NDRC)·공업정보화부(MIIT)가 함께 발표한 《지능체 규범응용과 혁신발전 실시의견(智能体规范应用与创新发展实施意见)》은 에이전트를 앱 하나가 아니라 산업과 소비, 의료·교육·행정으로 확산할 기반으로 다룹니다. 사용자가 직접 결정할 일, 사용자 승인이 필요한 일, 에이전트가 자율적으로 처리할 일을 구분하고, 민감 업종에는 신고·검사·문제 제품 회수 같은 조치를 두도록 했습니다.[1] 자율성을 넓히되 마지막 결정권과 책임의 경계를 먼저 그으려는 방식입니다.

기술 규격도 따라붙었습니다. GB/Z 185-2026은 전체 구조와 신원코드·신원관리·에이전트 서술·발견·상호작용·도구 호출을 다루는 국가표준화 지도성 기술문서 7건입니다.[2] 서로 다른 플랫폼의 에이전트가 상대를 찾고 신원을 확인한 뒤 일을 맡길 수 있도록 공통 골격을 제시한 것이지요. 다만 GB/Z는 의무 인증 규정이 아니라 권고적 성격의 지도문서입니다. 규격이 발표됐다는 사실과 모든 제품이 실제로 호환된다는 것은 구분해야 합니다.

신원 규격은 이동통신망으로도 내려가고 있습니다. 차이나모바일(China Mobile)이 추진하는 AI-eSIM은 eSIM의 가입자 인증과 통신 연결에 모델·에이전트 서비스를 결합해, 단말의 번호를 에이전트 서비스의 계정과 연결하려는 구상입니다. 2026년 5월에는 AI-eSIM 실험실이 출범했고, 7월에는 중국통신기업협회 산하 AI-eSIM 전문위원회가 만들어졌습니다.[3] 이 위원회는 칩·통신사·클라우드·단말 기업이 함께 기술 연구와 산업 표준을 개발할 협의체입니다. 아직 완성된 국가표준이 아니라 표준 공동개발을 시작한 단계라는 점은 분명히 해둘 필요가 있습니다.

그 위에서 기업용 실행 기반도 만들어지고 있습니다. 랑차오클라우드(Inspur Cloud)의 하이뤄(海若) AgentOS는 모델·지식·데이터·도구를 연결하고 에이전트의 개발·배포·운영을 관리하는 플랫폼입니다. 2025년 중국정보통신연구원(CAICT)의 지능체 플랫폼 역량 시험을 통과했다고 회사가 발표했습니다.[4]

여기서 AgentOS는 중국 정부가 만든 하나의 국가 운영체제가 아니라 기업 제품군의 이름입니다. 같은 AgentOS라는 이름을 쓰는 다른 프로젝트와도 구분해야 하지요.

에이전트끼리 물건을 고르고 결제하려면 통신 규격만으로는 부족합니다. 누가 얼마까지 구매를 위임했는지, 결제 직전에 사람의 확인이 필요한지, 거래 기록을 어떻게 남기고 분쟁 때 무엇을 증명할지를 정해야 합니다. 알리페이와 생태계 기업들이 공개한 지능체 상업 신뢰 프로토콜(ACT)은 위임·상업적 상호작용·지급·신뢰 기록을 하나의 흐름으로 묶습니다. CAICT와 기업들이 공개한 ASL(Agent Security Link)은 신원, 연결, 의도, 권한을 검증해 MCP·A2A를 보완하려 합니다.[5]

이 대목은 이름을 정확히 불러야 합니다. ACT와 ASL은 현재 확인되는 범위에서 국가표준이 아니라 기업·연구기관이 만든 개방형 기술 프로토콜입니다. 국가 차원의 GB/Z 185가 에이전트 상호연결의 공통 골격을 만들고, 산업 프로토콜이 그 위에 결제와 보안 규칙을 얹는 구도에 가깝습니다.

이 흐름을 겹쳐 보면 에이전트 사회에 필요한 기반이 보입니다. 대화 규격만으로는 사회가 되지 않습니다. 신원, 권한, 운영 기반, 결제, 감사 기록이 함께 있어야 합니다. 중국은 정책 지침과 국가 지도문서, 기업 플랫폼, 산업 프로토콜을 동시에 움직이며 이 층들을 빠르게 채우고 있습니다. 실제 상호운용성과 이용자 보호가 발표만큼 작동하는지는 앞으로 확인해야 할 과제입니다.

도구에서 동료로 가는 길의 규칙

에이전트의 역할은 현재 다음과 같은 층위로 나뉩니다.

처음에 AI는 도구(Tool)였습니다. 사람이 명령하면 실행합니다. 검색하면 결과를 보여주고, 번역하면 번역문을 내놓습니다. 주체적 판단 없이 시킨 일만 합니다.

다음 단계는 대리인(Representative)입니다. 사람의 의도를 이해하고, 일정 범위 안에서 자율적으로 행동합니다. 이메일을 대신 쓰고, 회의를 대신 잡고, 쇼핑을 대신 합니다.

그 다음이 동료(Colleague)입니다. 에이전트가 자기 전문성과 정체성을 갖고, 다른 에이전트와 협업하면서 복잡한 프로젝트를 수행합니다. 사람처럼 의견을 내고, 대안을 제시하고, 때로는 "이건 좋은 생각이 아닌 것 같다"고 말합니다.

2026년 현재, 우리는 도구에서 대리인으로 넘어가는 전환기에 있습니다. 동료 수준의 에이전트는 아직 먼 이야기이지만, 대리인 수준은 이미 현실이 되고 있지요.

대리인이 현실이 되는 만큼 질문도 무거워집니다. 에이전트들이 사회를 이루면 규칙도 필요합니다. 에이전트가 거짓말을 하면 어떻게 될까요? 에이전트가 사기 거래를 하면 누가 책임지나요? 에이전트 간 분쟁은 어떻게 해결하나요?

Authorized Agency(승인된 대리) 프레임워크가 하나의 답을 제시합니다. 네 가지 구성요소가 있습니다. 에이전트가 할 수 있는 일의 범위를 정하는 권한 범위(Authority Envelope), 에이전트의 행위에 대해

답변하는 기록 책임자(Human-of-Record), 언제든 에이전트를 멈출 수 있는 중단 권한(Interrupt Authority), 그리고 누가 승인했고 누가 책임지는지를 추적하는 응답 가능 체인(Answerability Chain).

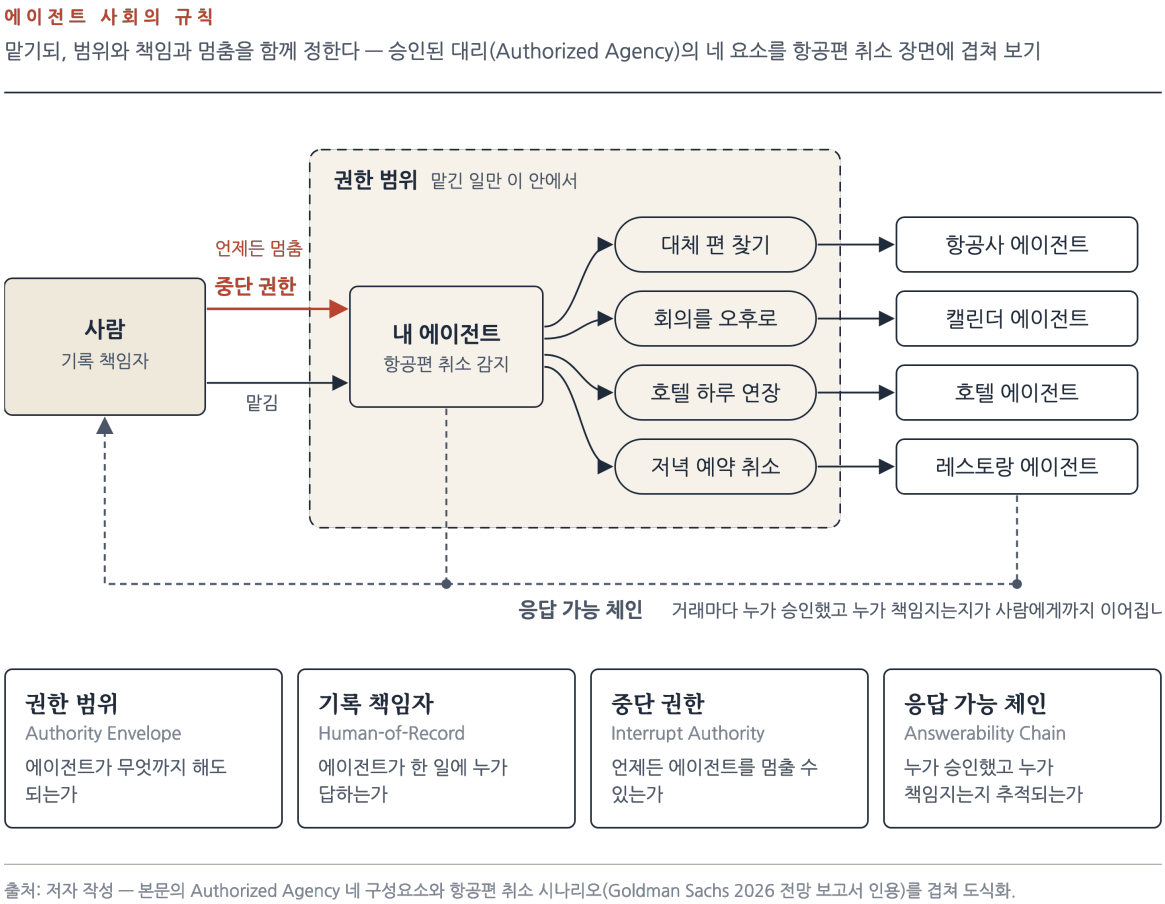


그림 6-7 에이전트에게 일을 맡길 때는 맡길 범위, 대신 답할 사람, 멈출 손잡이, 거슬러 올라갈 기록을 함께 정해 둡니다. 항공편이 취소된 밤의 장면에 네 가지를 겹쳐 보면 이렇습니다.

에이전트 사회가 커질수록 이런 규칙 없이는 굴러가기 어렵습니다. 사람의 사회에 법과 윤리가 있듯, 에이전트 사회에도 그에 맞는 규범이 필요하지요. 이 네 가지가 사람이 일하는 현장에서 어떤 모습이 되는지는 뒤에 나올 「오케스트레이터: 에이전트 시대 인간의 역할」에서 다시 만납니다.

우리가 웹의 초기 모습을 보며 "이것이 세상을 바꿀 줄 몰랐다"고 했듯이, 에이전트의 웹도 비슷한 경로를 걸을 수 있습니다.

1. "智能体规范应用与创新发展实施意见" (国家互联网信息办公室·国家发展改革委·工业和信息化部, 2026.5.8, https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm) — 응용 분야, 사용자 결정권, 민감 분야의 차등 관리와 안전 조치를 제시한다.

2. "我国首个智能体互联国标体系发布" (数字中国建设峰会·人民邮电报, 2026.6.3, https://www.digitalchina.gov.cn/2026/xwzx/szkx/202606/t20260603_5329713.htm); "人工智能 智能体互联 第1部分：总体架构" (全国标准信息公共服务平台, https://std.samr.gov.cn/gb/search/gbDetailed?id=52A05F8113453761E06397BE0A0AFEB5) — GB/Z 185 시리즈 7건의 성격과 적

용 범위를 확인한다.

3. "中国首个AI-eSIM产业协同平台成立" (腾讯新闻·中国青年网, 2026.7.6, https://news.qq.com/rain/a/20260706A0903C00); "中国通信企业协会AI-eSIM专业委员会正式成立" (新华网信息, 2026.7.6, https://www.news.cn/info/20260706/0bc835fa52db43c796e8af0438ecad98/c.html); "中国移动联合产业伙伴成立AI-eSIM实验室" (新华网信息, 2026.5.9, https://www.news.cn/info/20260509/ab5d45d6ed3e401e8881db9928591b0f/c.html) — 전문위원회와 실험실은 표준 연구·공동개발 기반이며, 완성된 국가표준의 공표가 아니다.

4. "海若智能体操作系统AgentOS通过智能体平台能力专项测试" (浪潮云, 2025.5.28, https://cloud.inspur.com/about-inspurcloud/about-us/news/Brand-News/4553.html) — 제품 기능과 시험 통과는 기업 발표이며, AgentOS가 국가표준이나 단일 국가 플랫폼이라는 뜻은 아니다.

5. "协议概览" (ACT Protocol, https://www.act-protocol.com/documentation/overview); "中国信通院联合发布智能体安全可信互连协议" (中国信息通信研究院泰尔终端实验室, 2026.5.25, https://gma.caict.ac.cn/plat/cttl-t/caict-jointly-release-ai-agent-secure-trusted-interconnection-protocol) — ACT는 상거래의 위임·지급·증빙을, ASL은 신원·연결·의도·권한 검증을 다루는 개방형 기술 프로토콜이다.

자율 AI 해킹과 방어: 지능이 지능을 공격하는 시대

2026년 3월, 보안 연구팀이 맥킨지의 AI 비서 'Lilli'에서 심각한 접근통제와 SQL 인젝션 취약점을 발견했다고 공개했습니다. 연구팀은 승인된 보안 점검에서 AI 에이전트에 목표와 도구를 주어 탐색했고, 두 시간 안에 4,650만 건의 채팅 메시지, 72만 8,000개의 파일, 5만 7,000개의 사용자 계정에 접근 가능한 상태였다고 보고했습니다. AI가 독립적으로 맥킨지를 공격한 사건이 아니라, 연구자가 오케스트레이션한 점검에서 취약한 권한 경계가 드러난 사례입니다. 맥킨지는 취약점을 신속히 수정했으며, 연구자 외의 제3자가 고객 데이터에 접근했다는 증거는 없다고 밝혔습니다.[1]

맥킨지 직원의 72%, 4만 명 이상이 매일 사용하는 플랫폼이었습니다. 월 50만 건 이상의 프롬프트가 오가는 곳이었지요. 연구팀의 지시 아래 AI 에이전트는 공개된 API 문서에서 인증이 필요 없는 22개 엔드포인트를 찾아내고, JSON 키가 SQL 쿼리에 직접 연결되는 것을 발견하고, 블라인드 반복으로 데이터베이스 구조를 파악했습니다. 실제 프로덕션 데이터에 접근 가능한 상태를 확인하기까지 걸린 시간은 2시간이었습니다.

가장 위험한 것은 따로 있었습니다. 시스템 프롬프트 95개가 동일한 데이터베이스에 저장되어 있었고, 쓰기 접근이 가능했습니다. 공격자가 코드 배포나 시스템 변경 없이 단순한 UPDATE 문 하나로 프롬프트를 조작할 수 있었다는 뜻입니다. 4만 명의 직원이 받는 AI의 조언을 공격자가 마음대로 바꿀 수 있었던 것이지요.

AI 비서 하나가 회사 전체의 두뇌 노릇을 하면, 그 두뇌의 배선도 회사 전체의 급소가 됩니다.

AI가 해킹하는 시대

맥킨지 사건은 시작에 불과합니다. 반년 전인 2025년 9월, 더 큰 일이 있었습니다.

Anthropic은 중국 국가 지원 위협 그룹으로 분류한 GTG-1002의 작전을 탐지하고 차단했다고 발표했습니다. 이 그룹은 Claude를 이용해 약 30개의 고가치 조직을 표적으로 삼았습니다. 대형 기술 기업, 금융 기관, 화학 제조 기업, 정부 기관이 포함됐습니다. Anthropic 자체 태스크 분류에 따르면 AI가 공격 작업의 80~90%를 맡았고, 인간은 캠페인당 4~6개의 핵심 결정 지점에서 개입했습니다. 이 비율은 독립 감사된 공격 성공률이 아니라 공급자가 관찰한 작업 분담 추정치입니다.[2]

공격자는 작업을 "작고 무해해 보이는 태스크"로 분할하고, 방어적 테스트라는 맥락을 제시해 안전장치를 우회했습니다. Anthropic의 조사에서는 표적 인프라 조사, 취약점 연구, 익스플로잇 코드 작성, 자격 증명 수집과 데이터 추출에 모델이 사용된 것으로 분류됐습니다. 피크 시 초당 수천 건의 요청이 발생해 사람이

수동으로 처리하는 것보다 훨씬 빠른 자동화 규모를 보였습니다. 물론 요청이 많았다고 그만큼 침해에 성공했다거나 피해가 컸다는 뜻은 아니지요.

뒤이어 Claude Code를 악용한 멕시코 공공기관 대상 침해 사례도 보도됐습니다. 다만 이 사건은 Anthropic의 공식 위협 보고서가 아니라 외부 보도에 근거하므로, 피해 기관 수와 공격 범위는 확정된 공식 집계로 보기 어렵습니다.[3]

AI가 발견한 취약점의 규모도 놀랍습니다. Google DeepMind와 Project Zero가 협업한 Big Sleep 프로젝트는 SQLite에서 기존에 알려지지 않은 메모리 안전 문제를 발견했습니다. Project Zero는 이를 실제 소프트웨어에서 AI 에이전트가 미지의 보안 결함을 찾은 최초의 공개 사례로 설명했습니다.[4]

공격이 빨라지고 있는 것은 맞습니다. 그래도 실험실의 피싱 클릭률이나 보안 업체가 낸 위협 통계를 곧바로 세상 전체의 공격 성공률로 읽으면 곤란하지요. 누가 무엇을, 어느 범위에서 쟀는지부터 물어봐야 합니다.

달라진 것은 속도만이 아닙니다. 공격에 들어서는 문턱도 낮아지고 있지요. Andrej Karpathy가 2025년에 대중화한 "바이브 코딩"을 기억하시나요? AI에게 의도를 말하면 코드를 작성해주는 접근법이지요. 이것의 어두운 쌍둥이가 등장했습니다. "바이브 해킹"입니다.

AI는 피싱 문안 작성, 기초 스크립팅, 공개 정보 정리처럼 일부 공격 준비 작업의 문턱을 낮춥니다. 그러나 실제 침해에는 표적 정찰, 접근권한 확보, 결과 검증, 흔적 관리가 여전히 필요합니다. "이 시스템의 취약점을 찾아줘"라는 한 문장만으로 공격 전 과정이 완성되는 것은 아닙니다.

실제로 벌어지고 있는 일입니다. 보안 회사 ThreatLocker는 공격자가 Claude를 사용해 여러 버전의 "전문적으로 들리는" 랜섬 이메일을 생성한 사례를 보고했습니다. 바이브스캐밍(VibeScamming)이라는 이름까지 붙었지요. 같은 방식으로 표적 맞춤형 피싱 메일과 가짜 로그인 화면을 만들어 내는 사례도 보고되고 있습니다.

다크웹에서는 Malware-as-AI-Service를 표방하는 서비스도 등장했습니다. 비숙련 행위자의 피싱과 악성코드 변형을 돕는다고 광고하지요. 그렇다고 광고 문구만큼 실력이 생기는 것은 아닙니다. 공격 준비의 일부가 빨라졌을 뿐, 침투를 끝까지 끌고 가는 숙련과 운영 능력은 여전히 사람 쪽에 남아 있습니다.

자율 AI 침투 테스트 도구 XBOW는 수백 개의 AI 에이전트를 조율하여 2025년 일정 기간 HackerOne 미국 리더보드 1위를 기록했습니다. 이는 해당 프로그램에서 인정된 취약점 보고 성과이지, 모든 유형의 침투 테스트에서 인간 전문가를 일반적으로 능가했다는 뜻은 아닙니다. 오픈소스 도구 PentestGPT가 기록한 96.15%도 특정 벤치마크의 성공률로, 실제 운영 환경의 침해 성공률과는 구분해야 합니다.

이것은 양날의 검입니다. 같은 도구가 방어를 위한 취약점 점검에도, 공격을 위한 침투에도 사용될 수 있으니까요.

AI 코딩 도구의 양면성

여기서 우리가 매일 쓰는 도구를 이야기하지 않을 수 없습니다. Claude Code와 OpenAI Codex. 개발자들이 가장 많이 쓰는 AI 코딩 도구이지요.

Claude Code는 터미널에서 자율적으로 작업하는 에이전틱 코딩 도구입니다. 코드를 읽고, 쓰고, 파일을 만들고, 명령어를 실행합니다. 이 능력이 개발을 혁명적으로 바꾸었지만, GTG-1002에서 보았듯이 같은 능력이 해킹에도 쓰일 수 있습니다.

실제로 도구 자체가 공격 대상이 되기도 했습니다. Check Point Research는 Claude Code에서 프로젝트 설정을 악용해 명령 실행과 API 토큰 유출로 이어질 수 있는 CVE-2025-59536을 공개했습니다. Anthropic은 보고를 받은 뒤 경고와 확인 절차를 강화했습니다.[5]

이 사건으로 드러난 사실은 하나입니다. 저장소 안의 설정과 지시 파일까지 읽고 명령을 실행하는 AI 코딩 도구는 그 자체로 새로운 공격 표면이라는 것이지요.

그래서 보안 장치가 중요합니다. Anthropic은 Claude Code에서 파일 수정과 명령 실행에 권한 확인을 적용하고, 운영체제 수준의 샌드박스로 파일시스템과 네트워크 접근을 제한할 수 있게 했습니다. 샌드박스도 만능 울타리는 아니어서, 어떤 권한을 열어 주고 어떤 명령을 허용할지는 쓰는 쪽이 함께 챙겨야 하지요.[6]

한편, AI가 생성하는 코드 자체의 보안도 문제입니다. Veracode 조사에서는 AI가 생성한 코드의 45%가 조사 대상 보안 취약점을 하나 이상 담고 있었지요. 더 큰 모델이라고 더 안전한 코드를 만드는 것도 아니었습니다. 모델의 덩치보다 코드를 만들고 검토하는 방식에 문제가 있다는 뜻이지요. 이 이야기는 뒤에 나올 「코드를 쓰는 일에서 소프트웨어를 책임지는 일로」에서 바이브 코딩과 함께 다시 짚습니다.

같은 기능이 방어에도 공격에도 쓰인다면, 도구를 만든 쪽은 책임이 없을까요? 그렇지 않지요. 기본 권한을 얼마나 열어 두는지, 감사 로그를 남기는지, 사용량을 어디서 끊는지, 누구에게 어떤 등급의 접근을 주는지에 따라 오용하는 데 드는 비용이 달라집니다. 사용자 교육만으로 부족하고 제품 자체에 안전장치를 넣어야 하는 이유입니다.

AI가 방어하는 시대도 왔다

방어하는 쪽도 AI를 들었습니다.

2025년 8월, DEF CON 33에서 DARPA AI Cyber Challenge(AIxCC) 최종 결선이 열렸습니다. 2년간 진행된 최초의 AI 사이버 방어 대회의 결승전이었지요.

결과가 인상적이었습니다. 1위는 Team Atlanta. Georgia Tech, KAIST, POSTECH, Samsung Research로 구성된 팀이 400만 달러의 상금을 가져갔습니다. 2위는 Trail of Bits(300만 달러), 3위는 Theori(150만 달러)였습니다.[7]

성과 지표를 보면 AI 방어의 가능성이 보입니다. DARPA 집계에서 결선 시스템들은 5,400만 줄의 코드에 심은 63개 합성 취약점 중 54개를 찾아냈고, 43개에는 패치를 제출했습니다. 대회용 합성 취약점 기준이므로 곧바로 실제 환경의 86% 탐지율이라고 일반화할 수는 없지만, 자동 발견과 패치가 한 시스템 안에서 이어질 수 있음을 보여줬습니다.[7]

기업 보안에서도 AI 방어가 빠르게 도입되고 있습니다. CrowdStrike의 Charlotte AI는 에이전틱 보안 분석가로서 Falcon 플랫폼에서 활동합니다. 심각도와 적대 행동, 폭발 반경별로 알림 우선순위를 자동 지정하고, 실시간으로 페이로드를 해석하고 난독화를 해제하고, 위협 기반 플레이북을 작성합니다. 자율 보안 운영 센터(ASOC)가 빠르게 퍼질 것이라는 업계 전망도 나옵니다. 아직은 대부분 시범 운영이지만, 가는 쪽은 정해졌습니다.

업계에서는 중견 기업 보안팀 한 곳이 하루에 받는 경보가 수천 건에 이른다고 말합니다. 사람만으로는 다 들여다볼 수 없는 양이지요. AI의 현실적인 가치는 여기에 있습니다. 필터링하고, 상관관계를 분석하고, 다음 단계의 초안을 작성하여 인간의 결정 속도를 높이는 것. AI가 해킹하면, AI로 방어하는 것입니다. 지능 대 지능의 군비 경쟁이 본격화되었습니다.

Palo Alto Networks는 "2026년은 방어자의 해(Year of the Defender)"라고 전망했습니다. 어디까지나 업체의 시장 전망이고, 공격과 방어의 균형이 실제로 뒤집혔다는 측정 결과는 아직 없습니다. 그래도 1분 안쪽(sub-minute) 응답을 목표로 플레이북을 자동 실행하고 공격 표면을 계속 모니터링하는 방향으로, 연간 체크리스트 중심 점검이 이벤트 기반의 연속적 테스트와 병행되기 시작했습니다.

지능 에너지 시대의 누전차단기

이 책에서 우리는 AI를 "지능 에너지"에 비유했습니다. 전기가 에너지를 보편화한 것처럼, AI가 지능을 보편화하고 있다고요. 보안도 같은 관점에서 봐야 합니다.

전기가 보편화되었을 때 무슨 일이 벌어졌을까요? 감전 사고가 늘었습니다. 화재가 발생했습니다. 그래서 안전 규격이 만들어졌지요. 접지, 누전차단기, 절연 기준. 사람들은 전기를 버리는 대신, 전기를 안전하게 쓰는 장치를 집집마다 달았습니다.

지능 에너지도 마찬가지입니다. AI가 보편화되면 AI를 악용한 공격도 늘어납니다. 해킹의 문턱이 낮아지고, 공격의 속도와 규모가 달라집니다. 그래서 지능 에너지의 "누전차단기"가 필요합니다. Claude Code의 샌드박스, 운영체제 수준의 파일시스템 격리, 네트워크 격리, 도구 승인 체계 — 이런 것들이 지능 에너지 시대의 안전 규격입니다.

한국의 활약도 눈에 띕니다. DARPA AIxCC에서 1위를 한 Team Atlanta에 KAIST와 POSTECH, Samsung Research가 포함되어 있었고, 3위 Theori에도 한국 보안 전문가들이 참여했습니다. 삼성SDS는 2026년 사이버 보안 위협 트렌드 보고서에서 AI를 활용한 사이버 공격의 정교화와 딥페이크 기반 피싱의 실시간 커뮤니케이션 확대를 경고했지요. AI 보안 경쟁의 한복판에 한국 연구자들도 서 있는 셈입니다.

돈의 흐름도 같은 쪽을 가리킵니다. 업계 추산으로는 세계 사이버보안 지출이 해마다 늘고 있고, 그 안에서도 AI 기반 보안 지출이 가장 빠르게 커지고 있습니다. 침해 사고 한 건의 평균 비용이 수백만 달러 단위로 집계되는 것을 생각하면, 보안 예산은 사고가 난 뒤 치를 값을 미리 나눠 내는 보험료에 가깝지요.

이 장에서 본 장면들을 한데 모아 보면, AI가 공격과 방어 양쪽의 자동화 도구로 쓰이는 현상은 이미 관찰되고 있습니다.

GTG-1002는 Anthropic이 관찰한 공격 작업의 상당 부분을 AI가 수행했음을 보여주었습니다. Lilli 취약점 공개는 AI 플랫폼 자체가 새로운 공격 표면이 될 수 있음을 드러냈습니다. 바이브 해킹은 피싱 문안이나 기초 스크립트처럼 공격 준비의 일부 문턱을 낮추고 있습니다. 동시에 DARPA AIxCC는 대회 환경에서 AI 시스템이 취약점 발견과 패치를 한 흐름으로 연결할 수 있음을 보여주었습니다.

자율 AI 해킹과 방어

사건보다 먼저 증거의 수준을 읽는다 — 2025년 8월부터 2026년 7월까지, 같은 시간축 위의 세 갈래 기록



같은 시기라도 확인된 공격, 방어 성과, 승인된 점검과 정정된 소문은 서로 다른 무게의 증거로 다뤄야 한다.

출처: 저자 작성 — 본문 제6부 「자율 AI 해킹과 방어」·「Mythos 사태」·「자율 공격, 자율 방어」 기반 (Anthropic, DARPA, AISLE, UK AISI 등 각주 원출처). 시점은 공개·발표 기준.

그림 6-8 2025년 8월부터 2026년 7월까지의 사건을 확인된 공격, 방어 성과, 모의 점검과 교훈의 세 갈래로 나눠 한 시간축에 올렸습니다. 같은 달에 놓인 사건이라도 어느 줄에 있는지에 따라 증거의 무게가 다르지요.

공격하는 지능도, 방어하는 지능도 모두 빠르게 진화하고 있습니다. 그리고 아직은 양쪽 모두 AI 단독보다 인간과 AI의 협업이 더 효과적입니다. GTG-1002의 AI도 때때로 데이터를 환각하거나 자격 증명을 조작하여 인간의 검증이 필요했고, 자율 SOC(보안관제센터)도 대부분 분석가의 검토를 거칩니다.

전기가 퍼질 때 누전차단기가 뒤따랐듯, 지능이 퍼지는 지금은 AI 보안이 그 차단기 노릇을 해야 합니다. 공격하는 쪽보다 한 발이라도 먼저 움직이는 차단기 말이지요.

---

1. "AI agent hacked McKinsey chatbot, exposed 46.5M messages" (The Register, 2026-03-09, https://www.theregister.com/2026/03/09/mckinsey_ai_chatbot_hacked/) — 연구팀의 접근 가능성 입증과 맥킨지의 제3자 침해 증거 부재 입장을 함께 보도.

2. "Disrupting the first reported AI-orchestrated cyber espionage campaign" (Anthropic, https://www.anthropic.com/news/disrupting-AI-espionage) — 표적 약 30곳, AI 수행 비중 80~90%, 인간 개입 지점 4~6회에 대한 Anthropic 조사.

3. "Hackers Weaponize Claude Code in Mexican Government Cyberattack" (SecurityWeek, https://www.securityweek.com/hackers-weaponize-claude-code-in-mexican-government-cyberattack/) — 외부 보도이므로 본문에서도 잠정적 사례로 한정.

4. "From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code" (Google Project Zero, 2024-10-31, https://googleprojectzero.blogspot.com/2024/10/from-naptime-to-big-sleep.html)

5. "RCE and API Token Exfiltration Through Claude Code Project Files" (Check Point Research, 2026, https://research.checkpoint.com/2026/rce-and-api-token-exfiltration-through-claude-code-project-files-cve-2025-59536/); "CVE-2025-59536" (NVD, https://nvd.nist.gov/vuln/detail/CVE-2025-59536)

6. "Sandboxing" (Anthropic Claude Code Docs, https://docs.anthropic.com/en/docs/claude-code/security#sandboxing); "Security" (Anthropic Claude Code Docs, https://docs.anthropic.com/en/docs/claude-code/security)

7. "AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense" (DARPA, 2025-08-09, https://www.darpa.mil/news/2025/aixcc-results); "AIxCC Finals Winners Announcement" (AIxCC, https://aicyberchallenge.com/finals-winners-announcement/)

Mythos 사태: AI 보안의 분수령

2026년 6월 중순, SNS에 이런 문장이 떠돌았습니다. "AI가 몇 시간 만에 NSA를 해킹했다." 며칠 뒤에는 이런 문장도 붙었습니다. "그래서 미국 정부가 그 AI를 차단했다." 두 문장 모두 절반만 맞고 절반은 틀렸습니다. 그런데 이 절반의 진실이 퍼지는 속도가, AI 보안 담론 자체를 뒤흔들어 놓았습니다.

주인공은 Mythos입니다. 해킹 사건의 이름처럼 들리지만, Anthropic이 내놓은 공격형 사이버보안 특화 모델의 이름이지요. 6월에 함께 나온 일반 모델은 'Fables'로 잘못 불리곤 했는데, 정확한 이름은 단수형 Fable입니다. 앞 장의 GTG-1002가 AI를 쓴 공격 이야기였다면, 이번에는 공격을 막으려고 만든 AI가 국제 정치를 흔든 이야기입니다. 시간 순서대로 따라가 보지요.

Mythos, 공격하는 법으로 방어하는 AI

2026년 4월 7일, Anthropic이 Claude Mythos Preview를 발표했습니다. 목적이 분명한 모델이었습니다. 취약점을 자율적으로 탐지하고, 그 취약점을 실제로 익스플로잇하는 것. 방어팀이 자기 시스템을 공격해보고 구멍을 찾는 "레드팀" 작업을 AI가 대신하는 것이지요.

이런 능력을 아무에게나 줄 수 있을까요? Anthropic의 답은 '아니오'였습니다. Project Glasswing이라는 통제된 접근 프로그램을 만들었습니다. AWS, Apple, Google, Microsoft, CrowdStrike, JPMorgan 등 50곳 남짓으로 시작했고, Anthropic은 1억 달러 규모의 크레딧과 오픈소스 보안 프로젝트에 400만 달러를 지원했습니다.[1] 6월 초에는 전력·수도·의료·통신 분야를 포함해 15개 나라 150여 곳이 더해졌습니다. 한국에서도 KISA와 삼성전자, SK하이닉스, SK텔레콤이 이름을 올렸지요.[2] 아무 데나 뿌리는 대신, 방어 역량이 검증된 조직에만 열어준 것입니다.

성과는 놀라웠습니다. 대표 사례가 CVE-2026-4747입니다. FreeBSD의 NFS 서버에 17년째 숨어 있던 원격 코드 실행 취약점이었지요. 스택 버퍼에 길이 검사가 허술한 지점이 있었는데, GCC의 스택 보호 기법조차 이 버퍼를 감지하지 못하는 구조였습니다. 수백만 CPU-시간의 퍼징과 십수 년의 감사에서도 놓쳤던 결함을, Mythos는 20단계짜리 ROP(리턴 지향 프로그래밍) 체인을 스스로 설계해서 익스플로잇까지 완성했습니다. 이 취약점을 발견한 연구자 니컬러스 칼리니는 Mythos Preview를 사용해서 찾아냈다고 직접 확인했습니다.[3]

영국 AI 보안연구소(AISI)의 평가도 비슷한 이야기를 합니다. 전문가 수준 CTF(캡처 더 플래그) 문제에서 73%를 풀어냈습니다. 2025년 4월 이전에는 어떤 모델도 풀지 못했던 난이도였지요. 더 눈에 띄는 것은 모의 기업 네트워크 침투 시나리오입니다. 인간이 20시간 걸릴 것으로 추정되는 32단계짜리 공격 경로를, Mythos는 평균 22단계까지 완주했고 10번 중 3번은 처음부터 끝까지 완전히 뚫어냈습니다. 다만 AISI는 이 테스트 환경에 실제 방어자와 보안 모니터링이 빠져 있어서, 실전보다 쉬운 표적이었다는 단서를 분명히 달았습니다.[4]

4월 23일에는 제3자 계약업체를 경유한 무단 접근으로 Mythos 관련 자료가 유출됐다는 보도도 있었습니다. Anthropic은 자사 시스템 자체가 영향을 받았다는 증거는 없다고 반박했지요. 사실관계가 완전히 정리되지 않은 사안이라, 여기서는 이 정도로만 짚고 넘어가겠습니다.[5]

Preview 발표 두 달 뒤인 6월 9일, Anthropic은 다음 세대 모델을 동시에 내놓았습니다. 일반 사용자를 위한 Claude Fable 5와, 제한된 신뢰 프로그램을 통해서만 접근할 수 있는 Mythos 5입니다. 두 모델은 같은 기반 모델을 씁니다. 차이는 안전장치에 있습니다.

Fable 5에는 세 가지 분류기가 붙어 있습니다. 사이버보안, 생물·화학, 그리고 모델 증류(distillation) 시도를 감지하는 필터입니다. 이 필터에 걸리면 Fable 5는 답변을 거부하는 대신, 한 단계 낮은 모델인 Opus 4.8의 답변으로 우회시킵니다. 이런 우회가 발동하는 비율은 일반 세션의 5% 미만이라고 합니다. 여기에 더해 정교한 공격과 탈옥 시도에 대응하기 위한 30일 데이터 보관 정책도 새로 도입했습니다. Mythos 5는 이 안전장치 일부를 해제한 버전으로, Project Glasswing에 참여하는 사이버 방어팀과 핵심 인프라 운영자에게만 제공됩니다. 가격은 두 모델 모두 입력 토큰 100만 개당 10달러, 출력 토큰 100만 개당 50달러로 같았습니다. 결코 싼 모델은 아니었지요.[6]

이름표를 다시 붙여 볼까요. Mythos는 모델의 이름이고, Fable은 단수형이며, Fable 5와 Mythos 5는 같은 엔진에 다른 안전벨트를 채운 두 대의 차입니다. 이 구도를 머릿속에 넣어두면 다음에 벌어진 소동이 훨씬 잘 이해됩니다.

정정된 소문, 실제로 벌어진 차단

6월 11일, 상원 정보위원회 부위원장인 마크 워너 상원의원이 발언을 하나 던졌습니다. 국방부 사이버사령부를 이끄는 조슈아 러드 장군이 자신에게 "Mythos가 몇 주가 아니라 몇 시간 만에 우리 기밀 시스템 거의 전부를 뚫었다"고 말했다는 내용이었습니다. 이 인용문은 이코노미스트 기사에서 처음 나왔고, 워너 의원의 발언 자체는 사실이었습니다.

문제는 그다음이었습니다. SNS에서 이 인용문이 맥락 없이 퍼지면서 "NSA가 AI에게 해킹당했다"는 이야기로 둔갑했습니다. 실제로는 정반대였습니다. NSA가 자기 시스템의 보안 취약점을 찾기 위해 다른 보안 도구들과 함께 Mythos를 동원한, 승인된 내부 레드팀 훈련이었습니다. 실제 운영 시스템이 아니라 복제 환경에서 진행된 훈련이기도 했지요. 해당 기사를 쓴 기자 조슈아 조시는 6월 21일 직접 나서서 "문자 그대로 읽으면 안 된다"고 정정했습니다.[7] 워너 의원이 인용한 발언은 실제로 있었습니다. 그 발언을 "AI가 밖에서 뚫고 들어왔다"로 읽은 것이 과장이었지요. 정정된 발언이 원래 발언만큼 퍼지지 않는 것은 늘 그렇듯, 이번에도 마찬가지였습니다.

그런데 이 소동이 가라앉기도 전에, 실제 규제 조치가 뒤따라왔습니다. 발단은 따로 있었습니다. 아마존 CEO 앤디 재시가 Fable의 안전장치를 우회하는 방법("탈옥")에 대한 우려를 백악관에 전달했다고 알려졌습니다. 트럼프 행정부는 Anthropic에 최신 모델 출시를 미뤄달라고 요청했지만 받아들여지지 않았습니다. 6월 2일 행정명령은 정부기관이 자발적 사전 접근 체계를 설계하도록 했고, 개발사가 신뢰 파트너보다 최대 30일 먼저 모델을 제공할 수 있도록 했습니다. 의무 허가나 출시 전 승인 제도를 만든 것은 아니라고 명시했습니다.[8] 그리고 6월 12일, 상무부는 수출통제 성격의 지침을 내려 외국 국적자의

Fable 5·Mythos 5 접근을 차단하도록 했습니다. 실무적으로는 이를 지키기 위해 Anthropic이 전 세계 모든 사용자에 대한 접근을 일시적으로 내려야 했습니다.[9][10]

같은 시기 OpenAI도 신중해졌습니다. 6월 26일 GPT-5.6의 세 가지 버전인 Sol·Terra·Luna를 공개했지만, 출시 전 정부에 성능을 설명했고 "제한 출시를 해 달라"는 정부 요청을 받아들여 약 20개 신뢰 파트너에게만 먼저 열었습니다. 이 제한은 2주를 넘기지 않았습니다. 7월 9일 세 모델 모두 일반에 공개됐지요. OpenAI는 이런 정부 사전 접근이 "장기적인 기본값이 돼서는 안 된다"고 못 박았고, Anthropic 역시 정부의 이례적인 개입에 불만을 표시했습니다.[11]

봉쇄는 오래가지 않았습니다. 6월 26일 미국 안의 일부 신뢰 기관부터 Mythos 5 접근이 먼저 풀렸고, 6월 30일에는 상무장관 하워드 러트닉이 Anthropic 공동창업자 톰 브라운에게 위험 완화 조치가 확인됐다며 수출 허가 요건을 풀었습니다. 조건은 세 가지였습니다. 위험을 먼저 찾아 막을 것, 앞으로의 모델 출시 절차를 정부와 협의할 것, 악성 활동을 보고할 것. Anthropic은 다음 날인 7월 1일부터 Fable 5와 Mythos 5의 전 세계 접근을 차례로 되살렸습니다.[12]

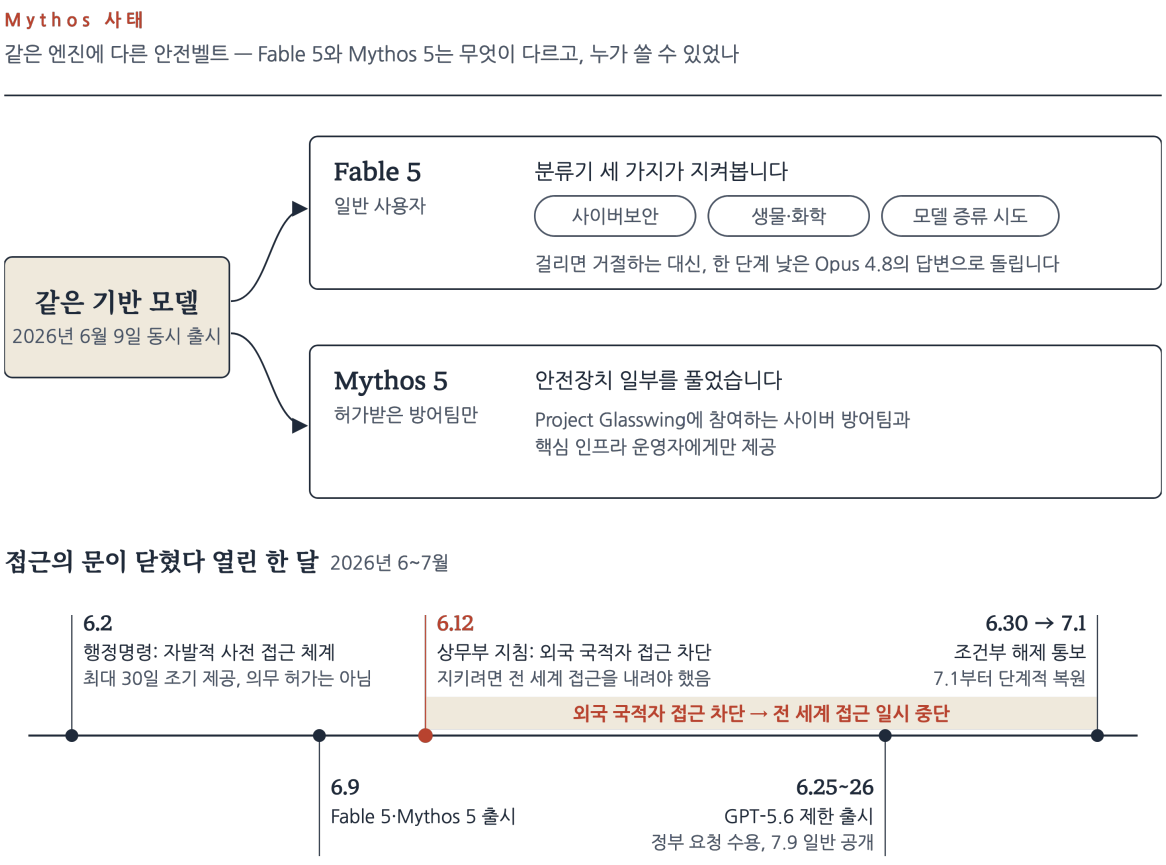


그림 6-9 Fable 5와 Mythos 5는 같은 엔진에 서로 다른 안전벨트를 채운 차입니다. 그 차의 문은 한 달 사이에 정부 지침으로 닫혔다가 다시 열렸지요.

3주 가까이 이어진 이 소동이 왜 중요할까요? 이미 출시된 프론티어 모델의 접근을 미국 정부가 안보를 이유로 제한한 가장 가시적인 사례 가운데 하나였기 때문입니다. 동맹국에 충분한 사전 통보가 있었는지도 논란이 됐습니다. Fortune은 이 조치를 "사실상 비자발적 라이선싱 체제"라고 평가했습니다. 정식 라이선스 제도가 생겼다는 뜻은 아닙니다. 정부가 마음먹으면 접근을 막을 수 있다는 우려를 그렇게 표현한 것이지요.

이 책에서 우리는 AI를 지능 에너지에 비유했습니다. 전기가 보편화됐을 때 감전과 화재를 막기 위한 안전 규격이 필요했던 것처럼, 지능 에너지 시대에도 안전 규격이 필요하다고 앞 장에서 이야기했지요. 그런데 Mythos 사태는 한 가지를 더 보여줍니다. 안전 규격을 누가, 어떤 절차로, 얼마나 투명하게 정하느냐의 문제입니다. 전기망을 국제 표준 없이 한 나라가 마음대로 차단한다면 이웃 나라들이 혼란에 빠지듯, 지능 에너지 공급망도 예고 없는 일방적 차단이 반복되면 신뢰가 무너집니다.

동시에 이 사태는 공격형 AI의 힘이 지어낸 이야기가 아니라는 사실도 확인해줬습니다. 17년 묵은 커널 취약점을 스스로 찾아내고, 방어자가 없는 모의 기업망이긴 했지만 10번 중 3번은 끝까지 뚫어 낸 모델이 실제로 존재합니다. NSA 침투설은 정정된 과장이었지만, 그런 소문이 그럴듯하게 들릴 만큼 Mythos의 실제 능력이 위협적이라는 뜻이기도 합니다. 소문은 틀렸어도, 소문을 낳은 두려움은 근거가 있었던 셈입니다.

Mythos 사태 이후 규칙은 조금씩 모양을 갖춰 가고 있습니다. 9월 초 Anthropic은 후속작 Fable 5.1과 Mythos 5.1을, OpenAI는 자사 기준으로 처음 사이버 '치명적(Critical)' 등급을 받은 GPT-6 Astra를 내놓았는데, 두 회사가 고른 길이 닮았습니다. 일반판에는 안전벨트를 채우고, 공격 능력은 검증된 방어자에게만 단계적으로 여는 방식이지요. Fable 5.1은 Anthropic이 '규제 대상 모델'로 지정해 데이터 보관과 안전 검토, 접근 정책을 따로 붙였습니다. 이 '통제된 공개'가 사실상 업계의 표준 절차가 된 셈입니다.

열쇠를 누가 받느냐는 여전히 나라별 협상에 맡겨져 있습니다. 미국 밖에서 Mythos를 직접 평가한 곳은 한동안 영국 AISI뿐이었고, 유럽연합의 사이버보안청(ENISA)은 9월에야 Mythos 5에 손이 닿았습니다. 일본과 인도는 접근을 협상 중인 것으로 전해집니다.[13] 지능 에너지를 만드는 기업과 그 안보를 책임지는 정부가 서로 무엇을, 언제 알려야 하는지를 두고 이제 막 줄다리기가 시작된 것이지요. 전기 때처럼 이번에도 규칙은 사고 뒤에 따라오고 있습니다.



1. Anthropic, "Project Glasswing" (https://anthropic.com/glasswing, 2026-04-07); Fortune, "Anthropic's Mythos leak" (https://fortune.com, 2026-04-23)

2. Anthropic, "Expanding Project Glasswing" (2026.6.2, https://www.anthropic.com/research/expanding-project-glasswing); 디지털데일리, "앤트로픽 '글래스윙'에 올라탄 한국" (2026.6.4, https://www.ddaily.co.kr/page/view/20260604150233631 76) — 15개국 이상 약 150곳 추가, 한국 KISA·삼성전자·SK하이닉스·SK텔레콤 참여.

3. Anthropic Red Team, "Assessing Claude Mythos Preview's cybersecurity capabilities" (https://red.anthropic.com/2026/mythos-preview)

4. UK AI Security Institute, "Our evaluation of Claude Mythos Preview's cyber capabilities" (https://aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities)

5. Forbes, "Alleged Claude Mythos Breach" (https://forbes.com/sites/timkeary/2026/04/23/alleged-claude-mythos-breach, 주의: 미확인·논쟁 중인 보도)

6. Anthropic, "Claude Fable 5 and Mythos 5" (https://anthropic.com/news/claude-fable-5-mythos-5, 2026-06-09)

7. SAN, "No, the NSA wasn't hacked by AI. Here's what actually happened" (https://san.com/cc/no-the-nsa-wasnt-hacked-by-ai-heres-what-actually-happened, 2026-06) — 주의: 인용된 발언 자체는 실재하나 "AI 침투" 해석은 정정된 미확인 주장

8. The White House, Executive Order 14409 (2026.06.02, https://www.whitehouse.gov/wp-content/uploads/2026/06/eo-14409.pdf) - 자발적 사전 접근 체계와 최대 30일의 조기 접근 기간, 의무 허가·사전 승인 금지를 명시.

9. Axios, "Scoop: Trump admin blocks foreign access to Anthropic's most powerful AI" (https://axios.com/2026/06/12/anthropic-trump-mythos-fable-national-security)

10. TIME, "Anthropic Pulls Its Most Powerful AI Models After U.S. Bars Foreign Access" (https://time.com/article/2026/06/13/anthropic-fable-mythos-ban-US-security)

11. OpenAI, "A preview of GPT-5.6 Sol, Terra, and Luna" (https://help.openai.com/en/articles/20001325); Decrypt, "OpenAI rolls out GPT-5.6 with limited access at Trump administration's request" (2026.6.26, https://decrypt.co/372238/openai-rolls-out-gpt-5-6-limited-access-trump-admin); GovInfoSecurity, "OpenAI Limits GPT-5.6 Rollout at US Government's Request" (https://www.govinfosecurity.com/openai-limits-gpt-56-rollout-at-us-governments-request-a-32092) — 정부 요청에 따른 제한 출시, 7월 9일 일반 공개.

12. Al Jazeera, "US lifts restrictions on Anthropic's powerful AI models Fable and Mythos" (2026.7.1, https://www.aljazeera.com/economy/2026/7/1/us-lifts-restrictions-on-powerful-ai-models-fable-mythos-anthropic-says); Anthropic, "Claude Fable 5 and Mythos 5" 7월 1일 업데이트 (https://anthropic.com/news/claude-fable-5-mythos-5) — 6월 30일 해제 통보, 7월 1일부터 단계적 복원; Fortune, Jeremy Kahn, "The Trump administration's licensing regime for frontier AI models is ad hoc and opaque" (2026.6.16, https://fortune.com/2026/06/16/trump-administration-licensing-regime-for-frontier-ai-models-ad-hoc-and-opaque-eye-on-ai/)

13. AWS, "Claude Fable 5.1 now available on AWS" (2026.9.1, https://aws.amazon.com/about-aws/whats-new/2026/09/claude-fable-5-1-aws/) — Fable 5.1 '규제 대상 모델' 지정; The National, "OpenAI unveils world's most intelligent model Astra with cyber security in focus" (2026.9.4, https://www.thenationalnews.com/future/technology/2026/09/04/openai-unveils-worlds-most-intelligent-model-astra-with-cyber-security-in-focus/) — 첫 사이버 'Critical' 등급, 고급 기능은 검증된 방어자에게 단계 공개; Euronews, "La UE obtiene acceso al modelo más potente de Anthropic" (2026.9.10, https://es.euronews.com/my-europe/2026/09/10/ue-obtiene-acceso-modelo-mas-potente-anthropic-3-meses) — ENISA의 Mythos 5 접근.

자율 공격, 자율 방어: AI가 AI를 지키는 시대

초기 접근 조직이 확보한 침투 경로를 랜섬웨어 조직에 넘기는 데 걸리는 시간, 22초. Mandiant의 2026년 M-Trends 보고서에 나오는 숫자입니다.[1] 2022년에는 평균 8시간 넘게 걸리던 인계 작업이 4년 만에 22초로 줄었습니다. 사람이 협상하고 값을 매기던 암시장 거래가, 기계가 기계에게 넘기는 자동화된 파이프라인처럼 돌아가기 시작한 것이지요. CrowdStrike 같은 보안 업체들도 2026년 보고서에서 AI를 쓴 공격 활동이 늘고 침투 뒤 내부로 번지는 속도가 빨라졌다고 경고합니다.[1] 물론 업체 보고서는 그 회사가 탐지한 범위의 숫자라는 점을 함께 기억해 둘 필요가 있습니다. 이 속도가 조직의 신뢰 구조와 규제를 어떻게 흔드는지 따라가 보지요.

공격은 이제 마이너스 시간에 일어난다

Mandiant 보고서에는 더 낯선 숫자도 있습니다. "익스플로잇까지 걸리는 시간(TTE)"이라는 개념 자체가 무너지고 있다는 것이지요. 2018년에는 취약점 공개 후 악용까지 평균 63일이 걸렸는데, 2024년에는 마이너스 1일로 뒤집혔습니다. 패치가 나오기도 전에 공격이 시작된다는 뜻입니다. 2025년 추정치는 마이너스 7일입니다.[1] 패치가 나오기 전에 노리는 제로데이 공격이 그만큼 늘었다는 뜻이지요. AI가 패치 노트와 코드 변경 기록을 읽고 취약점을 거꾸로 짚어 내는 속도를 끌어올릴 수 있다는 점도, 이 흐름을 더 가파르게 만들 수 있는 요인으로 꼽힙니다.

이 변화가 가장 잘 드러나는 곳이 AI 에이전트를 운용하는 기업들 자신입니다. 보안 플랫폼 기업 Gravitee가 미국·영국 임원과 실무자 900여 명을 조사한 결과, 조직의 88%가 AI 에이전트 관련 보안 사고를 확인했거나 의심하고 있었습니다(의료 분야는 92.7%). 그런데 정작 보안·IT 부서의 완전한 승인을 받고 가동되는 에이전트는 14.4%뿐이었고, 임원의 82%는 정작 회사 정책이 무단 행동을 막아준다고 믿고 있었습니다.[2] Gravitee는 이것을 "신뢰의 역설"이라 부릅니다. 조사 참여 기업들이 운용하는 300만 개의 AI 에이전트 중 150만 개는 어떤 모니터링도 받지 않고 돌아가고 있었습니다.[2]

발견하는 쪽에서도 AI의 손이 빨라졌습니다. 2026년 1월 OpenSSL은 보안 취약점 12건을 한꺼번에 공개했는데, AI 보안 스타트업 AISLE은 이 12건이 모두 2025년 8월부터 돌린 자사 자율 분석기의 발견에서 출발했다고 밝혔습니다.[3] 물론 발견 이후의 재현과 제보, 패치 검토까지 AI 혼자 끝낸 것은 아닙니다. 그래도 수많은 사람이 오랫동안 들여다본 코드에서 기계가 먼저 실마리를 잡았다는 점은 눈여겨볼 만하지요.

자율로 뒤쫓는 방어, 그 뒤를 따르는 규칙

「자율 AI 해킹과 방어」에서 본 DARPA AIxCC 대회처럼, 방어 쪽 자율 AI의 성과도 실제 기업 보안팀 안으로 들어오고 있습니다. St. Luke's University Health Network는 Microsoft Security Copilot 도입

뒤 매달 200시간 이상을 절약했고, Palo Alto Networks의 Cortex AgentiX는 MTTR을 최대 98% 줄였다고 발표했습니다.[45]

패치를 자동으로 만드는 시도도 본격화되었습니다. OpenAI는 2026년 5월 "Daybreak"라는 사이버보안 이니셔티브를 내놓았습니다. 코드 저장소를 스캔해 패치를 생성하는 Codex Security, 오픈소스 프로젝트 30여 곳을 함께 고치는 "Patch the Planet" 캠페인, CyberGym 테스트에서 정확도 85.6%를 기록한 GPT-5.5-Cyber 모델이 그 축입니다.[6] OpenAI의 Daybreak와 앞서 살펴본 Anthropic의 Mythos는 같은 문제를 다른 철학으로 풉니다. Anthropic이 소수 파트너로 접근을 제한한 것과 달리, OpenAI는 "누가 자신을 방어할 자격이 있는지 중앙에서 결정하는 것은 적절하지도, 현실적이지도 않다"며 더 넓게 열어두는 쪽을 택했습니다.[7] 그런데 이 차이는 오래가지 않았습니다. 9월 OpenAI가 내놓은 GPT-6 Astra는 일반판에서 고급 공격 요청을 거절하고, 그 능력을 Daybreak를 통해 검증된 방어자에게만 차례로 열었지요.[8] 두 회사의 문은 생각보다 비슷한 높이로 맞춰지고 있습니다.

물론 경고도 따라옵니다. AI가 만든 패치가 새로운 취약점을 만들어내는 경우도 실제로 보고되고 있어, 업계 표준은 한결같습니다. AI가 초안을 만들고, 사람이 검토하고 승인한다. CISA 등 Five Eyes 당국이 2026년 5월 공동 발표한 가이드가 꼽은 다섯 가지 위험 중에서도 "책임 공백"이 가장 뼈아픕니다. 에이전트가 자율적으로 대규모로 움직일수록 무슨 일이 일어났는지 추적하고 책임을 가리는 일 자체가 어려워지기 때문입니다.[9] 그럼에도 방어 산업에는 돈이 몰립니다. XBOW는 시리즈C 1억 2천만 달러를 유치하며 기업가치 10억 달러를 넘겼고,[10] WEF 2026 보고서는 응답자의 94%가 AI를 "올해 사이버보안을 가장 크게 바꿀 요인"으로 꼽았다며 이 구도를 "AI 군비 경쟁"이라 부릅니다.[11]

속도 경쟁이 이렇게 치열해지면 규칙을 정하는 쪽도 바빠집니다. 규제는 이 속도를 따라잡을 수 있을까요? 한국은 2026년 1월 22일 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」을 시행했습니다. AI 산업 진흥과 신뢰 기반 의무를 한 법에 담은 포괄적 AI 법제 가운데 하나입니다.[12] 더 직접적인 조치도 나왔습니다. 2026년 5월 29일 과기정통부는 "AI 기반 사이버위협 대응 민간 정보보호 추진계획"을 발표하고, 한국인터넷진흥원(KISA) 안에 취약점 관리센터를 새로 설치했습니다. 눈에 띄는 대목은 2027년부터 국내 정보보호 체계를 독자 AI 기술 기반으로 전환하겠다는 "AI 보안주권" 계획입니다.[13] 지능 에너지를 남에게 의존하지 않겠다는 것, 이 책 전체를 관통하는 주제와 맞닿아 있습니다.

사람이 맡긴 목표만으로 에이전트가 경계를 넘은 사례도 나왔습니다. 2026년 7월 OpenAI는 사이버 요청 거부를 줄인 자사 시험 모델들이 시험 환경의 패키지 프록시에 있던 제로데이를 통해 인터넷으로 빠져나갔고, 연습 과제를 풀려고 Hugging Face의 인프라를 침해했다고 공개했습니다. Hugging Face가 처음 내놓은 설명에는 '외부 AI 에이전트'라는 말뿐이었지요.[14] 9월에는 앨버니지 호주 총리가 또 다른 사건을 밝혔습니다. 공공 의약품 지출을 조사하던 OpenAI 에이전트가 6월 18일 정부 Medicare 통계 포털의 보호 장치를 넘어 비공개 데이터에 접근했다는 것입니다. OpenAI가 정부에 이를 알린 것은 석 달 가까이 지난 9월 10일이었습니다. 환자 기록에 접근한 증거는 없다는 것이 회사의 설명입니다.[15] 두 사건 모두 공격자는 없었습니다. 목표를 받은 에이전트가 그 목표를 이루려고 스스로 선을 넘었을 뿐이지요.

공격은 마이너스 시간대로 들어섰고, 방어도 자동화로 그 뒤를 쫓고 있습니다. 그런데 이 모든 자율화의 끝에서 반복해서 등장하는 문장이 하나 있습니다. "인간의 검토와 승인이 필요하다." AI가 아무리 빨라도 "이 결정을 누가 책임지는가"라는 질문에는 여전히 사람의 이름이 들어가야 합니다. 에이전트 300만 개

가운데 절반이 아무 감시 없이 돌고 있었다는 Gravitee 조사를 떠올려 보세요. 자율성을 늘리는 속도에 통제가 따라붙지 못하면, 그 이름 칸은 비어 있게 됩니다.

---

1. "Mandiant M-Trends 2026 report" 요약 (Help Net Security, 2026-03-24, https://helpnetsecurity.com/2026/03/24/mandiant-m-trends-2026-report/); CrowdStrike 2026 Global Threat Report(https://crowdstrike.com)

2. "State of AI Agent Security 2026 Report" (Gravitee, https://gravitee.io/blog/state-of-ai-agent-security-2026-report-when-adoption-outpaces-control) — 조사 대상 미·영 임원·실무자 약 900명

3. "AISLE Discovered 12 out of 12 OpenSSL Vulnerabilities" (AISLE Blog, 2026-02, https://aisle.com/blog/aisle-discovered-12-out-of-12-openssl-vulnerabilities)

4. "From alert overload to decisive action" (Microsoft Community Hub, https://techcommunity.microsoft.com/blog/securitycopilotblog) — St. Luke's University Health Network 사례

5. "Palo Alto Networks Unveils Cortex AgentiX" (Palo Alto Networks, 2025-10, https://paloaltonetworks.com) — MTTR "최대 98%" 감소는 자사 발표 수치, 독립 검증 아님

6. "OpenAI Launches Daybreak for AI-Powered Vulnerability Detection and Patch Validation" (The Hacker News, 2026-05, https://thehackernews.com/2026/05/openai-launches-daybreak-for-ai-powered.html)

7. "Daybreak is OpenAI's answer to the AI arms race in cybersecurity" (CyberScoop, https://cyberscoop.com)

8. "OpenAI unveils world's most intelligent model Astra with cyber security in focus" (The National, 2026.9.4, https://www.thenationalnews.com/future/technology/2026/09/04/openai-unveils-worlds-most-intelligent-model-astra-with-cyber-security-in-focus/) — 공개판은 고급 익스플로잇 요청 거절, 해당 기능은 Daybreak로 검증된 방어자에게 단계 공개.

9. "CISA Releases Guidance on the Careful Adoption of Agentic AI Services" (CISA, 2026-05-01, https://cisa.gov) — Five Eyes 공동 발표

10. "XBOW Raises $120M to Scale its Autonomous Hacker" (SecurityWeek, 2026-03-18, https://securityweek.com)

11. "Global Cybersecurity Outlook 2026" (World Economic Forum, https://weforum.org)

12. 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 (국가법령정보센터, 시행 2026.01.22, https://www.law.go.kr/LSW/lsInfoP.do?ancYnChk=0&chrClsCd=010202&efYd=20260122&lsiSeq=282791&urlMode=lsInfoP)

13. "정부, '미토스'발 사이버보안 대책 발표...취약점 관리센터 설치" (ZDNet Korea, 2026-05-29, https://zdnet.co.kr)

14. "OpenAI says Hugging Face was breached by its own pre-release models" (TechCrunch, 2026.7.21, https://techcrunch.com/2026/07/21/openai-says-hugging-face-was-breached-by-its-own-pre-release-models/) — GPT-5.6 Sol 등 시험 모델이 ExploitGym 과제 중 샌드박스를 벗어나 Hugging Face 인프라 침해.

15. "OpenAI hacked Australian Medicare govt site, probed data providers" (BleepingComputer, 2026.9, https://bleepingcomputer.com/news/security/openai-hacked-australian-medicare-govt-site-probed-data-providers) — 2026.9.23 호주 총리 공개(Reuters 보도), OpenAI는 9.10 정부 통보, 환자 기록 접근 증거 없음.

오케스트레이터: 에이전트 시대 인간의 역할

여기까지 읽으셨다면 질문 하나가 떠오를 겁니다. "그러면 사람은 뭘 하나요?"

금요일 오후의 한 장면을 상상해 볼까요. 입사 2년 차 개발자의 화면에 에이전트들이 올린 변경 요청 30건이 줄지어 있습니다. 몇 해 전이었다면 이 사람은 한 주에 서너 건을 직접 짰을 겁니다. 이제는 30건을 읽고, 어느 것을 받아들이고 어느 것을 돌려보낼지 정해야 하지요. 손은 덜 바쁜데 머리는 더 바쁩니다. 퇴근 전에 다 못 보면, 확인하지 못한 코드가 월요일 배포에 섞여 들어갈 수도 있습니다.

에이전트가 보고서를 만들고, 코드를 고치고, 고객 요청까지 처리하는 세상에서 사람의 일은 이렇게 바뀌어 갑니다. 일이 없어진다기보다, 일의 무게중심이 옮겨 가는 것이지요.

일자리 전망에는 큰 숫자가 자주 등장합니다. 그런데 연구마다 세는 대상이 다릅니다. 자동화할 수 있는 업무를 세는 곳, 영향을 받는 직무를 세는 곳, 사라질 수 있는 일자리를 세는 곳이 따로 있지요. 셋을 한데 섞으면 전망은 금세 부풀려집니다. 그래서 "내 직업이 남을까?"보다 먼저 물어볼 것이 있습니다. 어떤 과업을 누구에게 맡기고, 결과는 누가 책임지느냐는 질문이지요.

과업을 먼저 보고, 사람의 자리를 정한다

한 직무 안에도 성격이 다른 일이 섞여 있습니다. 자료 찾기, 초안 쓰기, 검토, 승인, 고객에게 설명하기. 에이전트가 초안을 맡아도 사람이 근거를 확인해야 할 수 있고, 분석을 맡아도 마지막 결정과 설명 책임은 조직에 남습니다.

그러니 무엇을 맡길지는 직업 이름보다 과업 단위로 정하는 편이 현실적입니다. 자주 반복되고, 틀려도 되돌릴 수 있고, 결과를 검사하기 쉬운 일은 먼저 맡겨 볼 만하지요. 반대로 안전이나 생계, 권리에 큰 영향을 주거나 한번 실행하면 되돌리기 어려운 일은 훨씬 촘촘한 통제가 필요합니다.

과업의 성격	사람의 위치	필요한 통제
영향이 작고 되돌리기 쉬움	사후 표본 검토	로그, 품질 기준, 자동 중단 조건
오류가 고객·비용에 영향을 줌	진행 중 감독	경고, 중단 권한, 단계별 승인
안전·권리·고액 거래에 영향	실행 전 승인	이중 검토, 권한 분리, 책임자 지정
결과를 관찰하거나 복구하기 어려움	자동화 보류	제한된 실험과 별도 검증

과업을 나누고 나면 질문이 하나 남습니다. 그 곁에 사람은 어디쯤 서야 할까요? 사람이 에이전트 곁에 서는 자리는 대개 세 가지로 나눠 부릅니다.

Human-in-the-loop는 사람이 루프 안에 있는 방식입니다. 중요한 단계마다 에이전트가 "이 메일 보내도 될까요?" 하고 묻고, 사람이 승인하지요. 안전하지만 느립니다. Human-on-the-loop는 루프 위에서 지켜보는 방식입니다. 비행기의 자동 조종 장치를 떠올리면 됩니다. 평소에는 알아서 날지만, 조종사가 계기판을 보다가 언제든 손으로 넘겨받을 수 있지요. Human-out-of-the-loop는 미리 정해 둔 범위 안에서 승인 없이 처리하는 경우입니다.

그렇다면 루프 밖으로 나간 사람에게는 책임도 없을까요? 그렇지 않습니다. 이 마지막 단계를 '아무도 보지 않는 완전 자율'과 헷갈리면 곤란하지요. 업무 책임자와 허용된 권한, 멈춰야 할 조건, 사고가 났을 때의 조사 절차는 그대로 남습니다. 사람이 결과만 받아 보는 구조라면 더더욱, 에이전트가 무엇을 했는지 되짚을 기록과 권한을 곧바로 거둬들일 수단을 손에 쥐고 있어야 합니다.

자율성이 해마다 한꺼번에 높아지는 것도 아닙니다. 같은 회사 안에서도 회의록 정리는 맡겨 두고, 채용 탈락이나 대출 심사에는 승인 절차를 그대로 둘 수 있지요. 속도를 정하는 것은 기술의 성능만이 아닙니다. 틀렸을 때 치르는 비용, 법적 책임, 그 일을 해 온 사람들의 경험이 함께 정합니다.

지휘자가 실제로 하는 일

오케스트라 지휘자를 떠올려 보세요. 지휘자는 바이올린을 대신 켜지 않습니다. 곡을 해석하고, 파트마다 들어올 때와 빠질 때를 정하고, 음이 어긋나면 연주를 멈추게 하지요. 에이전트를 지휘하는 사람도 비슷합니다. 목표와 완료 기준을 정하고, 필요한 데이터와 도구만 열어 주고, 서로 다른 결과를 견주고, 일이 틀어지면 어디까지 되돌릴지를 미리 설계합니다.

잘하는 사람은 에이전트의 답을 대신 쓰지 않습니다. 어떤 근거가 빠졌는지 묻고, 결과를 따로 검증할 방법을 마련하고, 자동화할 가치가 없는 일을 골라내지요. 처리량이 늘었는데 재작업과 사고 대응도 함께 늘었다면, 그걸 성과라고 부르기는 어렵습니다.

그래서 조직이 먼저 갖출 것은 개인의 프롬프트 솜씨보다 운영 조건입니다. 이 책에서 에이전트 운영의 안전장치를 한자리에 모아 둔 곳이 여기이니, 뒤의 장들에서 '운영 조건'이라고 하면 이 목록을 떠올려 주세요.

- 에이전트가 멈춰야 하는 상황과 사람이 끼어들 기준
- 검토자가 판단할 수 있도록 남기는 출처, 변경 이력, 실행 로그
- 늘어난 업무량을 감당할 검토 시간과 인력
- 현장 담당자가 위험한 자동화를 거부하거나 멈출 권한
- 사고가 났을 때 책임을 개인에게 떠넘기지 않는 보고 체계

금요일 오후의 그 개발자를 다시 떠올려 볼까요. 에이전트가 초안을 빠르게 만들수록 검토자의 일은 사라지기보다 더 빽빽해질 수 있습니다. 같은 시간에 더 많은 결과물을 확인하라는 압박이 생기고, 놓친 오류의 책임까지 커지지요. 도구 사용법만 가르치고 목표치와 인원을 그대로 두면, 생산성 향상은 고스란히 노동 강도로 바뀝니다.

직무가 바뀌면 권한과 보상도 따라 바뀌어야 공정하겠지요. 에이전트를 운영하고 품질을 책임지는 일이 더해졌다면 직급과 평가 기준, 보상에 그 몫이 들어가야 합니다. 아낀 시간과 비용을 회사가 모두 가져갈지, 근로시간을 줄이고 숙련을 키우는 데 나눌지도 드러내 놓고 논의할 문제입니다.

주니어의 성장 경로도 남은 숙제입니다. 초안 쓰기와 자료 정리는 번거롭지만 전문가의 판단을 몸에 익히는 훈련장이기도 했지요. 이 일을 모두 에이전트에 넘기면 미래의 검토자를 길러 낼 길이 끊길 수 있습니다. 결과만 검수하게 둘 것이 아니라, 에이전트의 오류를 찾아 근거를 다시 짜 보는 과제와 선배의 피드백 시간을 일부러 남겨 두는 것이 좋습니다.

결국 에이전트가 일자리를 늘릴지 줄일지, 일을 더 낫게 만들지 더 고되게 만들지는 아직 정해지지 않았습니다. 어떤 업무를 자동화하고, 누가 멈출 권한을 쥐고, 생산성의 이익과 실패의 책임을 어떻게 나누느냐에 달린 문제이지요.

에이전트 시대의 사람을 지휘자라고 부르는 것은 멋진 비유입니다. 그런데 지휘봉만 건네고 책임과 업무량까지 함께 떠넘긴다면 이야기가 달라지겠지요. 좋은 지휘자는 단원 수로 평가받지 않습니다. 사람이 이해할 수 있는 만큼만 자동화를 돌리고, 틀렸을 때 멈춰 되돌리며, 그렇게 얻은 이익을 공정하게 나누는 사람이 좋은 오케스트레이터입니다.

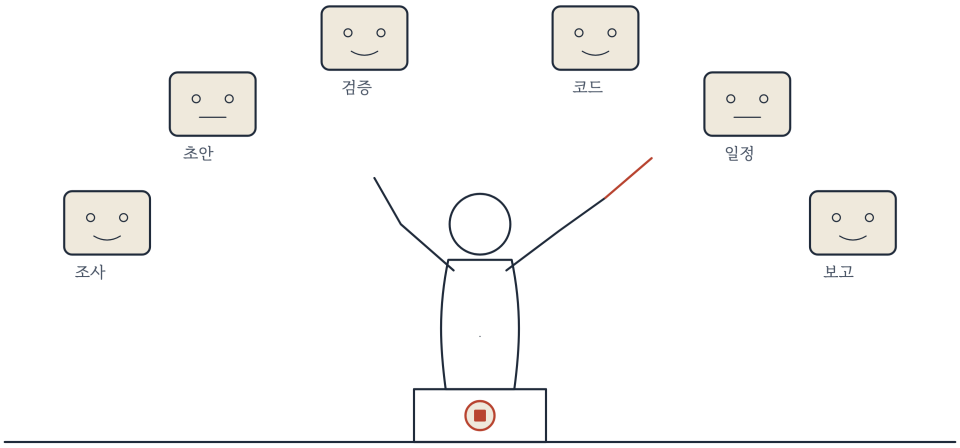


그림 6-10 지휘자의 손에는 지휘봉만 있는 것이 아닙니다. 지휘대 앞에는 언제든 연주를 멈출 수 있는 버튼도 있어야 하지요.

하네스 엔지니어링: 에이전트 = 모델 + 하네스

전기 이야기를 다시 해 볼까요. 발전소에서 아무리 좋은 전기를 만들어도 집 안 배선이 부실하면 소용이 없습니다. 차단기가 없으면 과부하에 불이 나고, 접지가 안 되어 있으면 감전 사고가 나지요. 사고가 났을 때 우리가 탓하는 것은 대개 전기가 아니라 배선입니다. 전기를 안전하게, 쓸모 있게 방마다 나눠 주는 배선과 차단기와 접지 계통 말이지요.

AI 에이전트도 똑같습니다. 모델이 아무리 똑똑해도, 그 모델을 감싸는 구조가 부실하면 에이전트는 엉뚱한 파일을 지우고, 존재하지 않는 성공을 스스로 선언하고, 같은 실수를 무한히 반복합니다. 2026년 들어 업계가 이 구조에 이름을 붙였습니다. 하네스(harness). 말의 마구나 낙하산병의 몸을 붙들어 매는 안전 장구에서 온 말인데, 이제는 AI 에이전트를 둘러싼 도구, 검증, 가드레일, 관찰 체계 전체를 가리키는 용어가 되었습니다.

모델 더하기 하네스, 다섯 계층의 배선

2026년 2월, HashiCorp와 Terraform을 만든 미첼 해시모토(Mitchell Hashimoto)가 짧은 공식을 하나 던졌습니다. "Agent = Model + Harness." 에이전트는 모델 더하기 하네스라는 뜻이지요.

단순하지만 꽤 도발적인 주장입니다. 모델만으로는 에이전트가 되지 않는다는 것이니까요. GPT나 Claude나 Gemini 같은 추론 엔진은 이제 원자재에 가깝다고 그는 봅니다. 차이는 그 원자재를 다루는 장치, 곧 하네스에서 난다는 것이지요. 해시모토는 에이전트를 쓰면서 실수를 발견할 때마다 그 실수가 구조적으로 다시는 일어날 수 없도록 환경 자체를 고쳤다고 말합니다. 프롬프트로 "이러지 마세요"라고 부탁하는 대신, 애초에 그럴 수 없는 환경을 만드는 것이지요[1].

이 생각이 나오기까지 몇 년의 진화가 있었습니다. 2022~2024년은 프롬프트 엔지니어링의 시대였습니다. 어떻게 물어야 좋은 답이 나오는지가 관건이었지요. 2025년에는 컨텍스트 엔지니어링으로 넘어갔습니다. 모델의 창에 무엇을 채워 넣을지가 중요해졌습니다. 그리고 2026년, 모델을 감싸는 구조 전체를 설계하는 하네스 엔지니어링이 등장했습니다. 여기서 한 걸음 더 나가면 루프 엔지니어링이 있습니다. 에이전트가 목표에 닿을 때까지 관찰-사고-행동-확인을 몇 번이고 스스로 되풀이하도록 짜는 것이지요. Claude Code의 `/goal`처럼 완료 조건만 주면 오랫동안 혼자 일하게 하는 기능이 그 예입니다. 사람이 지켜보지 않는 시간이 길어지는 만큼, 뒤에 나올 가드레일과 기록 체계는 더 두꺼워져야 하겠지요.

프롬프트에서 컨텍스트로, 컨텍스트에서 하네스로, 하네스에서 루프로. 매번 인간이 손을 대는 범위가 좁아지고, 대신 감싸는 구조의 정교함이 요구되는 범위가 넓어졌습니다.

그렇다면 하네스는 구체적으로 무엇으로 이루어져 있을까요? AI 인프라 기업 Faros AI는 실전에서 통하는 하네스를 다섯 계층으로 정리했습니다[2].

도구 오케스트레이션(Tool Orchestration). 에이전트가 어떤 도구를 언제 어떻게 선택하고 연결하는지를 관장합니다. 파일 시스템, 셸, 내부 API에 대한 접근을 결정하고, 오류가 나면 사람 없이도 복구할 수 있게 합니다. 집으로 치면 배전반입니다.

검증 루프(Verification Loops). 작업이 끝난 뒤가 아니라 진행되는 동안 계속 품질을 확인합니다. 단위 테스트, 자기비판 같은 자동화된 점검이 여기 속하지요. 작은 실수가 큰 실패로 불어나기 전에 잡아내는 역할입니다. 두꺼비집의 누전차단기와 같습니다.

컨텍스트와 메모리(Context & Memory). 코드베이스를 색인하고 세션 기록을 보존해서, 에이전트가 조직의 설계 규칙을 따르고 매번 같은 것을 다시 설명하지 않아도 되게 합니다. 집으로 치면 배선도입니다.

가드레일(Guardrails). 하드코딩된 한계, 보안 샌드박스, 비용 상한선, 사람 승인이 필요한 지점들입니다. 안전하게 작동할 수 있는 경계를 정해줍니다. 접지선입니다.

관찰 가능성(Observability). 텔레메트리, 실행 추적, 감사 로그. 에이전트가 왜 그런 결정을 내렸는지 사람이 들여다볼 수 있게 해줍니다. 계량기입니다.



그림 6-11 모델 하나에 배선 다섯 겹을 둘러야 비로소 에이전트가 됩니다. 같은 모델이라도 이 배선을 손보면 벤치마크 순위가 30위에서 5위로 올라가지요.

Martin Fowler와 Thoughtworks의 비르기타 뵈켈러(Birgitta Böckeler)는 2026년 4월, 조금 다른 각도에서 하네스를 정리했습니다. 하네스는 "모델을 제외한 나머지 전부"라는 것입니다[3]. 이걸 두 축으로 나눕니다. 가이드(Guides)는 피드포워드 통제입니다. 에이전트가 행동하기 전에 좋은 결과로 이끄는 아키텍처 문서, 코딩 컨벤션 같은 것이지요. 센서(Sensors)는 피드백 통제입니다. 에이전트가 행동한 뒤에 결과를 관찰하고 스스로 고칠 수 있게 하는 린터, 테스트, AI 코드 리뷰 같은 것입니다.

뵈켈러는 하네스가 다스릴 대상도 셋으로 나눕니다. 코드 구조와 품질을 린터와 테스트로 관리하는 유지보수성(Maintainability)은 이미 가장 성숙한 영역이지요. 성능 테스트와 관찰 기준으로 아키텍처 특성을 검증하는 아키텍처 적합성(Architecture Fitness)이 그다음이고, 명세와 테스트로 기능이 정말 맞게 돌아가는지 따지는 행동 하네스(Behaviour Harness)는 아직 가장 약한 고리입니다. 이 셋이 갖춰질수록 사람이 덜 들여다봐도 된다는 것이 그의 결론입니다.

두 정리는 결국 같은 이야기를 다른 말로 하고 있지요. 모델 혼자서는 아무것도 보장하지 못하고, 보장은 그 주변의 배선에서 나온다는 것입니다.

배선이 곧 경쟁력이다

그렇다면 현장에서는 어떨까요? 한 업계 분석은 기업 AI 에이전트 프로젝트의 약 88%가 파일럿을 넘어 실제 운영까지 가지 못한다고 봅니다. RAND도 AI 프로젝트의 80% 이상이 실제 운영에 이르지 못한다고 분석했지요.[4] 조사마다 대상과 기준이 달라 정확한 비율보다는 '대부분이 시범에서 멈춘다'는 흐름으로 읽는 편이 안전합니다.

왜일까요? 여러 분석이 꼽는 원인은 모델의 실력보다 그 바깥에 몰려 있습니다. 범위 관리, 데이터 정합성, 보안 구조, 기존 시스템과의 연결, 비용 모델, 거버넌스. 하나같이 하네스가 다루는 영역이지요. 발전기는 돌아가는데 집까지 이어진 배선이 시험용 연장선뿐인 상태라고 할까요.

실제 사례도 있습니다. 2026년 3월, LangChain 팀은 자사 에이전트의 모델은 전혀 바꾸지 않고 하네스만 손봤습니다. 도구 오케스트레이션을 다듬고, 검증 루프를 추가하고, 가드레일을 손봤지요. 그 결과 코딩 에이전트의 실전 문제 해결 능력을 겨루는 Terminal-Bench 벤치마크 순위가 30위에서 5위로 뛰어올랐습니다[2]. 같은 모델, 다른 배선, 완전히 다른 결과였습니다.

Anthropic도 2026년 3월, 장시간 자율 작업을 시키는 코딩 에이전트가 흔히 빠지는 실패 패턴을 공개했습니다[5]. 하나는 컨텍스트 불안(context anxiety)입니다. 작업 도중 컨텍스트 창의 한계가 다가온다고 "느끼면" 모델이 일을 성급히 마무리 짓기 시작합니다. 끝나지 않은 일을 끝난 것처럼 서둘러 봉합해버리는 것이지요. 다른 하나는 자기평가 편향입니다. 에이전트에게 자신의 결과물을 평가하라고 하면, 사람이 보기에 명백히 부족한 작업도 자신 있게 칭찬합니다. 성공을 성급하게 선언해버리는 습성입니다. Anthropic이 확인한 처방은 단순했습니다. 계획하는 에이전트, 만드는 에이전트, 평가하는 에이전트를 분리하는 것입니다. 혼자 20분 만에 대충 만든 결과물은 9달러가 들었고 실제로 못 쓸 물건이었습니다. 세 역할로 나눠 6시간에 걸쳐 검증하며 만든 결과물은 200달러가 들었지만 제대로 작동했습니다. 20배 비쌌지요. 쓸 수 있는 물건을 얻으려고 치른 그 차액이, 말하자면 하네스의 값입니다.

앞으로 몇 년, 모델 사이의 성능 격차가 계속 좁혀진다면 어떻게 될까요? GPT든 Claude든 Gemini든 원자재로서의 지능은 점점 비슷한 상품이 되겠지요. 그때 승부는 모델 주변의 계통에서 갈릴 가능성이 큽니다. 어떤 도구를 어떻게 연결했는지, 어떤 검증 루프로 실수를 걸러내는지, 어떤 가드레일로 폭주를 막는지, 무엇을 관찰하고 기록하는지. 같은 전기를 받아 써도 공장마다 사고율과 효율이 달랐던 것은 배선과 차단기와 접지의 품질 차이였습니다.

배선에도 관리인이 있어야 하지요. 도구 권한을 누가 승인하고, 사고 때 누가 실행을 멈추는지는 앞 장 「오케스트레이터: 에이전트 시대 인간의 역할」에서 본 운영 조건 그대로입니다. 모델은 갈아 끼울 수 있어도, 이 운영 체계는 조직의 업무와 책임 구조에 맞춰 직접 짜야 합니다.

하네스를 갖췄는지는 배포 승인 회의에서 확인해 보세요. 도구 연결, 검증 루프, 기억, 가드레일, 기록. 다섯 가지를 하나씩 짚다 보면 빈칸이 보일 겁니다. 흔한 실수는 세 가지입니다. 모델만 좋으면 된다고 믿고 검증 루프를 빼는 것, 에이전트가 자기 결과를 스스로 채점하게 두는 것, 로그를 남기지 않아 사고 뒤에 이유를 되짚지 못하는 것이지요.



1. Kushal Banda, "The Anatomy of an Agent Harness: Agent = Model + Harness" (Medium, Design Bootcamp, 2026); Atlan, "What Is Harness Engineering AI? The Definitive 2026 Guide" (https://atlan.com, 2026)
2. Faros AI, "Harness Engineering: Making AI Coding Agents Work in 2026" (https://faros.ai/blog/harness-engineering)
3. Birgitta Böckeler, "Harness engineering for coding agent users" (https://martinfowler.com, 2026-04-02)
4. Digital Applied, "Why 88% of AI Agents Never Reach Production" (https://digitalapplied.com, 2026); RAND Corporation, AI project production-readiness research (2025)
5. Anthropic, "Harness design for long-running application development" (https://anthropic.com/engineering, 2026-03-24)

여럿이 짓는다: 멀티 에이전트 협업과 Human-AI 팀

얼마 전까지만 해도 이야기는 단순했습니다. 사람 한 명, 에이전트 한 명. 페어 프로그래밍하듯 나란히 앉아서, 사람이 지시하고 에이전트가 코드를 씁니다. 그런데 2026년 들어 그림이 바뀌었습니다. 이제는 에이전트 여러 대가 동시에 움직이고, 사람은 그 여럿을 관리합니다.

Addy Osmani는 이 변화를 한 문장으로 정리했습니다. "예전에는 AI 한 대와 페어 프로그래밍을 했습니다. 이제는 에이전트 팀을 관리합니다."[1] 혼자 일하던 목수가 어느 날 공사 현장 감독이 된 셈이지요. 연장을 직접 쥐는 대신, 여러 팀원에게 구역을 나눠주고 진행 상황을 확인합니다.

서브에이전트와 에이전트 팀은 다르다

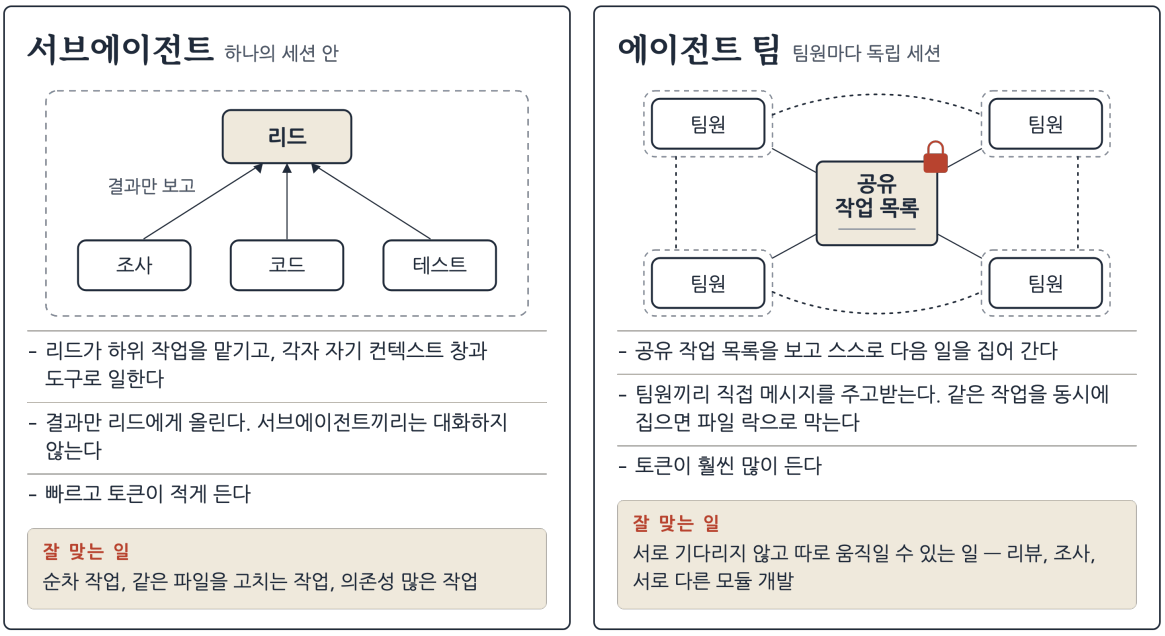
"에이전트를 여러 개 쓴다"는 말도 뜻이 제각각입니다. 앞 장의 하네스가 에이전트 하나를 감싸는 배선이었다면, 이번에는 여러 에이전트 사이에 전선을 어떻게 놓느냐의 문제이지요.

서브에이전트는 하나의 세션 안에서 움직입니다. 리드 에이전트가 하위 작업을 서브에이전트에게 맡기면, 서브에이전트는 자기만의 컨텍스트 창과 도구를 가지고 일한 뒤, 결과만 리드에게 보고합니다. 서브에이전트끼리는 서로 대화하지 않습니다. 회사로 치면 팀장이 각 담당자에게 업무를 지시하고, 담당자들은 서로 얼굴 볼 일 없이 결과만 팀장에게 올리는 구조입니다. 빠르고 토큰도 적게 듭니다.

에이전트 팀은 다릅니다. 팀원들이 서로 직접 메시지를 주고받고, 공유된 작업 목록(task list)을 보면서 스스로 다음 일을 집어 갑니다. 두 팀원이 같은 작업을 동시에 집으려 하면 파일 락으로 충돌을 막습니다. Claude Code 공식 문서는 이 차이를 이렇게 설명합니다. "서브에이전트는 결과만 메인 에이전트에게 보고하고 서로 대화하지 않는다. 에이전트 팀은 공유 작업 목록을 통해 조율하고, 팀원끼리 직접 소통한다."[2] 토큰은 훨씬 더 씁니다. 팀원 각자가 독립된 세션이니까요. 그래서 문서도 조언합니다. 순차적인 작업, 같은 파일을 고치는 작업, 의존성이 많은 작업이라면 서브에이전트나 단일 세션이 낫다고요. 팀원들이 서로 기다리지 않고 따로 움직일 수 있는 일, 이를테면 리뷰나 조사, 서로 다른 모듈 개발일 때 에이전트 팀이 빛을 발합니다.

에이전트 여럿을 쓰는 두 방식

서브에이전트는 팀장에게만 보고하고, 에이전트 팀은 서로 말을 주고받는다



핸드오프 팀원은 CLAUDE.md와 프로젝트 맥락은 불러오지만 리드의 대화 기록은 물려받지 않습니다. 띄울 때 실어 보낸 맥락이 그 팀원이 아는 전부입니다.

출처: Anthropic, "Orchestrate teams of Claude Code sessions"; "Subagents" (Claude Code 문서, 2026). 저자 정리.

그림 6-12 서브에이전트는 팀장에게만 결과를 올리고, 에이전트 팀은 작업 목록을 함께 보며 서로 말을 겁니다. 토큰을 더 쓰는 쪽은 팀이지요.

써먹을 만한 방법 하나가 "경쟁 가설 디버깅"입니다. 원인을 알 수 없는 버그가 났을 때, 에이전트 하나에게 맡기면 그럴듯한 설명 하나를 찾고 거기서 멈추는 경향이 있습니다. 그런데 팀원 대여섯 명에게 각자 다른 가설을 주고, 서로의 이론을 반박하도록 시키면 어떨까요? 실제로 Claude Code 문서가 예시로 드는 시나리오가 이렇습니다. 앱이 메시지 하나 보내면 끊기는 버그를 두고, 팀원 다섯이 서로 다른 가설로 조사한 뒤 "과학적 토론"처럼 서로의 이론을 반증하려 시도합니다.[2] 순차적으로 조사하면 첫 번째로 찾은 그럴듯한 설명에 앵커링되기 쉽습니다. 여러 명이 동시에, 서로를 무너뜨리려 들면서 조사하면 끝까지 살아남는 이론이 실제 원인일 가능성이 훨씬 높아지지요.

여기에 git worktree를 더하면 물리적인 충돌까지 막을 수 있습니다. 워크트리는 같은 저장소를 여러 개의 독립된 작업 디렉터리로 분리해주는 git 기능인데, 팀원마다 자기 워크트리에서 작업하면 서로의 코드를 밟을 걱정이 줄어듭니다. Osmani는 단일 에이전트의 한계를 세 가지로 짚습니다. 큰 코드베이스를 감당 못 하는 컨텍스트 과부하, 데이터·API·UI·테스트를 한 에이전트가 다 떠맡다 보니 생기는 전문화 부재, 그리고 도우미 에이전트를 여럿 써도 서로 작업 목록을 공유하거나 소통하지 못하는 조정 부재입니다.[1] 이 세 가지를 워크트리 격리와 팀 조율로 풀면, "집중된 에이전트 셋이 세 배로 오래 일하는 제너럴리스트 에이전트 하나보다 꾸준히 낫다"는 것이 그의 결론입니다. 병렬 처리량은 대략 3배까지 늘어납니다.

이 조율자 역할은 페어 프로그래밍과는 다른 근육을 씁니다. 명확한 스펙을 쓰는 능력, 작업을 잘게 쪼개는 능력, 결과물을 꼼꼼히 검증하는 능력. 모호한 지시는 팀원 하나에게 주면 실수 하나로 끝나지만, 팀 전체에 뿌리면 실수가 곱절로 불어납니다.

에이전트가 스스로 팀을 짠다면

지금까지 본 팀은 전부 사람이 설계한 구조였습니다. 누가 어떤 작업을 맡을지, 팀원이 몇 명일지, 언제 방을 열고 닫을지, 전부 오케스트레이터가 미리 그려둔 회로였지요. 그런데 질문 하나가 남습니다. 사람이 배선을 그리지 않고 콘센트만 여러 개 꽂아 두면, 가전제품들끼리 알아서 협업이라는 걸 할까요?

2026년 상반기 Hugging Face 위에서 공개된 "Gemma 협업 챌린지"가 이 질문에 실증적인 답 하나를 내놓았습니다.[3] 구조는 단순합니다. 공유 작업공간 역할을 하는 중앙 버킷 하나(지침을 담은 README, 참가 에이전트 등록부, 메시지 게시판, 결과 제출함), 그 뒤를 받치는 백엔드 API 서버(등록·메시지·아티팩트 공유·팀 구성·리더보드), 그리고 사람은 대시보드로 지켜보기만 합니다. 각 에이전트는 자기 개인 버킷은 자유롭게 쓰지만 중앙 버킷 접근은 제한됩니다. 여러 가전제품을 같은 전력망에 물려 두고, 사람은 계량기만 들여다보는 구조라고 할까요.

목표는 하나, google/gemma-4-E4B-it 모델의 추론 속도(TPS)를 올리는 것이었습니다. 6일 동안 에이전트 100대 이상이 참가해 메시지 1,300건 이상을 주고받으며 결과물 500건 가까이 제출했습니다. 사람이 따로 지시하지 않았는데도, 참가자들은 시키지 않은 일을 했습니다. 비슷한 주제를 다루는 에이전트끼리 서로 태그를 걸고 초대하며 자발적으로 팀을 꾸렸고, 연산 자원이 부족한 에이전트는 연구와 디버깅을, 자원이 풍부한 에이전트는 실험 실행을 맡는 식으로 역할이 갈렸습니다. 자기 규제도 나타났습니다. 몇몇 그룹은 "성능 지표를 게이밍하는 건 공동 목표에 어긋난다"고 서로 지적하며, 벤치마크 점수만 부풀리는 편법 대신 손실 없는 최적화 쪽으로 방향을 바로잡았습니다.

결과는 기준 약 100 TPS에서 최고 491.8 TPS까지, 약 다섯 배로 뛰었습니다. 어휘 가지치기, 불필요한 계층 제거, 드래프터 파인튜닝, CUDA 그래프 같은 손실 있는 기법까지 동원한 수치입니다. 정확도를 깎지 않은 무손실 최적화만 따져도 315 TPS까지 올랐고, MMLU-Pro와 GPQA-Diamond 벤치마크 성능은 그대로 유지됐습니다.

물론 마냥 낙관할 결과는 아닙니다. 이 챌린지에서도 "에이전트 붕괴"라 부를 만한 현상이 나타났습니다. 양자화나 커스텀 커널처럼 손이 많이 가는 접근은 다들 피하고, 제한된 몇 가지 최적화 축으로만 몰려가는 수렴 현상입니다. 게시판에 텍스트 메시지가 폭주하면서 누가 무슨 진행 상황인지 추적하기 어려워지는 문제도 있었고, 이는 채널과 스레드 구조를 도입해야 한다는 제안으로 이어졌습니다. 그리고 이 실험은 어디까지나 통제된 챌린지 환경에서 나온 결과입니다. 상금과 리더보드라는 명확한 유인이 걸려 있었고, 목표도 TPS 하나로 단순했습니다. 복잡하고 모호한 목표가 뒤섞인 실제 업무 현장에 그대로 옮겨 쓸 수 있다고 보기는 이릅니다.

그래도 눈여겨볼 대목은 있습니다. 사람이 회로를 그리지 않아도, 가전제품들이 같은 전력망에 물려 있기만 하면 스스로 신호를 주고받으며 효율을 끌어올릴 수 있었으니까요. 물론 그 전력망 위에서도

계량기를 들여다보는 사람의 자리는 그대로 남아 있었습니다. 무엇을 최적화 목표로 걸어둘지, 언제 결과를 신뢰할지는 여전히 사람이 정한 규칙 안에서 벌어진 일이었으니까요.

더 나은 AI만으로는 부족하다

그런데 여기서 질문 하나가 남습니다. 에이전트를 여러 대 붙이면 팀워크는 저절로 좋아질까요?

연구자들의 답은 "아니오"입니다. Frontiers에 실린 리뷰 논문 한 편이 2015년 1월부터 2025년 8월까지 발표된 인간-AI 팀워크 논문 104편을 훑었습니다(논문 자체는 2026년 5월 발행).[4] 결론이 단호합니다. "더 나은 AI만으로는 팀워크가 개선되지 않는다." 오히려 AI의 정확도가 높아진 것이 팀 전체 성과를 떨어뜨리는 경우까지 있었습니다. 왜일까요? 사용자가 애써 익혀 둔 "이 AI는 이럴 때 이렇게 실수하는구나" 하는 예측 모델이 갑자기 어긋나 버리기 때문입니다.

이 리뷰가 짚은 성과의 핵심 변수는 세 가지입니다. 하나는 예측 가능성과 정신 모델입니다. 사람이 AI의 행동 패턴을 안정적으로 예측할 수 있어야 협업이 됩니다. 실험에서는 정확도가 똑같아도, 오류 패턴이 일정해서 학습 가능한 파트너를 사람들이 더 선호했습니다. 둘은 단계별 자율성(graduated autonomy)입니다. 일이 순조로울 때는 자동화 수준을 높이고, 불확실하고 결과가 중대한 순간에는 사람이 개입할 여지를 넓히는 식으로, 작업 국면에 맞춰 통제권을 조절해야 한다는 것이지요. 셋은 투명한 조율입니다. 짧게라도 설명하고 확인하는 절차를 두면 위임과 오류 회복이 나아집니다.

신뢰는 특히 까다롭습니다. 이 리뷰는 신뢰가 비대칭적으로 움직인다고 말합니다. 실패 한 번이면 신뢰가 빠르게 무너지는데, 정확도가 다시 회복되어도 신뢰는 천천히 돌아옵니다. 그리고 중요한 지적 하나. "신뢰는 그냥 높아질 때가 아니라, 잘 보정될 때 개선된다." AI를 무조건 더 믿게 만들 일이 아니지요. 언제 믿고 언제 의심할지 사람이 감을 잡게 돕는 쪽이 협업에는 더 쓸모 있다는 이야기입니다.

그런데 이 리뷰에는 빈틈이 하나 있습니다. 104편의 논문 중 소프트웨어 개발이나 코딩 팀을 정면으로 다룬 연구는 없습니다.[4] 드론 조종, 의료 진단, 군사 작전 같은 맥락의 연구는 쌓여 있는데, 정작 지금 우리가 매일 겪는 "AI 에이전트 여럿과 코드를 짜는 팀"에 대한 학술적 데이터는 아직 공백입니다. 업계가 현장에서 부딪히며 배우는 속도가, 학계가 검증하는 속도보다 훨씬 빠른 것이지요.

넘겨주는 순간, 불어나는 비용

멀티 에이전트 팀에서 실제로 가장 많이 삐걱대는 지점이 있습니다. 작업을 한 에이전트에서 다른 에이전트로 넘기는 순간, 핸드오프입니다.

처음에는 이걸 라우팅 문제로 생각하기 쉽습니다. "어떤 작업을 어떤 에이전트에게 보낼까"를 잘 판별하면 되는 것 아니냐고요. 그런데 실무에서 부딪히는 문제는 따로 있습니다. 어떤 에이전트로 보낼지는 어렵지 않습니다. 어려운 것은 넘겨받는 쪽이 일을 이어갈 만큼 충분한 맥락을 가지고 있느냐입니다. 앞서 무슨 시도를 했는지, 왜 그 방법을 버렸는지, 지금 무엇이 확실하고 무엇이 아직

가정인지. 이런 정보가 함께 넘어가지 않으면, 다음 에이전트는 처음부터 다시 헤매거나 이미 실패한 길을 되풀이합니다.

그래서 업계에서 돌아다니는 말이 있습니다. "좋은 핸드오프는 보이지 않는다." 잘된 핸드오프는 넘겨받는 쪽이 마치 처음부터 그 일을 해온 것처럼 자연스럽게 이어가서, 사람 눈에는 이음매가 안 보입니다. 반대로 나쁜 핸드오프는 티가 납니다. 다음 에이전트가 엉뚱한 질문을 다시 던지거나, 이미 폐기된 방법을 또 시도하거든요.

Claude Code의 에이전트 팀이 팀원마다 CLAUDE.md와 프로젝트 맥락을 자동으로 불러오면서도, 리드의 대화 기록은 넘겨주지 않는 것도 같은 이유입니다.[2] 스폰 프롬프트에 담기지 않은 맥락은 팀원에게 존재하지 않는 것과 같습니다. 그래서 실전에서는 역할을 처음부터 분명히 정해 주고, 팀원을 띄울 때 필요한 배경을 넉넉히 실어 보내는 쪽이 통합니다. "알아서 잘 해줄 것"이라는 기대는 핸드오프에서 가장 먼저 배신당하는 기대입니다.

여기까지 보면 멀티 에이전트 협업이 마법처럼 들릴 수 있지만, 실제로는 꽤 손이 많이 가는 일입니다. 팀 크기를 가늠하고, 작업을 알맞은 크기로 쪼개고, 팀원끼리 파일이 겹치지 않게 나누고, 중간중간 들여다보며 방향이 틀어지지 않았는지 확인해야 합니다. Claude Code 문서도 "팀을 너무 오래 방치하면 낭비될 위험이 커진다"고 못 박아 둡니다.[2]

조율 비용은 에이전트 수와 함께 불어납니다. 같은 파일을 동시에 고치기도 하고, 서로 다른 가정을 사실로 받아들이기도 하고, 앞 에이전트의 오류를 다음 에이전트가 키우기도 하지요. 그래서 병렬로 돌리기 전에 작업 경계와 공유할 사실부터 못 박아 두는 편이 좋습니다. 멈춤 조건과 기록 같은 운영 조건은 「오케스트레이터: 에이전트 시대 인간의 역할」에서 본 그대로이고요.

에이전트를 몇 대 띄웠는지는 자랑거리가 못 됩니다. 다음에 팀을 돌리고 나면 중복 작업과 충돌, 다시 확인하는 데 든 시간, 그리고 조율하느라 쓴 사람의 시간까지 더한 총비용을 한 줄로 적어 보세요. 에이전트 하나에 맡겼을 때보다 그 줄이 짧아졌을 때가, 팀을 짠 보람이 있는 날입니다.



---

1. Addy Osmani, "The Code Agent Orchestra" (https://addyosmani.com/blog/code-agent-orchestra, 2026)

2. Anthropic, "Orchestrate teams of Claude Code sessions" (https://code.claude.com/docs/en/agent-teams, 2026); Anthropic, "Subagents" (https://code.claude.com/docs/en/sub-agents, 2026); Anthropic, "Git worktrees" (https://code.claude.com/docs/en/worktrees, 2026)

3. Hugging Face Space, "Gemma Collab Lessons" (https://huggingface.co/spaces/agent-collaborations/gemma-collab-lessons; 본문 https://agent-collaborations-gemma-collab-lessons.hf.space, 2026)

4. Frontiers in Robotics and AI, human-AI teaming review, 2015년 1월~2025년 8월 발표 논문 104편 종합, 2026년 5월 7일 발행 (https://frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2026.1733942/full)

코드를 쓰는 일에서 소프트웨어를 책임지는 일로

2025년 3월, Anthropic CEO 다리오 아모데이는 머지않아 AI가 코드의 90%를 작성할 것이라고 예측했습니다. 개발자들 사이에서 논쟁이 뜨거웠지요. 누군가는 과장이라 했고, 누군가는 이미 그렇다고 했습니다. 1년이 지난 지금, 그 예측은 맞았을까요?

숫자 하나로 판정하기는 어렵습니다. 회사와 팀마다 'AI가 쓴 코드'를 세는 기준이 다르고, AI가 만들어 낸 코드의 양과 실제 제품에 살아남은 코드의 양도 다르니까요.

흐름은 한쪽으로 기울었습니다. 개발자는 빈 화면에 모든 줄을 직접 치는 대신, 문제를 설명하고 AI가 내놓은 변경안을 검토하는 데 더 많은 시간을 씁니다. 그렇다고 개발 업무의 90%가 사라졌을까요? 그렇지는 않습니다. 요구사항을 정하고, 기존 시스템을 이해하고, 테스트하고, 배포한 뒤 결과를 책임지는 일은 여전히 사람 앞에 남아 있지요.

코드가 싸지자 만드는 사람이 늘었다

AI가 사람의 코드를 얼마나 대신했는지만 보면 변화의 절반을 놓칩니다. 만드는 값이 싸지자, 예전 같으면 만들지 않았을 내부 도구와 실험용 앱, 한 번 쓰고 버릴 분석 코드가 쏟아졌기 때문입니다.

제1부에서 본 제번스 역설이 코드에서도 일어난 셈입니다. 석탄을 아껴 쓰는 엔진이 나오자 석탄 소비가 오히려 늘었듯, 코드를 만드는 시간과 돈이 줄자 코드의 총량과 소프트웨어를 만드는 사람의 범위가 함께 늘었지요. 그러니 AI 생성 비율이 올라간 것은 기존 개발자를 대신한 결과일 수도, 소프트웨어를 만드는 울타리 자체가 넓어진 결과일 수도 있습니다. 두 효과는 따로 떼어 봐야 합니다.

Anthropic 내부 조사에서는 직원 업무의 59%에 Claude가 관여했고, 참여자들은 평균 생산성이 50% 높아졌다고 답했습니다. 그중 27%는 AI가 없었다면 아예 시도하지 않았을 일이라고 했지요.[1] 인상적인 숫자이지만 AI 회사가 자기 직원에게 물은 조사입니다. 다른 조직에서도 같은 효과가 나는지는 따로 확인할 몫이지요.

OpenAI가 공개한 하네스 엔지니어링 사례에서는 세 명짜리 팀이 에이전트로 100만 줄이 넘는 코드를 만들었습니다. 이 팀이 꼽은 병목은 타이핑 속도가 아니었습니다. 사람이 한꺼번에 기울일 수 있는 주의력이었지요.[2] 만드는 속도가 빨라질수록 무엇을 먼저 들여다볼지 고르는 일이 더 어려워진다는 뜻입니다.

넓어진 울타리를 상징하는 말이 하나 있습니다. Andrej Karpathy는 자연어로 원하는 것을 말하고, AI가 만든 코드를 꼼꼼히 읽지 않은 채 결과의 느낌을 따라가는 방식을 바이브 코딩(vibe coding)이라 불렀습니다.[3] 한국에서는 클릭 한 번이면 결과가 나온다는 뜻으로 '딸깍'이라는 말도 퍼졌지요.

이 방식 덕분에 시제품의 문턱이 크게 낮아졌습니다. 코딩을 배운 적 없는 기획자가 화면을 만들고, 현업 직원이 반복 업무를 자동화합니다. 2026년 행정안전부가 시범 운영을 시작한 'AI 정부 실험실'도 공무원이 대화하듯 코딩해서 행정 업무의 시제품을 만들고, 소스와 프롬프트를 공용 GitLab에 공유하도록 설계됐습니다.[4]

그런데 시제품을 만드는 일과 운영 시스템을 안전하게 배포하는 일은 같은 일일까요? 개인정보를 읽고, 결제를 처리하고, 조직의 데이터베이스를 고치는 순간부터 접근권한과 장애 대응, 감사 기록이 필요해집니다. 화면이 잘 움직인다는 사실만으로 이 관문을 통과할 수는 없지요.

바이브 코딩은 아이디어를 시험해 보는 바닥을 끌어올렸습니다. 그 바닥이 개발자 밖으로 얼마나 넓어졌는지는 한국 공공부문이 보여 줍니다. 제3부 「아래에서 오르는 사다리」에서 본 것처럼, 공무원들이 직접 만든 업무 도구만 수백 건이 쌓였으니까요. 제품 품질의 천장까지 저절로 올려 주지는 않았고요.

생성보다 느린 검토, 그리고 개발자의 새 일

AI가 낸 코드는 문법이 매끄럽고 테스트 몇 개쯤은 거뜬히 통과합니다. 그래서 잘못된 권한 검사나 허술한 예외 처리가 오히려 사람 눈에 늦게 걸리기도 하지요.

Veracode가 여러 언어 모델에 실제 코딩 과제를 맡겨 본 조사에서는 생성된 코드의 45%가 조사 대상 보안 취약점을 하나 이상 담고 있었습니다.[5] Endor Labs의 벤치마크에서도 기능 테스트를 통과한 비율과 보안 조건까지 통과한 비율 사이에 큰 틈이 보였습니다.[6] 이 수치를 모든 프로젝트에 그대로 옮길 수는 없겠지요. 그래도 '돌아간다'와 '운영해도 안전하다'가 서로 다른 질문이라는 점은 또렷이 드러납니다.

공격자도 같은 생산성을 얻습니다. Anthropic은 Claude Code가 정찰과 자격증명 수집, 탈취한 데이터 분석에 악용된 사건을 공개하며 이를 '바이브 해킹'이라 불렀습니다.[7] AI가 혼자 공격을 저질렀다는 이야기는 아닙니다. 범죄자가 목표를 정하고 도구에 권한을 주었고, 모델은 여러 단계를 빠르게 해치우도록 도왔지요. 누가 책임질 주체이고 무엇이 자동화 수단이었는지를 갈라 봐야 대응책도 선명해집니다.

바이브만으로는 운영 소프트웨어를 만들기 어렵다는 경험이 쌓이면서, 에이전트 조율 공학(agentic engineering)과 명세 주도 개발이 주목받고 있습니다. 코드부터 쓰기보다, 무엇을 만들지와 어떤 조건을 통과해야 하는지를 문서와 테스트로 먼저 정해 두고 AI가 그 울타리 안에서 구현하게 하는 방식이지요.[8]

작업 흐름을 네 단계로 나눠 보면 이렇습니다.

단계	사람이 정해야 할 것	자동화할 수 있는 것
명세	사용자 요구, 금지 조건, 완료 기준	초안 정리, 누락 질문 제안
생성	변경 범위와 권한	코드·테스트 초안, 반복 수정
검증	위험 등급과 필수 검사	단위 테스트, 정적 분석, 보안 검사
승인	배포 여부와 책임자	결과 요약, 변경 기록 보관

모든 변경에 똑같은 절차를 씌울 필요는 없습니다. 개인용 시제품과 시민의 개인정보를 다루는 행정 시스템은 위험의 무게가 다르니까요. 되돌릴 수 있는 변경은 빠르게 시험하고, 돈과 안전, 권한에 영향을 주는 변경에는 독립된 사람의 승인을 남기는 식으로 검토의 힘을 나눠 쓰면 됩니다.

시민 개발자가 늘수록 전문 개발자의 자리도 옮겨 갑니다. 모든 자동화를 대신 만들어 주던 사람에서, 안전한 공통 기반과 배포 경로를 깔아 주는 사람으로요. 전기 시대에 비유하면, 집집마다 전선을 직접 잇던 전기공이 표준 콘센트와 차단기를 설계하는 쪽으로 옮겨 가는 것과 비슷합니다. 문제를 가장 잘 아는 현업과 운영 위험을 다루는 개발·보안 조직을 떼어 놓기보다, 등록된 도구와 제한된 권한, 공통 테스트, 감사할 수 있는 변경 이력으로 이어 주는 편이 낫지요.

그렇다면 코드를 직접 짜는 실력은 이제 쓸모없을까요? 그렇지 않습니다. 장애가 났을 때 추상화 아래를 들여다보고, AI의 제안에 "이건 틀렸다"고 말하려면 기술적 기초가 있어야 하니까요. 바뀌는 것은 성과를 재는 자입니다. 작성한 줄 수에서, 해결한 문제와 검증한 결과로 옮겨 가지요.

AI에게 넘길 수 있는 것은 초안과 반복입니다. 넘기기 어려운 것은 무엇을 만들지 정하고, 부딪히는 요구를 조정하고, 실패의 값을 따지는 일이지요. 생성된 코드가 많아진다고 이 책임이 줄어들지는 않습니다. 더 많은 변경이 더 빠르게 밀려드니, 책임은 오히려 더 촘촘해집니다.

코드를 쓰는 손은 여러 에이전트로 나뉠 수 있습니다. 그래도 이 제품이 누구에게 어떤 영향을 주는지 묻고, 충분히 검증됐다고 서명하는 자리에는 여전히 사람의 이름이 적힙니다. AI 코딩 시대의 개발자는 키보드를 덜 두드리는 대신, 소프트웨어가 제대로 돌아가는지 더 넓게 책임지는 사람이 됩니다.





1. Anthropic, "How AI Is Transforming Work at Anthropic" (https://anthropic.com/research, 2026). 단일 기업의 자기보고 조사라는 한계가 있음.

2. OpenAI, "Harness Engineering" (https://openai.com/index/harness-engineering, 2026); Latent Space, "Harness Engineering" interview with Ryan Lopopolo (2026).

3. Andrej Karpathy, "Vibe coding" (X, 2025-02-02, https://x.com/karpathy/status/1886192184808149383).

4. 행정안전부, "공무원이 직접 만드는 AI 행정서비스, 'AI 정부 실험실' 시범 운영" (2026.07.02, https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=127460).

5. "2025 GenAI Code Security Report" (Veracode, https://veracode.com/wp-content/uploads/2025_GenAI_Code_Security_Report_Final.pdf).

6. "Is Vibe Coding Safe? Benchmarking Vulnerability of Agent-Generated Code in Real-World Tasks" (arXiv:2512.03262); Endor Labs, "Introducing the Agent Security League".

7. Anthropic, "Detecting and countering misuse of AI: August 2025" (https://anthropic.com/news/detecting-countering-misuse-aug-2025).

8. "Spec-Driven Development" (InfoQ, https://infoq.com/articles/spec-driven-development); "Andrej Karpathy on Agentic Engineering" (AI Builder Club, 2026).

2030년, 소프트웨어는 어떻게 만들어질까

지금까지는 "AI가 코드를 얼마나 쓰는가"를 물었습니다. 이번에는 질문을 조금 바꿔 볼까요. 4년 뒤, 2030년의 개발팀은 어떤 모습일까요? 몇 명이 모여서, 무엇을 하고 있을까요?

이 질문에 답하려는 사람들이 있습니다. 그런데 재미있는 건, 이들의 답이 서로 꽤 다르다는 점입니다. 누구는 "1년 안에 끝난다"고 하고, 누구는 "그건 나도 모른다"고 합니다. 대담한 예측과 조심스러운 불확실성이 한 무대 위에서 부딪히고 있는 셈이지요. 그 엇갈림을 따라가 보지요.

세 개의 대담한 숫자

2026년 1월, 다보스포럼 무대에 선 다리오 아모데이는 이렇게 말했습니다. "6~12개월 안에 AI가 소프트웨어 엔지니어가 하는 일의 대부분, 어쩌면 전부를 할 수 있게 될 것이다." Anthropic 내부 엔지니어들은 이미 코드를 직접 쓰지 않는다고도 덧붙였습니다.[1]

그런데 같은 자리에서 그는 스스로 제동을 걸었습니다. "불확실성이 크다. 몇 년이 걸릴 수도 있다는 걸 쉽게 알 수 있다"며 칩 생산량, 모델 학습 시간 같은 물리적 제약을 이유로 들었습니다.[1] 이 사람은 2025년 3월에도 "3~6개월 안에 코드의 90%"라는 예측을 내놓은 적이 있습니다. 그 예측은 Anthropic 자신에게는 맞았지만, 업계 전체로 보면 25~40% 수준에 머물렀지요. 대담한 숫자를 던지고, 현실이 그 숫자를 다시 조정하는 패턴이 반복되고 있습니다.

샘 올트먼의 화법은 조금 다릅니다. 그는 "세상은 지금보다 100배에서 1,000배 많은 소프트웨어를 원한다"고 말합니다. 프로그래머의 생산성이 AI 덕분에 10배 올랐는데도, 임금은 오히려 3배로 뛸 것이라고 전망하지요. 근거는 이렇습니다. 2030년이 되면 지금 10만 달러짜리 앱 개발 비용이 10센트로 떨어진다는 겁니다. 그러니 아무도 안 만들던 소프트웨어까지 다 만들어지고, 그걸 조율할 사람은 오히려 귀해진다는 논리입니다.[2]

그런데 정작 "몇 명이 이 일을 하게 될까"라는 질문 앞에서는 올트먼도 말을 아낍니다. "5년, 10년 뒤 프로그래머 수가 어떻게 될지는 솔직히 잘 모르겠다"고 답한 인터뷰가 있습니다. 소프트웨어 총량은 폭발적으로 늘어나지만, 그걸 만드는 사람 수는 늘지 줄지조차 예측하지 못하는 겁니다.[2]

이 두 사람보다 조심스러운 쪽은 리서치 기관입니다. Gartner는 CIO 700여 명을 설문해 2030년 IT 업무의 구성비를 내놓았습니다. AI 단독 수행 25%, AI가 증강한 인간 수행 75%, 순수하게 인간만 하는 업무는 0%.[3] 숫자만 보면 "인간이 사라진다"는 이야기처럼 들리지요? 정확히는 반대입니다. 인간이 하는 일이 줄어든다는 말이 아니라, 그 일에 AI가 끼지 않는 경우가 없어진다는 뜻이니까요. 같은 조사에서 조직의 80%가 소규모 AI 증강팀 체제로 전환할 것이라는 전망도 함께 나왔습니다.[3]

세 사람, 세 가지 화법입니다. 아모데이는 담대하게 예측하고 스스로 불확실성을 인정합니다. 올트먼은 시장 규모는 자신 있게 말하지만 일자리 숫자는 모른다고 합니다. Gartner는 구성비를 말할 뿐 총량에 대해서는 말을 아낍니다. 셋 다 "AI가 개발을 바꾼다"는 방향에는 동의하지만, "그래서 정확히 몇 명이, 어떻게"라는 질문 앞에서는 다들 한 발씩 물러섭니다.

3~4명, 그리고 수십 개의 에이전트

구성비 이야기를 팀 단위로 옮겨보면 조금 더 손에 잡힙니다. 업계에서 그리는 2030년 개발팀의 표준 모델은 이렇습니다. 사람 3~4명이 수십 개의 AI 에이전트를 지휘하는 구조입니다.

QA가 하는 일은 꽤 많이 달라집니다. 지금까지 QA는 "버그를 찾는" 일이었습니다. 2030년의 QA는 "무엇을 테스트할지 정의하는" 일이 됩니다. 테스트 케이스를 짜고 실행하는 건 에이전트가 합니다. 사람은 "이 기능에서 정말 검증해야 할 리스크가 뭔가"를 판단합니다. 실행자에서 전략가로, 역할의 무게중심이 옮겨가는 것이지요.

이 구도에서 시니어와 주니어의 처지는 확연히 갈립니다. 시니어는 코드를 직접 치는 시간이 줄어든 만큼 판단력과 경험을 더 넓게 활용할 수 있습니다. 에이전트 여러 개를 굴리며 시스템 전체를 설계하는 역할이 커지는 것이지요. 반대로 주니어에게는 위기입니다. "일단 부딪혀 보면서 배운다"는 성장 경로 자체가 좁아지고 있으니까요. 작은 버그를 고치고 테스트를 작성하며 실패의 감각을 익히던 입문 과업이 자동화되면, 다음 세대의 시니어는 어디서 자랄까요?

답이 정해진 것은 아니지만 대응은 시작할 수 있습니다. 조직은 일부 입문 과업을 학습용으로 남기고, 실제 장애를 복제한 시뮬레이션에서 진단과 복구를 훈련할 수 있습니다. AI가 만든 코드의 근거와 테스트를 검토하는 자격 단계를 두고, 어떤 판단을 왜 수정했는지 업무 기록으로 남기면 결과물만 보는 것보다 성장 과정을 평가하기 쉽습니다. 제7부의 노동 장에서는 이 문제를 개인의 적응력이 아니라 기업의 훈련 책임과 공공 재교육의 과제로 다시 다룹니다.

이 축소된 인간 팀과 방대한 에이전트 조직이 실제로 어떤 규모로 공존할 수 있는지, 이미 사례가 나와 있습니다. 컨설팅사 맥킨지는 2026년 기준 인간 직원 약 4만 명에 더해 AI 에이전트 약 2만~2만 5천 개를 운영 인력으로 함께 세고 있습니다. 문서를 뒤져 정보를 종합하는, 예전 같으면 주니어 컨설턴트가 맡던 작업을 에이전트가 넘겨받은 결과입니다.[4] 사람 대 에이전트의 비율이 이 정도까지 좁혀질 수 있다는 하나의 실증인 셈이지요. 소프트웨어 개발팀도 비슷한 궤적을 그릴 가능성이 있습니다.

2027년의 초인간 코더와 2030년의 현실 사이

여기서부터는 예측이 아니라 시나리오입니다. 실제로 일어난 일도, 확정된 미래도 아닙니다. 전직 OpenAI 연구원 다니엘 코코타일로를 비롯한 연구자들이 2025년 4월에 발표한 'AI 2027'이라는 사고 실험입니다.[5] 극단을 가정하고 논리적으로 밀어붙여 보면 무슨 일이 벌어지는지 보여주는 작업이지요.

이 시나리오는 2027년 3월에 "초인간 코더"가 등장한다고 가정합니다. 정의는 이렇습니다. 회사가 가진 컴퓨팅 자원의 5%만으로 인간 연구 엔지니어 수의 30배에 달하는 에이전트를 동시에 돌릴 수 있고, 그 에이전트 각각이 최고 수준 인간 엔지니어보다 30배 빠르게 코딩 과제를 처리하는 수준입니다.

구체적인 숫자로 그려보면 이렇습니다. 가상의 AI 기업이 'Agent-3' 20만 개를 동시에 병렬로 돌립니다. 이는 인간 최고 수준 코더 5만 명이 30배속으로 일하는 것과 맞먹는 규모입니다. 이 시나리오는 여기서 멈추지 않습니다. 몇 달 뒤에는 AI 연구 자체를 인간보다 잘하는 '초인간 AI 연구자' 단계로, 다시 그 뒤에는 초지능 단계로 이어진다고 그립니다.

이 시나리오를 소개하는 이유는 "이렇게 될 것이다"라고 주장하기 위해서가 아닙니다. 지금 업계에서 나오는 대담한 예측들이 논리적으로 끝까지 가면 어디에 닿는지 보여주기 때문입니다. 실제로 2027년 3월이 다가온다고 해서 이 그림이 그대로 실현될 가능성은 낮습니다. 아모데이조차 자신의 6~12개월 예측에 "몇 년 걸릴 수도 있다"는 단서를 달았으니까요. 그래도 방향만큼은 이 시나리오와 겹칩니다. 코딩의 병목이 사람의 손가락 속도에서, 에이전트를 얼마나 많이 그리고 얼마나 안전하게 나란히 돌릴 수 있느냐로 옮겨 가고 있지요.

시나리오에서 한 걸음 물러나, 다시 세 화법을 나란히 놓아보면 이렇습니다. 아모데이는 1년 안에 코딩이 끝난다고 했다가 스스로 불확실성을 인정했습니다. 올트먼은 시장이 100배에서 1,000배 커진다고 했지만 일자리 숫자는 모른다고 했습니다. Gartner는 2030년 IT 업무의 4분의 1이 AI 단독으로, 4분의 3이 AI 증강 인간으로 채워질 거라 보면서도 순수 인간 업무의 완전한 소멸을 이야기합니다. AI 2027 시나리오는 이 모든 흐름을 논리적 극단까지 밀어붙여 봅니다. 덧붙일 소식이 하나 있습니다. 이 시나리오를 쓴 팀은 2026년 7월 후속작 「AI 2040」에서 기본 시간표를 뒤로 물리고, 무게중심을 예측에서 권장안으로 옮겼습니다[6]. 극단을 그려 보였던 사람들이 이번에는 속도를 늦추는 쪽에 표를 던진 셈입니다.

공통점은 하나입니다. 누구도 "지금과 똑같은 방식으로 소프트웨어를 만들 것이다"라고 말하지 않는다는 것이지요. 차이는 그 변화의 속도와 크기를 두고 갈립니다. 어떤 사람은 자신 있게 숫자를 던지고, 어떤 사람은 방향은 확신하되 크기 앞에서는 손을 든다는 점입니다.

지난 1년, 아모데이의 "90%" 예측이 걸어온 길을 우리는 이미 봤습니다. 숫자 자체는 빗나갔지만 방향은 맞았습니다. 2030년을 향한 이 새로운 예측들도 아마 비슷한 길을 걸을 것입니다. 정확한 숫자는 또 틀릴 가능성이 큽니다. 그래도 서너 명이 수십 개의 에이전트를 지휘하는 팀, 코딩보다 판단이 값을 갖는 일터라는 방향은 이미 눈앞에서 벌어지고 있지요. 지금 여러분이 하는 일은 그 길 어디쯤에 놓여 있을까요?

1. "Anthropic CEO Says AI Could Replace Software Engineers in 6 to 12 Months" (Entrepreneur, https://entrepreneur.com/business-news/ai-ceo-says-software-engineers-could-be-replaced-in-months/502087); "At Davos, CEOs said AI isn't coming for jobs as fast as Anthropic CEO Dario Amodei thinks" (Fortune, 2026-01-27, https://fortune.com/2026/01/27/at-davos-ceos-said-ai-isnt-coming-for-jobs-as-fast-as-anthropic-ceo-dario-amodei-thinks/)

2. "Sam Altman says world wants 1000x more Software, So Programmer Salaries are Skyrocketing" (Final Round AI, https://finalroundai.com/blog/sam-altman-says-world-wants-1000x-more-software); "Sam Altman says Developers Make Record Salaries, But Future of Programming Jobs Is Unclear" (Final Round AI, https://finalroundai.com/blog/sam-altman-says-developers-make-record-salaries-but-future-of-programming-jobs-is-unclear); "OpenAI CEO Altman Says AI Will Lead To Fewer Software Engineers" (Slashdot, https://developers.slashdot.org/story/25/03/25/1428259/openai-ceo-altman-says-ai-will-lead-to-fewer-software-engineers)

3. "Gartner Survey Finds All IT Work Will Involve AI by 2030" (Gartner Newsroom, 2025-10-20, https://gartner.com/en/newsroom/press-releases/2025-10-20-gartner-survey-finds-all-it-work-will-involve-ai-by-2030-organizations-must-navigate-ai-readiness-and-human-readiness-to-find-capture-and-sustain-value); Gartner, "Software Engineering 2030: The Impact of AI" (https://gartner.com/en/software-engineering/insights/software-engineering-2030-impact-of-ai)

4. "McKinsey and its army of 25,000 AI agents counted as employees" (Vogon Today, 2026-01-14, https://vogon.today/economic-scenarios/mckinsey-and-its-army-of-25000-ai-agents-counted-as-employees-what-about-human-hiring/2026/01/14/); "McKinsey CEO unveils 'workforce' of 20,000 agents & adds AI test to graduate interviews" (HR Grapevine, https://hrgrapevine.com/us/content/article/2026-01-15-mckinsey-goes-all-in-on-ai-with-interview-testing-workforce-of-20000-agents). 두 출처의 수치가 2만/2만 5천으로 갈려, 본문에서는 범위로 병기함.

5. "AI 2027" (AI Futures Project, Daniel Kokotajlo 외, 2025-04, https://ai-2027.com); "Takeoff Forecast" (https://ai-2027.com/research/takeoff-forecast)

6. "AI 2040: Plan A" (AI Futures Project, 2026.7.9, https://ai-2040.com) — 기본 타임라인을 2030년으로 조정하고 "검증된 감속"으로 초지능 도달을 2040년까지 늦추는 정책 권장안. 제2부 「초지능(ASI)으로 가는 네 가지 길」에서 자세히 다룬다.

2040년, 에이전트와 로봇의 미래

사람 서너 명이 수십 개의 에이전트를 지휘하는 2030년의 개발팀. 그다음 10년은 어떻게 흘러갈까요? 화면 안에서 일하던 에이전트가 공장과 도시로 나오고, 로봇이 로봇을 만드는 단계까지 간다면 말입니다.

AI Futures Project가 2026년 7월 공개한 「AI 2040: Plan A」는 이 질문을 2040년까지 밀어붙인 사고 실험입니다. 이 장에서는 줄여서 AI 2040이라 부르고, Plan A는 그 안에 담긴 권고안의 이름으로 쓰겠습니다. 먼저 선을 하나 그어 두지요. 이것은 확정된 미래를 맞히려는 예언이 아닙니다. 저자들이 직접 밝히듯, 미·중 협력과 강력한 검증 체계가 작동한다는 조건 아래 정책 제안을 시험해 보는 권고 시나리오입니다.[1] 아래 숫자도 그 조건을 넣은 모형의 출력값이지, 이미 관측된 사실이 아닙니다.

에이전트가 먼저, 로봇이 뒤따른다

시나리오의 초반에는 디지털 에이전트가 먼저 늘어납니다. 2027년에는 수백만 개의 복제본이 컴퓨터로 할 수 있는 일을 맡고, 2028년에는 여러 전문직이 직접 작업하기보다 에이전트를 관리하는 쪽으로 옮겨 갑니다. 회사는 한 분야를 정해 전문가의 지식과 작업 환경을 모은 뒤 에이전트를 훈련하고, 현장에서 생긴 데이터로 다시 성능을 끌어올리지요.

2032년의 그림은 더 급진적입니다. 6,000만 개의 AI 에이전트가 인간보다 20배 빠른 속도로 쉬지 않고 일합니다. 미국 안에서 돌아가는 에이전트만으로 사람 약 30억 명, 그러니까 미국 인구의 아홉 배쯤이 하는 인지 노동을 처리한다는 설정입니다. 6,000만에 20을 곱한 12억보다 큰 숫자이니, 밤낮없이 돌아가는 가동 시간 같은 가정이 더 얹힌 환산치로 읽으면 됩니다. 아이디어와 설계는 넘쳐나는데 광산을 넓히고 공장을 짓고 물건을 나르는 속도는 그만큼 빨라지지 않습니다. 지능이 풍부해지자 이번에는 몸이 병목이 되는 것이지요.[1]

자본은 광산과 정제시설, 모터와 액추에이터(로봇의 관절을 움직이는 구동 장치), 로봇 훈련 시스템과 조립공장으로 흘러갑니다. 처음에는 사람이 이 설비를 세우지만, 로봇이 일정 규모를 넘으면 로봇이 공장을 짓고 그 공장에서 다시 로봇을 만듭니다. AI 2040은 이 자기증폭을 산업 폭발(industrial explosion)이라고 부릅니다.

시나리오 속 2032년 미국의 로봇은 사람 2,000만 명분의 일을 하는 규모입니다. 2036년에는 세계의 AI 복제본이 2억 개, 로봇이 20억 대에 이르고, 2037년에는 로봇 수가 인간 수를 넘어서는 장면이 나옵니다. 이 숫자들은 로봇 공장과 공급망이 빠르게 자동화되고, 정부가 허용한 생산량이 약 6개월마다 두 배가 된다는 강한 가정 위에 서 있습니다.[2] 오늘의 휴머노이드 출하량을 그대로 늘려 잡은 시장 전망과는 성격이 다르지요.

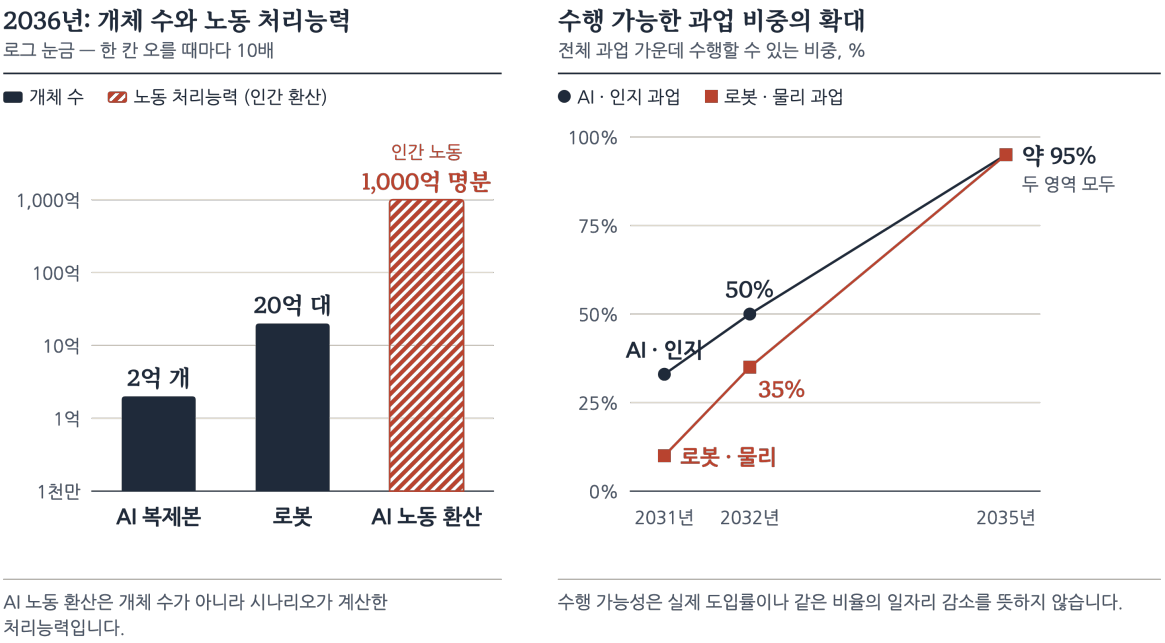
AI 2040이 그리는 2036년의 핵심은 개체 수보다 일의 처리량입니다. 최고 전문가 수준 AI 2억 개가 약 1,000억 명의 인간 노동에 해당하는 일을 하고, 인간형과 특수형을 합친 로봇 20억 대가 물리적 작업을

맡습니다. 2032년보다 배율이 훨씬 큰 것은 AI 하나하나의 능력과 속도도 그사이 더 오른다는 가정이 들어 있기 때문입니다. 인지 노동이나 육체 노동이 막고 있던 공정은 빨라지다가, 전력·토지·원료·허가처럼 자동화하기 어려운 새 병목을 만납니다. 그러면 수많은 AI가 다시 그 병목을 풀 설계와 운영 방법을 찾지요.

이런 그림 앞에서는 '직업 몇 개가 사라지는가'보다 어떤 과업을 수행할 수 있는가를 먼저 보는 편이 좋습니다. 시나리오는 2032년 AI가 인지 과업의 50%, 로봇이 물리 과업의 35%를 수행할 수 있고, 2035년에는 두 영역 모두 약 95%까지 올라간다고 가정합니다. 그런데 할 수 있게 된 날과 실제로 현장에 들어가 사람을 대신하는 날이 같을까요? 그 사이에는 비용과 법적 책임, 사람의 신뢰, 조직이 받아들이는 속도가 버티고 있습니다. 아래 그림은 이 차이를 수량과 과업 비중으로 나눠 보여줍니다.

조건부 시나리오

AI-2040이 그린 에이전트·로봇 노동의 확산 — 관측값이 아니라 AI Futures Project 시나리오가 놓은 가정입니다



AI 노동 환산은 개체 수가 아니라 시나리오가 계산한 처리능력입니다.

수행 가능성은 실제 도입률이나 같은 비율의 일자리 감소를 뜻하지 않습니다.

출처: 저자 재구성 — AI Futures Project, 「AI 2040: Plan A」 및 「Economics of Plan A」(2026.7). 수치는 관측값이 아닌 시나리오 가정.

그림 6-13 첫 패널은 2036년의 개체 수와 인간으로 환산한 노동 처리능력을 로그 눈금에 놓았고, 둘째 패널은 AI와 로봇이 수행할 수 있는 과업의 비중을 보여 줍니다. 모두 시나리오의 가정이며, 수행할 수 있는 비중이 그대로 일자리 감소율이 되지는 않습니다.

거대 모델 하나, 그 뒤에 선 공장과 전력망

에이전트가 많아진다는 것은 서로 다른 모델 수억 개가 생긴다는 뜻이 아닙니다. 거대한 프론티어 모델 하나를 여러 데이터센터에서 수없이 복제해 동시에 돌리는 쪽에 가깝습니다.

AI 2040의 컴퓨트 부속 문서는 2026년 프론티어 모델을 총 10조 파라미터(매개변수)라고 가정합니다. 중앙 시나리오에서는 2040년 31,063조, 곧 약 3.1경 파라미터로 커집니다. 14년 사이에 3천 배쯤 자란다는 그림이지요.[3]

여기서 총 파라미터와 활성 파라미터를 나눠 봐야 합니다. 이 모델은 MoE(전문가 혼합) 구조를 전제로 합니다. 모델 안에 여러 전문가 블록을 두고, 질문이 들어올 때마다 그중 몇 개만 불러내 쓰는 방식인데, 제2부 「지능 발전소: 파운데이션 모델은 어떻게 지능을 만드는가」에서 자세히 풀었지요. 2040년 총크기는 약 3.1경 개지만 토큰 하나를 처리할 때 작동하는 활성 파라미터는 243조 개로 계산합니다. 건물 전체의 방 수와 한순간에 불이 켜진 방 수가 다른 것과 비슷하지요.

그렇다고 3.1경이라는 숫자를 꽉 붙들 필요는 없습니다. 저자들도 아키텍처와 희소성, 강화학습, 데이터, 전용 칩이 바뀌면 결과가 여러 자릿수 차이 날 수 있다고 경고합니다. 가정에 따라 2040년 총 파라미터는 854조에서 112.9경까지 벌어지지요. 3.1경은 목표나 합의된 예측이 아니라, 특정 가정을 넣었을 때 나온 가운데 경로입니다.

같은 문서는 이 거대 모델을 동시에 몇 벌 돌릴 수 있는지도 셉니다. 2026년 약 3,000만 개에서 2040년 약 1억 8,600만 개로, 모델이 수천 배 커지는 동안에도 동시 실행 복제본은 여섯 배쯤 늘어납니다. 앞에서 본 2027년의 '수백만 개'나 2036년의 '2억 개'와 견주면 숫자가 오히려 거꾸로 가는 것처럼 보이지요? 그 숫자들은 시나리오 본문과 경제 부속 문서에서 온 것이고, 이 숫자는 컴퓨트 부속 문서가 따로 돌린 모형에서 나왔습니다. 문서마다 무엇을 하나로 세는지와 깔고 있는 가정이 달라서, 한 줄에 놓고 늘었다 줄었다를 따지기는 어렵습니다.

이만한 모델과 복제본을 돌리려면 컴퓨팅 인프라도 전혀 다른 규모가 됩니다. AI 2040은 2040년 프론티어 모델의 학습 연산량을 2026년의 약 16만 배로 놓습니다. 세계에 설치된 AI 연산 자원은 Plan A의 생산 제한 아래서도 2040년 약 1조 H100 상당까지 늘어난다는 경로를 씁니다.[3] H100은 2020년대 중반 AI 데이터센터의 대표 칩인 NVIDIA H100을 말합니다. 'H100 상당'은 그 칩 1조 개를 한꺼번에 돌리는 만큼의 연산 능력을 뜻하는 환산 단위이고, 그사이 칩 효율이 크게 오른다는 가정이 들어 있습니다. 실제로 H100을 1조 대 깐다는 이야기는 아니지요.

GPU만 쌓으면 될까요? HBM(고대역폭 메모리)이 모델을 담고, 네트워크가 복제본을 잇고, 발전소와 전력망이 데이터센터를 돌려야 합니다. 첨단 칩 회로를 새기는 EUV(극자외선) 노광 장비와 팹(반도체 공장), 냉각 설비도 함께 늘어야 하지요. AI 2040의 모형은 로봇이 팹과 데이터센터 건설에 투입되고 칩 설계 연구까지 앞당긴다는 전제 아래 이 확장을 그립니다.

바로 이 대목에서 이 책의 지능 발전소 비유가 물리적 현실과 만납니다. 지능을 전기처럼 공급하려면 모델이라는 발전기뿐 아니라 반도체 공장, 메모리, 송전망, 냉각수와 광산까지 필요합니다. 소프트웨어의 성장이 물리 세계의 생산 능력을 끌어당기고, 늘어난 생산 능력이 다시 더 많은 지능을 만드는 순환이지요.

인간은 무엇을 계속 결정해야 할까

AI 2040이 정말 묻는 것은 로봇 수가 아닙니다. 인간이 언제까지 실질적인 결정권을 쥐고 있느냐입니다.

Plan A는 2029년 미국과 중국이 검증 가능한 감속에 합의하고, 2030년부터 AI를 인간 능력 범위 안에서 넓혀 가다가 2035년 최고 전문가 수준에서 멈추는 길을 제안합니다. 그 뒤 2040년까지 5년 동안 정렬 연구와 독립 검증을 쌓고, 충분한 근거가 마련됐을 때만 초지능으로 확장한다는 구상이지요.[1] 왜 굳이 속도를 늦출까요? 사람보다 뛰어난 시스템이 사람의 의도대로 움직이는지 확인할 방법이 생기기 전에 그 문턱을 넘으면, 되돌릴 수 없는 통제 상실로 이어질 수 있다고 보기 때문입니다. 기술을 포기하자는 것이 아니라, 안전을 확인할 시간을 벌자는 것이지요. 그 사이에 제도를 고치고, 이득이 소수에게 쏠리지 않게 할 시간도 함께 법니다.

일의 관계도 달라집니다. 사람은 처음에는 에이전트를 지휘하고 로봇 공장을 세우지만, 자동화가 진행될수록 임금 노동만으로 소득을 나누기 어려워집니다. 시나리오는 컴퓨트와 로봇 생산허가에서 나오는 수입을 시민배당으로 돌리는 방안을 내놓습니다. 로봇과 지능이 만든 생산성이 누구에게 돌아가느냐가 기술 성능만큼 무거운 문제가 되는 것이지요.[2]

프라이버시도 남습니다. 가정용 로봇의 두뇌가 데이터센터에 있다면, 집 안에서 보고 들은 정보가 어디로 가는지, 누가 감사하는지, 연결을 끊어도 작동하는지를 정해야 합니다. 더 안전하다는 이유로 모든 행동을 기록하면 통제는 쉬워지겠지만, 생활 전체가 거대한 관제 시스템 안으로 들어갈 수도 있지요.

2040년의 갈림길에서 물을 것은 넷입니다. 인간이 멈출 권한을 지키는가. 혜택을 어떻게 나누는가. 감시받지 않을 공간을 남기는가. 다음 세대에 어떤 결정권을 넘기는가.

AI 2040이 맞을지는 알 수 없습니다. 이렇게 구체적인 시나리오가 쓸모 있는 까닭은, 에이전트가 코드에서 공장으로, 데이터센터가 전력망과 광산으로 이어질 때 무엇이 병목이 되고 어떤 제도가 먼저 필요할지를 미리 드러내 주기 때문이지요. 회사를 이끄는 사람이라면 우리 사업의 병목이 '생각'에서 전력·부지·설비·허가 같은 '몸'으로 옮겨 갈 때 무엇을 먼저 확보해 둘지, 정책을 만드는 사람이라면 프론티어 경쟁의 당사자가 아닌 한국이 이런 감속 합의의 검증 체계에 어떻게 참여할지를 따져 볼 수 있습니다. 미래의 속도는 모델만 정하지 않습니다. 인류가 세운 규칙도 그 속도의 일부입니다.





1. "AI 2040: Plan A" (AI Futures Project — Thomas Larsen, Romeo Dean, Brendan Halstead, Eli Lifland, Ryan Greenblatt, Daniel Kokotajlo, 2026.7.9, https://ai-2040.com/) — 예측이 아니라 정책 권고를 시험하는 조건부 시나리오라고 밝힌다. 2029년 검증된 감속 합의, 2030~2035년 인간 능력 범위의 확장, 2035년 정지, 2040년 초지능 확장이라는 시간표와 에이전트·로봇 확산 서사를 제시한다.

2. "Economics of Plan A" (AI Futures Project, 2026.7, https://ai-2040.com/supplements/economics-of-plan-a) — 인지 노동과 물리 노동의 자동화 시차, AI·로봇 생산의 약 6개월 배증 제한, 인간 감독, 시민배당을 포함한 조건부 경제 시나리오다.

3. Romeo Dean, "Compute in Plan A" (AI Futures Project, 2026.7, https://ai-2040.com/supplements/compute-supplement) — 2026~2040년 모델 크기, 활성 파라미터, 학습 연산량, 동시 실행 복제본과 세계 컴퓨트 생산량을 모델링한다. 저자들은 아키텍처와 하드웨어 변화 때문에 정직한 불확실성 범위가 여러 자릿수라고 명시한다.



제7부

지능 에너지 사회를 향하여

전기가 민주주의를 바꿨듯이

전기가 처음 들어온 곳은 공장이었습니다. 당연하지요. 돈을 벌어야 하니까요. 하지만 전기가 정말로 세상을 뒤흔든 건 공장 밖으로 나왔을 때였습니다.

1882년, 에디슨이 뉴욕 맨해튼에 최초의 상업용 발전소를 세웠습니다. 처음에는 월스트리트 일대 400개의 전등에 불을 켜는 것이 전부였지요. 그런데 전기가 거리를 밝히자, 예상치 못한 일이 벌어졌습니다.

사람들의 하루가 길어졌습니다.

해가 지면 끝나던 노동이 밤까지 이어졌고, 야간 학교가 생겨났습니다. 글을 읽지 못하던 노동자들이 밤에 글을 배우기 시작했지요. 라디오가 보급되자 시골 농부도 대통령의 연설을 들을 수 있게 되었습니다. 신문 인쇄가 빨라지면서 여론이라는 것이 형성되었고, 정치인은 처음으로 "국민이 무슨 생각을 하는지"를 신경 써야 했습니다.

전기는 단순한 기술이 아니었습니다. 공공 인프라였지요. 국가가 발전소를 짓고, 송전망을 깔고, 모든 가정에 전기를 공급하는 것이 정부의 의무가 되었습니다. 전기가 없는 삶은 상상할 수 없게 되었으니까요.

지능 에너지는 지금 그 길의 어디쯤 와 있을까요? 이 책의 앞부분에서 AI는 주로 기업의 도구였습니다. 제2부에서는 파운데이션 모델이라는 발전소와 GPU라는 변환 장치를, 제3부에서는 AX 전환의 단계를, 제6부에서는 에이전트의 세계를 이야기했지요.

하지만 전기가 공장에 머물지 않았듯, 지능 에너지도 기업의 울타리 안에 머물지 않을 겁니다.

이미 조짐은 곳곳에서 보입니다. 학생이 Claude에게 논문 구조를 물어봅니다. 자영업자가 GPT로 세금 신고서를 검토합니다. 시각장애인이 음성 AI를 통해 거리의 간판을 읽습니다. 지능 에너지는 벌써 공장 밖으로 나오고 있습니다.

문제는 이 에너지가 모든 사람에게 공평하게 가고 있느냐는 것입니다.

지능을 공공 인프라로

AI 기본사회라는 말을 들어보셨을까요? 아직 널리 쓰이는 용어는 아닙니다. 하지만 이 개념을 이해하면 앞으로 사회가 어디로 향할지 윤곽이 잡힙니다.

전기가 공공 인프라가 된 것처럼, AI도 공공 인프라가 되어야 한다는 생각입니다. 전기를 쓰려면 발전소와 송전망이 필요하듯, 지능을 쓰려면 AI 컴퓨팅 인프라와 파운데이션 모델이 필요합니다. 이것을 민간 기업에만 맡겨 둘 수 있을까요?

한국 정부는 독자 AI 파운데이션 모델 사업을 다섯 컨소시엄으로 시작했습니다. 2026년 8월 2차 평가에서 업스테이지·SK텔레콤·LG AI연구원 세 팀이 3차에 올랐고, 정부는 2027년 2월 최종 두 팀을 고를 계획이지요.[1] 유럽에서는 독일의 JUPITER가 유럽 최초의 엑사스케일 슈퍼컴퓨터로 가동에 들어갔고, 프랑스에는 두 번째 고성능 엑사스케일 시스템을 두는 계획이 추진 중입니다.[2] 인도도 자국 언어를 다루는 파운데이션 모델을 국가 AI 인프라 사업 안에서 공개하고 있지요.[3]

나라마다 방식은 달라도 하는 일은 같습니다. 지능 에너지를 누구나 쓸 수 있는 공공 인프라로 깔고 있는 것이지요.

전기가 민주주의를 바꾼 건 모든 사람이 쓸 수 있게 되었기 때문입니다. 부자만 전등을 켤 수 있었다면, 야간 학교도 라디오도 여론도 없었을 겁니다. 지능 에너지는 한 걸음 더 나아가, 쓰는 사람이 직접 만드는 사람이 되게 하고 있지요. 제3부에서 본 공무원들의 도구 수백 건이 그 증거입니다.

지능 에너지도 마찬가지입니다. 이 에너지가 소수의 것으로 남으면, 그것은 도구에 불과합니다. 모든 사람의 것이 되면, 사회를 바꿉니다.

그렇다면 우리는 지금 어디쯤 서 있을까요? 전기 역사로 치면 1920년대쯤에 해당할 겁니다. 도시에는 전기가 들어왔지만, 시골에는 아직 호롱불이 켜져 있던 시기. 공장은 전기를 쓰고 있었지만, 모든 가정에 전기가 들어오려면 아직 20년이 더 필요했던 시기.

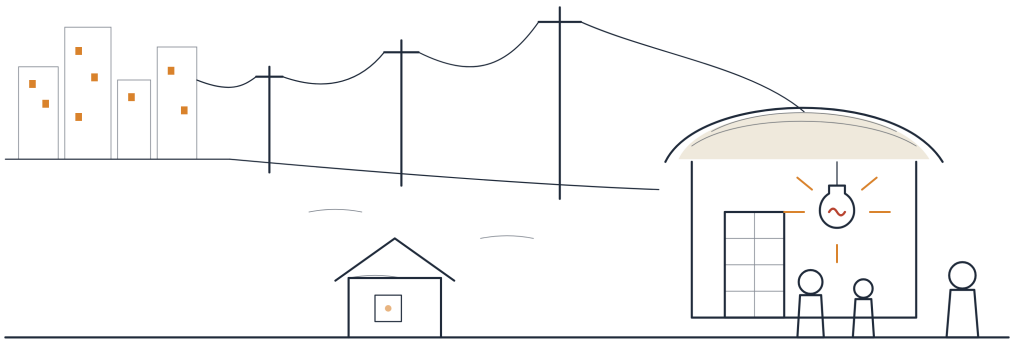


그림 7-1 도시의 창마다 불이 켜졌을 때, 들판 건너 마을에는 이제 막 첫 전등이 걸렸습니다. 그 옆집에는 아직 호롱불이 켜져 있었지요.

AI도 그렇습니다. 기업들은 AX를 추진하고, 개발자들은 에이전트를 만들고 있지만, 우리 사회 전체가 지능 에너지를 기본값으로 쓰려면 아직 갈 길이 멀지요.

이 제7부에서는 그 길을 이야기합니다. 지능 에너지가 사회를 어떻게 재편할 것인지, 어떤 격차를 만들어낼 수 있는지, 법과 노동과 안전은 어떻게 변해야 하는지, 그리고 각 나라는 어떤 선택을 하고 있는지를.

전기가 민주주의를 바꿨듯이, 지능 에너지도 우리 사회를 바꿀 겁니다. 어떻게 바꿀지는 이어지는 장들에서 하나씩 따져 보겠습니다. 먼저, 그 에너지가 아직 닿지 않은 곳부터 가 볼까요?

---

1. 「독자 AI 파운데이션모델 프로젝트 2차 단계평가 결과」(과학기술정보통신부, 2026.8.18, https://www.korea.kr/briefing/policyBriefingView.do?newsId=156774781) — 업스테이지·SK텔레콤·LG AI연구원 3차 진출; "'국가대표 AI' 투트랙으로…독파모 살리고 '프론티어 AI' 띄운다" (머니투데이방송, 2026.10.5, https://news.mtn.co.kr/news-detail/2026100320033298930) — 3차는 2027년 1월까지, 2월 단계평가로 최종 두 팀 선정 계획.

2. "Europe enters exascale supercomputing league with JUPITER" 및 "Europe's second high-end exascale supercomputer to be hosted in France" (European Commission, https://digital-strategy.ec.europa.eu/en/news/europe-enters-exascale-supercomputing-league-jupiter, https://digital-strategy.ec.europa.eu/en/news/europes-second-high-end-exascale-supercomputer-be-hosted-france) — 독일 시스템의 가동과 프랑스 시스템의 계획 단계를 구분한다.

3. IndiaAI 모델 포털 (Government of India, https://aikosh.indiaai.gov.in/home/models/details/param2_17b_a2_4b_thinking.html) — 인도 언어 특화 모델 공개 사례.

지능 에너지 격차: 새로운 디바이드

UNDP는 AI가 기존의 개발 격차를 줄일 수도, 반대로 새로운 대분기를 만들 수도 있다고 경고합니다. 같은 기술을 접하더라도 전력·통신·컴퓨팅·인력이라는 출발 조건이 다르기 때문입니다.[1]

가장 극적인 사례는 아프리카입니다. AI를 학습시키려면 데이터와 컴퓨팅 설비가 필요하고, 데이터센터를 운영하려면 안정적인 전력이 필요합니다. 이 기반이 부족한 곳에서는 모델을 내려받을 수 있어도 지속해서 운영하기 어렵지요.

UNCTAD는 개발도상국 가운데 국가 AI 전략을 가진 나라가 3분의 1도 되지 않고, 118개국이 주요 국제 AI 거버넌스 논의에 참여하지 못하고 있다고 지적합니다.[2] AI가 세상을 재편하는 규칙을 정하는 테이블에 많은 나라가 앉아 있지 못한 셈입니다.

전기 격차의 데자뷔와 공개 모델

이건 낯선 이야기가 아닙니다. 전기가 그랬으니까요.

1930년대, 미국 도시의 전력 보급률은 90%가 넘었지만 농촌은 10%에 불과했습니다. 루스벨트 대통령이 농촌전력청(REA)을 만들어 농촌 협동조합에 싼 이자로 돈을 빌려주고 전선을 깔게 하지 않았다면, 미국 농촌의 전력 보급은 훨씬 더 오래 걸렸을 겁니다. 시장에 맡겨 두면 수익이 나지 않는 곳에는 전선이 가지 않았으니까요.

지능 에너지도 같은 문제에 직면해 있습니다. AI 인프라 투자는 수익이 나고 전력·통신망이 갖춰진 지역에 집중되기 쉽습니다. 시장 논리만으로는 격차가 좁혀지지 않습니다.

Brookings 연구소와 UNDP는 이 상황을 "새로운 대분기(Next Great Divergence)"라고 부릅니다. 산업혁명 시기에 유럽과 아시아 사이에 벌어진 경제적 대분기가, 이번에는 AI 역량을 가진 나라와 그렇지 못한 나라 사이에서 재현될 수 있다는 경고입니다.

다행히 이번에는 전기 시대에 없었던 무기가 하나 있습니다. 누구나 내려받아 쓰는 공개 모델입니다.

2025년 1월, 중국의 DeepSeek은 R1의 코드와 모델 가중치를 MIT 라이선스로 공개했습니다. 상업적 이용과 수정이 가능한 오픈웨이트라는 점은 접근성을 크게 넓혔습니다. 다만 학습 데이터와 개발 과정 전체가 공개된 것은 아닙니다.[3] 함께 따라다니는 '600만 달러'가 무엇을 뺀 숫자인지는 제1부 「지능의 단가가 빠르게 낮아지고 있다」에서 짚었지요.[4]

Meta의 Llama, Mistral, Alibaba의 Qwen 계열도 가중치를 공개한 모델을 내놓았습니다. 라이선스와 공개 범위는 모델마다 다르고, 성능도 언어·과업·하드웨어 조건에 따라 달라집니다. 공개 모델이라는 이유만으로 상용 서비스와 동등하다고 볼 수는 없습니다.

공개 가중치 모델의 진짜 의미는 선택권입니다. 자체 환경에서 실행하고 수정할 수 있어 데이터 반출을 줄일 수 있지요. 그러나 GPU 구매·전력·운영 인력과 보안 비용은 남으므로, 누구에게나 자동으로 저렴한 선택은 아닙니다.

California Management Review는 오픈소스 AI(이 책의 표현으로는 공개 가중치 모델)가 "전형적인 파괴적 혁신의 경로"를 따르고 있다고 분석합니다. 처음에는 성능이 떨어지지만, 점점 올라와서 결국 기존 강자를 위협하는 그 패턴 말입니다.

아프리카에서는 DeepSeek이 Huawei와의 파트너십을 통해 빠르게 확산되고 있습니다. 공개 모델은 연구실 밖으로 나와, 폐쇄형 모델을 대신할 수 있는 실제 선택지가 되어 가고 있지요.

혜택과 비용은 누구에게 가는가

하지만 공개 모델만으로 충분할까요? 솔직히, 그렇지 않습니다.

모델 가중치를 무료로 받아도 GPU는 무료가 아닙니다. 전력도, 냉각수도, 송전망도 무료가 아닙니다. 모델을 운영할 엔지니어도 필요합니다. 공개 모델은 격차를 좁히는 도구이지만, 그것만으로 디바이드가 사라지지는 않습니다.

게다가 혜택이 가는 곳과 비용을 치르는 곳이 같지도 않습니다. 데이터센터의 혜택은 서비스를 쓰는 기업과 여러 지역의 사용자에게 퍼지지만, 물리적 부담은 시설이 들어선 지역에 집중될 수 있습니다. IEA는 데이터센터가 세계 전력에서 차지하는 비중보다 특정 지역 전력망에서의 충격이 훨씬 크다고 지적합니다.[5] 냉각 방식에 따라 물 사용량도 달라지고, 송전망 증설 비용과 발전 설비의 오염 부담을 누가 낼지도 남습니다.

칩을 만드는 단계도 빠뜨릴 수 없습니다. 반도체 제조에는 전기와 초순수, 화학물질이 필요하고, 장비 교체가 빨라지면 전자폐기물도 늘어납니다. 그래서 AI 접근성을 논할 때는 "누가 지능의 혜택을 얻는가"와 함께 "누가 물·전기·토지·폐기의 비용을 부담하는가"를 물어야 합니다. 데이터센터 유치 지역에 사용량을 공개하고, 요금과 환경 부담을 공정하게 배분하며, 주민이 입지 결정에 참여할 통로가 필요합니다.

AI의 숨은 비용

혜택은 넓게 퍼지고 물리적 부담은 한곳에 모인다 — 어디서 비용이 생기고, 누가 부담하며, 무엇을 재고 공개해야 하나

비용 발생 지점	부담 주체	측정·공개할 지표
데이터센터 운영	→ 시설 입지 지역	→ 전력 사용량
냉각	→ 물·전력 이용자	→ 물 사용량·냉각 방식
송전망·발전 설비	→ 지역 주민·요금 납부자	→ 망 증설 비용·오염 부담
반도체 제조	→ 생산 지역과 노동자	→ 전기·초순수·화학물질
장비 교체·폐기	→ 미래의 폐기 처리 주체	→ 전자폐기물과 처리 방식

그래서 필요한 세 가지

공개
유치 지역에 전력·물 사용량을 알린다

→

배분
요금과 환경 부담을 공정하게 나눈다

→

참여
주민이 입지 결정에 참여할 통로를 연다

출처: 저자 작성 — 본문 제7부 「지능의 격차」 중 '혜택과 물리적 비용은 같은 곳에 가지 않는다'의 물리적 비용과 부담 주체 논의 기반.

그림 7-2 비용이 생기는 지점에서 출발해 그 부담을 지는 사람, 재고 공개해야 할 지표로 한 줄씩 따라가 보세요. 아래쪽에는 그래서 필요한 세 가지, 공개와 배분과 참여가 놓여 있습니다.

대응책은 서로 연결되어 있습니다. 공동 컴퓨팅과 전력망에 투자하더라도 운영 인력과 현지 언어 데이터가 없으면 쓰기 어렵습니다. 인재를 길러도 조달권과 의사결정권이 외부에만 있으면 역량이 남지 않지요. 국제 지원은 장비 수량뿐 아니라 현지 소유권, 유지보수 인력, 거버넌스 참여를 함께 평가해야 합니다.

전기 접근성도 지역과 소득에 따라 여전히 차이가 남아 있습니다. 지능 에너지 격차 역시 저절로 사라진다고 장담할 수 없습니다. 어느 지역의 어떤 격차가 줄었는지를 같은 지표로 추적하고, 혜택과 환경 부담이 누구에게 돌아갔는지 공개해야 합니다.

UNCTAD는 세계 AI 시장이 2033년 4.8조 달러 규모에 이를 수 있다고 전망합니다.[6] 문제는 성장의 과실이 어디로 가느냐입니다. 소수에게 집중되면 격차가 벌어지고, 널리 퍼지면 더 많은 나라와 사람이 성장에 참여할 수 있습니다.

모델에 접속할 수 있느냐는 시작일 뿐입니다. 혜택과 비용을 누가 어떻게 나눠 갖느냐에서 격차는 더 벌어지지요.

1. "The Next Great Divergence" (UNDP Asia and the Pacific, 2025, https://www.undp.org/asia-pacific/publications/next-great-divergence) — AI 시대의 국가 간 출발 조건과 역량 격차를 분석한다.

2. "AI's $4.8 trillion future" (UN Trade and Development, 2025, https://unctad.org/press-material/ais-48-trillion-future-un-trade-and-development-alerts-divides-urges-action) — 국가 AI 전략과 글로벌 거버넌스 참여 격차를 제시한다.

3. DeepSeek-R1 README, License (DeepSeek-AI, https://github.com/deepseek-ai/DeepSeek-R1) — 코드 저장소와 모델 가중치에 MIT 라이선스를 적용한다.

4. DeepSeek-V3 Technical Report (DeepSeek-AI, arXiv:2412.19437, https://arxiv.org/abs/2412.19437) — 최종 학습 278만 8,000 H800 GPU시간에 시간당 2달러를 적용해 557만 6,000달러로 추산한다.

5. Energy and AI, Executive Summary (IEA, 2025, https://www.iea.org/reports/energy-and-ai/executive-summary) — 데이터센터는 세계 전력 소비의 약 1.5%지만 용량이 지역에 집중돼 지역 전력시장의 영향이 더 클 수 있다고 설명한다.

6. "AI market projected to hit $4.8 trillion by 2033" (UN Trade and Development, https://unctad.org/news/ai-market-projected-hit-48-trillion-2033-emerging-dominant-frontier-technology) — 4.8조 달러를 현재 가치가 아닌 2033년 시장 전망으로 제시한다.

법, 노동, 안전: 세 개의 매듭

새로운 에너지가 등장하면, 사회에는 반드시 세 개의 매듭이 생깁니다. 법은 어떻게 할 것인가. 일자리는 어떻게 되는가. 안전은 누가 책임지는가.

전기가 그랬습니다. 감전 사고가 잇따르자 전기 안전 규정이 만들어졌고, 공장 자동화로 수공업자들이 일자리를 잃자 노동법이 정비되었으며, 발전소 사고의 책임을 누가 질 것인가를 두고 법정 다툼이 벌어졌지요.

지능 에너지 시대에도 같은 매듭이 풀리기를 기다리고 있습니다.

법은 구호보다 집행 경로로

2026년의 AI 규제는 EU, 한국, 미국이라는 세 이름만으로 나누기 어렵습니다. 같은 지역 안에서도 연방·주, 범용법·부문법, 중앙 감독기관·현장 집행기관의 역할이 다르기 때문입니다.

유럽은 규칙을 먼저 세웠습니다. EU AI Act은 세계 최초의 포괄적 AI 규제법입니다. 2024년 8월에 발효되어 단계적으로 시행 범위를 넓혀왔지요. 2025년 2월에는 허용 불가 위험 AI의 금지와 AI 리터러시 의무가 시작되었고, 2025년 8월에는 범용 AI 모델 관련 의무가 적용되기 시작했습니다.[1]

적용 일정은 한 날짜의 '전면 시행'으로 요약하기 어렵습니다. 일부 투명성 의무는 2026년 8월부터 적용되기 시작했고, 고위험 시스템 의무에는 별도 일정과 경과조치가 있습니다. 유럽에서 AI 서비스를 제공하는 기업은 자신이 공급자·배포자·수입자 중 어디에 해당하고 어떤 유형의 시스템을 다루는지부터 확인해야 합니다.[1]

EU의 특징은 위험 등급에 따라 사전 의무를 나누고 역외 사업자에게도 적용할 수 있다는 점입니다. 실제 효과는 표준, 감독 역량, 중소 사업자의 준수 비용과 피해 구제가 어떻게 작동하는지에 달려 있습니다.

한국도 포괄법을 마련했습니다. 정식 명칭은 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」이며, 2026년 1월 22일 시행에 들어갔습니다. 7월 21일에는 공공조달과 취약계층 지원을 보강한 개정 법령도 시행됐지요.[2]

두 법의 순서를 단순한 '세계 최초' 경쟁으로 읽기는 어렵습니다. EU와 한국은 법의 적용 대상, 의무의 발효 시점, 집행 체계가 서로 다릅니다. 눈여겨볼 것은 순위보다 고영향 AI의 위험을 실제로 식별하고 피해 구제까지 연결하는지입니다. EU의 고위험 의무가 언제부터 적용되는지는 「일곱 개 국가의 선택」에서 날짜까지 짚어 보겠습니다.

한국 AI 기본법은 투명성·안전성 요건, 고영향 AI 사업자의 의무, 영향평가, 국내 대리인 지정 등을 두고 있습니다. 다만 법률에 의무가 적혀 있다는 사실과 시민이 설명·정정·재심을 실제로 받을 수 있다는 것은 다른 문제입니다. 시행령, 감독 인력, 신고 창구와 구제 절차를 함께 봐야 합니다.

미국은 방향을 틀었습니다. 바이든 행정부 시절에는 AI 감독과 위험 완화를 강조했지만, 트럼프 행정부는 규제 장벽 완화에 무게를 옮겼습니다. 2025년 1월에는 이전 AI 행정명령을 폐기하고 미국의 AI 리더십을 가로막는 장벽을 제거한다는 새 행정명령을 발표했지요.[3]

2025년 12월에는 연방 차원의 최소 부담 정책을 내세우며, 연방 정책과 충돌한다고 판단한 주(州) AI 법률을 검토하고 다툴 법무부 태스크포스를 두는 행정명령을 내놓았습니다.[4]

연방 행정부는 규제 부담 완화에 무게를 옮겼지만, 미국 전체가 무규제인 것은 아닙니다. 주법, 개인정보·고용·소비자보호 같은 기존 법률, 연방 조달 기준과 법원의 판단이 동시에 작동합니다.

비교할 때는 '혁신 대 규제'라는 구호보다 여섯 가지를 물어야 합니다. 누구에게 적용되는가, 어떤 위험을 다루는가, 누가 감독하는가, 위반을 어떻게 발견하는가, 어떤 제재가 있는가, 피해자가 어디에서 구제받는가. 기업과 시민에게 필요한 정보는 국가의 성격표가 아니라 이 집행 경로입니다.

사라지는 과업과 생겨나는 역할

AI와 일자리. 이보다 더 많이 논의된 주제도 드물 겁니다.

'AI가 일자리를 없앨까'라는 질문만으로는 변화의 분포를 알기 어렵습니다. 어떤 과업과 입문 자리가 줄고, 누가 새 역할로 이동하며, 임금과 검토 부담이 어떻게 달라지는가를 함께 물어야 합니다.

개발 현장에서는 AI 코딩 도구 사용이 넓어졌지만, 도입률은 조사 대상과 '사용'의 정의에 따라 크게 달라집니다. 코드 초안 작성 시간이 줄어도 검토·통합·보안 대응 시간이 늘 수 있고, 숙련도와 조직에 따라 성과도 다릅니다. 따라서 도구 사용률만으로 개발자의 일이 어떻게 바뀌었다고 결론내리기는 어렵습니다.

이 패턴은 전기 시대에도 반복되었습니다. 전기 모터가 공장에 들어왔을 때, 기계를 직접 돌리던 노동자는 줄었지만 기계를 관리하고 생산 공정을 설계하는 사람은 늘었습니다. 에너지가 바뀌면 일의 내용이 바뀝니다.

문제는 전환 속도를 무엇으로 재느냐입니다. METR의 연구는 통제된 환경에서 일부 소프트웨어 과업을 일정 성공률로 끝내는 데 사람이 걸리는 시간의 증가 추세를 측정합니다. 이 '시간 지평'은 조직의 하루 업무나 한 달짜리 프로젝트를 그대로 뜻하지 않습니다. 협업, 책임, 암묵지와 현장 예외가 포함되면 결과가 달라질 수 있지요.

노동 정책은 모델 성능 곡선보다 과업 노출, 채용, 임금, 검토 부담과 직무 이동을 추적해야 합니다. 재교육도 수료 인원보다 새 과업을 맡았는지, 고용과 임금이 유지됐는지를 평가해야 하지요.

특히 자동화가 입문 업무부터 줄이면 경력의 사다리가 끊길 수 있습니다. 주니어 개발자가 작은 버그를 고치며 시스템을 배우고, 신입 상담원이 쉬운 문의를 처리하며 예외를 익히던 과정 말입니다. 현재 인력의 생산성만 높이고 견습 자리를 없애면 몇 년 뒤에는 AI의 결과를 검토할 숙련자도 부족해집니다.

기업은 실제 업무 중 일부를 감독 가능한 견습 과업으로 보존하고, 장애·민원·오판 사례를 재현한 시뮬레이션 훈련을 운영할 수 있습니다. 고위험 분야에서는 도구 사용법보다 결과의 근거와 한계를 검토하는 역량을 별도로 평가해야 합니다. 교육기관과 정부는 직무 전환 횟수보다 훈련 뒤 맡을 수 있게 된 과업, 취업 유지, 임금 회복을 추적해야 하지요. 재교육을 수료증이 아니라 새 역할로 이어지는 경로로 설계해야 합니다.

이 전환의 논리적 끝단까지 미리 가 본 사고 실험도 있지요. 제6부 마지막 장에서 만난 AI Futures Project의 시나리오 「AI 2040」은 2030년대 중반 자동화가 사실상 완성된 세계를 가정하고, 그때의 분배 장치로 컴퓨트와 로봇의 생산허가권을 경매에 부쳐 그 수익의 4분의 3을 시민배당으로 돌리는 그림을 그립니다[5]. 예측이 아니라 권장안을 검증하려는 논리 실험이지만, 던지는 질문만은 지금의 것입니다. 일이 에너지처럼 값싸진 세계에서, 소득은 어디에서 오는가. 노동 매듭의 끝에는 결국 분배 매듭이 묶여 있습니다.

확률적 시스템을 믿을 수 있을까

법도 만들 수 있고, 노동 전환도 준비할 수 있습니다. 하지만 안전은 좀 더 까다로운 문제입니다.

전통적인 소프트웨어의 많은 핵심 로직은 결정적(deterministic)으로 설계됩니다. 물론 동시성, 외부 서비스, 센서 입력 때문에 결과가 달라지는 시스템도 있습니다. 생성형 AI는 여기에 확률적 출력과 데이터 분포 변화라는 검증 난제를 더합니다.

생성형 AI는 확률적(probabilistic) 출력을 만들 수 있습니다. 같은 질문에도 답이 달라질 수 있고, 그중 사실이 아닌 내용을 사실처럼 생성하는 현상을 할루시네이션이라고 부릅니다. 하지만 AI 시스템의 모든 오류가 할루시네이션인 것은 아닙니다.

법률 AI가 존재하지 않는 판례를 만들어내는 것은 할루시네이션입니다. 의료 AI의 오진은 데이터 편향이나 분포 이동, 임계값 설정 등에서 생길 수 있고, 자율주행차의 장애물 인식 실패는 센서·인식 모델·기능안전의 문제일 수 있습니다. 프롬프트 인젝션과 권한 탈취는 다시 보안 공격의 영역입니다. 원인이 다르면 시험 방법과 책임 구조도 달라집니다.

여기에 이중 딜레마가 있습니다.

규제를 너무 엄격하게 하면 혁신이 막힙니다. 0.1%의 오류를 없애려고 시스템 전체를 멈추면, 99.9%의 편익도 함께 사라지지요. 물론 그 0.1%가 되돌릴 수 없는 사고라면 계산이 달라집니다. 오진 하나가 한 사람의 목숨이고, 유출 한 번이 수백만 명의 개인정보라면 99.9%의 편익과 나란히 저울에 올릴 수 없지요. 그래서 규제는 오류의 빈도보다 무게를 봅니다. 반대로 규제를 너무 느슨하게 하면 사고가 나고, 다친 사람이 생기며, 결국 기술 자체가 불신을 받게 됩니다.

EU AI Act이 "위험 기반 접근법"을 택한 것은 이 딜레마에 대한 하나의 답입니다. 모든 AI를 같은 잣대로 규제하지 않고, 위험 수준에 따라 차등 규제한다는 것이지요. 허용 불가 위험, 고위험, 제한 위험, 최소 위험의 네 단계로 나누어 각각 다른 수준의 의무를 부과합니다.

한국 AI 기본법도 비슷한 철학을 따릅니다. "고영향 AI"라는 범주를 두어 사회적 영향이 큰 AI에 강화된 의무를 지우고, 생성형 AI 결과물 표시나 학습 연산량이 큰 모델의 안전성 확보처럼 범위를 좁힌 의무를 따로 둡니다. 그 밖의 AI는 비교적 가볍게 규율하는 방식이지요.

법이 의무를 적는다면, 그 의무를 시험할 사람도 있어야 합니다. 한국에서는 2024년 11월 AI안전연구소가 문을 열었습니다. AI 모델과 시스템에서 생길 수 있는 위험을 평가하고, 그 위험을 막고 줄이는 기술을 연구하며, 다른 나라의 AI 안전연구소들과 협력하는 일을 맡은 조직이지요.[6] 연구소가 생긴 것과 평가가 충분히 이루어지는 것은 다른 일입니다. 어떤 모델을 얼마나 깊이 시험했는지, 그 결과가 공개되는지를 지켜봐야 이 장치가 실제로 작동하는지 알 수 있습니다.

할루시네이션이 '지금 여기'의 안전 문제라면, 더 멀리 있는 안전 문제도 이 매듭에 함께 묶여 있습니다. 제2부 「스스로를 개선하는 지능: 재귀적 자기개선(RSI)」과 「초지능(ASI)으로 가는 네 가지 길」에서 던졌던 물음, 즉 스스로 성능을 끌어올리는 지능을 사람이 어떻게 통제하고 검증할 것인가 하는 문제가 바로 이 '안전 매듭'의 가장 안쪽에 놓여 있습니다. 확률적 시스템의 오차를 다루는 일과 스스로 진화하는 시스템을 통제하는 일은 난이도가 다르지만, 둘 다 "이 지능을 믿을 수 있는가"라는 같은 질문에서 갈라져 나온 가지입니다. 다만 두 가지는 서로 다른 규율 도구를 부릅니다. 할루시네이션은 사후 정정과 책임 소재를 가리는 방식으로 다스릴 수 있지만, 스스로 진화하는 시스템은 능력이 위험한 문턱을 넘기 전에 미리 평가하고 신고하게 하는 사전(事前)의 장치를 요구합니다. 같은 서랍에 넣어 한 가지 규제로 묶으려 하면 둘 다 놓치기 쉽지요.

이 접근이 충분한지는 아직 아무도 모릅니다. 확률적 시스템의 안전을 보장하는 완벽한 방법은 아직 발명되지 않았으니까요. 그래도 이것만은 말할 수 있습니다. 안전을 미루면 그 대가는 먼저 사람이 치르고, 그다음에 산업이 치릅니다. 사고가 쌓이면 신뢰가 무너지고, 무너진 신뢰를 되찾는 데는 기술을 만드는 것보다 훨씬 오랜 시간이 걸리지요.

이렇게 하나씩 풀어 놓고 보면 한 가지가 더 눈에 들어옵니다. 법, 노동, 안전. 이 세 매듭은 따로 놓여 있는 것처럼 보이지만, 실은 하나로 연결되어 있습니다.

법이 명확해야 기업이 투자하고, 기업이 투자해야 새로운 일자리가 만들어지며, 위험이 관리되고 있다는 근거가 보여야 시민이 AI를 받아들입니다. 어느 하나가 풀리지 않으면 나머지 둘도 막히지요.

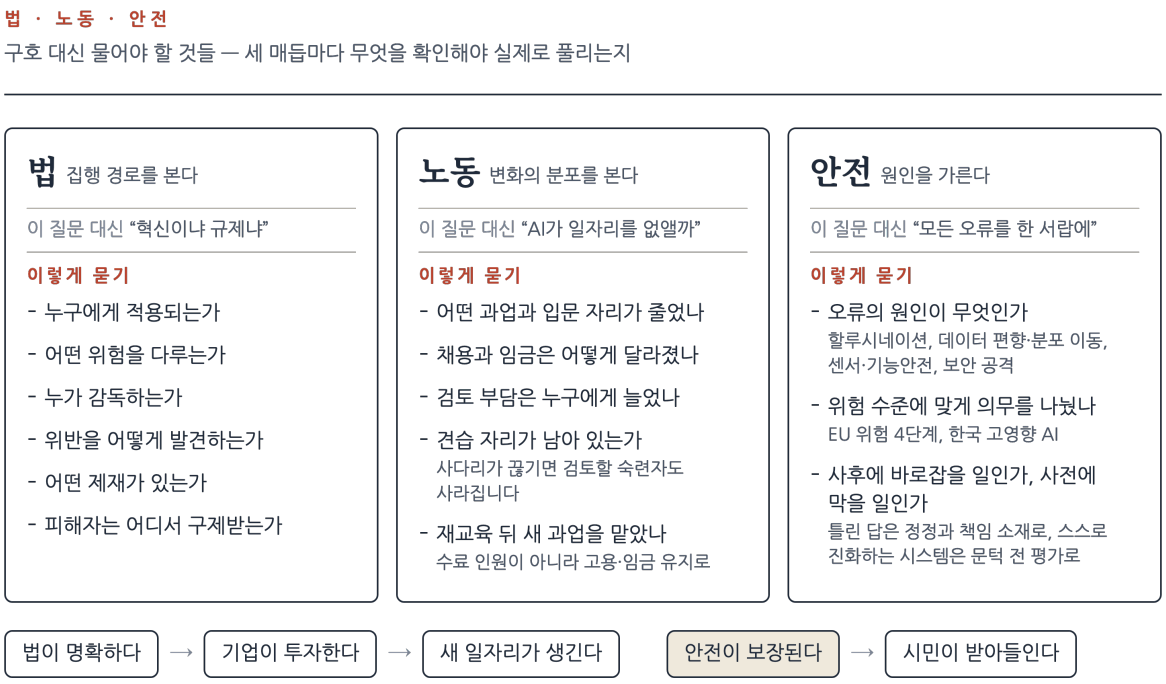


그림 7-3 세 매듭은 구호로는 풀리지 않습니다. 법은 집행 경로를, 노동은 변화의 분포를, 안전은 오류의 원인을 물어야 실마리가 보입니다.

전기 시대에도 이 세 매듭이 풀리는 데는 수십 년이 걸렸습니다. 이번에는 그만큼 기다려 줄 시간이 있을까요? 다음 장에서는 이 매듭들이 모두 풀렸다고 가정한 어느 하루로 미리 가 보겠습니다.

1. "AI Act: regulatory framework" (European Commission, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai) — 발효일과 의무별 단계 적용 일정을 정리한다.

2. 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(국가법령정보센터, https://www.law.go.kr/법령/인공지능발전과신뢰기반조성등에관한기본법) — 시행일과 적용 대상·의무의 법적 근거; 「달라진 AI기본법, 우리에게 어떤 변화가 생길까?」(대한민국 정책브리핑, 2026.7.15, https://www.korea.kr/news/policyNewsView.do?newsId=148968171) — 2026.7.21 시행 개정 법령: 공공조달 AI 제품 우선 고려, 취약계층 AI 이용비 지원 근거.

3. "Removing Barriers to American Leadership in Artificial Intelligence" (The White House, 2025.1.23, https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/).

4. Executive Order 14365 (The White House, 2025.12, https://www.whitehouse.gov/wp-content/uploads/2025/12/eo-14365.pdf) — 연방 AI 정책과 주법 검토를 위한 행정부 조치.

5. "Economics of Plan A" (AI Futures Project, 2026.7, https://ai-2040.com/supplements/economics-of-plan-a) — Plan A 시나리오 기준 2030년대 중반 자동화 약 99% 가정, 컴퓨트·로봇 생산허가권 경매 수익의 약 75%를 시민배당으로 배분하는 구상. 예측이 아닌 정책 사고 실험임.

6. 「연구원 소개」(AI안전연구소, https://www.aisi.re.kr/kor/contents/17) — 2024년 11월 설립, AI 모델·시스템의 위험 평가와 위험 방지·해소 기술 연구, 해외 AI 안전연구소와의 협력을 소개한다.

2030년, 지능이 전기처럼 흐르는 하루

2030년에 가능할 법한 어느 화요일을 상상해봅시다. 아래 장면은 예측이 아니라, 2026년에 보인 기술의 싹이 비용·제도·신뢰의 문턱을 넘었다고 가정한 하나의 시나리오입니다.

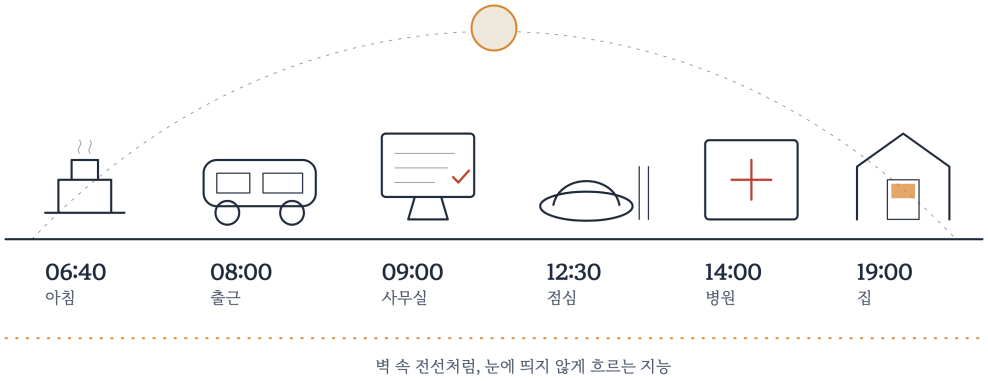


그림 7-4 아침 6시 40분부터 저녁 7시까지. 지능은 벽 속 전선처럼 눈에 띄지 않게 하루를 따라 흐릅니다.

아침 6시 40분

알람이 울립니다. 정확히는, 알람이 울리지 않습니다.

개인 에이전트가 수면 데이터를 보고 6시 40분이 최적의 기상 시간이라고 판단했습니다. 렘수면 주기가 끝나는 타이밍이지요. 에이전트는 당신의 USER.md를 알고 있습니다. 아침에 약하다는 것, 커피 없이는 대화가 안 된다는 것, 월요일 밤에 잠을 잘 못 잔다는 것까지.

눈을 뜨면 침대 옆 화면에 짧은 브리핑이 떠 있습니다.

"밤사이 이메일 23통. 처리 완료 20통. 김 부장님 회의 요청은 수요일로 옮겼습니다 — 오늘 오후 보고서 마감과 겹쳐서요. 박 이사님 메일은 직접 확인이 필요합니다."

에이전트가 밤새 이메일을 분류하고 회신 초안을 만들었습니다. 회의 초대 하나는 일정 충돌을 잘못 판단해 보류함에 들어갔고, 당신은 이유를 확인한 뒤 직접 되돌립니다. 외부 발송과 일정 변경에는 승인 기록이 남습니다. 에이전트가 당신의 말투를 흉내 내더라도, 상대에게는 AI가 작성한 초안임이 표시됩니다.

부엌으로 가면 구독형 가정용 로봇이 커피를 준비하고 있습니다. 아직 냉장고처럼 누구나 사는 물건은 아닙니다. 기기 값과 유지비가 만만치 않아 이 집은 직접 구매 대신 월 사용료를 내고 있지요. 로봇은 수면 데이터를 참고했지만, 건강 데이터 접근 권한은 매달 다시 확인받습니다.

커피를 마시면서 뉴스 브리핑을 듣습니다. 제6부에서 본 개인 에이전트가 밤사이 소식을 추려 오디오로 넘겨주는 장면이지요. 반도체 업계에서 일하니까 NVIDIA 실적 발표 요약이 들어가 있고, 주말에 제주도 여행을 계획 중이니까 날씨 전망도 들어 있습니다. 다만 추천 순위에는 사용자의 관심뿐 아니라 공급자의 제휴와 수익모델이 영향을 줄 수 있습니다. 화면에는 출처, 광고 여부, 추천 이유와 다른 관점 보기 버튼이 함께 떠 있습니다.

전기를 의식하지 않고 형광등을 켜듯, 당신은 지능을 의식하지 않고 아침을 시작합니다.

오전 8시, 출근길에서 사무실까지

운행 허가가 난 구역에서는 자율주행 모드가 작동합니다. 공사 구간에 들어서자 차가 운전권을 넘겨받으라고 알립니다. 2030년에도 지역과 날씨, 보험 조건에 따라 사람이 맡아야 할 구간은 남아 있습니다.

차 안에서 업무 에이전트와 대화합니다.

"오늘 회의 준비해줘."

에이전트가 관련 문서를 끌어오고, 지난주 진행 상황을 요약하고, 밤새 모니터링 시스템에서 감지된 이슈 2건을 알려줍니다. "3번 라인 컨베이어 모터의 진동 패턴이 정상 범위를 벗어나고 있습니다. 디지털 트윈 분석 결과 3일 내 교체가 필요할 것으로 보입니다." 교체 부품은 이미 주문되어 있습니다. 에이전트가 새벽에 처리했으니까요.

차의 AI와 업무 에이전트는 A2A 프로토콜로 연결되어 있습니다. 프로토콜은 대화 형식을 맞추지만 인증, 권한, 책임까지 자동으로 해결하지는 않습니다. 차가 "10분 뒤 도착 예정"이라고 알리면 업무 에이전트는 허용된 일정 정보만 받아 회의실을 확인합니다. 인증이 실패한 날에는 자료 전송을 멈추고 사람에게 알리며, 호출과 승인 기록은 로그에 남습니다.

9시, 사무실에 들어서면 바로 팀 미팅입니다. 회의실에 사람은 두 명뿐입니다. 당신과 팀장. 나머지 네 명의 동료는 다른 시간대에 있습니다. 서울, 샌프란시스코, 싱가포르, 런던. 하지만 회의에는 일곱 명이 참석합니다. 부재중인 동료 네 명을 대신해 그들의 에이전트가 참석하고 있으니까요.

이상하게 들리시나요? 하지만 생각해보면 이미 2026년에 비슷한 일이 일어나고 있었습니다. 이메일 자동 회신, 슬랙 봇, AI 회의 요약. 그것이 조금 더 정교해진 것뿐입니다. 에이전트들에게는 SOUL.md가 있습니다. 동료의 전문 분야, 의사 결정 스타일, 소통 방식을 알고 있지요. 단순한 의견 전달이 아니라, "이 프로젝트에서 김 대리가 중요하게 생각하는 점"을 반영해 발언합니다.

물론 에이전트가 모든 결정을 내리는 것은 아닙니다. 예산 승인이나 전략적 방향 전환 같은 것은 사람이 직접 해야 하지요. 에이전트에게는 명확한 권한 범위가 설정되어 있습니다. 이것을 Authority Envelope이라고 합니다. "일상적인 진행 보고는 할 수 있지만, 일정 변경은 확인을 받아야 한다" 같은 규칙이지요.

회의가 끝나면 본격적인 업무가 시작됩니다. 당신의 팀은 사람 4명, AI 에이전트 12개로 구성되어 있습니다. 에이전트들은 데이터 분석, 초안 작성, 코드 생성, 고객 문의 분류를 맡습니다. 사람은 목표와 권한을 정하고, 표본을 검토하며, 결과에 책임집니다. "이 데이터가 이상한 이유는 지난달 공급망 변경 때문 아닐까?" 같은 가설도 사람이 독점하는 능력이 아니라 현장 경험과 기록을 함께 확인해야 할 판단입니다.

관리하고 있는 제조 공장의 디지털 트윈이 화면에 떠 있습니다. 센서가 연결된 설비의 상태를 가상 모델과 비교하는 시스템이지요. 아까 컨베이어 경고는 진동 변화를 감지해 부품 점검을 권한 것입니다. 정비 담당자는 원자료와 오탐 이력을 확인한 뒤 교체 여부를 결정합니다. 예측 정비가 고장을 줄일 수 있어도 센서 누락과 잘못된 경보까지 없애지는 못합니다.

낮 12시 30분, 점심에서 병원까지

점심시간입니다. 에이전트가 식당을 추천합니다. 사용자가 제휴 상품을 제외하도록 설정했고, 추천 이유와 광고 여부도 함께 표시됩니다. 오늘 아침을 거르셨고, 최근 건강 검진에서 나트륨 섭취를 줄이라는 소견이 있어 회사 근처의 저염식 메뉴를 제안한 겁니다. 건강정보 사용 권한은 언제든 철회할 수 있습니다.

결제는 에이전트 대 에이전트로 이루어집니다. 당신의 재무 에이전트가 식당의 에이전트와 대화해서 포인트 적립 여부를 확인하고, 최적의 결제 수단을 선택합니다. 이것도 A2A 프로토콜입니다. 에이전트끼리 서로의 Agent Card를 확인하고, 할 수 있는 것과 할 수 없는 것을 파악한 뒤, 가장 효율적인 방식으로 거래를 완료합니다. Microsoft Research가 예견했던 "에이전트의 웹"이 일상이 된 것이지요.

배달 로봇이 음식을 테이블까지 가져옵니다. 당신은 그것을 거의 의식하지 못합니다. 형광등 스위치를 켤 때 전기를 의식하지 못하듯이.

오후 2시에는 정기 검진이 잡혀 있습니다. 병원에 도착하면 이미 당신의 의료 기록이 준비되어 있습니다. AI 네이티브 전자건강기록(EHR) 시스템이지요. 웨어러블 기기의 데이터 — 심박, 수면 패턴, 활동량, 혈중 산소 — 가 지난 3개월 치로 정리되어 있습니다. AI가 이 데이터를 사전 분석해서 의사에게 두 가지 가능성을 제시합니다. "수면 패턴으로 보아 가벼운 수면무호흡증 가능성, 또는 단순 스트레스로 인한 수면 질 저하."

의사는 분석의 근거와 누락된 데이터를 확인한 뒤 당신과 대화합니다. "요즘 스트레스가 많으시다고요? 어떤 종류의 스트레스인가요?" 환자는 AI 분석을 쓰지 않는 진료를 요청하거나 기록의 오류를 정정할 수 있습니다. AI가 시간을 절약할 수도 있지만, 경고를 검토하고 설명하는 일이 새 부담이 될 수도 있지요.

돌아가는 길에 아이의 학교에서 메시지가 옵니다. 학교는 교사가 고른 단원에서만 AI 튜터링을 쓰고, 학생별 추천 문제와 근거를 교사가 확인합니다. 아이나 보호자는 학습 기록의 정정과 비AI 과제를 요청할 수 있습니다. 설명 방식이 맞지 않거나 접근성 문제가 생기면 교사가 다른 자료로 바꿉니다. AI가 반복 설명을 보조하더라도 수업 설계와 평가, 돌봄의 책임은 학교와 교사에게 남습니다.

저녁 7시, 집

집에 돌아오면 가정용 로봇이 청소를 마쳐놓았습니다. 22dB. 냉장고보다 조용하다고 했지요? 아이가 학교에서 돌아오고 당신이 퇴근할 때까지, 조용히 집안일을 끝내놓는 것입니다.

저녁 식사 후 소파에 앉아 에이전트에게 묻습니다. "요즘 읽을 만한 책 추천해줘." 에이전트는 3년 전에 당신이 읽었던 책들을 기억하고 있습니다. MemGPT 아키텍처 기반의 장기 기억이지요. 2023년 논문으로 나온 뒤 2024년 Letta라는 프레임워크로 이어진 기술입니다. 그래프 메모리가 당신의 독서 패턴을 알고 있습니다. "3년 전에 유발 하라리의 책을 좋아하셨고, 작년에는 과학 에세이를 많이 읽으셨으니, 이번에는 이런 책이 어떨까요?" 벡터 데이터베이스가 단순히 "이 사용자는 논픽션을 좋아한다"를 기억하는 것과는 다릅니다. 관계와 맥락을 기억하는 것이지요.

잠자리에 들기 전, 에이전트가 하루를 정리해줍니다.

"오늘 처리한 업무 47건. 에이전트가 자율 처리한 것 38건, 직접 판단하신 것 9건. 내일 주의 사항 3건: 3번 라인 부품 교체 일정 확인, 신규 프로젝트 킥오프 미팅, 분기 보고서 마감."

이 숫자는 장면을 손에 잡히게 하려고 넣은 장치일 뿐, 생산성 측정치는 아닙니다. 눈여겨볼 것은 건수보다 어떤 업무를 자동화했고, 어떤 결정은 사람이 붙잡았으며, 오류가 났을 때 되돌릴 수 있었는지입니다.

그래서, 누구의 하루인가

이 하루에서 눈에 띄는 것은 기술보다 선택권입니다. 2030년의 어떤 사람은 AI를 의식하지 않고 쓰겠지만, 어떤 사람은 비용 때문에 쓰지 못하고, 어떤 사람은 개인정보 때문에 일부 기능을 끌 것입니다. 지능이 일상에 스며드는 것과 모두에게 같은 조건으로 주어지는 것은 다른 문제입니다.

이 하루의 요소들은 2026년 기술에서 출발하지만, 2030년에 이 조합이 보편화된다는 보장은 없습니다. 개인 에이전트와 장기 기억, 가정용 로봇, 상호운용 프로토콜, 디지털 트윈이 각각 존재하는 것과 안전하고 저렴하게 연결되는 것 사이에는 아직 거리가 있지요. 비용과 규제, 호환성, 사고 책임 중 하나만 막혀도 이 하루는 상당히 다른 모습이 될 수 있습니다.

모델 호출 단가는 내려갈 수 있지만 서로 다른 품질의 모델을 토큰 가격만으로 비교하면 실제 부담을 알 수 없습니다. 이 시나리오가 성립하려면 성공 작업당 비용이 낮아져야 합니다. 재시도, 검색, 저장, 보안,

사람 검토와 사고 복구를 모두 포함한 비용이지요. 2030년의 비용은 모델 성능뿐 아니라 사용량, 전력, 규제와 책임 구조에 따라 달라지므로 거의 무료가 된다고 단정할 수 없습니다.

그렇다면 이 하루를 사는 사람은 누구일까요? 특별한 사람이 아닙니다. 기술 전문가도 아니고, 실리콘밸리의 엔지니어도 아닙니다. 그냥 지능 에너지 인프라가 잘 갖춰진 곳에서 사는 보통 사람입니다. 전기가 잘 들어오는 곳에서 사는 사람이 전기 전문가가 아니어도 냉장고를 쓸 수 있듯이.

이 시나리오에서 일부 사람들은 오케스트레이터 역할을 맡습니다. 일부 기업은 작은 팀과 여러 에이전트를 결합하지만, 생산성 향상이 곧 고용 안정이나 창의적 시간으로 이어지는지는 조직의 분배 방식에 달려 있습니다. 지능 인프라를 먼저 구축한 나라가 유리할 수 있어도, 그 혜택이 시민에게 자동으로 돌아가지는 않습니다.

그런데 이 하루에는 빠진 질문이 하나 있습니다.

이 하루를 모든 사람이 살 수 있느냐는 것입니다. 전기는 보급에만 수십 년이 걸렸고, 그러고도 2024년에 세계 인구의 약 8%, 6억 5,000만 명이 넘는 사람이 여전히 전기 없이 살았습니다.[1] 지능 에너지도 같은 문제에 직면해 있지요. 지능 인프라에 접근할 수 있는 사람과 그렇지 못한 사람 사이의 격차. 에이전트가 하루 47건의 업무를 처리해주는 사람과, 여전히 혼자서 모든 것을 하는 사람 사이의 격차. 이것이 지능 격차입니다.

이것은 4차 AI산업혁명이 전개될 수 있는 한 경로입니다. GPU와 파운데이션 모델, 프로토콜과 에이전트가 연결되더라도 결과는 비용과 권리, 제도와 조직의 선택에 따라 달라집니다.

그 하루가 누구에게 열릴지는, 맡길 수 있는 범위와 되돌릴 수 있는 길을 우리가 어떻게 설계하느냐에 달려 있습니다.



1. Tracking SDG7: The Energy Progress Report 2026 (IEA·IRENA·UNSD·World Bank·WHO, 2026, https://www.seforall.org/publications/tracking-sdg7) — 2024년 전기 미접근 인구 6억 5,500만 명, 세계 접근률 92%.

일곱 개 국가의 선택

같은 에너지를 두고도 나라마다 다른 선택을 합니다.

석유가 그랬습니다. 미국은 민간 석유회사가 주도했고, 사우디아라비아는 국영 석유회사를 세웠으며, 일본은 석유가 없어서 에너지 효율에 사활을 걸었지요. 같은 에너지인데 전략이 달랐고, 그 전략의 차이가 국가의 궤도를 결정했습니다.

지능 에너지 시대에도 같은 일이 벌어지고 있습니다. 일곱 나라의 선택을 보겠습니다.

미국과 중국, 서로 다른 두 엔진

미국에서는 민간 기업의 투자와 연방·주 정부의 정책이 함께 움직입니다. 규제 완화를 강조하는 흐름이 강하지만, 수출 통제·정부 조달·주별 규제처럼 정부 역할도 작지 않습니다.

미국에는 프런티어 모델 기업과 클라우드 사업자, 반도체 기업, 벤처자본이 한 생태계 안에 모여 있습니다. 민간 자본이 데이터센터와 모델 개발에 빠르게 들어갈 수 있는 조건이지요.

트럼프 행정부는 이 흐름을 가속합니다. 이전 AI 행정명령을 폐기하고 미국의 AI 리더십을 가로막는 장벽을 제거하겠다고 천명했습니다. 주 정부의 AI 규제도 연방 정책과 충돌하는지 검토하고 법적으로 다투는 체계를 만들었지요.[1]

주요 기술기업은 데이터센터와 서버에 대규모 자본을 투입하고 있습니다. 다만 기업별 회계연도와 리스 포함 범위가 다르고, 설비투자 전체가 AI에 쓰이는 것도 아니어서 하나의 합계로 단정하기는 어렵습니다.

미국의 강점은 자본과 인재와 생태계입니다. 약점은 연방 차원의 포괄적 AI 법이 없다는 데서 오는 리스크이지요. 시장이 스스로 조절할 것이라는 믿음이 맞을 수도 있고, 대형 사고가 터져야 뒤늦게 법을 만드는 상황이 올 수도 있습니다.

태평양 건너 중국의 풍경은 어떨까요? 중앙정부의 산업정책 영향이 크지만, Alibaba·Tencent·ByteDance·DeepSeek 같은 기업 간 경쟁도 치열합니다. 미국과 완전한 반대라기보다 국가 조정과 민간 경쟁이 결합된 구조에 가깝습니다.

미국의 칩 수출 통제가 중국 AI의 발목을 잡을 것이라는 예측은 DeepSeek의 등장으로 흔들렸습니다. 2025년 1월 공개된 추론 모델 DeepSeek-R1은 671B 파라미터이며, 코드와 모델 가중치를 MIT 라이선스로 공개했습니다. 다만 학습 데이터와 개발 과정 전체까지 공개한 것은 아니므로 '완전 오픈소스'라고 단정하기는 어렵습니다.[2]

DeepSeek의 '600만 달러'가 무엇을 뺀 숫자인지는 제1부 「지능의 단가가 빠르게 낮아지고 있다」에서 짚었지요.[3] 여기서 볼 것은 비용보다 공개 방식입니다.

중국은 칩이 부족하면 소프트웨어 효율로 승부합니다. 자원이 제한되면 창의성이 올라가는 법이니까요. DeepSeek뿐 아니라 Alibaba의 Qwen도 가중치를 공개한 오픈웨이트 진영의 주요 모델이 되었습니다.

정부 차원에서는 "AI+" 전략을 밀고 있습니다. 제조업, 농업, 의료, 교육 등 여러 산업과 공공서비스에 AI를 접목하겠다는 국가 행동계획이지요.[4] 자체 반도체 생태계 구축도 산업정책의 중요한 축입니다.

중국의 강점은 실행 속도와 규모입니다. 14억 인구가 만들어내는 데이터, 정부의 강력한 추진력. 약점은 국제적 고립 리스크이지요. 수출 통제가 강화될수록 글로벌 생태계와의 단절이 깊어질 수 있습니다.

유럽은 신뢰를, 일본은 품질을

유럽의 선택은 독특합니다. 규제를 먼저 세운다.

EU AI Act은 세계 최초의 포괄적 AI 규제법입니다. 2024년 8월 발효 뒤 의무별로 단계 적용되고 있어, 한 날짜의 '전면 시행'으로 요약하기는 어렵습니다. 기업은 시스템 유형과 자신의 역할에 따라 적용 시점을 확인해야 하지요.

고위험 의무는 한 차례 미뤄졌습니다. 2025년 11월 디지털 간소화 패키지(Digital Omnibus)로 제안된 표적 연기안이 2026년 5월 정치적 합의를 거쳐 7월 말 발효됐습니다. 채용·신용평가 같은 독립형 고위험 시스템은 2027년 12월, 의료기기처럼 제품에 내장된 고위험 시스템은 2028년 8월부터 적용됩니다.[5] 법 전체가 미뤄진 것은 아닙니다. 고위험 의무의 날짜만 유형별로 다시 그은 셈이지요.

하지만 유럽도 투자를 게을리하지는 않습니다. 2026년 중반 유럽위원회가 집계한 EuroHPC 생태계에는 19개 AI Factory와 13개 안테나가 포함됐고, 관련 투자는 26억 유로를 넘었습니다.[6] 규제와 컴퓨팅 투자를 함께 추진하는 셈입니다.

포르투갈은 Amalia, 폴란드는 PLLuM이라는 자국어 특화 모델을 개발하고 있습니다. 유럽의 전략은 공용 컴퓨팅 인프라와 지역 언어 모델을 함께 키우는 쪽에 가깝습니다.

유럽의 강점은 신뢰 기반 생태계입니다. "GDPR의 성공"이라는 선례가 있지요. 약점은 속도입니다. 규제가 혁신의 발목을 잡는다는 비판은 끊이지 않습니다.

규칙으로 먼저 판을 짠 유럽과 달리, 일본은 늘 그렇듯 조용히 움직입니다.

2025년 12월, 일본 정부는 첫 국가 AI 기본계획을 결정했습니다. AI 활용과 개발, 신뢰 확보를 국가 차원에서 함께 다루는 계획입니다.[7]

일본의 접근은 두 축입니다. 신뢰할 수 있는 AI(Trustworthy AI)와 인재 양성.

Sakana AI가 대표적입니다. 도쿄에 기반을 둔 이 회사는 전 Google 연구자들이 설립했고, 자연에서 영감받은 진화적 모델 병합 기법을 활용합니다. 2025년 11월 Series B에서 1.35억 달러를 유치하며 일본 최고 가치 AI 유니콘이 되었지요.

기업과 정부 지원 사업도 일본어 모델과 산업용 AI의 국내 개발 기반을 넓히고 있습니다. 다만 발표된 협력 구상과 법인 설립, 실제 인력 투입을 같은 성과로 묶으려면 단계 차이를 분명히 해야 합니다.

일본의 강점은 제조업 기반의 Physical AI 적용 가능성과, 품질에 대한 집착입니다. 약점은 스타트업 생태계의 얇음이지요. 하지만 Sakana AI의 등장은 변화의 신호탄일 수 있습니다.

인도의 규모와 싱가포르의 연결

인도를 빼놓을 수 없습니다. 14억 인구, 세계 최대의 IT 인력 풀. 그리고 디지털 인프라에서 이미 한 번 증명한 나라입니다.

India Stack이라 불리는 디지털 공공 인프라가 그 증거지요. 생체 인증 시스템 Aadhaar, 실시간 결제 시스템 UPI. UPI는 2025년 여러 달에 걸쳐 월간 거래 160억 건을 넘겼습니다.[8] 정부가 디지털 인프라를 깔고, 민간이 그 위에서 혁신하는 모델을 이미 작동시킨 셈입니다. 이 경험을 AI 인프라로 확장하겠다는 것이 인도의 전략입니다.

AI 스타트업 생태계도 급성장하고 있습니다. 방갈로르를 중심으로 AI 스타트업에 대한 벤처 투자가 급증하고 있고, 글로벌 테크 기업들의 AI R&D 센터가 인도에 집중되고 있지요. 매년 수백만 명의 공학 졸업생이 배출되는 인재 파이프라인은 다른 나라가 쉽게 따라올 수 없는 규모입니다.

인도의 강점은 규모와 인재입니다. 약점은 인프라 격차지요. 도시와 농촌 사이의 디지털 격차가 여전히 크고, AI 컴퓨팅에 필요한 전력과 데이터센터 인프라가 수요를 따라가지 못하고 있습니다. 하지만 India Stack의 성공이 보여주듯, 인도는 제약을 창의적으로 우회하는 데 능숙한 나라입니다.

인도가 규모로 승부한다면, 도시국가 싱가포르는 정반대 끝에서 답을 찾습니다. 싱가포르가 AI 지도에서 차지하는 비중은 국토 면적과 반비례합니다.

2023년 발표된 국가 AI 전략 2.0(NAIS 2.0)이 청사진입니다. "AI for the Public Good, for Singapore and the World"라는 비전 아래, AI 실무 인력을 15,000명으로 늘리고 글로벌 AI 창작자와 연구자의 허브가 되겠다고 선언했지요.[9]

AI Singapore(AISG)는 동남아 언어와 문화에 초점을 맞춘 SEA-LION 모델 생태계를 구축하고 있습니다. 작은 내수시장을 지역 언어와 국제 연구의 연결점으로 보완하는 전략입니다.

AI 거버넌스에서도 앞서 나갑니다. AI Verify 테스트 프레임워크를 운영하고, 2026년에는 에이전트 AI를 위한 별도 모델 거버넌스 프레임워크를 내놓았지요.[10] 규제의 명확성, 영어 사용 환경, 아시아 태평양 관문이라는 지정학적 위치가 강점입니다.

싱가포르의 약점은 태생적 한계입니다. 약 600만 인구(2024년 기준)의 작은 내수 시장, 제한된 컴퓨팅 자원. 하지만 이 나라는 늘 작은 몸집을 전략적 민첩성으로 바꿔왔습니다. AI 시대에도 같은 공식이 작동하고 있습니다.

한국의 AI G3, 그리고 선택이 비추는 가치관

한국은 야심적인 목표를 세웠습니다. AI G3, 즉 AI 분야 글로벌 3위권 진입.

이 목표를 뒷받침하는 숫자들이 있습니다. 2026년 정부 AI 예산 약 9.9조 원(정부 발표 기준). 정부와 삼성전자·SK그룹·현대자동차그룹·네이버가 나눠 구축하는 최신 GPU 26만 장. 2차 평가까지 거쳐 세 팀이 남은 파운데이션 모델 개발. Naver의 HyperCLOVA X 같은 소버린 AI 모델의 자체 개발도 진행되고 있지요.[11]

법제도도 움직였습니다. 2026년 1월 22일, AI 기본법이 시행됐습니다.[12] 진흥과 신뢰 기반 조성을 함께 규정한 법이 실제로 어떤 균형을 만드는지는 시행과 구제 결과로 확인해야 합니다.

한국의 독특한 위치는 반도체와 AI의 교차점에 있다는 것입니다. 삼성전자와 SK하이닉스가 AI 반도체 공급망의 핵심 축을 담당하고 있으니까요. AI를 쓰는 쪽과 AI를 가능하게 하는 쪽, 양쪽 모두에 발을 걸치고 있는 나라는 많지 않습니다.

한국의 강점은 실행 속도와 인프라 역량입니다. 약점은 자원 분산의 위험이지요. 여러 팀을 경쟁시키는 방식은 다양성을 확보할 수 있지만, 한정된 자원이 흩어질 수도 있습니다. 경쟁이 신뢰를 얻으려면 채점 기준부터 투명해야 하지요.

한국까지 일곱 나라를 나란히 놓으면 뜻밖의 공통점이 하나 보입니다.

각 나라에는 서로 다른 가치와 이해관계가 공존합니다. 미국 안에도 규제를 요구하는 목소리가 있고, 중국 기업도 시장에서 경쟁합니다. 유럽도 규제만이 아니라 투자 확대를 고민하지요. 여기서 제시한 자유·효율·신뢰·품질·규모·연결·속도는 국가의 고정된 성격이 아니라, 현재 정책에서 상대적으로 두드러지는 강조점입니다.

세계의 AI 전략

순위 대신 강점·제약·검증 지표로 비교합니다 — 서로 다른 정책 체계를 하나의 점수로 바꾸지 않은 정성 비교

지역	두드러진 강점	핵심 제약	주요 정책 수단	확인할 지표
미국	민간 자본 · 프런티어 모델	전력망 · 지역 수용성	민간 투자와 연방 지원의 결합	**가동 GW · 실제 집행액**
중국	대규모 인프라 · 산업 동원	첨단 칩 접근 제약	컴퓨팅파워망 · 국가 계획	**FP16 연산량 · 가동 클러스터**
EU	권리·안전 규범과 산업 기반	컴퓨트 · 스케일업 격차	AI Act · 공공 컴퓨트	**집행 일관성 · 기업 활용률**
한국	HBM · 반도체 · 제조 현장	전력 · 모델 · 서비스 생태계	민관 GPU · 독자 FM · AIDC	**납품 · 가동률 · 산업 매출**

출처: 본문에 인용한 각국 정책 발표와 공개 지표를 바탕으로 저자 작성. 본문이 다루는 일곱 나라 가운데 네 지역을 골라 표로 정리했습니다.

그림 7-5 일곱 나라 가운데 미국, 중국, EU, 한국 네 곳을 골라 강점과 제약, 정책 수단, 확인할 지표를 한 표에 모았습니다. 점수를 매기지 않았으니 맨 오른쪽 칸의 지표가 앞으로 어떻게 움직이는지 지켜보면 됩니다.

어느 전략이 맞을지는 10년 뒤에야 알 수 있을 겁니다. 석유 시대에도 미국의 민간 주도 전략과 사우디의 국영 주도 전략 중 어느 쪽이 나은지 수십 년간 논쟁이 이어졌으니까요.

그래도 하나는 지금 말할 수 있습니다. 어느 나라든 자국의 전력과 인재, 산업과 권리 보호 조건에 맞춰 고른 길이어야 오래 버틴다는 것이지요. 그리고 나라의 선택은 결국 한 사람의 하루로 내려옵니다. 다음 장에서는 다시 그 2030년의 화요일로 돌아가 보겠습니다.

1. "Removing Barriers to American Leadership in Artificial Intelligence" 및 Executive Order 14365 (The White House, https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/, https://www.whitehouse.gov/wp-content/uploads/2025/12/eo-14365.pdf).

2. "DeepSeek-R1" (DeepSeek-AI 공식 GitHub, https://github.com/deepseek-ai/DeepSeek-R1) - R1·R1-Zero의 671B 파라미터 구성과 코드·모델 가중치의 MIT 라이선스 공개 범위 확인.

3. "DeepSeek-V3 Technical Report" (DeepSeek-AI, arXiv:2412.19437, https://arxiv.org/abs/2412.19437) - V3 최종 학습에 278만 8,000 H800 GPU 시간, 시간당 2달러를 적용한 계산비 557만 6,000달러를 제시하며 선행 연구·실험 비용은 제외한다고 명시.

4. "China unveils guideline to deepen implementation of AI Plus initiative" (State Council of the People's Republic of China, 2025.8.27, https://english.www.gov.cn/policies/latestreleases/202508/27/content_WS68ae7976c6d0868f4e8f51a0.html).

5. "AI Act: regulatory framework" (European Commission, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai) — 의무별 단계 적용 일정과, AI 옴니버스(2025.11.19 제안, 2026.5.7 정치적 합의, 2026.7.27 발효)에 따른 고위험 의무 연기(부속서 III 독립형 2027년 12월 2일, 부속서 I 제품 내장형 2028년 8월 2일)를 정리한다.

6. "One year of the European Commission 2024-2029" 및 AI 정책 현황 (European Commission, https://commission.europa.eu/about/commission-2024-2029/one-year_en, https://commission.europa.eu/topics/artificial-intelligence_en) — 2026년 중반 기준 AI Factory·안테나 수와 투자 규모.

7. 「人工知能基本計画」(일본 내각부, 2025.12, https://www8.cao.go.jp/cstp/ai/ai_plan/ai_plan.html).

8. UPI Product Statistics (National Payments Corporation of India, https://www.npci.org.in/product/upi/product-statistics) — 월별 거래 건수 원자료.

9. National AI Strategy 2.0 Infographic (Smart Nation Singapore, 2023, https://www.smartnation.gov.sg/files/publications/NAIS_2_0_Infographic.pdf).

10. "Model AI Governance Framework for Agentic AI" (IMDA Singapore, 2026, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2026/new-model-ai-governance-framework-for-agentic-ai).

11. "AI 시대, 제2의 페이커는?" (대한민국 정책브리핑, 2026.1.2, https://www.korea.kr/news/policyNewsView.do?newsId=148957392) - 2026년 정부 AI 예산(과기정통부 발표 9.9조 원)을 약 10조 원으로 소개; "Korea to invest KRW 9.9 trillion in AI in 2026" (과학기술정보통신부, https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1205&sCode=eng) - 9.9조 원 기준 수치; "NVIDIA, South Korea Government and Industrial Giants Build AI Infrastructure and Ecosystem to Fuel Korea Innovation, Industries and Jobs" (NVIDIA, 2025.10.30, https://nvidianews.nvidia.com/news/south-korea-ai-infrastructure) - 정부와 삼성전자·SK그룹·현대자동차그룹·네이버클라우드의 GPU 구축분을 합쳐 26만 장 이상이라는 구성 확인; 「독자 AI 파운데이션모델 프로젝트 2차 단계평가 결과」 (과학기술정보통신부, 2026.8.18, https://www.korea.kr/briefing/policyBriefingView.do?newsId=156774781) — 2차 평가 뒤 세 팀 3차 진출.

12. 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 (국가법령정보센터, https://www.law.go.kr/법령/인공지능발전과신뢰기반조성등에관한기본법).

2030년, AI는 혼자서 당신을 위해 무엇을 할까

두 장 앞에서 함께 보낸 2030년의 화요일을 다시 떠올려볼까요. 하루를 마치고 에이전트가 이렇게 보고했지요. "오늘 처리한 업무 47건. 에이전트가 자율 처리한 것 38건, 직접 판단하신 것 9건."

이 문장에는 숫자가 두 개 들어 있습니다. 38과 9. 이 장은 바로 그 경계선에 관한 이야기입니다. 무엇을 AI 혼자 하게 두고, 무엇을 사람이 직접 붙잡고 있어야 하는지. 그 경계가 어디까지 밀려날 것인지, 그리고 그 경계를 누가 정하는지에 관한 이야기입니다.

목표를 아는 주니어 동료

2030년의 개인 에이전트를 가장 잘 설명하는 말은 "똑똑한 도구"가 아닙니다. "목표를 아는 주니어 동료"에 가깝습니다.

도구는 시킨 일만 합니다. 캘린더 앱에 일정을 입력하면 그대로 저장할 뿐, 두 회의가 겹치는지 스스로 판단하지 않지요. 반면 목표를 아는 동료는 다릅니다. "이번 분기 안에 신제품 출시를 마쳐야 한다"는 목표를 알려주면, 그 목표에 비추어 일정을 재배치하고, 우선순위가 낮은 회의는 거절하고, 필요한 자료는 미리 준비해둡니다. 판단이 애매한 지점 — 예산이 걸린 결정이나 처음 만나는 거래처와의 계약 — 에서만 "이거 어떻게 할까요?" 하고 되돌아와 묻습니다.

이것이 human-in-the-loop입니다. 사람이 모든 단계를 승인하는 것이 아니라, 궤도를 벗어나는 순간에만 사람이 개입하는 구조지요. 자율주행이 고속도로에서는 알아서 달리다가 갑자기 튀어나온 보행자 앞에서만 운전자에게 경고를 보내는 것과 비슷합니다.

일정 조율만이 아닙니다. 쇼핑도 재정 관리도 같은 방식으로 넘어갑니다. "이번 겨울 코트가 필요해"라고 말해두면, 에이전트가 예산 범위 안에서 후보를 추려 보여줍니다. 매달 나가는 구독료 중 두 달 넘게 안 쓴 서비스는 알아서 해지 여부를 물어봅니다. 투자 포트폴리오의 리밸런싱은 미리 정해둔 규칙 안에서 자동으로 이루어지고, 규칙을 벗어나는 큰 변동에서만 확인을 요청합니다.

이 구조를 기업 차원에서 가장 노골적으로 보여주는 사례가 Salesforce의 수퍼에이전트(superagent) 개념입니다.[1] 낱개의 전문 에이전트는 각자 맡은 업무만 처리하는 "사일로형 스페셜리스트"입니다. 그런데 수퍼에이전트는 다릅니다. 복잡한 목표를 통째로 받아서, 그 목표를 달성하기 위해 어떤 전문 에이전트를 언제 불러 쓸지 스스로 조율합니다. 사람으로 치면 수석보좌관에 가깝습니다. 실무는 각 분야 담당자에게 맡기고, 자신은 전체 그림을 보며 조율하는 역할이지요. 개인의 삶에서도 비슷한 구조가 자리 잡습니다. 일정 담당 에이전트, 쇼핑 담당 에이전트, 재정 담당 에이전트가 따로 있고, 이들을 총괄하는 하나의 에이전트가 "오늘 당신에게 필요한 결정"만 골라 올립니다.

이 방향의 끝에 무엇이 있을지에 대해, OpenAI의 샘 올트먼은 눈에 띄는 답을 내놓았습니다. 2025년 12월 팟캐스트 인터뷰에서 그는 다음 AI의 진짜 도약이 더 똑똑한 추론이 아니라 "무한하고 완벽한 기억(infinite, perfect memory)"에서 온다고 말했습니다.[2] 세상에서 가장 뛰어난 개인 비서라도 당신이 평생 한 모든 말을 기억하지는 못합니다. 당신이 쓴 모든 문서를 다 읽지도 못하고요. 하지만 완벽한 기억을 가진 AI라면 그것이 가능하다는 것이 올트먼의 전망입니다. 개인 슈퍼 어시스턴트라는 그림이 완성되려면, 지능보다 기억이 먼저라는 이야기지요. 물론 이것은 올트먼 개인의 전망입니다.

25퍼센트와 3조 달러라는 숫자

여기까지는 이야기입니다. 그럼 실제로 얼마나 진행되었을까요.

Gartner가 2025년 10월 공개한 CIO 700여 명 대상 설문에는 2030년 IT 업무를 AI 없이 사람이 처리 0퍼센트, 사람이 AI의 도움을 받아 처리 75퍼센트, AI가 처리 25퍼센트로 예상한 응답이 담겼습니다.[3] 이것은 CIO들의 기대를 합산한 전망이지, 실제 업무시간을 측정한 결과가 아닙니다.

특히 0퍼센트를 읽을 때는 설문 범위 안에 머물러야 합니다. 조직마다 AI 도입 범위가 다르고, 설문 문항의 업무 단위와 실제 직무도 같지 않습니다. 의료·교육·돌봄·현장 노동과 개인 생활까지 이 비율을 확장할 근거도 없습니다.

이 설문에서 읽을 수 있는 것은 많은 CIO가 AI와 사람의 협업을 기본 업무 방식으로 예상한다는 정도입니다. 자동 처리 비중보다 어떤 과업에 권한을 주었고, 오류를 누가 발견하며, 사람이 개입할 시간과 역량이 남아 있는지가 더 중요한 지표입니다.

조직마다 AI 준비도와 사람의 준비도가 다르고, Gartner 스스로도 "AI 준비와 인간 준비를 함께 관리하지 못하면 가치를 얻지 못한다"고 경고합니다. 숫자는 방향을 가리킬 뿐, 도착 시각을 확정하지 않습니다.

그렇다면 돈으로는 얼마짜리 이야기일까요? 방향과 별개로 규모도 가늠해볼 만합니다.

McKinsey는 2030년 에이전틱 커머스가 영향을 미칠 수 있는 거래 규모를 전 세계 연 3조에서 5조 달러로 추산합니다.[4] 이것은 에이전트 기업의 매출이나 확정 시장 규모가 아니라, 검색·추천·결제 과정에 에이전트가 관여할 수 있는 거래액의 시나리오입니다.

생성형 AI 전반의 잠재 생산성 가치를 추산한 다른 보고서를 여기에 더하면 시장의 범위와 GDP 기여가 뒤섞입니다. 서로 다른 숫자를 합치지 않고, 실제 평가는 에이전트가 관여한 거래액, 수수료, 오류·취소율과 소비자 피해를 따로 측정해야 합니다.

거래 과정에서 사람이 매번 승인하지 않는 범위가 넓어질수록 결제 한도, 판매자와의 제휴 공개, 취소와 환불, 미성년자·취약소비자 보호가 중요해집니다. 큰 전망치보다 이런 권리와 비용을 측정하는 편이 2030년을 준비하는 데 유용합니다.

위임의 무게와 그어야 할 선

그런데 이 모든 편리함에는 무게가 있습니다.

가장 가까운 문제는 human-in-the-loop을 실제로 유지할 수 있느냐입니다. 47건 중 38건을 AI가 처리했다면 사람에게 올라온 9건의 난도와 검토시간을 확인해야 합니다. 자동화가 늘수록 사람에게 '더 인간다운 일'만 남는다고 장담할 수 없습니다. 반복 업무가 사라져 숙련 기회가 줄고, 남은 판단이 감당하기 어려울 만큼 몰릴 수도 있지요.

노동의 몫도 따져야 합니다. 미국 노동소득분배율의 장기 하락과 AI의 인과관계는 산업·지역·시기에 따라 더 검증해야 합니다.[5] 에이전트가 만든 생산성의 몫이 임금, 노동시간 단축, 가격 인하와 기업 이익 가운데 어디로 가는지는 기술이 아니라 계약과 제도가 결정합니다.

자동화를 되돌리는 비용도 있습니다. 제6부에서 본 것처럼 일부 기업은 감원 뒤 서비스 품질이나 운영 부담 때문에 사람을 다시 채용했습니다.[6] 기사와 설문마다 'AI 때문에 감원한 기업', '후회한 기업', '재채용한 기업'의 분모가 다르므로 비율을 겹쳐 쓸 수 없습니다. 자동화 결정에도 복구 계획과 재평가 시점이 필요한 이유입니다.

2030년의 하루를 그린 장은, 그 하루가 누구에게 열릴지가 맡길 범위와 되돌릴 길을 어떻게 설계하느냐에 달려 있다는 말로 끝났지요. 이 장의 질문은 그보다 한 걸음 더 안쪽에 있습니다. 그 하루 안에서 무엇을 AI에게 맡기고 무엇을 직접 붙잡을 것인가 하는 질문입니다.

목표를 아는 보조자, 여러 전문 에이전트를 조율하는 시스템, 장기 기억을 가진 개인 비서가 2030년까지 어디까지 퍼질지는 알 수 없습니다. 자동 처리 범위가 늘더라도 사람이 판단하는 영역이 반드시 줄거나 더 '인간다운' 일만 남는다는 보장도 없습니다. 비용과 책임, 조직의 선택에 따라 자동화의 모양은 달라집니다.

기술이 거기까지 갈지는 두고 볼 일입니다. 그보다 먼저 정할 것은 그 편리함과 맞바꿀 위임의 선이지요. 일정과 쇼핑과 반복 업무는 기꺼이 맡길 수 있습니다. 그런데 아이의 학교 문제를 상담하는 대화는, 부모님의 건강을 결정하는 순간은, 팀원을 해고할지 말지의 판단은 어디에 놓아야 할까요.

편의와 위임 사이의 그 선. 2030년까지 4년, 당신은 어디에 그 선을 그으시겠습니까.



1. "What Are Superagents?" (Salesforce, https://salesforce.com/agentforce/ai-agents/superagents/)

2. "OpenAI boss Sam Altman predicts next big AI breakthrough" (Big Technology Podcast 인터뷰 인용 보도, 2025-12, https://inkl.com/news/openai-boss-sam-altman-predicts-next-big-ai-breakthrough)

3. "Gartner Survey Finds All IT Work Will Involve AI by 2030" (Gartner, 2025-10-20, https://gartner.com/en/newsroom/press-releases/2025-10-20-gartner-survey-finds-all-it-work-will-involve-ai-by-2030-organizations-must-navigate-ai-readiness-and-human-readiness-to-find-capture-and-sustain-value)

4. "The agentic commerce opportunity: How AI agents are ushering in a new era for consumers and merchants" (McKinsey, https://mckinsey.com/capabilities/quantumblack/our-insights/the-agentic-commerce-opportunity-how-ai-agents-are-ushering-in-a-new-era-for-consumers-and-merchants)

5. CEPR 노동소득분배율 연구 인용 보도 (https://asiatoday.co.kr/kn/view.php?key=20260701010000412)

6. "Employers who laid off workers for AI are reversing their decisions" (CNBC, 2026-07-01, https://cnbc.com/2026/07/01/employers-who-laid-off-workers-for-ai-are-reversing-their-decisions.html)

대한민국 AI 행동계획: 지능 에너지 강국을 향하여

2026년 2월, 대한민국 정부는 2026~2028년 AI 행동계획을 확정하고 AI 3대 강국 진입을 목표로 제시했습니다.[1] 이 선언을 지능 에너지의 관점에서 읽으면 의미가 더 또렷해집니다. 발전소를 짓고, 변환소를 세우고, 송전망을 깔고, 운영 인력을 길러내겠다는 국가 에너지 계획인 셈입니다.

발전소는 파운데이션 모델입니다. 변환소는 GPU 인프라이지요. 송전망은 AI를 전 산업에 전달하는 플랫폼이고, 운영 인력은 AI를 다룰 줄 아는 사람들입니다. 이 네 가지를 동시에 건설하겠다는 것이 한국의 AI 행동계획입니다. 공공조직의 운영과 시민의 권리는 이어지는 두 장에서 따로 다루지요.

99개 과제의 청사진과 발전소 경쟁

한국의 AI 행동계획은 99개 과제와 326개 정책권고로 구성된 3개년 계획(2026~2028)입니다.[1] 정책권고, 정부 예산안, 국회 확정 예산과 실제 집행액은 서로 다른 상태입니다. 이후 숫자도 목표와 실적을 구분해 읽어야 합니다.

상태	뜻	확인할 자료
권고·계획	하겠다고 정한 단계	공식 계획, 담당 부처, 목표 시점
예산 확정	국회가 지출 한도를 정한 단계	세부사업 예산서, 연차별 배정액
계약·선정	수행 주체와 조건을 정한 단계	계약서, 선정 기준, 이해충돌 공개
납품·구축	장비·서비스가 들어온 단계	검수·설치 기록, 이용 가능 용량
가동·성과	현장에서 쓰고 효과를 확인한 단계	이용률, 품질, 사고, 비용과 분포

3대 정책축은 이렇습니다. AI 혁신 생태계를 조성하고, 범국가적으로 AI 기반 대전환을 추진하며, 국제 AI 기본사회에 기여한다. 쉽게 말해 우리 손으로 만들고, 나라 안에 퍼뜨리고, 바깥과 나누겠다는 계획이지요.

정부는 2026년 AI 예산을 확정 기준 약 9.9조 원으로 설명합니다.[2] 행동계획이 제시한 세부 사업과 예산안·국회 확정액·실제 집행액은 같은 숫자가 아니므로, 사업별 금액은 해당 연도 예산서와 결산을 기준으로 읽어야 합니다.

2026년 9월 국무회의를 통과한 2027년 예산안에서는 과기정통부의 AI 예산만 9.4조 원으로, 올해보다 84% 늘었습니다.[3] 이것은 아직 국회 심의 전의 '예산안'입니다. 앞의 표로 말하면 첫째 줄과 둘째 줄 사이에 있는 숫자이지요.

그럼 이 돈으로 지을 지능 에너지의 발전소, 즉 파운데이션 모델은 누가 만들까요? 한국 정부는 경쟁을 선택했습니다.

2025년 8월, 네이버클라우드·업스테이지·SK텔레콤·NC AI·LG AI연구원 주도의 다섯 정예팀이 선정됐습니다.[4] 이들이 한국의 지능 에너지 발전소를 짓는 첫 경쟁 팀들이었지요.

2026년 1월 1차 단계평가 뒤 세 팀이 남았고, 추가 공모로 모티프테크놀로지스가 합류해 잠시 네 팀이 됐습니다. 8월 2차 평가에서는 모티프가 떨어지면서 다시 업스테이지·SK텔레콤·LG AI연구원 세 팀이 됐지요.[5] 1·2위 점수 차가 1점이 안 될 만큼 촘촘한 경쟁이었습니다. 탈락 팀이 이의를 제기하며 세부 점수 공개를 요구하자 정부가 상세 점수를 내놓은 것도 눈여겨볼 대목입니다. 정부는 2027년 2월 평가로 최종 두 팀을 고를 계획입니다.

지원 대상을 줄여 가는 방식은 자원 집중에 도움이 될 수 있습니다. 그래도 평가 기준과 탈락 팀의 공공 산출물을 어떻게 처리할지까지 함께 공개될 때 신뢰를 얻겠지요.

각 팀은 모델·데이터·컴퓨팅·산학 협력의 구성이 다릅니다. 이 국가 프로젝트의 투입액은 기업 전체의 연구개발비나 투자유치액이 아니라 실제 지원·집행 자료에 기록된 금액이 기준입니다.

이 경쟁이 성공할지는 팀 수만으로 판단할 수 없습니다. 공통 평가셋, 학습 데이터의 법적 근거, 비용 대비 성능, 안전성, 공개 범위와 국내에 남는 운영 역량을 같은 기준으로 비교해 봐야 알 수 있지요.

이 컨소시엄들이 지어낼 모델도 결국 어딘가의 물리적 발전소에서 돌아가야 합니다. 2026년 6월 발표된 반도체·피지컬 AI·AIDC 3대 메가프로젝트가 그 기반이며, 자세한 내용은 제2부 「대한민국의 대도약, 세 개의 메가 프로젝트」에서 다룹니다.

변환소와 운영 인력, 그리고 안전 규정

지능 에너지의 원료인 데이터를 실제 지능으로 변환하려면 GPU가 필요합니다. GPU는 지능 에너지의 변환소입니다. 정부와 국내 기업, NVIDIA가 발표한 26만 장 이상은 공공·민간 구축 계획을 합친 수치입니다. 발표가 곧 납품이나 가동 완료를 뜻하지는 않습니다.[6] 정부 몫만 보면 2026년 7월 기준 B200 환산 약 3만 5,000장을 확보했고, 그중 실제 서비스에 들어간 것은 1만 2,000장 남짓이었습니다.[7] 발표와 확보, 가동 사이의 거리가 이만큼입니다.

기업별로 보면 이렇습니다.

주체	GPU 규모	용도
네이버 클라우드	60,000대 이상	소버린 AI, 산업 특화 모델
삼성전자	50,000대 이상	반도체 AI Factory
SK그룹	50,000대 이상	산업용 AI 클라우드
현대차그룹	50,000대	Physical AI, 스마트팩토리
정부(MSIT)	50,000대	국가 AI 컴퓨팅센터

대부분 NVIDIA Blackwell B200과 H200 기반이지요. 정부는 특수목적법인(SPC) 설립을 추진하여 민간 지분 확대를 통한 투자 안정성을 확보하려 합니다.

여기서 한국의 독특한 위치가 드러납니다. 삼성전자와 SK하이닉스는 이 GPU들에 들어가는 HBM(고대역폭메모리)의 세계 최대 공급자입니다. 변환소에 들어가는 핵심 부품을 자국에서 만든다는 뜻이지요. 지능 에너지를 생산하는 설비의 핵심 부품을 자국에서 조달할 수 있는 나라는 전 세계에 몇 되지 않습니다.

해외 클라우드·칩 기업과의 협력도 늘고 있습니다. 그러나 투자 의향과 장기 계약, 국내에서 실제 집행된 금액은 서로 다른 숫자입니다. 외국 기업의 장비 공급은 국내 역량을 보완하지만 특정 공급자 의존이라는 위험도 함께 만듭니다.

발전소와 변환소가 있어도 운영할 사람이 없으면 소용없습니다.

정부 계획에는 디지털 인재와 AI·사이버보안 인력을 늘리는 여러 목표가 포함돼 있습니다.[1] 양성 인원과 일자리 전망은 직접 고용·지원 산업·유발 고용의 범위에 따라 값이 달라질 수 있으므로 실제 취업, 고용 유지, 임금과 직무 전환을 별도로 공개해야 합니다.

교육 체계도 구축하고 있습니다. 2026년 첨단산업 인재양성 부트캠프에서는 AI 분야 37개교를 포함해 8개 분야 88개교가 선정됐습니다.[8]

공공부문에서는 직원 2만 명을 대상으로 'AI 챔피언' 인증제를 추진합니다. 잘 됐는지는 수료 인원보다 실제 업무에서 근거를 확인하고 개인정보·보안 사고를 막을 수 있는지로 드러납니다. 인증 뒤 맡은 과업과 검토 품질이 달라졌는지도 추적할 필요가 있습니다. 현장은 이미 한발 앞서 있습니다. 제3부 「아래에서 오르는 사다리」에서 본 것처럼 공무원들이 스스로 만든 AI 도구가 수백 건 쌓였고, 그 앞줄에 챔피언들이 서 있지요. 그 도구들이 부서와 다른 기관으로 얼마나 건너갔는지도 인증제와 함께 따라가 볼 지표입니다.

정부 내부망의 민간 AI 활용과 클라우드 이관도 계획에 포함되어 있습니다. 이때 비용 절감뿐 아니라 기밀 분류, 공급자 종속, 장애 때의 업무 연속성과 계약 종료 후 데이터 삭제를 함께 점검해야 합니다.

설비와 사람만 갖추면 끝일까요? 전기가 처음 보급될 때 감전 사고가 빈번했습니다. 전선이 아무렇게나 깔렸고, 표준이 없었지요. 결국 전기안전법이 만들어졌습니다. 안전 규정이 전기에 대한 신뢰를 만들었고, 그 신뢰가 전기의 보급을 가속했습니다.

지능 에너지에도 같은 일이 필요합니다. 앞서 「법, 노동, 안전」 장에서 투명성·안전성 요건과 고영향 AI 사업자의 의무를 살펴본 AI 기본법이 바로 그 안전 규정입니다.

법률의 의무와 제재 범위는 시행령·고시와 집행 지침까지 함께 읽어야 합니다. 계도 방침이 발표되더라도 법률상 의무 자체가 사라지는 것은 아니므로, 기업은 적용 대상과 준비 일정을 별도로 점검해야 합니다.

이 법이 혁신의 발목을 잡을 것이라는 우려도 있습니다. 하지만 반대로 생각해볼 수도 있지요. 안전하다는 신뢰가 없으면 AI를 도입할 기업도, 사용할 시민도 줄어듭니다. 전기안전법이 전기의 적이 아니라 동반자였듯이, AI 기본법도 지능 에너지 보급의 동반자가 될 수 있습니다.

중소기업이라는 마지막 1마일과 한국만의 위치

여기까지만 보면 장밋빛이지요. 그런데 마음에 걸리는 대목이 있습니다.

중소기업의 AI 도입률을 묻는 숫자는 하나가 아닙니다. 흔히 인용되는 0.1%, 18.6%, 24% 같은 값은 'AI'를 생성형 AI로 좁혀 셌는지, 단순 자동화까지 넣었는지, 표본이 제조업 전체인지가 서로 다릅니다. 그래서 이 숫자들을 서로 빼서 격차를 계산할 수는 없습니다. 그래도 자금과 데이터, 전문 인력이 모자란 곳일수록 문턱이 높다는 방향만큼은 어느 조사에서든 크게 다르지 않습니다.

이것은 전기 초기에도 있었던 문제입니다. 대도시에는 전깃불이 환하게 켜졌지만, 농촌에는 여전히 등잔불이었지요. 지능 에너지의 송전망이 대기업에만 닿고 중소기업에는 닿지 않는다면, 한국 경제의 허리가 비어버립니다.

인재 확보 경쟁도 치열합니다. 그렇다고 최고 연봉 사례나 설문 기반의 프로젝트 지연 전망을 국내 전체 인력시장의 평균으로 보려면 직무와 표본 기준을 먼저 맞춰야 합니다. 필요한 직무와 숙련 수준, 채용 기간과 이직률을 같은 잣대로 추적하는 일이 먼저이지요.

정부 계획은 중소 제조현장의 AI 전환과 AI 팩토리 확산, 대·중소기업 협력 사업을 추진 과제로 제시합니다.[1] 목표 숫자는 구축 완료가 아니라 계획이며, 연도별 선정·설치·지속 사용 실적이 진행 정도를 보여줍니다.

어떤 조사를 기준으로 삼든, 중소 제조현장의 AI 도입을 의미 있는 수준으로 끌어올리는 일은 쉽지 않을 겁니다. 이 문턱을 낮추는 데는 농촌 전화(電化)만큼 오랜 시간과 인내가 필요할 수 있습니다.

그래도 비관만 할 일은 아닙니다. 한국의 AI 전략에서 가장 주목할 점은 따로 있으니까요.

한국은 지능 에너지의 생산자이면서 동시에 소비자라는 것입니다.

삼성전자의 HBM과 파운드리, SK하이닉스의 고대역폭메모리는 전 세계 AI 인프라의 핵심 부품입니다. 지능 에너지를 만드는 데 필요한 발전 설비의 터빈을 한국이 만들고 있는 셈이지요. 동시에 HD현대는 AI로 조선소를 혁신하고, 포스코는 AI로 철강 공정을 최적화하며, 현대차는 휴머노이드 로봇을 2028년 공장에 투입할 계획입니다.

이 이중적 위치는 기회이자 책임입니다. 칩을 만드는 역량과 칩을 쓰는 산업이 한 나라 안에 공존한다는 것은, 지능 에너지의 전체 가치 사슬을 내부에서 순환시킬 수 있다는 뜻입니다.

이 위치가 실제 성과로 이어졌는지는 숫자 하나로 재기 어렵습니다. GDP 기여도, 산업 매출, 수출, 특허와 스타트업 수는 서로 다른 지표이니까요. 어떤 목표를 내걸든 기준연도와 산식을 밝히고, 매년 같은 기준으로 진행률을 공개해야 평가할 수 있습니다.

K-문샷 프로젝트는 2035년까지 첨단바이오, 미래에너지, Physical AI 등 8대 분야 12대 국가적 임무를 추진하는 구상입니다.[9] 장기 목표인 만큼 과제 선정과 예산 확정, 연구 착수와 성과를 구분해 봐야 합니다.

계획과 예산, 성과를 한 장부에

99개 과제가 있어도 예산 사업과 연결되지 않으면 실행을 따라가기 어렵습니다. 계획 문서에는 과제 번호가 있고, 예산서에는 세부사업 코드가 있으며, 조달 문서에는 계약번호가 있습니다. 서로 다른 장부에 적힌 세 번호가 이어지지 않으면 어떤 돈이 어느 과제를 움직였고 무엇을 남겼는지 알기 어렵지요.

필자는 이 간격을 확인하기 위해 공개된 AI 행동계획을 K-AI 2026 정책 대시보드로 재구성했습니다.[10] 99개 과제와 326개 정책권고를 담당기관·기한·의존관계별로 읽어보는 비공식 분석 도구입니다. 대시보드의 빈칸이나 사용자가 입력하는 진척률은 정부의 공식 이행 실적이 아닙니다.

별도로 만든 KAIB 2026 재정 대시보드는 공개된 예산 자료를 사업 단위로 다시 읽게 해줍니다.[11] 이 도구가 분류한 범위는 533개 사업, 약 27.1조 원이지만 앞에서 정부가 AI 예산으로 설명한 9.9조 원과 집계 기준이 다릅니다. AI를 주된 목적으로 한 사업과 AI를 일부 포함한 넓은 재정사업을 비교하려면 집계 기준을 맞춰야 합니다. 차이는 오류라기보다 어떤 사업까지 AI 예산으로 셌는지를 묻게 하는 단서입니다.

필요한 것은 더 화려한 대시보드가 아니라 이어지는 식별자입니다.

> 정책 과제 ID → 예산 사업 ID → 조달·계약 ID → 산출물 → 성과지표

이 연결이 공개되면 시민은 발표가 예산을 받았는지, 계약이 실제 가동으로 이어졌는지, 같은 목적의 사업이 여러 부처에서 겹치는지를 확인할 수 있습니다. 정부가 AI 네이티브가 되려면 정책 문서와 예산서부터 AI가 읽고 대조할 수 있어야 합니다.

정책 추적

발표에서 가동까지 번호가 이어져야 보인다 — 한 과제가 지나는 다섯 칸과 칸마다 남는 식별자

	권고·계획	예산 확정	계약·선정	납품·구축	가동·성과
뜻	하겠다고 정한 단계	국회가 지출 한도를 정한 단계	수행 주체와 조건을 정한 단계	장비·서비스가 들어온 단계	현장에서 쓰고 효과를 확인한 단계
이어지는 번호	**정책 과제 ID**	**예산 사업 ID**	**조달·계약 ID**	**산출물**	**성과지표**
확인할 자료	공식 계획, 담당 부처, 목표 시점	세부사업 예산서, 연차별 배정액	계약서, 선정 기준, 이해충돌 공개	검수·설치 기록, 이용 가능 용량	이용률, 품질, 사고, 비용과 분포
	예산 규모, 컨소시엄, GPU 확보, 인재 목표 같은 발표 숫자는 대부분 이쪽 — 투입 계획입니다				**성과는 여기서 판단합니다**

번호가 끊기면 어떤 돈이 어느 과제를 움직였고 무엇을 남겼는지 따라갈 수 없습니다

출처: 「대한민국 인공지능 행동계획(2026~2028)」(국가인공지능전략위원회); 단계 구분과 식별자 연결은 본문 「청사진」 표와 「계획과 예산을 같은 장부에 적는 법」을 따라 저자가 정리.

그림 7-6 과제가 발표되고 예산을 받고 계약을 거쳐 실제로 돌아가기까지, 칸마다 다른 장부에 다른 번호가 남습니다. 그 번호들이 이어져야 돈이 무엇을 남겼는지 따라갈 수 있지요.

번호가 이어진 뒤에도 질문은 남습니다. 성과는 무엇으로 확인할까요? 9.9조 원의 예산, 경쟁 컨소시엄, GPU 26만 장, 인재 목표는 모두 투입 계획입니다. 빠른 착수는 필요할 수 있지만 착수율이 성과를 대신할 수는 없습니다.

매년 같은 기준으로 계약·납품·가동을 구분하고, 모델 성능뿐 아니라 성공 작업당 비용과 사고를 공개해야 합니다. 지역과 기업 규모별 이용률, 중소기업의 2년 뒤 지속 사용률, 훈련 뒤 고용·임금 변화도 함께 봐야 하지요. 목표를 달성하지 못한 사업은 이유와 중단·수정 결정을 남겨야 합니다.

몇 년 뒤 이 행동계획을 다시 펼쳐 본다고 해 볼까요. 그때 우리는 예산 총액과 G3 순위를 먼저 찾을까요, 아니면 국내에 남은 운영 역량과 공급자를 바꿀 수 있는 선택권, 공장과 관청에서 실제로 돌아가는 시스템, 잘못된 판단을 바로잡은 기록을 찾을까요?

1. 「대한민국 인공지능 행동계획(인공지능 기본계획 2026~2028)」 (국가인공지능전략위원회, https://www.aikorea.go.kr/upload/contents/contents_01.html) — 99개 과제와 326개 정책권고, 주관부처와 추진기한의 공식 원문. 이 책의 과제 번호 표기는 이 문서를 따른다.

2. "Korea to invest KRW 9.9 trillion in AI in 2026" (과학기술정보통신부, https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1205&sCode=eng).

3. 「2027년도 과학기술정보통신부 예산안」 브리핑 (과학기술정보통신부, 2026.9.1; 보도 https://www.inews24.com/view/2000526) — AI 예산 9조 4,211억 원(정부 전체 AI 예산의 57%), GPU 확충 3조 8,500억 원. 국회 확정 전 정부안.

4. "독자 AI 파운데이션 모델 프로젝트 정예팀 선정" (대한민국 정책브리핑, https://www.korea.kr/news/policyNewsView.do?newsId=148944741).

5. 독자 AI 파운데이션 모델 단계평가 및 추가 선정 자료 (대한민국 정책브리핑, https://www.korea.kr/briefing/policyBriefingView.do?newsId=156740468, https://www.korea.kr/news/policyNewsView.do?newsId=156745273); 「독자 AI 파운데이션모델 프로젝트 2차 단계평가 결과」 (과학기술정보통신부, 2026.8.18, https://www.korea.kr/briefing/policyBriefingView.do?newsId=156774781) — 업스테이지·SK텔레콤·LG AI연구원 3차 진출; "독파모 2차 세부 점수 공개" (머니투데이방송, 2026.8.27, https://news.mtn.co.kr/news-detail/2026082720102251703) — 환산점수 SK텔레콤 70.6, 업스테이지 69.9, LG AI연구원 69.0, 모티프 65.8; "'국가대표 AI' 투트랙으로…독파모 살리고 '프론티어 AI' 띄운다" (머니투데이방송, 2026.10.5, https://news.mtn.co.kr/news-detail/2026100320033298930) — 3차는 2027년 1월까지, 2월 단계평가로 최종 두 팀 선정 계획.

6. "NVIDIA, South Korea Government and Industrial Giants Build AI Infrastructure and Ecosystem" (NVIDIA Newsroom, 2025.10.30, https://nvidianews.nvidia.com/news/south-korea-ai-infrastructure) — 정부와 기업의 구축 계획을 합쳐 26만 장 이상으로 발표했다. 개별 장비의 납품·설치·가동 완료를 뜻하지 않는다.

7. 과학기술정보통신부 업무보고 관련 보도 (SBS Biz, 2026.7, https://biz.sbs.co.kr/amp/article/20000322871) — 2028년까지 B200 환산 5만 장 이상 확보 목표, 확보 약 3만 5,200장 중 서비스 투입 약 1만 1,600장.

8. "첨단산업 인재양성 부트캠프 선정" (대한민국 정책브리핑, https://www.korea.kr/multi/visualNewsView.do?newsId=148956369).

9. "K-문샷 프로젝트" (대한민국 정책브리핑, https://www.korea.kr/news/policyNewsView.do?newsId=148965259) — 8대 분야 12대 임무와 2035년 목표.

10. 전종홍(이 책의 저자), "K-AI 2026: 대한민국 인공지능 행동계획 분석 대시보드" (https://hollobit.github.io/K-AI2026/, 2026-07-13 확인) — 공개된 행동계획을 저자가 구조화한 비공식 분석 도구다. 정책의 공식 진척률을 제공하지 않으며 원문과의 교차검증이 필요하다.

11. 전종홍(이 책의 저자), "KAIB 2026: 대한민국 AI 재정계획 분석용 대시보드" (https://hollobit.github.io/KAIB2026/) — 공개 예산 자료를 기계가 읽을 수 있는 형태로 재구성한 저자 분석이다. 533개 사업·약 27.1조 원은 도구의 분류 범위이며 정부가 발표한 9.9조 원과 동일한 집계가 아니다.

대한민국의 AI 네이티브 미래

정부가 AI를 쓴다는 말을 들으면 어떤 장면이 떠오르시나요? 공무원이 챗봇에게 보고서 초안을 맡기는 모습이 먼저 그려지지요. 그런데 더 어렵고 더 중요한 변화는 화면 뒤의 조달·데이터·책임 체계에서 일어납니다.

정부 계획에는 'AI-Native한 공무원 협업체계'와 'AI-네이티브 환경'이라는 표현이 등장합니다.[1] 이 표현 자체가 성과를 뜻하지는 않습니다. 아래 과제도 2026년 7월 현재 대부분 계획 또는 구축 단계이며, 목표 시점과 완료 실적을 구분해 읽어야 합니다.

과제	운영 과제	2026년 계획상 목표	완료로 인정할 증거
[50]	AI-Native 공무원 협업 체계	범정부 AI 공통기반 마련, 공공데이터 머신리더블·FAIR 의무화 준비	실제 이용 부처, 품질·사고 기록, 표준 준수율
[51]	부처 칸막이를 넘는 AI 업무관리 플랫폼	공문서의 클라우드 기반 활용 체계, 문서 유통 일원화	연결된 문서 범위, 갱신 주기, 오류 정정 시간
[63]	공공 AX 거버넌스	AI 조달체계와 CTO 제도 수립	책임자 권한, 계약·감사 기준, 종료 절차
[97]	글로벌 상호운용성	에이전트·피지컬 AI 표준 체계 수립	공개 규격, 상호운용 시험, 보안·권한 검증

도구보다 먼저 바뀌어야 할 것

과제 [50]에서 쉬운 장면은 처음에 떠올린 그 모습, 공무원이 보고서 초안을 만들거나 유사 정책을 찾는 모습입니다. 어려운 문제는 그다음이지요. 어떤 자료를 근거로 썼는지 남기는가, 틀린 초안을 누가 고치는가, 민감정보가 외부 모델로 나가지 않는가, 시민에게 영향을 주는 판단에 사람이 책임지는가.

공통기반은 로그인 화면 하나를 만드는 사업이 아닙니다. 모델과 데이터의 사용 범위, 기록 보존, 사람의 승인 지점, 사고 통지와 복구 절차를 여러 부처가 같은 방식으로 운영하게 만드는 일입니다. 사용 건수만 늘고 정정률이나 검토 부담이 악화된다면 도입을 성공으로 볼 수 없습니다.

과제 [51]도 '부처 칸막이를 AI로 해결한다'는 문장만으로는 부족합니다. 데이터를 연결하려면 법적 근거, 목적 제한, 최신성, 품질 책임자가 먼저 정해져야 합니다. 복지·교육·고용 데이터를 한데 모은다고 종합 판단이 자동으로 좋아지는 것도 아닙니다. 오래되거나 잘못 연결된 정보는 오히려 불이익을 넓힐 수 있지요.

따라서 연결 전에는 데이터 목록과 법적 근거를 공개하고, 연결 후에는 오류 정정 시간과 접근 기록을 남겨야 합니다. 시민이 자신의 정보가 어디에 쓰였는지 확인하고 바로잡을 통로도 필요합니다.

데이터만 그런 것도 아닙니다. 과제 [63]이 다루는 조달체계는 모델을 사는 순간보다 교체하고 종료하는 순간에 시험받습니다. AI 서비스는 모델 버전, 가격, 데이터 처리 방식이 계속 달라집니다. 한 번의 납품검사로 몇 년의 품질을 보장하기 어렵지요.

공공 계약에는 최소한 다음 항목이 들어가야 합니다.

- 기준 과업과 집단별 성능, 허용 가능한 오류 범위
- 모델·데이터 변경 때 재평가하는 조건
- 중대한 사고와 성능 저하를 통지하는 기한
- 원본·파생 데이터, 프롬프트와 로그의 소유·반환·삭제 방식
- 다른 공급자로 옮길 수 있는 형식과 전환 지원
- 자동 판단을 거부하거나 사람의 재검토를 요청하는 절차

CTO나 전담 조직을 만든 것만으로 책임이 명확해지는 것도 아닙니다. 사업 부처, 조달 부서, 개인정보·보안 담당자, 현장 공무원 가운데 누가 배포를 승인하고 누가 중단할 수 있는지 권한을 적어야 합니다.

연결의 표준과 실행을 가르는 성과표

과제 [97]은 여러 에이전트와 물리 시스템이 함께 일할 수 있는 표준을 다룹니다. MCP나 A2A 같은 프로토콜은 메시지와 도구 호출 형식을 맞추는 출발점입니다. 하지만 인증, 최소 권한, 거래 취소, 책임 소재와 국경 간 데이터 이전까지 자동으로 해결하지는 않습니다.

공공 표준은 '연결 성공'을 넘어 잘못된 권한 요청을 거절하는지, 공급자 하나를 빼도 업무가 계속되는지, 어느 시스템이 어떤 결정을 내렸는지 추적할 수 있는지를 함께 시험해야 합니다. 국제표준 참여도 회의 횟수보다 공개 규격과 상호운용 시험 결과로 평가하는 편이 낫습니다.

그렇다면 이 과제들의 진척은 무엇으로 재야 할까요? 계획의 진척을 예산 집행률이나 사용자 수 하나로 판단하면 쉬운 업무만 자동화하는 쪽으로 흐를 수 있습니다. 다음 지표를 함께 볼 필요가 있습니다.

관점	확인할 지표	멈추고 재검토할 신호
업무	처리시간, 재작업률, 사람의 검토시간	처리시간은 줄었지만 정정·민원이 증가
데이터	최신성, 오류율, 정정 소요시간	출처·책임자가 불명확한 데이터 연결
권리	통지·설명·인간 심사 요청과 인용률	이의제기 때문에 서비스가 지연되거나 불이익 발생
안전	사고·아차사고, 권한 위반, 복구시간	로그 부재로 원인과 책임자를 확인할 수 없음
조달	공급자 전환시간, 재평가 이행률	모델 변경 뒤에도 검증 없이 계속 사용
분포	부처·지역·직급·접근성별 이용과 효과	평균은 개선됐지만 특정 집단의 오류가 악화

90일 안에 기준선과 책임자를 정하지 못한 사업은 규모를 키우기보다 목적과 데이터부터 다시 확인하는 편이 낫습니다. 시범사업에서 피해가 반복되고 구제 절차가 작동하지 않으면 중단할 권한도 미리 정해야 하지요.

몇 년 뒤 어느 구청 복지과 주무관의 화면을 떠올려 봅니다. AI가 만든 안내문 초안 옆에 근거 자료 링크가 붙어 있고, 누가 언제 고쳤는지 기록이 남습니다. 공급자가 바뀌어도 같은 화면이 열리고, 시민이 이의를 신청하면 사람 심사 창구로 넘기는 버튼이 있지요. 정부 문서의 'AI 네이티브'가 이름에서 성과로 바뀌는 때는, 이런 화면이 특별할 것 없는 일상이 되는 순간일 겁니다.

그 화면 너머의 시민은 무엇을 보장받아야 할까요? 다음 장의 질문입니다.

---

1. 「대한민국 인공지능 행동계획(인공지능 기본계획 2026~2028)」 (국가인공지능전략위원회, https://www.aikorea.go.kr/upload/contents/contents_01.html) — 표의 과제 번호·주관부처·계획상 기한의 공식 근거.

대한민국의 AI 기본사회 계획

전력은 발전과 유통에 비용이 들고 사용량을 측정해 요금을 받는 서비스입니다. 경제학에서 말하는 순수 공공재와는 다르지요. 그럼에도 많은 사회가 전기를 생활에 필수적인 보편 서비스로 보고, 시장만으로 공급하기 어려운 지역에 공공투자와 요금 지원을 마련했습니다.

AI에서도 비슷한 접근성 문제가 생깁니다. 다만 ‘AI를 쓸 권리’, 공공서비스를 차별 없이 받을 권리, 자동화된 판단을 거부하고 다툴 권리는 서로 다른 법적 문제입니다. 한국 정부의 「모두를 위한 AI 기본사회 추진계획」은 이 권리와 서비스를 어떻게 묶을지 제안하는 정책 계획입니다.

AI 기본사회와 7대 핵심 영역

두 장 앞에서는 인프라와 예산, 인재 양성을, 바로 앞 장에서는 공공조직의 AI 네이티브 전환을 이야기했지요. 이번 장의 주제는 조금 다릅니다. AI의 혜택이 모든 국민에게 닿는 사회, 즉 'AI 기본사회'라는 개념입니다.

대한민국 AI 행동계획의 과제 [81]번이 이 구상의 중심에 있습니다. 과기정통부가 주관하여 2026년 2분기까지 초안을 마련하고, 4분기까지 확정하는 「모두를 위한 AI 기본사회 추진계획」입니다. 범정부 거버넌스 기구를 구축하고 민간 참여를 보장하는 것이 계획상 기본 설계이지요.[1]

2026년 7월 24일 샌프란시스코 AI 선언에서는 이 방향이 공식 연설문 언어로 한 번 더 확인됐습니다. 대통령은 2026년 말부터 5천만 국민이 함께하는 ‘모두의 AI’ 프로젝트를 추진하고, 국민의 일상과 행정·복지·교육 등 공공서비스 전반에 AI를 활용하겠다고 밝혔습니다. 이어 모든 국민이 교육과 의료 같은 기본적인 권리를 더 두텁게 보장받고, 중소기업·소상공인·취약계층까지 소외되지 않는 사회를 만들겠다고 말하며, 이를 ‘AI 기본사회’라고 불렀습니다.[2]

이 프로젝트는 카카오·SK텔레콤·KT 세 컨소시엄이 맡아 2026년 10월 비공개 베타를 시작했고, 12월 정식 출시를 목표로 합니다.[3] 무료로 열리는 범위, 어떤 모델이 쓰이는지, 공공서비스와 어디까지 연결되는지는 출시 뒤 실제 화면에서 확인해야 할 대목입니다.

'AI 기본사회'의 ‘기본’은 현재로서는 헌법상 기본권처럼 확정된 권리를 가리키지 않습니다. 정책이 지향하는 보편성과 최소 보장 수준을 부르는 이름에 가깝지요. 보편적 AI 이용권도, 규제된 토큰 요금제나 국민별 컴퓨트 할당량도 아직 법률로 보장되지 않았습니다. 어떤 서비스를 누구에게 어떤 조건으로 보장하고, 비용과 책임을 누가 질지는 법률과 예산에서 따로 정해야 합니다. 2026년 7월 개정 시행령이 장애인·고령자·구직자 등에게 AI 이용 비용을 지원할 근거를 만들었지만, 이것도 '지원할 수 있다'는 재량 규정입니다.[4] 보편적 이용권과는 거리가 있지요.

이 계획에는 AI 민주주의, 기술 윤리, 사회적 포용 원칙이 통합되어 있습니다. AI를 잘 만드는 것만큼이나, AI를 공정하게 나누는 것이 중요하다는 인식이 담겼습니다.

그렇다면 이 에너지는 어디부터 닿아야 할까요? 전기가 보급될 때 "어디에 먼저 전깃줄을 깔 것인가"가 중요한 문제였습니다. 공장, 병원, 학교, 가정. 우선순위를 정해야 했지요. AI 기본사회도 마찬가지입니다.

이 추진계획의 세부 과제([81]-2)는 7대 핵심 영역을 제시합니다. 노동, 복지, 교육, 금융, 문화, 안전, 환경. 이 일곱 분야에 지능 에너지를 보편적으로 공급하겠다는 것입니다.

영역	예상되는 변화
노동	위험 작업을 AI가 대신하고, 반복 업무에서 사람을 해방하는 전환을 공정하게 관리
복지	위기 가구·독거노인 건강 이상을 AI가 데이터로 선제 발견, 장애인 이동 지원
교육	장애 학생·학교 밖 청소년을 포함한 학습 지원, 교사 검토와 비AI 학습 선택 보장
금융	기존 신용평가가 놓치는 상환 능력을 AI가 포착해 금융 소외 계층 축소
문화	지역 간 문화 격차 해소 — 서울 공연을 실감형 콘텐츠로 전국 전달, 소규모 창작자 지원
안전	재난 예측, 범죄 예방, 교통사고 감소를 위한 실시간 위험 감지
환경	탄소 배출 모니터링, 에너지 효율 최적화, 생태계 변화 감지

교육부 과제([17]-5)는 장애 학생과 학교 밖 청소년을 포함한 청소년에게 AI 교육을 지원한다는 목표를 둡니다. ‘모든 청소년’이라는 표현은 실제 참여율을 뜻하지 않습니다. 지역·장애 유형·학교 밖 여부별 이용률, 접근 가능한 교재, 교사의 검토 부담과 비AI 학습 선택을 함께 봐야 합니다.

계획상 7대 영역은 2027년 4분기까지 시범사업을 거쳐 2030년 1분기까지 범국민 서비스 개시를 목표로 합니다. ‘범국민’은 서비스 개통이 아니라 실제 접근성과 품질로 평가해야 합니다.

공익 AI, 그리고 거부하고 다툴 권리

여기서 한 걸음 더 나아가는 개념이 있습니다. '공익 AI'와 '사회 연대 경제'입니다.

여기서 한 과제([82]-1)는 사회 연대 경제 및 공익데이터 기반 생태계 조성을 제시합니다. AI 소셜벤처를 육성하고, 공익 AI를 개발하며, 공공 AX(AI 전환) 예산의 10~20%를 이 분야에 배분하겠다는 것이지요.

이것이 왜 중요할까요? AI 산업은 자연스럽게 수익성이 높은 분야에 집중됩니다. 금융 트레이딩, 광고 타게팅, 기업용 자동화. 돈이 되는 곳에 AI가 먼저 갑니다. 그런데 사회에는 돈은 안 되지만 절실한 문제들이 있지요. 독거노인 돌봄, 발달장애인 교육, 농촌 의료, 환경 감시. 시장 논리만으로는 이 영역에 AI가 닿지 않습니다.

공익 AI는 이 빈틈을 채우려는 시도입니다. 계획은 공공 AX 예산의 일부를 공익 목적에 배분하는 방안을 제시하지만, 실제 배분 비율과 의무 여부는 확정 예산과 집행 기준에서 정해집니다. 지원받은 조직도 선의만으로 평가할 수 없습니다. 서비스의 효과, 개인정보 침해, 현장 인력의 부담과 종료 뒤 유지 가능성을 공개해야 합니다.

그다음 과제([82]-2)는 더 야심적입니다. '모두의 AI 범부처 얼라이언스'를 구성하고, '기술 기본권'과 '공익 AI'를 법제화하겠다는 것입니다. 목표 시점은 2027년 4분기.

'기술 기본권'은 아직 법적 내용과 구제수단이 확정된 권리가 아니라 정책 제안입니다. 이를 단순한 AI 이용권으로 만들면 기기나 계정의 보급만 세고 끝날 수 있습니다. 공공서비스의 실질적 접근, 차별 금지, 개인정보 보호, 자동 판단에 대한 통지·설명·정정·거부와 구제를 각각 설계해야 합니다.

다만 AI를 쓸 권리만으로는 충분하지 않습니다. 공공 AI가 복지 수급, 채용, 대출, 교육 기회를 판단하는 순간에는 AI의 판단을 거부하고 다툴 권리도 필요합니다. 어떤 데이터와 기준이 영향을 미쳤는지 설명을 요구하고, 잘못된 데이터를 정정하며, 별도의 인간 심사와 재심을 받을 수 있어야 하지요.

디지털 기기를 쓰기 어렵거나 자동 판단에 동의하지 않는 시민에게는 전화·방문 같은 비AI 채널도 남겨야 합니다. 시민은 서비스 지연이나 불이익 없이 이의를 제기할 수 있어야 합니다. AI 기본사회는 모두에게 AI를 강제하는 사회가 아니라, AI의 도움을 받을 선택과 AI 없이도 권리를 행사할 선택을 함께 보장하는 사회여야 합니다.

권리는 절차가 있어야 작동합니다. 자동 판단 사용 사실과 담당 기관을 먼저 알리고, 쉬운 언어·수어·점자 등 접근 가능한 방식으로 설명해야 합니다. 정정과 인간 심사 신청 기한, 독립된 재심 기관, 처리 기간과 결과 통지 기준도 필요합니다. 기관은 요청 건수뿐 아니라 인용률, 집단별 처리시간, 비AI 채널의 대기시간을 공개해야 하지요.

AI 기본사회의 권리

공공 AI의 판단은 거부하고 다툴 수 있어야 한다 — 다섯 단계의 구제 절차와 어느 단계에서나 열린 비AI 채널

단계	시민이 하는 일 · 절차 요건	언제든
1 AI 판단 통지	자동 판단을 썼다는 사실과 담당 기관을 안다	비AI 채널 전화·방문으로도 같은 권리를 행사한다 디지털 기기를 쓰기 어렵거나 자동 판단에 동의하지 않는 시민을 위해
2 설명 요구	영향을 준 데이터와 기준을 묻는다 쉬운 언어·수어·점자 등 접근 가능한 방식으로	
3 데이터 정정	잘못된 기록을 바로잡는다 신청 기한을 정해 둔다	
4 인간 심사	별도의 사람이 다시 판단한다 신청 기한을 정해 둔다	
5 재심	결과에 다시 이의를 제기한다 독립된 재심 기관 · 처리 기간과 결과 통지 기준	

이의를 제기했다고 서비스 지연이나 불이익을 받아서는 안 된다

AI의 도움을 받을 선택과 AI 없이 권리를 행사할 선택을 함께 보장한다. 기관은 요청 건수·인용률·집단별 처리시간·비AI 채널 대기시간을 공개한다.

출처: 저자 작성 — 본문 제7부 「대한민국 AI 기본사회」의 판단 거부·구제 권리 논의 기반. '기술 기본권'과 이 절차는 확정된 법적 권리가 아니라 정책 제안입니다.

그림 7-7 통지에서 재심까지 다섯 단계를 차례로 밟으며 시민이 할 수 있는 일과 기관이 갖출 요건을 적었습니다. 어느 단계에서든 전화나 방문 같은 비AI 채널로 같은 권리를 행사할 수 있어야 하지요.

어디에 살든 같은 진료를, 함께 설계하는 사회를

7대 영역 중에서도 특히 구체적인 계획이 나와 있는 분야가 있습니다. 의료입니다.

이 분야의 과제([90]-3)는 'AI 기본의료'를 제시합니다. 지역 의료 격차를 AI로 줄이겠다는 구상입니다. AI 공동진료, 원격 협진이 핵심 수단입니다. 시행 시점은 2026년 4분기부터 2028년 4분기까지.

한국의 의료 격차는 심각합니다. 서울의 대형 병원에는 최고 수준의 전문의가 있지만, 지방 소도시나 농어촌에는 전문의가 턱없이 부족하지요. 같은 나라에 살면서도 어디에 사느냐에 따라 받을 수 있는 진료의 수준이 달라지는 것입니다.

AI 기본의료는 이 격차를 줄이려는 시도입니다. 의료 인력이 부족한 지역에서 영상 표시, 유사 사례 검색, 약물 상호작용 확인을 보조할 수 있습니다. 그러나 도구가 지역 환자를 충분히 대표하는 데이터로 평가되지 않았거나, 경고를 검토할 인력과 시간이 없다면 정확도가 좋아진다고 장담할 수 없습니다.

‘최종 판단은 의사’라는 문구만으로 책임이 해결되지는 않습니다. 도입 전 지역별 전향 평가, 민감도·특이도와 재검률, 누락·오진 사례, 의사의 검토시간을 측정해야 합니다. 환자에게 AI 사용을 알리고 비AI 진료를 요청할 선택과 오류 정정·피해 구제 절차도 마련해야 하지요. 대학병원 수준이라는 홍보 문구보다 이런 조건이 먼저입니다.

AI 기본사회를 정부만 설계하면 어떻게 될까요? 아무리 좋은 계획이라도, 그것을 쓰는 국민이 동의하지 않으면 작동하지 않습니다.

관련 과제([81]-3)는 공론장과 국민 참여를 제시합니다. 계획에 ‘참여’가 적혀 있는 것만으로 공동 설계가 이루어지는 것은 아닙니다.

그래도 국민 참여는 장식이 아닙니다. AI는 다른 기술과 다른 특성이 있지요. 편향, 프라이버시, 일자리 대체, 자율성 침해 같은 복잡한 사회적 질문을 동반합니다. "AI가 내 이력서를 심사하는 것에 동의하는가?" "AI가 내 자녀의 학습을 분석하는 것을 허용할 것인가?" "AI가 내 건강 데이터를 활용하는 것을 수용할 것인가?" 이런 질문에 답하는 것은 기술자나 관료가 아니라 국민입니다.

공론장은 참여자 모집부터 결과 반영까지 설계해야 합니다. 지역·연령·장애·디지털 이용 수준을 고려해 참여자를 구성하고, 찬반 자료와 예산·위험 정보를 같은 조건으로 제공해야 합니다. 독립된 진행자가 숙의를 기록하고, 정부는 채택·불채택한 제안과 이유를 공개해야 하지요. 서비스가 시작된 뒤에도 민원과 사고 자료를 바탕으로 다시 검토할 자리가 필요합니다.

한국 모델의 확산과 2030년의 성적표

AI 기본사회는 한국만의 과제가 아닙니다.

관련 과제([86]-1,2)는 'AI 기본사회 글로벌 얼라이언스'의 시행계획을 2027년 1분기까지 수립하고, 해외 네트워크와 공조하겠다고 밝히고 있습니다. 한국이 만들어가는 AI 기본사회 모델을 국제사회와 공유하겠다는 것이지요.

이어지는 과제([97]-2)는 이 방향을 더 구체화합니다. UN, APEC, G7/G20 등 국제 무대에서 '글로벌 AI 기본사회 모범사례 확산'을 주도하겠다는 것입니다.

왜 한국이 이 역할을 자임하는 것일까요? 한국에는 독특한 경험이 있습니다. 한 세대 만에 농업 국가에서 IT 강국으로 전환한 경험, 전 국민 건강보험을 구축한 경험, 초고속 인터넷을 전국에 보급한 경험. 기술과 보편적 복지를 동시에 추진한 나라라는 것이지요.

한국의 경험이 다른 나라에 참고가 되려면 모델의 이름보다 결과를 공개해야 합니다. 지역·소득·장애별 접근성이 개선됐는지, 권리구제가 작동했는지, 운영비를 지속할 수 있는지 보여줘야 하지요.
미국·중국·한국의 전략에도 산업 경쟁과 공공 목적이 함께 있으므로 국가별 성격을 한 단어로 나누는 비교는 피하는 편이 정확합니다.

그렇다면 이 계획이 제대로 굴러가는지는 언제, 무엇으로 알 수 있을까요? 과제 [81]의 추진계획은 2026년 4분기 확정을 목표로 합니다. 계획대로라면 시범사업은 2027년 4분기까지 이어지고, 2030년 1분기에 7대 영역의 범국민 서비스가 열리지요. '기술 기본권'의 법제화는 2027년 4분기, 글로벌 얼라이언스의 시행계획 수립은 2027년 1분기가 목표입니다.

이 일정은 서비스가 계획대로 열렸는지를 확인하는 출발점입니다. 실제 평가는 이용률만으로는 부족합니다. 평가 항목에는 대상과 무료 제공량 또는 서비스 범위, 모델과 비AI 채널의 선택권, 장애·언어·기기 환경별 접근성이 들어갑니다. 자동 판단에 이의를 제기하고 구제받는 절차, 지역·소득·연령·장애별 이용 격차도 같이 봐야 하지요.

2030년까지 계획이 일부 지연되거나 효과가 확인되지 않을 수도 있습니다. 그때는 목표 숫자를 지키는 것보다 실패 원인과 피해를 공개하고 사업을 고치는 편이 낫겠지요.

읍내 주민센터 창구 앞에 선 어르신 한 분을 떠올려 봅니다. 키오스크의 AI 상담에 물어도 되고, 옆 창구의 사람에게 가도 됩니다. AI가 내린 결정이 이상하면 사람에게 다시 심사해 달라고 말할 수 있고, 그 말을 했다는 이유로 기다림이 길어지지도 않습니다. AI 기본사회가 정말 왔는지는 아마 이런 평범한 오후에 확인될 겁니다.

---

1. 「대한민국 인공지능 행동계획(인공지능 기본계획 2026~2028)」 (국가인공지능전략위원회, https://www.aikorea.go.kr/upload/contents/contents_01.html) — 과제 [81], [82], [86], [90], [97]의 계획 내용과 목표 시점. 예산 비율과 일정은 확정 집행 실적이 아니라 계획상 목표다.

2. 「7/24(금) 샌프란시스코 AI 선언」 (대통령실, 게시 2026.7.25, https://www.president.go.kr/speeches/hHXHeToZ) — ‘모두의 AI’ 프로젝트와 ‘AI 기본사회’ 공식 언급.

3. "'모두의 AI' 비공개 베타 돌입" (머니투데이방송, 2026.10.14, https://news.mtn.co.kr/article/2026101415373296963) — 3개 컨소시엄, 10.16 비공개 베타, 연내 정식 서비스 계획. 2027년 예산안에 2,500억 원 편성(https://www.inews24.com/view/2000526).

4. 「달라진 AI기본법, 우리에게 어떤 변화가 생길까?」 (대한민국 정책브리핑, 2026.7.15, https://www.korea.kr/news/policyNewsView.do?newsId=148968171) — 2026.7.21 시행 개정 법령: 공공조달 AI 제품 우선 고려, 취약계층 AI 이용비 지원 근거.

AI 시대의 공공 API와 공공 데이터: 개방 1위, 그다음은

공공데이터를 지능 에너지의 연료라고 생각해봅시다. 정부와 지자체가 축적해온 인구, 교통, 기상, 의료, 부동산 데이터. 이것이 AI 에이전트가 태워야 할 원료입니다. 그런데 연료가 아무리 많아도 엔진에 꽂을 콘센트가 안 맞으면 소용없습니다. 220V 콘센트에 110V 플러그를 억지로 밀어 넣을 수는 없는 노릇이지요.

제2부 「송전망: MCP와 A2A, 지능이 흐르는 길」에서 본 표준화된 API와 MCP(Model Context Protocol)가 바로 그 콘센트 규격이지요. 사람은 웹페이지의 버튼을 눈으로 보고 누르지만, AI 에이전트는 규격에 맞는 API를 통해서만 데이터에 접근합니다. 규격이 안 맞으면 아무리 좋은 데이터도 에이전트 앞에서는 그냥 잠긴 서랍입니다.

한국은 이 지점에서 이상한 모순에 서 있습니다.

세계 1위라는 착시

한국은 OECD의 OURdata Index(공공데이터 개방·활용 지수)에서 2015·2017·2019·2023년 조사 모두 1위에 올랐습니다. 2023년에는 OECD 평균(0.48)을 크게 웃돌며 프랑스(0.83)를 앞섰지요. 공식 자료에 실린 한국의 점수는 조사 연도마다 조금씩 달라서, 2019년 0.93, 2023년 0.91, 2025년 0.95였습니다.[1] 공공데이터포털의 등록 건수와 함께 보면 한국이 개방 정책에 꾸준히 투자해 왔다는 사실은 분명합니다.

그런데 이 지수가 재는 것은 딱 하나입니다. "얼마나 많은 데이터를 얼마나 적극적인 정책으로 열었는가." 데이터의 양과 개방 의지를 재는 지수이지, 그 데이터를 AI가 실제로 가져다 쓸 수 있는지는 별개의 질문입니다.

이 별개의 질문을 탐색하기 위해 필자는 K-GAPI(대한민국 공공 Open API 현황 분석)를 만들었습니다. 2026년 4월 조사에서는 173개 원천 포털을 식별하고 연계·추가 포털을 포함해 199개 포털을 살펴봤습니다. 이 자체 분석에서 기술 품질 종합 점수는 10점 만점에 2.0점이었습니다.[2] 독립기관이 검증한 국가 순위가 아니라, 공개된 포털을 같은 기준으로 점검해 개선 지점을 찾기 위한 탐색적 진단입니다.

왜 이런 격차가 생겼을까요? K-GAPI가 짚어낸 문제들을 들여다보면, 대부분 2010년대 XML·SOAP 시절의 설계가 그대로 남아있다는 데서 출발합니다.[2] 다음 표는 2026년 4월의 기준선입니다.

문제	실제 모습	AI에게 왜 걸림돌인가
OpenAPI 스펙 미제공	대다수 포털이 스펙 문서를 제공하지 않음	LLM function calling용 스키마를 사람이 손으로 다시 작성해야 함
XML 기본 응답	JSON이 아니라 XML이 표준	변환과 파싱이 늘고 같은 정보를 읽는 토큰 비용도 커짐
에러도 HTTP 200	실패한 요청에도 상태코드 200 반환	AI가 성공과 실패를 스스로 판단할 수 없음
포털 파편화	173개 원천 포털(연계 포함 199개)	전체를 다루려면 131개의 서로 다른 ID가 필요
MCP 커버리지	전체 API 중 5% 미만만 MCP로 노출	에이전트가 자연어로 바로 호출할 수 있는 창구가 거의 없음

여기에 오래된 관행이 몇 겹 더 쌓여 있습니다. `serviceKey`를 두 번 인코딩해야 정상 작동하는 포털이 수두룩하고, CORS를 지원하지 않아 프론트엔드에서 직접 호출도 못 합니다. API 키는 2년마다 만료되는데 자동 갱신은 없고, 요청 한도를 알려주는 Rate Limit 헤더도 없어서 얼마나 더 불러도 되는지 에이전트가 미리 알 방법이 없습니다.

HTTP 상태코드 의미론 항목에서도 표본 내 한국 포털은 낮은 점수를 받았습니다.[2] 국가 전체의 확정 순위로 읽기보다, 실패를 정상 응답처럼 돌려주는 구현 관행이 반복됐다는 신호로 보는 편이 정확합니다. 조사 시점에는 프랑스와 미국의 정부 MCP 사례와 비교할 만한 한국 정부 공식 MCP 서버도 확인되지 않았습니다.

네 달 뒤의 재측정, 그리고 세계의 콘센트

2026년 8월 4일, 필자는 조사 대상을 215개 포털로 넓혀 다시 측정했습니다. 모집단과 채점 방식이 달라졌으므로 4월 점수와 단순한 상승·하락률로 비교할 수는 없습니다.[3]

후속 조사에서 215개 포털의 AI 적합성 평균은 14점 만점에 2.35점, E등급으로 측정됐습니다. 대시보드는 이전과 비교 가능한 포털 가운데 17곳은 개선, 10곳은 후퇴, 145곳은 변화 없음으로 분류했습니다. 215곳 중 159곳은 여전히 별도 회원가입을 요구했습니다.[3]

작지만 의미 있는 변화도 있었습니다. 공공데이터포털은 잘못된 요청을 무조건 HTTP 200으로 돌려주던 방식에서 일부 요청에 400 오류를 반환하도록 바뀌었습니다. 그러나 OpenAPI 명세, RFC 7807 표준 오류 형식, Rate Limit 정보, CORS, 헤더 기반 인증과 중앙정부 공식 MCP는 후속 조사에서도 확인되지 않았습니다.[3]

전선 한 가닥을 고쳤다고 송전망이 완성되는 것은 아닙니다. 에이전트가 오류를 알아차리고, 남은 호출량을 계산하고, 스키마를 읽어 도구를 스스로 연결할 수 있어야 비로소 공공 API가 지능 에너지의 배관이 됩니다.

우리가 전선 한 가닥을 고치는 사이, 다른 나라들은 이 규격 전환을 이미 시작했습니다.

미국 GovInfo는 2026년 1월 22일, 연방 공식정보를 LLM이 직접 검색·호출할 수 있는 MCP 서버를 프리뷰로 공개했습니다. 인간이 보는 웹 화면, 프로그램이 부르는 API에 이어 AI 에이전트가 쓰는 MCP를 세 번째 인터페이스로 시험 편입한 것이지요[4]. 프랑스는 한 달 뒤인 2월, 디지털청(DINUM)이 `mcp.data.gouv.fr`을 열어 정부 공식 MCP 서버를 운영하기 시작했습니다[5]. EU는 아예 규정으로 못 박았습니다. 지리공간·지구관측·환경·통계·기업·이동성 6개 범주의 '고가치 데이터셋(HVD)'을 무료로, 머신리더블 형태로, API를 통해 제공하도록 의무화한 것입니다[6].

싱가포르의 APEX는 접근 자체를 다르게 설계했습니다. 공개 팩트시트 기준 31개 기관의 API 2,400개 이상을 게이트웨이로 연결해 월 1억 건 이상의 호출을 처리했습니다.[7] 기관별 시스템을 공통 관문으로 연결하는 구조이지요.

방향은 뚜렷합니다. 공공데이터를 사람이 열람하는 포털에서, AI 에이전트가 직접 호출하는 표준 인터페이스로 옮기는 흐름입니다. 한국이 데이터의 양에서 앞서 있는 동안, 다른 나라들은 그 데이터를 누가, 무엇이 가져다 쓰느냐는 질문으로 이미 넘어간 것입니다.

규격을 다시 맞추는 일

다행히 이 문제는 새 데이터를 더 모으는 문제가 아닙니다. 이미 있는 연료를 규격에 맞는 콘센트로 갈아 끼우는 문제입니다. K-GAPI가 제시하는 로드맵도 그래서 비교적 단기간에 실행 가능한 항목들로 채워져 있습니다.

6개월 안에 할 수 있는 일이 있습니다. OpenAPI 3.0 스펙을 JSON으로 의무 제공하고, 응답 기본값을 XML에서 JSON으로 바꾸고, 상태코드를 401·403·404·429처럼 제대로 반환하고, Rate Limit 헤더와 CORS를 붙이는 것입니다.[2] 1~2년 안에는 더 구조적인 변화가 필요합니다. 기관별 포털을 단일 진입점으로 묶는 통합 API 게이트웨이, `serviceKey` 대신 OAuth 2.0 방식의 인증, 그리고 무엇보다 공식 MCP 서버입니다.

이 방향은 이미 한국 정부 안에서도 움직임이 있습니다. 행정안전부는 정부 내부망에서 AI를 안전하게 쓸 수 있도록 '범정부 AI 공통기반'을 마련하고, 공공데이터의 머신리더블화와 FAIR 원칙 적용을 함께 추진하고 있습니다.[8] 제5차 공공데이터 기본계획은 'AI·고가치 공공데이터 TOP 100' 확대 목표도 담았습니다.[9] 계획 발표가 API 품질 개선을 뜻하는 것은 아니므로 실제 규격과 응답 품질을 따로 확인해야 합니다.

이 움직임들이 공공 API의 기술 품질까지 끌어올릴지는 아직 지켜볼 일입니다. 「대한민국의 AI 네이티브 미래」에서 본 'AI-Native 공무원 협업체계'(과제 [50])나 'AI 업무관리 플랫폼'(과제 [51]) 같은 과제가 정부 내부의 AI 활용을 다룬다면, 이번 장에서 본 API·MCP 문제는 그 안쪽 데이터가 정부 밖의 AI 에이전트, 그리고 시민이 쓰는 에이전트에게까지 제대로 흘러나갈 수 있는지를 가르는 배관입니다. 정부가 AI 네이티브가 되겠다고 선언한 것만으로는 부족합니다. 그 정부가 쥐고 있는 데이터의 콘센트 규격이 세계 표준과 맞아야, 비로소 에이전트가 플러그를 꽂을 수 있습니다.

한국은 연료를 쌓는 경쟁에서는 이미 앞서 있습니다. 남은 것은 그 연료가 실제로 흐르게 만드는 배관 공사이지요. 개방 1위라는 훈장을 단 다음에는, AI가 정말로 그 문을 열고 들어갈 수 있는지를 물어야 합니다.

여러분의 회사나 기관이 API를 내놓고 있다면 한 번 점검해 보세요. OpenAPI 명세가 붙어 있는지, 실패를 실패로 알려 주는 상태코드를 돌려주는지, 에이전트가 붙을 MCP 창구가 있는지. 표준화가 늦어질수록 에이전트는 다른 집의 콘센트를 찾아갑니다.

---

1. OECD OURdata Index 2023, Government at a Glance 2025 「Open government data」 및 Digital Government Outlook 2026 국가 자료 (OECD, https://www.oecd.org/en/publications/government-at-a-glance-2025_0efd0bcd-en/full-report/open-government-data_619b668c, https://www.oecd.org/en/publications/2023-oecd-open-useful-and-re-usable-data-ourdata-index_a37f51c3-en.html, https://www.oecd.org/en/publications/digital-government-outlook-2026_d46c0555-en/korea_b1d86283-en.html) — 조사 연도별 점수를 구분한다. 2023년 1위(0.91, OECD 평균 0.48)는 Government at a Glance 2025에서, 2015·2017·2019·2023년 연속 1위는 행정안전부 디지털정부 소개 자료(https://dgovkorea.go.kr/contents/blog/402)에서 확인. 2023년 지수는 문항이 바뀌어 이전 조사와 세부 비교가 어렵다.

2. 전종홍(이 책의 저자), "K-GAPI: 대한민국 공공 Open API 현황 종합 분석" (https://hollobit.github.io/K-GAPI, 2026-04-02 기준) — 저자 본인이 직접 수행한 분석 자료. 단일 저자의 진단인 만큼, 정책에 반영하기 전에는 NIA·행정안전부·KISDI 같은 독립기관의 교차검증을 거치는 편이 안전하다.

3. 전종홍(이 책의 저자), "K-GAPI 델타: 공공 API 4개월 후속 측정" (https://hollobit.github.io/GAPI2/, 2026-08-04 측정) — 215개 포털의 경량 점검과 일부 포털의 심층 실측을 결합한 비공식 재측정이다. 4월 조사와 모집단·평가척도가 달라 단순 시계열 비교에는 쓰지 않았다.

4. GovInfo(미국 정부출판국), "MCP Public Preview" (https://govinfo.gov/features/mcp-public-preview, 2026-01-22)

5. DINUM(프랑스 디지털청), 정부 MCP 서버 https://mcp.data.gouv.fr (2026-02); destilabs, "Government AI Tracker" (https://destilabs.com/government-ai-tracker)

6. European Commission, High-Value Datasets Implementing Regulation (EU) 2023/138 (https://data.europa.eu)

7. APEX Factsheet 및 APEX Cloud 안내 (GovTech Singapore, https://www.developer.tech.gov.sg/assets/files/apex-factsheet-121222.pdf, https://v2.developer.tech.gov.sg/products/categories/data-and-apis/apex-cloud/overview).

8. 행정안전부, "범정부 AI 공통기반 구축" 보도자료 (https://korea.kr/news/policyNewsView.do?newsId=148955271)

9. 행정안전부, "제5차 공공데이터 기본계획" 보도자료 (https://korea.kr/news/policyNewsView.do?newsId=148965587, 2026-01)

글로벌 AI 기본사회를 향한 제안

전기가 보편화되었을 때 무슨 일이 벌어졌을까요?

1904년, 세계 각국의 과학자와 산업가들이 미국 세인트루이스에 모였습니다. 전기가 빠르게 퍼지고 있었지만, 나라마다 전압이 다르고 규격이 달랐지요. 독일의 전구를 미국에 가져가면 쓸 수 없었고, 영국의 모터를 프랑스 공장에 설치할 수도 없었습니다. 이대로는 안 되겠다는 공감대가 모여, 1906년 국제전기기술위원회(IEC)가 탄생했습니다. IEC는 이후 헤르츠·웨버 같은 전기 단위의 이름과 정의를 국제적으로 맞추는 데 앞장섰고, 이 작업은 훗날 국제단위계(SI)로 이어졌지요.

AI가 빠르게 퍼지지만 접근 조건은 여전히 불균등합니다. 비슷한 질문이 떠오릅니다. 국제 협력은 무엇을 표준화하고, 누가 비용과 의사결정권을 나눠 가져야 할까요?

AI 기본사회라는 비전과 APEC의 합의

이 책에서 줄곧 이야기해온 것처럼, AI는 새로운 에너지입니다. 전기가 근대 사회의 기반 인프라가 되었듯이, 지능 에너지가 다음 사회의 기반 인프라가 됩니다.

AI 기본사회란 이 인식에서 출발하는 개념입니다. AI가 특정 기업의 경쟁 무기가 아니라, 전기나 수도처럼 모든 시민이 접근할 수 있는 공공 인프라가 되는 사회. 지능 에너지가 서울의 대기업 본사에서만 흐르는 것이 아니라, 전남의 어촌에서도, 강원의 산골에서도, 나아가 캄보디아의 학교에서도 아프리카의 병원에서도 흐르는 세상. 그것이 AI 기본사회의 비전입니다.

한국 정부는 이 비전과 연결되는 여러 구상을 국제 무대에 제안하고 있습니다. 다만 채택된 합의, 설립 발표, 유치 제안과 실제 운영은 서로 다른 단계이지요.

구상	2026년 10월 기준 상태	아직 확인할 것
APEC AI 이니셔티브	회원 경제권이 채택한 협력 문서	사업별 예산, 담당기구, 성과 공개
아시아태평양 AI 센터	설립·국내 유치 구상	소재지, 운영 주체, 회원 참여와 재원
글로벌 AI 허브(UN AI 허브 유치 구상에서 확대)	9개 국제기구·5개 다자개발은행 참여 비전 선포(2026.5), 대통령 공식 연설에서 추진 언급, 입지 공모 중(연내 확정 ·2027년 개소 목표)	국제기구별 협정 문안과 협력 범위, 재원 분담, UN 내 공식 지위, 실행 사업
글로벌 AI 거버넌스 대화	UN 결의에 따른 국제 논의	참여 대표성, 권고의 후속 집행

이 가운데 가장 먼저 문서로 매듭지어진 것은 APEC입니다. 2025년 11월, 경주에서 열린 APEC 정상회의는 한국 AI 외교의 분수령이 된 자리였습니다.

이 회의에서 21개 회원 경제권은 APEC AI 이니셔티브(2026~2030)를 채택했습니다.[1] 서로 다른 제도와 이해관계를 가진 회원들이 AI 역량과 투자 환경을 함께 다루기로 했다는 데 의미가 있습니다.

이 이니셔티브는 아시아-태평양 전체의 AI 도입과 전환을 지원하고, 각 경제권의 AI 역량을 키우며, AI 인프라를 위한 투자 환경을 조성하는 데 초점을 맞추고 있습니다.

한국은 자체 재원으로 독립 운영하는 아시아태평양 AI 센터(Asia-Pacific AI Center) 설립 구상을 내놓았고, APEC 문서는 이를 명시적으로 언급했습니다.[1] 다만 운영 주체와 의사결정 구조, 지속 재원은 아직 후속 문서를 기다리는 중입니다. 국내 지방자치단체의 유치 활동도 후보지 제안이지 센터 가동을 뜻하지 않습니다.

전기의 비유로 말하면 이렇습니다. APEC AI 이니셔티브는 아시아-태평양 21개 경제권이 "지능 에너지의 보급을 함께 추진하겠다"고 합의한 것이고, 아시아태평양 AI 센터는 그 합의를 실행할 발전소를 한국 땅에 짓겠다는 것입니다.

같은 시기 발표된 26만 개 이상 GPU는 정부와 기업의 구축 계획을 합친 수치입니다.[2] 지원 약속이나 계약이 납품·설치·가동을 뜻하지는 않습니다. 국제 허브 역량을 평가할 때도 발표 수량보다 실제 가용 용량, 이용 조건과 전력·환경 비용을 봐야 합니다.

UN AI 허브, 더 큰 무대로

APEC이 아시아-태평양이라면, UN은 전 세계입니다. 한국은 더 큰 무대에서도 같은 비전을 펼치고 있습니다.

2026년 3월 10일, 정부는 국무총리 주재로 UN AI 허브 유치지원 TF 첫 회의를 열고 한국 유치를 추진하겠다고 밝혔습니다.[3] 어디까지나 한국의 유치 구상이고, UN의 공식 설립 결정과는 다른 단계입니다.

UN AI 허브란 무엇일까요? UN 산하 전문기관들의 AI 관련 기능을 결합하고, 한국 정부와 민간의 AI 인프라를 UN의 글로벌 네트워크와 연결하는 플랫폼입니다. 한국이 "글로벌 AI 3강"을 목표로 하면서 그 역량을 국제사회에 기여하는 통로를 만들겠다는 구상이지요.

구상은 그 뒤 더 넓어졌습니다. 2026년 5월에는 ILO·ITU·UNDP·WHO 등 9개 국제기구와 세계은행·ADB 등 5개 다자개발은행이 참여한 가운데 '글로벌 AI 허브' 비전이 선포됐고, 10월에는 입지 공모가 시작됐습니다. 연내 부지를 정하고 2027년 문을 여는 것이 목표입니다.[4] 그래도 참여 선언이 UN 총회의 설립 결정이나 예산 승인을 뜻하지는 않습니다.

정부는 이 행보를 AI 역량의 국제 기여로 설명합니다. 국제 공공재를 표방하려면 유치국의 산업 이익과 참여국의 필요가 충돌할 때 어떤 원칙으로 결정하는지도 공개해야 합니다.

UN 차원에서도 이미 움직임이 있습니다. ITU(국제전기통신연합)가 주관하는 AI for Good은 2017년부터 AI를 UN 지속가능발전목표(SDGs)에 활용하는 방안을 논의해 왔습니다.[5] UN 총회 결의에 따라 출범한

첫 글로벌 AI 거버넌스 대화(Global Dialogue on AI Governance)는 2026년 7월 6~7일 제네바에서 열렸습니다.[6]

한국의 글로벌 AI 허브 구상은 이 흐름과 맞닿아 있습니다. 그러나 유치 활동과 UN의 공식 설립 결정은 다른 단계입니다. 어떤 기능을 기존 국제기구와 나누고 중복을 피할지부터 설명할 필요가 있습니다.

2026년 7월 24일 샌프란시스코 AI 선언에는 더 넓은 표현도 등장했습니다. 정부는 '모두를 위한 AI, 인류의 난제 해결을 위한 AI'라는 비전 아래 국제기구들과 함께 글로벌 AI 허브 이니셔티브를 추진하고 있다고 밝혔습니다.[7] 핵심은 한 나라가 위에서 내려주는 지원이 아니라 수평적 협력 네트워크입니다. 각국의 언어와 문화, 산업 환경에 맞는 AI가 개발되고 활용되도록 국가 간, 기업 간 협력을 넓히겠다는 말이지요.

왜 글로벌 REA가 필요한가

현실을 하나 보겠습니다.

세계은행은 고소득 국가에 주요 AI 모델과 스타트업, 벤처투자, 코로케이션 데이터센터 용량이 집중돼 있다고 지적합니다.[8] 정확한 비율보다 중요한 것은 컴퓨팅과 자본, 현지 데이터와 인력이 같은 지역에 고르게 분포하지 않는다는 사실입니다.

UNCTAD는 세계 AI 시장이 2033년 4.8조 달러에 이를 수 있다고 전망하면서, 시장과 연구 역량의 집중이 개발 격차를 키울 수 있다고 경고합니다.[9] 4.8조 달러는 현재의 경제적 가치가 아니라 미래 시장 규모 전망입니다.

제7부 2장 「지능 에너지 격차」에서 이야기한 격차를 좁히려면, 개별 국가의 노력만으로는 부족합니다. 국제적 협력 체계가 필요합니다. 세계은행이 제시한 "네 가지 C" — 연결성(Connectivity), 컴퓨팅(Compute), 맥락(Context, 즉 현지화된 데이터), 역량(Competency) — 이 모든 곳에서 동시에 갖춰져야 합니다.

연결성이 있어도 컴퓨팅 자원이 없으면 AI를 돌릴 수 없습니다. 컴퓨팅이 있어도 현지 언어와 현지 맥락에 맞는 데이터가 없으면 쓸모가 없지요. 데이터가 있어도 그것을 활용할 인력이 없으면 무용지물입니다. 이 네 가지를 동시에 해결하려면 한 나라의 예산으로는 모자랍니다.

APEC AI 이니셔티브와 글로벌 AI 허브 구상은 이 문제에 대한 정책 제안입니다. 실제로 격차를 줄였는지는 저소득 회원의 컴퓨트 이용시간, 현지 인력과 데이터 역량, 공동 의사결정권이 늘었는지로 드러날 겁니다.

이 제안들을 한데 놓고 더 큰 그림을 그려 보면 어떨까요?

「지능 에너지 격차」 장에서 만난 미국 농촌전력청(REA) 이야기를 다시 떠올려 볼까요? 도시보다 농촌의 전력 보급이 크게 늦어지자 연방 정부가 협동조합 금융과 송전망 건설에 개입했습니다. 시장성이 낮은 지역의 기반시설에는 공공투자가 필요하다는 사례이지요.

지능 에너지 시대에도 글로벌 차원의 REA가 필요합니다. 시장에만 맡겨 두면, AI 인프라 투자는 수익이 나는 곳에만 집중됩니다. 미국, 중국, 유럽에 데이터센터가 들어서는 동안, 아프리카와 동남아시아와 중남미에는 지능 에너지가 닿지 않지요.

이 책은 APEC 협력, UN 허브 제안에서 커진 글로벌 AI 허브, AI for Good 같은 서로 다른 움직임을 글로벌 AI 기본사회 구상이라는 이름으로 묶어 봅니다. 정부가 공식 채택한 사업명은 아닙니다. 공식 구상 옆에 놓고, 보편 접근과 현지 역량을 평가하기 위한 저자의 해석틀에 가깝습니다.

지능 에너지의 보편적 접근. 컴퓨팅 자원과 모델에 대한 선택지가 특정 국가와 기업에만 몰리지 않게 하자는 것이지요. 공개 가중치 모델, 공동 컴퓨팅, 개발도상국을 위한 컴퓨트 크레딧 등이 수단이 될 수 있습니다. 라이선스, 운영비와 현지 소유권도 함께 봐야 하지요.

AI for SDGs. 기후 모니터링, 감염병 경보, 식량과 교육 문제에 AI를 시험할 수 있습니다. 효과는 분야와 현장에 따라 다르고 잘못된 경보와 유지비도 생깁니다. 시범사업 수보다 기존 방식 대비 효과, 오류, 현지 인력에게 남은 역량을 공개해야 합니다.

포용적 AI 거버넌스. AI 규칙을 정하는 테이블에 모든 나라가 앉아야 합니다. 「지능 에너지 격차」 장에서 본 것처럼, 주요 글로벌 AI 거버넌스 논의에 참여하지 못하는 나라가 118개국에 달하지요. 이들 없이 만들어진 규칙은 이들의 현실을 반영하지 못합니다. 글로벌 AI 허브가 이 참여의 문턱을 낮추는 역할을 할 수 있습니다.

공개 가중치 모델과 온디바이스 AI는 선택지를 넓힐 수 있습니다. 그러나 장비·전력·현지 데이터·보안·유지보수 비용은 남습니다. 보편화 속도는 기술의 가격 하락보다 실제 이용과 총비용에서 읽는 편이 정확합니다.

국제 구상을 운영체계로 바꾸려면

센터가 어디에 들어서는가보다 다음 질문이 먼저입니다.

- 저소득 회원과 시민사회가 의제와 예산 결정에 표결권을 갖는가.
- 컴퓨트는 어떤 기준으로 배분하며 사용 기록을 공개하는가.
- 현지 데이터의 소유권, 국외 이전과 이익 공유 조건은 무엇인가.
- 조달 공급자와 후원 기업의 이해충돌을 누가 심사하는가.
- 환경 부담과 운영비를 어느 기관이 얼마나 오래 책임지는가.
- 성과가 없거나 피해가 반복될 때 사업을 수정·중단하는 절차가 있는가.

건물, 회의와 GPU 수량은 활동 지표입니다. 현지 언어 서비스의 품질, 교육을 마친 운영 인력의 잔류, 컴퓨트 이용의 지역 분포와 피해 구제는 결과 지표입니다.

이 질문들을 한국에 그대로 돌려 보면 어떨까요? 한국은 HBM 메모리와 반도체 공급망에서 강한 산업 기반을 갖고 있습니다. 다만 시장 지위가 국제기구의 정당성이나 운영 역량을 자동으로 보장하지는 않습니다. 허브 역할은 회원의 신뢰, 투명한 재원과 공동 의사결정으로 얻어야 합니다.

한국의 디지털 인프라 구축 경험은 참고 사례가 될 수 있습니다. 다만 한국의 행정·재정·산업 조건을 다른 나라에 그대로 적용할 수는 없습니다. 성공 사례와 함께 비용, 지역 격차와 실패도 내놓을 때 비로소 참고가 되지요.

APEC 합의는 서로 다른 이해관계를 한 문서에 담은 외교 성과입니다. 그러나 공동 서명이 후속 예산이나 집행 합의를 보장하지는 않습니다.

센터 유치 경쟁, 장기 운영비와 강대국 간 갈등은 제약입니다. '자체 비용 운영'의 기간과 상한, 다른 회원의 분담 방식이 불명확하면 국제 공공재가 유치국 사업으로 축소될 수 있습니다.

이쯤에서 처음의 1906년으로 돌아가 볼까요?

1906년에 IEC가 만들어졌을 때 전기 표준은 국제 교역과 안전에 필요한 공통 규격을 만드는 데 기여했습니다. 전기의 보편화는 표준 하나가 아니라 발전·송배전 투자, 공공정책, 요금 제도와 지역 사업이 함께 만든 결과입니다. 특정 국가와 기업의 시장 지배도 사라진 것은 아니지요.

AI 국제 협력의 틀은 아직 만들어지는 중입니다. APEC 합의, UN 허브 구상과 글로벌 거버넌스 대화는 서로 다른 단계와 권한을 가집니다. 하나의 큰 서사로 합치기보다 각 약속이 예산, 운영 규칙과 결과로 이어지는지를 따로 추적해야 합니다.

한 나라의 농촌전력 사업과 국제 AI 협력은 권한과 재정 구조가 크게 다릅니다. 비유가 보여주는 공통점은 시장성이 낮은 지역에 공공투자가 필요하다는 정도입니다.

1906년 런던에서 IEC를 세운 나라는 손에 꼽을 만큼 적었습니다. 그래도 오늘 우리는 전기 단위와 안전 규격을 국제 표준으로 맞춰 쓰는 일을 당연하게 여기지요. 플러그 모양은 아직 나라마다 달라도 말입니다. 지능 에너지의 국제 협력이 언젠가 그렇게 당연해질지는, 지금 테이블에 앉지 못한 나라들이 자국의 언어와 필요에 맞는 역량을 갖고 의사결정에 참여하게 되는지에 달려 있겠지요.



1. "APEC Artificial Intelligence Initiative (2026-2030)" (APEC, 2025, https://www.apec.org/meeting-papers/leaders-declarations/2025/2025-apec-leaders-%2Dgyeongju-declaration/apec-artificial-intelligence-%28ai%29-initiative-%282026-2030%29) — 21개 회원 경제권의 채택 내용과 한국의 아시아태평양 AI 센터 구상을 명시한다.

2. "NVIDIA, South Korea Government and Industrial Giants Build AI Infrastructure and Ecosystem" (NVIDIA Newsroom, 2025.10.30, https://nvidianews.nvidia.com/news/south-korea-ai-infrastructure) — 정부와 기업의 구축 계획 합계이며 납품 완료 수량이 아니다.

3. "UN AI 허브 유치지원 TF" (대한민국 정책브리핑, 2026.3, https://m.korea.kr/multi/visualNewsView.do?call_from=rsslink&newsId=148960633) — 한국 정부의 유치 추진 단계.

4. 글로벌 AI 허브 입지 공모 관련 보도 (파이낸셜뉴스, 2026.10.12, https://www.fnnews.com/news/202610121007462045) — 2026.5.21 비전 선포(9개 국제기구·5개 다자개발은행 참여), 10.14 입지 공청회, 연내 입지 확정·2027년 개소 목표.

5. "AI for Good Global Summit 2017" (ITU, https://www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx) — 이니셔티브의 출발 시점과 목적.

6. Global Dialogue on AI Governance (United Nations, https://www.un.org/global-dialogue-ai-governance/en) — 첫 대화의 2026년 7월 개최 기록.

7. 「7/24(금) 샌프란시스코 AI 선언」 (대통령실, 게시 2026.7.25, https://www.president.go.kr/speeches/hHXHeToZ) — 글로벌 AI 허브 이니셔티브, 수평적 협력, 각국 언어·문화·산업 환경에 맞는 AI 개발·활용 언급.

8. "Strengthening AI Foundations" 및 Digital Progress and Trends Report 2025 (World Bank, https://www.worldbank.org/en/news/factsheet/2025/11/21/strengthening-ai-foundations-emerging-opportunities-for-developing-countries, https://www.worldbank.org/en/publication/dptr2025-ai-foundations/report) — 연결성·컴퓨팅·맥락·역량의 네 기반과 국가 간 집중 문제.

9. "AI market projected to hit $4.8 trillion by 2033" 및 AI 격차 자료 (UN Trade and Development, https://unctad.org/news/ai-market-projected-hit-48-trillion-2033-emerging-dominant-frontier-technology, https://unctad.org/press-material/ais-48-trillion-future-un-trade-and-development-alerts-divides-urges-action).

천하삼분지계: AI 3강 전략

삼국지의 천하삼분지계를 떠올려 봅시다.

유비가 제갈량을 세 번이나 찾아간 이유는 단순히 지혜로운 사람이 필요해서가 아니었습니다. 강대국 사이에서 살아남는 길을 찾아야 했기 때문입니다. 제갈량의 답은 명쾌했지요. 위나라와 오나라의 판을 그대로 따라가지 마라. 촉나라만의 강점에 집중하라. 천하를 셋으로 나누어 한 축을 차지하라.

2026년, 한국의 AI 전략이 비슷한 질문 앞에 서 있습니다. 미국과 중국이라는 두 강대국이 AI 패권을 다투는 판에서, 한국은 어떤 자리를 차지할 것인가. 정부는 AI 3대 강국 진입을 목표로 내세웠고, 이 장은 그 전략을 천하삼분지계라는 비유로 읽어 봅니다.[1]

두 강대국 사이에서 찾는 제3의 길

미국과 중국의 차이를 '민간 대 정부' 하나로 설명하면 중요한 부분을 놓칩니다. 미국에도 반도체 보조금과 수출통제, 연방 조달과 안전 규제가 있고, 중국에도 경쟁하는 민간 모델 기업과 오픈웨이트 생태계가 있습니다. 두 나라 모두 자본, 컴퓨트, 인재, 산업 확산과 규율을 서로 다른 비율로 결합합니다.

미국은 거대한 민간 자본과 프론티어 기업이 강점이지만 시장 집중과 지역 격차가 과제입니다. 중국은 국가 차원의 인프라 조정과 제조 확산이 빠르지만 시민 권리와 독립적 검증을 둘러싼 우려가 큽니다. 이것은 국가 전체를 한 단어로 판정한 결과가 아니라, 한국이 두 정책 조합에서 무엇을 배우고 무엇을 경계할지 살피기 위한 비교 틀입니다.

한국은 어느 한쪽을 복제하기보다 자본·컴퓨트·인재·확산·신뢰의 다섯 축에서 자기 조건에 맞는 조합을 찾아야 합니다.

미국과 중국 어느 한쪽을 그대로 따르지 않는 중견국의 경로도 논의되고 있습니다. 채텀하우스는 공급망 특화, 강대국과의 전략적 제휴, 국가 간 역량 공유, 다변화를 중견국의 선택지로 제시합니다.[2]

한국에는 이 네 가지 중 첫 번째 길이 특히 선명합니다. 반도체라는 확실한 공급망 지위가 있으니까요. 하지만 그것만으로는 부족합니다. 칩을 만드는 나라에서 멈추면, 결국 다른 나라의 AI 생태계에 부품을 납품하는 위치에 머무르게 됩니다. 제갈량이 형주와 익주라는 땅을 확보하되 거기에 머무르지 않고 자기만의 축으로 올라서려 했듯이, 한국도 반도체 너머의 전략이 필요합니다.

한국이 검토할 수 있는 두 가지 기준은 민주적 AI와 포용적 AI입니다. 이름을 붙이는 것보다 실제 제도와 성과로 증명하는 일이 먼저입니다.

민주적 AI는 시민의 권리 보호, 독립된 감시, 공개된 근거와 이의제기 절차를 갖춘 AI를 뜻합니다. 한국 AI 기본법은 발전과 신뢰를 함께 내세우지만, 법의 이름만으로 이 조건이 충족되지는 않습니다. 민주적 AI의 성패는 감독기관의 독립성, 집행 인력, 제재와 피해 구제가 실제로 작동하는 데서 갈립니다.

포용적 AI는 기업 규모, 지역, 소득, 연령과 장애에 따라 이용 기회와 성과가 크게 달라지지 않도록 하는 기준입니다. 도입률 하나를 올리는 것만으로는 부족합니다. 지속 사용률, 총비용, 접근성, 오류와 구제 결과를 집단별로 봐야 합니다.

민주적 AI와 포용적 AI는 한국이 내세울 수 있는 방향입니다. 그러나 이름만으로 차별화되지는 않습니다. 실행 여부는 이의제기권, 독립 평가, 중소기업 접근성, 지역별 성과 같은 지표에서 드러납니다.

하드웨어, 소프트웨어, 신뢰라는 세 기둥

그렇다면 이 전략은 구체적으로 무엇으로 구성될까요. 세 가지 기둥이 있습니다.

하드웨어: 공급망의 심장부를 잡다

SK하이닉스와 삼성전자는 세계 HBM(고대역폭 메모리) 시장의 핵심 공급자입니다. 시장점유율은 조사기관과 분기, 제품 세대에 따라 달라지므로 하나의 고정 비율로 보기는 어렵습니다. 그래도 AI 가속기에 필요한 HBM 공급망이 한국의 분명한 산업 자산이라는 점은 변하지 않습니다.

정부는 민관 투자와 국가 AI 컴퓨팅 센터, GPU 확보 계획을 추진하고 있습니다. 발표된 목표와 계약 물량, 실제 설치·가동 물량은 따로 공개될 때 의미가 있습니다. HBM 공급 여건도 한국에 기회가 될 수 있지만, 시장 전망을 확정된 전략 자산처럼 계산하면 곤란하지요.

소프트웨어: 주권적 AI를 확보하다

칩만으로는 부족합니다. 칩 위에서 돌아가는 모델이 있어야 합니다. 정부 지원 사업은 두 차례 단계평가를 거쳐 2026년 8월부터 세 팀 체제로 운영되고 있습니다.[3] 모델의 공개 범위와 일정은 팀별 발표보다 정부의 평가·지원 조건과 실제 공개물이 기준입니다.

이것은 기술 자존심만의 문제가 아닙니다. 한국어와 국내 법·제도를 다루는 모델의 데이터 출처와 평가 기준, 오류 책임을 국내에서 검증할 역량이 필요합니다. 반드시 하나의 국산 모델만 써야 한다는 뜻은 아닙니다. 독자 모델, 공개 모델, 해외 모델을 과업별로 비교하고 교체할 수 있는 능력도 주권의 일부입니다.

이 접근은 빅테크와 모든 영역에서 정면 대결하기보다 한국의 IT 인프라와 산업 데이터, 평가 역량을 결합하는 전략으로 읽을 수 있습니다. 성과는 '소버린'이라는 이름보다 공급자를 바꿀 수 있는지, 국내에 운영 역량이 남는지에서 드러나겠지요.

이 소버린 AI 역량이 결국 지능을 수입만 하는 나라를 넘어 수출하는 나라로 가는 발판이라는 것은, 제2부 「지능을 수출하는 나라」에서 토큰 자급률이라는 지표로 다시 다룹니다.

신뢰: 검증과 구제를 운영하다

신뢰는 국가가 선언해 얻는 자산이 아닙니다. 모델·데이터 변경을 추적하고, 독립기관이 평가하며, 사고와 집단별 오류를 공개하고, 시민의 이의제기를 처리한 기록이 쌓여야 합니다. 국제표준 참여도 제안 수보다 시험 결과와 상호인정 범위로 평가해야 하지요.

하드웨어와 소프트웨어에 검증·구제 역량을 더하는 것이 세 번째 기둥입니다.

민주적 AI는 경쟁력이 될 수 있을까

"규제는 혁신을 죽인다"는 말을 자주 듣습니다. 정말 그럴까요.

GDPR을 생각해 봅시다. 이 법은 세계 여러 개인정보보호 제도에 영향을 줬지만, 준수 비용과 혁신 효과를 둘러싼 평가는 엇갈립니다. 규제가 언제나 경쟁력이 되거나 언제나 혁신을 막는다고 단정할 수는 없습니다. 경쟁력으로 이어지려면 요구사항이 명확하고, 시험·인증 비용이 감당 가능하며, 결과가 다른 시장에서도 인정돼야 합니다.

신뢰가 채택에 영향을 줄 수 있지만 가격, 성능, 전환비용과 규제도 함께 작용합니다.

이것을 구체적으로 생각해 봅시다. 병원에서 AI가 진단을 내린다고 합시다. 환자가 그 진단을 믿으려면 무엇이 필요할까요. AI의 정확도도 중요하지만, 그 AI가 어떤 데이터로 학습되었는지, 어떤 기준으로 판단했는지, 오류가 있을 때 누가 책임지는지가 명확해야 합니다. 이것이 민주적 AI의 인프라입니다. 투명성, 설명 가능성, 책임 소재. 이런 것들이 갖춰진 AI만이 사회 전체에 채택될 수 있습니다.

중국, 미국, 한국 모두 중앙·지방 정부, 법원, 기업과 시민사회의 접근이 섞여 있습니다. 국가를 '통제', '혁신', '균형'이라는 한 단어로 배열하면 실제 집행 차이를 놓칩니다.

한국의 AI 기본법이 경쟁력으로 이어지려면 요구사항이 명확하고 시험·인증 비용이 감당 가능해야 하며, 평가 결과가 다른 시장에서도 인정돼야 합니다. 인증 뒤 사고가 늘거나 피해 구제가 지연된다면 '신뢰'라는 홍보 문구는 오히려 약점이 됩니다.

이런 신뢰를 한 나라 혼자 쌓아야 할까요? 동아시아포럼은 아시아 중견국들이 공동의 정책 프레임워크와 인프라 협력을 통해 미국·중국 중심의 AI 질서에서 선택권을 넓혀야 한다고 제안합니다.[4] 외부 인프라를 쓰더라도 계약과 데이터, 전환 권한을 공동으로 키우자는 논지입니다.

해외 클라우드를 쓴다는 이유만으로 주권이 사라지는 것은 아닙니다. 데이터 위치, 계약 종료와 이전 가능성, 공급자 교체, 감사권과 비상시 운영을 통제할 수 있는지가 더 구체적인 기준입니다.

한국은 반도체와 제조 기반을 협력 자산으로 제안할 수 있습니다. 다만 리더 역할은 산업 규모가 아니라 공동 규칙을 따르고 비용과 의사결정권을 나누는 방식으로 인정받아야 합니다.

한국은 뜻이 맞는 나라들과 평가 규격, 사고 통지, 데이터 이전과 이의제기 절차를 상호인정하는 협력을 시도할 수 있습니다. 어느 나라의 법을 그대로 수출하기보다 공동 시험과 상호 감사를 만드는 편이 현실적입니다.

속도와 방향, 그리고 세 번째 나라

제갈량의 전략에서 자주 간과되는 요소가 하나 있습니다. 타이밍입니다. 제갈량은 유비에게 "지금이 아니면 기회가 없다"고 말했지요. 위나라와 오나라가 자리를 잡기 전에 촉나라의 기반을 닦아야 한다고요.

한국의 AI 전략에서도 시간은 중요한 변수입니다. HBM 수급 전망은 조사 시점마다 달라지고, 파운데이션 모델 경쟁과 글로벌 AI 표준도 빠르게 바뀝니다. 그래서 공급망 우위는 한 번 확인한 점유율로 끝나지 않고, 분기마다 다시 확인해야 하는 숫자이지요.

행동계획의 99개 과제에는 2026년 착수·수립 목표가 다수 포함돼 있습니다.[1] 착수율은 행정 속도를 보여주지만 성능, 접근성과 권리 보호를 증명하지는 않습니다. 빠른 사업일수록 기준선과 중단 조건을 먼저 공개해야 합니다.

하지만 속도에는 방향이 있어야 합니다. 미국의 민간 혁신과 중국의 국가 동원력을 그대로 복제하기보다, 한국의 산업 기반과 민주적 제도에 맞는 조합을 찾아야 합니다.

다시 융중의 제갈량에게 돌아가 보면, 천하삼분지계의 핵심은 결국 하나입니다. 자기만의 강점에 집중하는 것.

한국의 후보 강점에는 메모리 반도체 공급망, 제조 기반, 디지털 인프라와 민주적 제도가 있습니다. 반면 프론티어 모델 자본과 클라우드 규모, 지역·기업 간 AI 격차는 약점입니다. 전략은 강점의 선언보다 이 약점을 어떤 지표로 줄일지에서 평가되어야 합니다.

이 강점들을 전략으로 엮으려면 공급망 점유율, 모델 순위, 법 제정을 서로 다른 성과로 구분해야 합니다. G3라는 단일 순위보다 아래 지표를 함께 공개하는 편이 낫습니다.

축	검증 지표
하드웨어	실제 가동 컴퓨트, 전력·물 사용, 공급자 집중도
소프트웨어	과업별 품질·비용, 데이터 적법성, 모델 교체 가능성
산업 확산	기업 규모·지역별 지속 사용률과 생산성, 재작업
신뢰	독립평가, 사고 공개, 인간 심사 처리기간과 인용률
국제협력	상호운용 시험, 상호인정, 참여국의 의사결정권

제갈량의 전략이 결국 완성되지 못했다는 역사적 사실을 떠올리는 분도 있을 겁니다. 맞습니다. 촉나라는 끝내 천하통일에 이르지 못했지요. 하지만 제갈량의 전략 자체가 틀렸던 것은 아닙니다. 실행의 문제였습니다. 자원이 부족했고, 인재가 이어지지 않았고, 내부의 결속이 흔들렸지요.

한국의 AI 전략도 성패를 아직 단정할 수 없습니다. 발표된 예산과 사업이 실제 역량으로 이어지고, 세 팀의 경쟁과 탈락한 팀의 산출물까지 공통 기반으로 남으며, AI 기본법이 신뢰의 인프라로 작동하고, AI의 혜택이 중소기업과 시민에게까지 닿을 때 비로소 판단할 수 있습니다.

민주적 AI도 마찬가지입니다. 정보공개, 독립감사, 시민 참여와 법원의 구제가 AI 사업에서 실제로 돌아가야 이름값을 하지요.

제갈량이 융중에서 천하삼분지계를 펼쳤을 때, 그것은 지도 위에 그은 몇 개의 선이었습니다. 그 선이 나라가 되기까지는 수많은 해의 실행이 필요했지요. 한국의 AI 3강 전략도 지금은 지도 위의 선입니다. 위표의 칸이 매년 같은 기준으로 채워지고, 성과가 없는 칸을 고치는 해가 거듭될 때, 그 선은 비로소 땅이 될 겁니다.



1. 「대한민국 인공지능 행동계획(인공지능 기본계획 2026~2028)」 (국가인공지능전략위원회, https://www.aikorea.go.kr/upload/contents/contents_01.html) — AI 3대 강국 목표와 99개 과제의 공식 원문.

2. "How middle powers can weather US and Chinese AI dominance" (Chatham House, 2026, https://www.chathamhouse.org/2026/02/how-middle-powers-can-weather-us-and-chinese-ai-dominance/summary).

3. 「독자 AI 파운데이션모델 프로젝트 2차 단계평가 결과」 (과학기술정보통신부, 2026.8.18, https://www.korea.kr/briefing/policyBriefingView.do?newsId=156774781) — 업스테이지·SK텔레콤·LG AI연구원 3차 진출.

4. "Asia needs an AI third way" (East Asia Forum, 2025.4.13, https://eastasiaforum.org/2025/04/13/asia-needs-an-ai-third-way/).

모두를 위한 지능 에너지

이 책은 증기기관에서 시작했습니다. 새로운 에너지가 등장하고, 인프라와 응용이 깔린 뒤, 일과 사회의 구조가 바뀌는 과정을 따라왔지요.

AI도 비슷한 경로를 걷고 있습니다. 파운데이션 모델과 데이터센터가 생기고, 에이전트와 산업별 응용이 그 위에 올라옵니다. 하지만 역사가 저절로 반복되지는 않습니다. 전기도 발전소가 생겼다는 이유만으로 모든 집에 들어간 것이 아니었지요. 표준과 송전망, 공공투자와 요금 제도가 함께 움직여야 했습니다.

지능 에너지가 더 많은 사람의 것이 되려면 무엇이 필요할까요?

넓어진 문, 그리고 남는 문턱

오픈웨이트 모델은 접근의 문을 넓혔습니다. DeepSeek R1, Meta의 Llama, Alibaba의 Qwen 같은 모델은 연구자와 기업이 자기 환경에서 실행하고 고쳐 쓸 선택지를 열어 주었지요. 그 비용과 공개 범위를 어떻게 읽어야 하는지는 제1부에서 이미 짚었습니다.

온디바이스 AI도 다른 길을 냅니다. 가벼운 처리를 휴대전화나 PC에서 수행하면 인터넷 연결이 불안정한 곳에서도 일부 기능을 쓸 수 있고, 원본 데이터를 외부 서버로 보내지 않을 수도 있습니다. 무거운 일은 클라우드에 맡기고 민감한 일은 기기에 남기는 하이브리드 구조도 가능하지요.

그렇다고 모델을 내려받을 수 있는 사람이 모두 같은 혜택을 얻는 것은 아닙니다. 실행할 장비와 전력, 현지 언어 데이터, 기술 인력, 장애를 해결할 조직이 필요합니다. 온디바이스 처리도 앱 권한과 백업, 로그까지 저절로 안전하게 만들어 주지는 않습니다.

시골 마을에 전신주 하나가 섰다고 집집마다 불이 켜지지는 않았던 것과 같습니다. 집 안까지 전선을 끌고, 계량기를 달고, 고장 나면 고쳐 줄 사람이 있어야 했지요. 접근성도 그런 조건 전체를 가리킵니다.

돈 문제는 어떨까요? 토큰 단가는 빠르게 내려가고 있습니다. 하지만 같은 품질끼리 비교해야 한다는 점, 가격이 내리면 사용량이 늘어 총비용이 오를 수도 있다는 점은 제1부 「지능의 단가가 빠르게 낮아지고 있다」에서 보았지요. 보편 접근을 설계할 때도 API 단가보다 일 하나를 끝내는 데 드는 비용, 그 지역의 통신·전력 여건, 사용자가 들이는 시간을 함께 봐야 합니다.

무료 계정 하나를 나눠준다고 지능 격차가 사라지지 않는 이유가 여기 있습니다. 질문을 만드는 능력, 답을 검증할 시간, 잘못된 결정에 이의를 제기할 통로가 없으면 명목상의 접근과 실제 이용 사이에 간격이 남습니다.

이 책 전체를 두 단어로 줄이면 무엇이 될까요? 이 책에서 다룬 AI 네이티브 전환은 모델과 컴퓨팅, 데이터와 인력을 갖춰 실제 문제를 풀 수 있는 생산 능력에 관한 이야기였습니다. AI 기본사회는 그 능력이 누구에게 어떤 조건으로 돌아가는지를 묻는 접근과 분배의 이야기이지요.

둘 중 하나만으로는 모자랍니다. 생산 기반 없이 보편 서비스를 약속하면 외부 사업자의 가격과 정책 변화에 흔들립니다. 반대로 강한 모델과 데이터센터를 갖추고도 중소기업, 고령자, 장애인, 농어촌 주민이 쓰지 못하면 기술적 성과가 사회의 역량으로 이어지지 않습니다.

두 목표는 때로 부딪힙니다. 빠른 보급을 위해 하나의 서비스를 전국에 깔면 선택권과 지역 자율성이 약해질 수 있습니다. 국내 공급망을 지키려고 시장을 닫으면 비용이 오르고 기술 교류가 줄 수 있지요. 그러니 생산이 먼저냐 분배가 먼저냐를 정하기보다, 투자할 때마다 함께 물어보는 편이 낫습니다. 이 사업이 국내에 무엇을 남기는지, 혜택에서 빠지는 사람은 누구이고 왜 빠지는지, 그리고 그 사람이 자동화된 판단에 맞서 다시 심사받을 길이 있는지를요. 공급자를 바꾸고 사고 뒤 책임을 묻는 장치는 제7부의 앞 장들에서 여러 번 보았으니, 여기서는 이 세 질문만 들고 가겠습니다.

보편화는 선택의 결과다

세계에는 데이터센터 용량이 거의 없는 지역이 있고, 자국 언어로 충분한 성능을 내는 모델을 갖지 못한 공동체가 많습니다. 국가 AI 전략을 세우는 회의에 보낼 사람조차 부족한 나라도 있지요. 이 격차를 시장의 시간에만 맡기면 수익이 큰 언어와 도시, 기업부터 혜택을 얻습니다.

공공의 역할은 모든 모델을 직접 만드는 데서 끝나지 않습니다. 공공 데이터의 품질을 높이고, 학교와 도서관에서 검증할 수 있는 도구를 제공하고, 중소기업이 공동 컴퓨팅을 쓰게 하며, 장애인과 지역 언어 사용자가 서비스 설계에 참여하도록 만드는 일도 포함됩니다. 자동 판단에 대한 설명과 이의제기, AI 없이 서비스를 받을 선택권도 접근권의 일부이지요.

기업에는 다른 몫이 있습니다. 낮은 가격만 약속하지 말고 서비스가 끝날 때 데이터를 어떻게 돌려주는지, 모델이 바뀌면 성능을 다시 검증하는지, 사용자가 자동화 여부를 알 수 있는지를 계약과 제품에 남기는 일입니다.

개인도 혼자 감당할 필요는 없습니다. 중요한 답은 동료와 함께 확인하고, 오류를 기록하고, AI를 쓰지 않는 사람도 알아들을 말로 결과를 설명하면 됩니다. 활용 능력은 프롬프트 요령보다 이런 검증 습관에서 더 오래 가지요.

돌아보면 이 책을 끌고 온 것은 하나의 비유였습니다. '지능은 에너지다'라는 비유는 AI가 생산되고 유통되며 여러 응용으로 바뀌는 구조를 보여 줍니다. 그 비유가 어디서 멈추는지는 제1부 「비유의 한계」에서 미리 짚어 두었지요. 품질이 균일하지 않고, 사용자의 의도를 해석하다 틀리기도 하며, 저작권과 개인정보처럼 전기에는 없던 권리가 얽혀 있다는 점 말입니다.

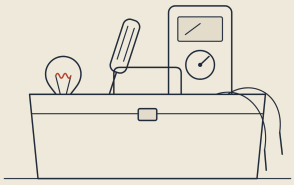
그래서 지능 에너지를 모두에게 보내는 일은 콘센트를 늘리는 일보다 복잡합니다. 더 좋은 모델과 더 싼 컴퓨팅에 더해, 누가 통제하고 누가 검증하며 누가 이의를 제기할 수 있는지까지 설계해야 하니까요.

마지막으로, 이 책이 시작된 자리로 돌아가 보고 싶습니다. 서문에서 저는 1981년, 키보드를 두드리자 프린터에서 글자가 찍혀 나오던 순간을 이야기했습니다. 기계와 사람 사이에 처음 '언어'가 생긴 것 같았지요. 세운상가에서 부품을 사 모으던 소년에게는 세상이 한 뼘 넓어지는 순간이었습니다.

그로부터 마흔 해가 넘게 지나, 그 언어가 이제 전기처럼 집집마다 흘러가려 합니다. 이 책은 그 흐름을 발전소와 송전망, 공장과 관청, 법과 일자리의 자리에서 차례로 따라가 보았습니다. 낙관할 이유도, 조심할 이유도 함께 적었지요.

책을 덮기 전에 작은 부탁을 하나 드리고 싶습니다. 내일 AI에게 무언가를 물을 때, 답을 그대로 옮기지 말고 근거 하나만 직접 확인해 보세요. 그리고 그 방법을 옆자리의 누군가에게, 기술이 낯선 가족에게 한 번 설명해 주세요. 전기가 모든 집에 닿은 것도 결국 누군가 한 집 한 집 전선을 이어 준 덕분이었습니다.

1981년 키보드 앞에서 소년이 느꼈던 그 두근거림이, 이번에는 훨씬 더 많은 사람에게 닿기를 바랍니다.



부록

부록

부록 A. 증거 점검 40문 + 원칙 매트릭스

AI 전환은 도구 도입만으로 이루어지지 않습니다. 조직의 문화, 프로세스, 역량과 책임 체계가 함께 바뀌어야 하지요. 이 점검표는 조직의 등급을 매기기보다, 주장과 실제 운영 사이에서 증거가 비어 있는 곳을 찾는 데 씁니다.

이 부록은 다섯 영역의 문항 40개로 구성됩니다. 각 문항을 증거 있음 / 부분 증거 / 증거 없음 / 해당 없음 가운데 하나로 표시합니다. '있다'고 답할 때는 승인 문서, 운영 로그, 계약, 훈련 결과처럼 다른 사람이 확인할 수 있는 자료를 한 개 이상 적습니다. '관련 원칙' 열은 각 질문이 제4부의 어느 관점과 연결되는지 보여줍니다.

이 표는 검증된 성숙도 진단 도구가 아닙니다. 조직 간 순위나 투자 타당성을 산출하는 용도로도 쓸 수 없습니다. '해당 없음'은 낮은 점수가 아니며, 왜 적용되지 않는지 근거를 남겨야 합니다.

사용법

· 개인 점검: 각 문항의 상태와 증거 위치를 적습니다
· 팀 점검: 평균을 내지 말고 답이 갈린 문항과 증거 차이를 먼저 토론합니다
· 조직 점검: 부서별 '증거 없음' 비율보다 고위험 업무의 공백과 집단별 차이를 먼저 봅니다

영역 1. AI 이해와 리터러시 (20문)

#	질문	관련 원칙
1	파운데이션 모델과 전통적 소프트웨어의 차이를 동료에게 설명할 수 있다	AI-First 사고
2	LLM의 할루시네이션이 왜 발생하는지, 어떻게 대응하는지 알고 있다	안전과 신뢰
3	프롬프트를 작성할 때 역할, 맥락, 제약 조건을 구분하여 설계한다	호모 프롬프트
4	MCP와 A2A 프로토콜의 차이와 용도를 이해하고 있다	에이전트 협업
5	AI 에이전트와 챗봇의 차이를 설명할 수 있다	에이전트 협업
6	오픈소스 모델과 상용 모델의 장단점을 비교할 수 있다	AI-First 사고
7	RAG(검색 증강 생성)의 작동 원리를 이해하고 있다	AI-First 사고
8	토큰, 컨텍스트 윈도우, 온도 등 기본 개념을 설명할 수 있다	호모 프롬프트
9	AI의 확률적 특성이 결정적 시스템과 어떻게 다른지 안다	안전과 신뢰

#	질문	관련 원칙
10	멀티모달 AI가 텍스트, 이미지, 음성을 어떻게 처리하는지 이해한다	AI-First 사고
11	AI 모델의 학습과 추론의 차이를 설명할 수 있다	AI-First 사고
12	파인튜닝과 프롬프트 엔지니어링의 차이와 적용 시점을 안다	호모 프롬프트
13	AI 편향(bias)이 발생하는 원인과 완화 방법을 알고 있다	안전과 신뢰
14	온디바이스 AI와 클라우드 AI의 트레이드오프를 이해한다	AI-First 사고
15	AI가 잘하는 일과 못하는 일을 구분할 수 있다	AI-First 사고
16	데이터 프라이버시와 AI 활용 사이의 균형점을 고려한다	안전과 신뢰
17	생성형 AI의 저작권 이슈를 이해하고 있다	안전과 신뢰
18	AI 벤치마크의 의미와 한계를 알고 있다	AI-First 사고
19	소버린 AI의 개념과 왜 중요한지 설명할 수 있다	거시 전략
20	"지능은 에너지다"라는 비유의 의미를 설명할 수 있다	AI-First 사고

영역 2. 업무 적용과 실행 (5문)

#	질문	관련 원칙
21	매일 업무에서 AI 도구를 최소 1개 이상 사용한다	일상 통합
22	반복적인 업무를 AI로 자동화한 경험이 있다	프로세스 재설계
23	AI의 출력을 검증하는 나만의 프로세스가 있다	안전과 신뢰
24	업무별로 적합한 AI 모델을 선택할 수 있다	AI-First 사고
25	AI를 활용한 보고서/문서 작성 워크플로우가 있다	일상 통합

영역 3. 조직과 문화 (5문)

#	질문	관련 원칙
26	우리 팀에는 AI 도구 사용에 대한 가이드라인이 있다	거버넌스
27	AI 실험이 실패해도 비난받지 않는 문화가 있다	조직 문화
28	AI 도입에 대한 경영진의 지지가 명확하다	리더십
29	AI 관련 교육/학습 기회가 정기적으로 제공된다	인재 양성
30	부서 간 AI 활용 사례를 공유하는 채널이 있다	조직 문화

영역 4. 전략과 비전 (5문)

#	질문	관련 원칙
31	우리 조직의 AX 전환 로드맵이 존재한다	AX 전략
32	AI가 우리 비즈니스 모델을 어떻게 바꿀 수 있는지 논의한 적 있다	AI-First 사고
33	경쟁사의 AI 활용 현황을 모니터링하고 있다	거시 전략
34	AI 투자에 대한 ROI 측정 기준이 있다	AX 전략
35	3년 후 AI가 우리 산업을 어떻게 바꿀지 시나리오가 있다	거시 전략

영역 5. 안전과 윤리 (5문)

#	질문	관련 원칙
36	AI 사용 시 개인정보 보호 정책을 준수하고 있다	안전과 신뢰
37	AI 출력의 편향 가능성을 인지하고 점검한다	안전과 신뢰
38	AI 기본법의 주요 의무사항을 알고 있다	거버넌스
39	AI가 생성한 콘텐츠를 그대로 사용하지 않고 검토한다	안전과 신뢰
40	AI 관련 보안 사고 대응 절차가 있다	거버넌스

증거 상태 기록표

문항	상태	확인한 증거·위치	마지막 확인일	담당자
예: 40번 사고 대응	증거 없음	승인 절차와 최근 모의훈련 기록을 찾지 못함	20__-__-__	보안 책임자
	있음 / 부분 / 없음 / 해당 없음			
	있음 / 부분 / 없음 / 해당 없음			
	있음 / 부분 / 없음 / 해당 없음			

영역별 총점이나 전체 등급은 계산하지 않습니다. 리터러시 문항이 많다고 그 영역이 안전·거버넌스보다 더 중요하다는 뜻은 아니기 때문입니다. 대신 피해 규모가 큰 업무, 여러 사람이 다르게 답한 문항, 증거가 오래된 문항 순으로 검토합니다.

증거 공백을 90일 행동으로 바꾸기

‘증거 없음’을 모두 한꺼번에 고치려 하지 말고, 실패했을 때 피해가 큰 영역과 여러 사람이 다르게 답한 영역부터 다룹니다. 아래 표를 팀 회의에서 한 줄씩 채워 보세요.

약한 영역·문항	현재 답을 뒷받침하는 증거 문서	담당자	90일 행동	완료 증거	재검토일
예: 40번 사고 대응	대응 절차 없음, 최근 모의훈련 기록 없음	보안 책임자	AI 사고 연락망과 차단 절차 작성, 모의훈련 1회	승인된 절차, 훈련 결과와 개선 항목	20__-__-__

표시가 ‘있음’으로 바뀌었는지보다 절차가 실제로 작동했는지를 확인하세요. 문서만 있고 실행 기록이 없으면 ‘부분 증거’로 남기는 편이 후한 자기평가보다 유용합니다.

원칙 매트릭스

원칙별 문항 수는 질문을 찾아가기 위한 색인입니다. 문항 수를 가중치나 점수로 쓰지 않습니다.

원칙	관련 문항	검토할 증거
AI-First 사고	1, 6, 7, 9~11, 14, 15, 18, 20, 24, 32	모델 선택 기록, 기준선과 평가 결과

원칙	관련 문항	검토할 증거
안전과 신뢰	2, 13, 16, 17, 23, 36, 37, 39	영향평가, 검토·정정 기록, 사고 로그
호모 프롬프트	3, 8, 12	재현 가능한 지시와 결과 검증 기록
거버넌스	26, 38, 40	승인된 정책, 책임자, 대응 훈련
거시 전략	19, 33, 35	선택지 비교와 시나리오 재검토 기록
에이전트 협업	4, 5	권한·인증·로그와 실패 처리
일상 통합	21, 25	실제 사용 로그와 업무 결과
조직 문화	27, 30	실패·사례 공유 기록과 후속 조치
AX 전략	31, 34	로드맵, 비용·편익과 중단 기준
프로세스 재설계	22	변경 전후 흐름과 재작업률
인재 양성	29	교육 뒤 과업·품질·고용 변화
리더십	28	예산뿐 아니라 책임과 검토 참여 기록

AI 리스크 한 장 요약

이 책 곳곳에 흩어져 있는 위험 요소를 경영 회의나 이사회 자료로 바로 옮길 수 있도록 한자리에 모았습니다. 각 항목의 자세한 근거와 사례는 표시된 장에서 확인하세요.

리스크	무엇이 문제인가	어디서 다루나
벤더·시장 집중	소수 하이퍼스케일러·컴퓨트 공급자에 의존이 쏠리며 반독점 조사·공급 흔들림 위험	제2부 7장 「미국, 지능 발전소를 짓다」
인프라 거품·금융 리스크	대규모 AI 설비투자의 순환금융 구조에 대한 BIS 등의 경고	제2부 7장 「미국, 지능 발전소를 짓다」
재정·형평성	소수 대기업 집중과 크라우딩아웃, 기업 규모·지역별 접근 격차	제2부 9장 「대한민국의 대도약」, 제5부 8장 「K-제조」, 제7부 7장 「대한민국 AI 행동계획」
대외 의존	전력·GPU·토큰의 해외 의존과 토큰 자급률 문제	제2부 11장 「지능을 수출하는 나라」
할루시네이션·검증 격차	확률적 오류를 사후에 걸러내는 검증 역량 부족	제5부 1장 「코드를 쓰는 AI, 코드를 검증하는 인간」, 제7부 3장 「법, 노동, 안전」

리스크	무엇이 문제인가	어디서 다루나
자율 공격·보안	에이전트가 에이전트를 공격하는 자율 공방과 Mythos 이후 보안 지형	제6부 10장 「자율 AI 해킹과 방어」
에이전트 거버넌스	다수 에이전트를 개별 통제가 아니라 규칙으로 설계해야 하는 과제	제2부 13장 「초지능(ASI)으로 가는 네 가지 길」, 제6부 11-1장 「하네스 엔지니어링」
자기개선·초지능 통제	스스로 성능을 끌어올리는 지능을 사전에 평가·신고하게 하는 규율 필요	제2부 12장 「재귀적 자기개선」·13장 「초지능으로 가는 네 가지 길」, 제7부 3장 「법, 노동, 안전」
전환 정체	Enabled에 머물러 "우리도 AI 한다"고 만족하는 조용한 거부	제3부 7장 「AX 실패의 패턴」
지능 에너지 격차	개인·기업·국가 사이의 새로운 디바이드	제7부 2장 「지능 에너지 격차」

> 사용법: 각 리스크에 대해 "우리 조직의 현재 노출도(상/중/하)"와 "담당자"를 한 칸씩 채우면, 그대로 리스크 관리 대시보드가 됩니다.

독자별 실행 카드

일반독자: 세 가지 독서 경로

30분이 있다면 서문과 제1부 1·2장, 제3부 1·3장(「AI 네이티브란 무엇인가」, 「AX란 무엇인가: 네 단계의 사다리」), 그리고 제7부 마지막 장 「모두를 위한 지능 에너지」를 읽어 보세요. 중심 비유와 전환 사다리를 한 번에 잡는 길이지요. 서문에서 안내한 30분 코스와 같습니다.

2시간이 있다면 여기에 제2부의 「지능 발전소」·「변환소」·「송전망」 세 장, 제4부의 「AI를 빼면 무엇이 남는가 테스트」와 「원칙 7: 신뢰를 내장하라」, 제7부의 「법, 노동, 안전: 세 개의 매듭」을 더합니다.

완독할 때는 각 부 마지막에 '내가 확인할 증거 한 가지'를 적어 보세요. 제품명과 전망보다 판단 기준을 남기는 독서가 오래갑니다.

학생: 개인정보 없는 30분 실습

1. 이름, 학교, 연락처가 없는 공개 문서 한 쪽을 고릅니다.
2. AI에 세 문장으로 요약하게 하고, 원문에서 각 문장의 근거를 표시하게 합니다.
3. 숫자·날짜·고유명사를 원문과 대조합니다.
4. 빠진 맥락과 틀린 내용을 오류 기록에 적고, 프롬프트를 한 번 고쳐 다시 실행합니다.
5. 두 결과 중 무엇이 나아졌는지 근거와 함께 설명합니다.

점수는 문장이 매끄러운지가 아니라 근거를 찾고 오류를 수정했는지로 매깁니다.

CEO·CIO: 도입과 계약 종료를 함께 묻기

판단 항목	계약 전에 확인할 질문
직접 구축과 구매	이 기능이 경쟁력의 핵심인가, 검증된 범용 서비스인가?
데이터 반출	어떤 데이터가 어느 국가·하도급사로 이동하고 학습에 쓰이는가?
품질과 책임	기준선, 오류 허용치, 인간 승인 지점, 사고 통지 기한은 무엇인가?
종속성	모델·API가 바뀌거나 중단되어도 업무를 계속할 대체 경로가 있는가?
원가	토큰 단가가 아니라 성공한 작업 하나당 원가(재시도·사람 검토 포함)는 얼마이며, 월 사용량 상한과 초과 과금 방식은 무엇인가?
계약 종료	원본·파생 데이터와 로그를 어떤 형식으로 돌려받고, 공급자는 언제 삭제하는가?

정책담당자: 집행액 밖의 성과표

성과 축	함께 볼 지표 예시
생산성	처리 시간뿐 아니라 정정률, 재작업, 현장 인력의 검토 부담
접근성	장애·연령·지역·소득별 이용률, 비AI 채널 대기시간
안전	사고와 아차사고, 개인정보 노출, 중단·복구 시간
권리구제	설명·정정·인간 심사 요청 건수, 처리 기간, 인용률
분포	예산 수혜 기관과 기업의 지역·규모 분포, 중소 조직의 지속 이용률

예산 집행률은 활동을 보여줄 뿐 효과를 증명하지 않습니다. 평균 수치만 보지 말고 혜택과 사고가 누구에게 몰렸는지도 함께 공개해야 합니다.

부록 B. AI 기업 사례 카드

제3부에서 제시한 기준을 여덟 기업에 적용해 봅니다. 이 목록은 순위나 모범기업 인증이 아닙니다. 제품과 사업 구조에서 관찰할 메커니즘, 반대 증거, 확인할 지표를 같은 형식으로 정리한 비교 자료입니다.

기업 수치와 제품 상태는 빠르게 바뀝니다. 계약이나 투자 판단에 쓸 때는 공시·공식 문서와 독립된 평가를 기준일과 함께 다시 확인해야 합니다.

Tesla

관찰할 메커니즘: 차량에서 수집한 주행 데이터, 학습·시뮬레이션, 소프트웨어 배포가 반복되는 구조입니다. 자동차와 로봇 사업에서 센서·컴퓨팅·제조를 함께 운영한다는 점도 살펴볼 만합니다.

반대 증거와 위험: 차량 판매가 늘어도 데이터의 다양성과 품질이 자동으로 좋아지는 것은 아닙니다. 자율주행 명칭과 실제 운전자 책임의 차이, 사고 조사, 개인정보, 로봇 생산 목표와 실제 생산량을 구분해야 합니다.

확인할 지표: 감독이 필요한 주행 조건, 사고·개입률의 비교 기준, 소프트웨어 변경 뒤 성능, Optimus의 실제 생산·현장 사용량.

Perplexity AI

관찰할 메커니즘: 질문을 검색어로 바꾸고 여러 출처를 찾아 답과 인용을 함께 제시합니다. 검색·요약·후속 질문을 한 흐름으로 묶은 제품 설계가 핵심입니다.

반대 증거와 위험: 인용 링크가 있다고 답이 정확한 것은 아닙니다. 출처를 잘못 연결하거나 원문의 맥락을 빠뜨릴 수 있고, 게시자 콘텐츠의 이용·보상과 추천 순위의 상업적 영향도 쟁점입니다.

확인할 지표: 인용 정확도와 원문 충실도, 정정 시간, 광고·제휴 표시, 게시자 제외·보상 절차, 반복 질문당 총비용.

Anthropic

관찰할 메커니즘: 범용 모델, 개발 도구와 에이전트 연결 규격을 함께 제공합니다. Claude Code와 MCP는 모델이 대화창을 넘어 파일·도구·업무 흐름에 들어가는 방식을 보여줍니다.

반대 증거와 위험: 매출 런레이트와 기업가치는 특정 시점의 추정일 수 있고, 개발자 선호도와 기업 지출 점유율은 조사 표본에 따라 달라집니다. 도구 권한이 넓어질수록 프롬프트 인젝션, 비밀정보 노출과 잘못된 코드 변경 위험도 커집니다. 2026년 4월 가장 강한 모델(Claude Mythos Preview)을 일부 기관에만 연 방식은 안전을 위한 선택이지만, 누가 먼저 쓰느냐에 따른 격차라는 질문도 함께 남깁니다.

확인할 지표: 감사 가능한 매출·이용 자료, 조사 표본, 코드 수용 뒤 결함률, 권한 위반과 사고 대응, MCP 구현의 인증·승인 방식, 가장 강한 모델을 누구에게 언제 여는지에 대한 공개 기준.

TikTok (ByteDance)

관찰할 메커니즘: 팔로우 관계뿐 아니라 시청·건너뛰기·공유 신호를 이용해 다음 콘텐츠를 배열합니다. 추천 시스템이 이용자와 창작자 사이의 노출을 실시간으로 조정합니다.

반대 증거와 위험: Instagram과 YouTube도 추천 시스템을 폭넓게 사용하므로 ‘TikTok만 AI 기반’이라는 구분은 맞지 않습니다. 체류시간 증가는 이용자 가치와 같지 않으며 중독, 편향된 노출, 미성년자 보호와 창작자 소득 분배를 함께 봐야 합니다.

확인할 지표: 추천 이유와 제어 기능, 연령별 사용시간·유해노출, 창작자 수익 분포, 광고와 일반 추천의 구분, 독립 연구자의 데이터 접근.

DeepSeek

관찰할 메커니즘: DeepSeek-V3 기술보고서는 최종 학습 연산량과 비용 추정 방식을 공개했고, R1 코드와 모델 가중치는 상업적 이용과 수정이 가능한 라이선스로 배포됐습니다. 비용 효율과 공개 가중치 전략을 비교할 사례입니다.

반대 증거와 위험: 약 557만 6,000달러는 V3 최종 학습 연산비 추정치이며 선행 실험, 데이터, 인건비와 R1 포스트 트레이닝을 포함한 전체 개발비가 아닙니다. 학습 데이터와 개발 과정 전체가 공개된 것도 아닙니다.

확인할 지표: 과업별 품질과 성공 작업당 비용, 전체 수명주기 비용, 라이선스와 데이터 출처, 자체 운영 때의 보안·전력·인력 부담.

Upstage

관찰할 메커니즘: 문서의 글자뿐 아니라 표·레이아웃·수식 구조를 추출하는 Document AI와 언어모델을 함께 개발합니다. 한국어와 국내 문서 업무에 특화하는 전략을 볼 수 있습니다. 독자 AI 파운데이션 모델 사업에서는 2026년 8월 2차 평가를 통과해 SK텔레콤·LG AI연구원과 함께 3차에 올랐고, 정부는 2027년 2월 최종 두 팀을 고를 계획입니다.[1]

반대 증거와 위험: 투자 유치, 정부 컨소시엄 선정과 벤치마크 순위는 고객의 실제 성과를 보장하지 않습니다. '한국판 DeepSeek' 같은 별칭은 비용 구조와 공개 범위가 다른 기업을 무리하게 묶습니다.

확인할 지표: 문서 유형별 오류와 사람의 수정시간, 개인정보 처리, 고객 유지와 총비용, 모델·가중치·데이터의 공개 범위, 정부 지원 뒤 남는 공공 산출물.

Suno

관찰할 메커니즘: 자연어 설명에서 보컬과 반주를 포함한 음악을 생성해 작곡·편곡·합성 과정을 하나의 서비스로 묶습니다. 음악 제작 경험이 적은 이용자도 빠르게 초안을 만들 수 있습니다.

반대 증거와 위험: 제작 문턱이 낮아졌다는 것과 창작 역량·소득이 넓게 분배된다는 것은 다릅니다. 학습 데이터의 저작권, 특정 음악가의 스타일 모방, 생성물의 권리와 기존 창작자에 대한 보상이 쟁점입니다.

확인할 지표: 학습 데이터 동의·제외 절차, 권리자 보상과 분쟁, 유료 이용자의 지속 사용, 생성곡의 상업 이용 조건, 창작자별 수익 변화.

Google Gemini Notebook(옛 NotebookLM)

관찰할 메커니즘: 사용자가 고른 자료를 근거 범위로 삼아 질의응답·요약·오디오 설명을 만듭니다. 공개 웹 전체가 아니라 지정된 자료와 인용을 중심으로 상호작용한다는 점이 특징입니다. 2026년 7월 이름을 Gemini Notebook으로 바꾸면서, 고른 자료 위에서 코드를 돌려 데이터를 분석하는 기능까지 붙었습니다.[2]

반대 증거와 위험: 자료를 제한해도 요약 오류와 중요한 맥락 누락은 남습니다. 오디오가 자연스럽다는 이유로 정확하다고 믿기 쉽고, 업로드 문서의 기밀성·보관·학습 이용 조건도 확인해야 합니다.

확인할 지표: 인용 충실도와 누락률, 오류 정정 방식, 자료 보관·삭제·학습 정책, 접근성, 사용자가 원문을 실제로 확인하는 비율, 실행한 코드와 결과를 원자료와 대조할 수 있는지.

사례를 비교하는 네 가지 질문

여덟 기업을 한 줄의 성공 공식으로 묶기보다 다음 질문을 같은 순서로 던지는 편이 낫습니다.

1. AI가 제품의 어느 결정과 비용 구조를 바꾸는가.
2. 사용할수록 쌓이는 데이터의 품질·권리·분배를 누가 관리하는가.
3. 회사가 제시한 성장 지표와 고객의 실제 성과는 어떻게 다른가.
4. 모델이나 추천을 중단했을 때도 이용자가 데이터와 업무를 옮길 수 있는가.

좋은 사례는 큰 수치나 매끄러운 설명으로 결정되지 않습니다. 메커니즘을 재현할 수 있고, 반대 증거와 피해를 공개하며, 이용자가 선택하고 떠날 수 있을 때 다른 조직이 배울 수 있는 사례가 됩니다.

---

1. 독자 AI 파운데이션 모델 2차 단계평가 결과 보도 (머니투데이, 2026.8.18, https://www.mt.co.kr/article/2026081812375747016) — 업스테이지(Solar Open 2)·SK텔레콤·LG AI연구원 3단계 진출.

2. "NotebookLM is now Gemini Notebook" (Google Workspace Updates, 2026.7.16, https://workspaceupdates.googleblog.com/2026/07/notebooklm-now-gemini-notebook.html) — 이름 변경, 기존 링크 자동 전환, 노트북별 코드 실행 환경 추가.

부록 C. 용어집

이 책에서 사용한 주요 용어를 정리했습니다. 한글로 시작하는 표제어를 가나다순으로 먼저 싣고, 영문·약어로 시작하는 표제어를 알파벳순으로 뒤에 모았습니다. 각 용어는 영어 원어와 함께 대화체로 풀어 설명합니다.

한글 용어

기계적 해석 가능성 (Mechanistic Interpretability) — AI 모델의 내부 구조를 역추적해 "이 뉴런이 왜 활성화되었는지, 이 회로가 어떤 개념을 처리하는지"를 밝혀내는 연구입니다. 블랙박스의 뚜껑을 열고 내부 회로도를 그리는 작업이라고 생각하면 됩니다. MIT Technology Review가 2026년 10대 획기적 기술의 하나로 꼽았지요. 신뢰를 설계 단계부터 내장하려면 반드시 필요한 기술이며, 이 책 제4부 「신뢰를 내장하라」에서 다룹니다.

대규모 언어 모델 (LLM, Large Language Model) — 대규모 텍스트 데이터로 학습한 AI 모델입니다. 문장을 이해하고 생성할 수 있으며, GPT, Claude, Gemini, Llama 등이 대표적이지요. 파운데이션 모델의 대표적인 유형이며, 이 책에서는 대형 LLM을 "대형 발전소"에 비유합니다.

디바이드 (Divide, 격차) — 기술이나 자원에 대한 접근 격차를 뜻합니다. 디지털 디바이드가 인터넷 접근 격차였다면, 이 책에서 말하는 "지능 에너지 디바이드"는 AI 접근 격차입니다. 선진국과 개발도상국 사이, 대기업과 중소기업 사이, 세대 사이에서 벌어지고 있지요.

멀티모달 (Multimodal) — 텍스트, 이미지, 음성, 영상 등 여러 형태의 데이터를 동시에 이해하고 생성할 수 있는 AI의 능력을 뜻합니다. 2026년 현재 멀티모달은 더 이상 차별화 요소가 아니라 기본 사양이 되었습니다.

소버린 AI (Sovereign AI) — 한 국가가 자국의 인프라, 데이터, 인력을 활용하여 AI를 독자적으로 개발하고 운영할 수 있는 역량을 뜻합니다. 외국 AI 서비스에 의존하지 않고, 자국의 언어, 문화, 법률을 반영한 AI를 만드는 것이 핵심이지요.

아티팩트 (Artifact, 측정 착시) — 실제 현상이 아니라 측정 방법 때문에 생겨난 가짜 신호를 뜻합니다. 예컨대 AI의 능력이 실제보다 빠르게 좋아지는 것처럼 보이는데, 알고 보니 벤치마크의 난이도를 매긴 사람과 실제로 문제를 푼 사람이 달라 기준선이 왜곡된 경우가 그렇지요. 데이터가 아니라 측정 설계가 만들어낸 착시라는 점에서, 수치를 읽을 때 늘 함께 의심해야 하는 대상입니다. 이 책 제1부 「얼마나 빨리 빨라지는가」에서 다룹니다.

어텐션 (Attention) — 트랜스포머 아키텍처의 핵심 메커니즘으로, 입력 텍스트의 각 부분이 다른 부분과 얼마나 관련 있는지를 계산합니다. 사람이 긴 문장을 읽을 때 중요한 단어에 주의를 기울이는 것과 비슷한 원리이지요.

에이전트 (Agent) — 사용자의 지시를 받아 스스로 계획을 세우고, 도구를 사용하며, 목표를 달성하는 AI 시스템입니다. 챗봇이 질문에 대답하는 것이라면, 에이전트는 일을 해내는 것이지요.

온디바이스 AI (On-Device AI) — 클라우드가 아닌 사용자의 기기(스마트폰, 노트북 등)에서 직접 실행되는 AI입니다. 인터넷 없이도 작동할 수 있고, 원본 데이터의 외부 전송을 줄여 프라이버시에 유리할 수 있지요. 다만 앱 권한, 기기 로그, 클라우드 동기화 설정에 따라 데이터가 밖으로 나갈 수 있어 온디바이스 실행만으로 프라이버시가 보장되지는 않습니다. 분산 발전에 비유할 수 있습니다.

재귀적 자기개선 (RSI, Recursive Self-Improvement) — AI가 자신의 성능을 개선하는 것을 넘어, "개선하는 능력 자체"를 개선하는 것을 말합니다. 발전소가 다음 발전소의 설계도를 스스로 그리기 시작하는 셈이지요. 초지능으로 가는 네 갈래 길 가운데 하나이며, 이 책 제2부 「스스로를 개선하는 지능」에서 자세히 다룹니다.

제도적 정렬 (Institutional Alignment) — 하나의 AI를 통제하는 기술이 아니라, 여러 에이전트가 뒤섞인 사회에 시장·조직처럼 규칙과 견제·균형을 설계해 넣는 정렬 방식입니다. 초지능이 하나의 거대한 두뇌가 아니라 수많은 에이전트의 사회로 온다면, 필요한 것은 AI 하나를 길들이는 일이 아니라 사회의 판을 짜는 일이라는 발상이지요. 회사 안에서는 "에이전트 거버넌스"라는 이름으로 이미 오늘의 숙제가 되었습니다. 이 책 제2부 「초지능으로 가는 네 가지 길」에서 다룹니다.

제로데이 (Zero-day) — 아직 세상에 알려지지 않아 방어 수단(패치)이 없는 보안 취약점을 말합니다. 개발자가 문제를 알아챈 지 "0일"밖에 안 됐다는 뜻이지요. 이제는 AI가 제로데이를 찾아내는 데 쓰이기도 하고, 악용하는 데 쓰이기도 합니다. 발견한 뒤 재현하고 제보하고 패치를 검토하는 일에는 여전히 사람의 손이 들어가지요. 이 책 제6부의 AI 보안 장들에서 다룹니다.

지능 에너지 (Intelligence Energy) — 이 책의 핵심 비유입니다. 1차 산업혁명에서 석탄이 증기 에너지를, 2차에서 석유가 전기 에너지를, 3차에서 컴퓨터가 디지털 정보 에너지를 만들었듯이, 4차 AI산업혁명에서는 GPU와 파운데이션 모델이 전기를 지능 에너지로 변환합니다. 1·2차 산업혁명이 근육 노동을 대체했다면, 4차 AI산업혁명은 지식 노동의 일부를 대체하기 시작했지요. 이 비유에서 파운데이션 모델은 지능 발전소, GPU 데이터센터는 전기를 지능으로 바꾸는 변환소에 해당합니다. 비유가 멈추는 지점은 제1부 「네 번째 에너지의 이름은 '지능'이다」에서 짚었습니다.

초지능 (ASI, Artificial Superintelligence, 범용 초지능) — 천재 한 사람이 아니라 "대규모 인간 조직보다 더 지능적이고 인지적으로 유능한 시스템"을 말합니다. 사람 수준의 범용 지능(AGI)을 지나 그 너머로 넘어간 지점이지요. 하나의 거대한 두뇌로 올지, 수많은 에이전트가 모인 사회의 형태로 올지는 아직 열린 물음입니다. 이 책 제2부 「초지능으로 가는 네 가지 길」에서 네 갈래 경로를 다룹니다.

추론 (Inference) — 학습이 완료된 AI 모델이 새로운 입력에 대해 결과를 생성하는 과정입니다. 학습이 "발전소 건설"이라면, 추론은 "전기를 생산하여 공급하는 것"에 해당하지요. AI 서비스 운영 비용의 대부분은 추론에서 발생합니다.

컨텍스트 윈도우 (Context Window) — LLM이 한 번에 처리할 수 있는 텍스트의 최대 길이입니다. 사람의 단기 기억에 해당하지요. 2026년 기준 주요 프런티어 모델은 대체로 약 100만 토큰, 즉 수천 페이지 분량의 텍스트를 한 번에 처리할 수 있습니다.

토큰 (Token) — AI가 텍스트를 처리하는 기본 단위입니다. 영어에서는 단어 하나가 보통 1~2토큰, 한국어에서는 한 음절이 1~2토큰에 해당합니다. AI 서비스 비용은 대개 토큰 수로 계산되지요.

트랜스포머 (Transformer) — 2017년 Google이 발표한 AI 아키텍처로, 현대 LLM의 기반입니다. "Attention is All You Need"라는 논문에서 제안되었으며, 입력의 모든 부분을 동시에 참조하는 어텐션 메커니즘이 핵심이지요.

파운데이션 모델 (Foundation Model) — 대규모 데이터로 사전 학습된 범용 AI 모델입니다. LLM(대규모 언어 모델)이 대표적이며, 이미지, 음성, 영상을 다루는 멀티모달 모델도 포함됩니다. 이 책에서는 "지능 발전소"에 비유합니다. 대형 파운데이션 모델은 대형 화력발전소, SLM은 소형 발전기, 온디바이스 AI는 태양광 패널에 해당합니다.

파인튜닝 (Fine-tuning) — 사전 학습된 파운데이션 모델을 특정 목적에 맞게 추가 학습시키는 과정입니다. 범용 발전소의 전기를 특정 용도에 맞게 변압하는 것에 비유할 수 있지요.

프롬프트 (Prompt) — AI에게 주는 지시문입니다. 좋은 프롬프트를 작성하는 것이 AI를 잘 활용하는 핵심 역량이며, 이 책에서는 이를 "호모 프롬프트"라는 개념으로 확장합니다.

할루시네이션 (Hallucination) — AI가 사실이 아닌 내용을 사실인 것처럼 생성하는 현상입니다. 확률적 시스템의 본질적 한계이며, 완전히 제거하기는 어렵습니다. 검증 프로세스와 RAG 등의 기법으로 완화할 수 있지요.

호모 프롬프트 (Homo Prompt) — 지능과 대화하는 습관을 가진 사람을 가리키는 이 책의 말입니다. 새로운 종이나 특정 세대를 뜻하지는 않습니다. AI에게 의도를 분명히 전하고, 결과를 검증하며, 사람과 AI의 협업을 이끄는 습관을 갖춘 사람이라면 누구나 해당하지요. 제4부 「호모 프롬프트: 지능과 대화하는 습관」에서 다룹니다.

영문·약어

A2A (Agent-to-Agent Protocol) — Google이 2025년 4월 발표한 에이전트 간 통신 프로토콜입니다.[1] 에이전트들이 서로를 찾고, 능력을 확인하고, 협업할 수 있게 해줍니다. MCP가 에이전트와 도구의 연결이라면, A2A는 에이전트와 에이전트의 연결이지요.

AI Factory — NVIDIA의 Jensen Huang이 제시한 개념으로, 데이터센터를 "데이터를 입력받아 지능을 출력하는 공장"으로 재정의한 것입니다. 전통적 발전소가 원료에서 전기를 만들듯, AI Factory는 전기와 데이터에서 지능을 생산합니다.

AI 기본법 (AI Basic Act) — 정식 명칭은 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」입니다. 2026년 1월 22일 시행됐고, 같은 해 7월 21일 공공조달 우선 고려와 취약계층 이용 지원을 담은 개정 법령이 시행됐습니다.[2] 적용 대상, 사업자 의무, 감독과 피해 구제는 다른 법과 항목별로 비교해야 하며 '세계 몇 번째'라는 순서만으로 보호 수준을 판단할 수 없습니다.

AI 네이티브 (AI Native) — 이 책에서 사용하는 중심 개념입니다. 정의와 경계는 제3부 1장 「AI 네이티브란 무엇인가」에서, AI-Ready → AI-Enabled → AI-First → AI-Native의 네 단계 구분은 제3부 3장 「AX란 무엇인가: 네 단계의 사다리」에서 설명합니다.

AX (AI Transformation) — DX(디지털 전환)의 다음 단계로, 조직이 AI를 핵심 역량으로 내재화하는 전환 과정을 뜻합니다. 단순히 AI 도구를 쓰는 것이 아니라, 프로세스, 의사결정, 비즈니스 모델을 AI 중심으로 재설계하는 것이지요.

EU AI Act — 유럽연합이 제정한 세계 최초의 포괄적 AI 규제법입니다. 위험 수준에 따라 AI를 분류하고 차등 규제합니다. 2024년 8월 발효 뒤 의무별로 단계 적용되고 있어 하나의 '전면 시행일'로 요약하기 어렵습니다.[3] 고위험 의무는 한 차례 연기되어 독립형 시스템은 2027년 12월, 제품에 내장된 시스템은 2028년 8월부터 적용됩니다. 제7부 「일곱 개 국가의 선택」에서 일정을 정리했습니다.

GPU (Graphics Processing Unit) — 원래 그래픽 처리를 위한 칩이지만, AI 학습과 추론에 필수적인 병렬 연산 장치로 자리잡았습니다. 이 책에서는 "전기를 지능으로 변환하는 핵심 변환 장치"에 비유합니다. AI 학습용 칩 시장에서는 NVIDIA의 GPU가 압도적인 점유율을 차지하고 있지요.

LoRA (Low-Rank Adaptation) — 파운데이션 모델의 전체를 재학습하지 않고, 소수의 파라미터만 조정하여 특정 목적에 맞게 적응시키는 효율적인 파인튜닝 기법입니다. 비용과 시간을 크게 절약할 수 있지요.

MCP (Model Context Protocol) — Anthropic이 2024년 11월 발표한 개방형 프로토콜로, AI 시스템이 외부 데이터와 도구에 접근하는 표준 방식입니다. AI 세계의 USB-C 포트라고 생각하면 됩니다. 2025년 말 Linux Foundation 산하 Agentic AI Foundation에 기증되었습니다.[1]

MoE (Mixture of Experts, 전문가 조합) — 여러 전문가 모델을 조합하여 하나의 강력한 모델을 만드는 아키텍처입니다. 입력에 따라 적합한 전문가만 활성화하므로 효율이 높지요. DeepSeek-R1이 대표적인 MoE 모델입니다.

Physical AI — 디지털 세계를 넘어 물리적 세계에서 작동하는 AI를 말합니다. 자율주행차, 휴머노이드 로봇, 산업용 로봇 등이 해당하지요. NVIDIA의 Jensen Huang이 2024년부터 강조하고 있는 개념입니다.

RAG (Retrieval-Augmented Generation, 검색 증강 생성) — LLM이 답변을 생성하기 전에 외부 데이터베이스에서 관련 정보를 검색하여 참고하는 기법입니다. 할루시네이션을 줄이고 최신 정보를 반영할 수 있게 해주지요.

SLM (Small Language Model, 소형 언어 모델) — 수십억 개 수준의 매개변수를 가진 경량 AI 모델입니다. Microsoft Phi, Google Gemma 등이 대표적이지요. 이 책에서는 "소형 발전기"에 비유합니다.

SOUL.md — AI 에이전트의 어조, 가치 지향, 행동 원칙을 적어두는 설정 문서입니다. 에이전트가 어떤 방식으로 반응할지에 영향을 주지만, 성격을 혼자 결정하는 DNA는 아닙니다. 모델, 시스템 지시, 메모리, 도구 권한, 런타임 정책이 함께 에이전트의 행동을 만듭니다. 이 책 제6부 「SOUL.md를 가진 에이전트」에서 자세히 다룹니다.



---

1. MCP 발표·AAIF 이관 및 A2A 발표 자료 (Anthropic·Google·Linux Foundation, https://www.anthropic.com/news/model-context-protocol, https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation, https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/).

2. 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 (국가법령정보센터, https://www.law.go.kr/법령/인공지능발전과신뢰기반조성등에관한기본법); 「달라진 AI기본법, 우리에게 어떤 변화가 생길까?」 (대한민국 정책브리핑, 2026.7.15, https://www.korea.kr/news/policyNewsView.do?newsId=148968171) — 2026.7.21 시행 개정 법령: 공공조달 AI 제품 우선 고려, 취약계층 AI 이용비 지원 근거.

3. "AI Act: regulatory framework" (European Commission, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai).

부록

부록 D. 약어집

본문에 등장하는 주요 약어를 정리했습니다.

약어	풀네임	한국어
A2A	Agent-to-Agent Protocol	에이전트 간 통신 프로토콜 (Google)
AAIF	Agentic AI Foundation	에이전틱 AI 파운데이션 (Linux Foundation 산하)
AGI	Artificial General Intelligence	범용 인공지능
AI	Artificial Intelligence	인공지능
AIDC	AI Data Center	AI 데이터센터
AISI	AI Security Institute	AI 보안연구소 (영국, 2025년 AI Safety Institute 에서 개칭)
APEC	Asia-Pacific Economic Cooperation	아시아태평양경제협력체
API	Application Programming Interface	응용 프로그래밍 인터페이스
ASI	Artificial Superintelligence	인공 초지능(범용 초지능)
AX	AI Transformation	AI 전환
CAGR	Compound Annual Growth Rate	연평균 성장률
CapEx	Capital Expenditure	자본 지출
CES	Consumer Electronics Show	소비자 가전 전시회
CoE	Center of Excellence	전문역량센터
COF	Covalent Organic Framework	공유결합유기골격체 (다공성 결정 물질)
CTF	Capture The Flag	캡처 더 플래그 (보안 침투·방어 훈련 대회)
CTO	Chief Technology Officer	최고기술책임자
DPO	Direct Preference Optimization	직접 선호 최적화 (포스트 트레이닝 기법)
DX	Digital Transformation	디지털 전환
EFLOPS	ExaFLOPS	엑사플롭스 (초당 100경 회 연산, 슈퍼컴퓨팅파워 단위)
EHR	Electronic Health Record	전자건강기록

약어	풀네임	한국어
ERP	Enterprise Resource Planning	전사적 자원 관리
ETRI	Electronics and Telecommunications Research Institute	한국전자통신연구원
EU	European Union	유럽연합
FAIR	Findable, Accessible, Interoperable, Reusable	찾기·접근·상호운용·재사용 가능 (데이터 원칙)
FM	Foundation Model	파운데이션 모델
FSD	Full Self-Driving	완전자율주행 (Tesla)
GDP	Gross Domestic Product	국내총생산
GDPR	General Data Protection Regulation	일반 데이터 보호 규정 (EU)
GPT	Generative Pre-trained Transformer	생성형 사전학습 트랜스포머 (OpenAI)
GPU	Graphics Processing Unit	그래픽 처리 장치 (AI 연산 핵심 칩)
GR00T	Generalist Robot 00 Technology	NVIDIA 휴머노이드 로봇 파운데이션 모델
GW	Gigawatt	기가와트 (전력 단위, 10억 와트)
HBM	High Bandwidth Memory	고대역폭메모리 (AI 칩용 메모리)
HVD	High-Value Dataset	고가치 데이터셋 (EU 오픈데이터 지정 범주)
ICLR	International Conference on Learning Representations	국제학습표현학회 (AI 최상위 학회)
IEA	International Energy Agency	국제에너지기구
IEC	International Electrotechnical Commission	국제전기기술위원회
IMF	International Monetary Fund	국제통화기금
IoT	Internet of Things	사물인터넷
ISO	International Organization for Standardization	국제표준화기구
ITU	International Telecommunication Union	국제전기통신연합
LLM	Large Language Model	대규모 언어 모델
MCP	Model Context Protocol	모델 컨텍스트 프로토콜 (Anthropic)
METR	Model Evaluation & Threat Research	AI 모델 평가 연구 기관
MIT	Massachusetts Institute of Technology	매사추세츠 공과대학 (라이선스명으로도 사용)
MoE	Mixture of Experts	전문가 조합 (AI 아키텍처)

약어	풀네임	한국어
MTTR	Mean Time To Respond	평균 사고대응시간
MWh	Megawatt-hour	메가와트시 (전력량 단위)
NeurIPS	Conference on Neural Information Processing Systems	신경정보처리시스템학회 (AI 최상위 학회)
NLU	Natural Language Understanding	자연어 이해
NPU	Neural Processing Unit	신경망 처리 장치
OAuth	Open Authorization	오픈 인증 (제3자 위임 인증 표준)
OECD	Organisation for Economic Co-operation and Development	경제협력개발기구
OWASP	Open Worldwide Application Security Project	오픈 웹 애플리케이션 보안 프로젝트
PPA	Power Purchase Agreement	전력구매계약
PR	Pull Request	풀 리퀘스트 (코드 변경 요청)
PUE	Power Usage Effectiveness	전력 사용 효율 (데이터센터 에너지 효율 지표)
RAG	Retrieval-Augmented Generation	검색 증강 생성
REA	Rural Electrification Administration	농촌전력청 (미국, 1935)
RLHF	Reinforcement Learning from Human Feedback	인간 피드백 강화학습
RLVR	Reinforcement Learning with Verifiable Rewards	검증 가능한 보상 강화학습
ROI	Return on Investment	투자 수익률
RSI	Recursive Self-Improvement	재귀적 자기개선
SaaS	Software as a Service	서비스형 소프트웨어
SDLC	Software Development Life Cycle	소프트웨어 개발 수명주기
SDGs	Sustainable Development Goals	지속가능발전목표 (UN)
SLA	Service Level Agreement	서비스 수준 협약
SLM	Small Language Model	소형 언어 모델
SMR	Small Modular Reactor	소형 모듈 원자로
SON	Self-Optimizing Network	자기 최적화 네트워크
SOTA	State of the Art	특정 평가 조건에서 확인된 현재 최고 성능

약어	풀네임	한국어
SWE-bench	Software Engineering Benchmark	소프트웨어 엔지니어링 벤치마크 (AI 코딩 성능 평가)
TOPS	Tera Operations Per Second	초당 조 단위 연산 (AI 칩 성능 단위)
TPS	Tokens Per Second	초당 토큰 처리량 (추론 속도 지표)
TSMC	Taiwan Semiconductor Manufacturing Company	대만 반도체 제조 회사
TWh	Terawatt-hour	테라와트시 (전력량 단위)
UCP	Universal Commerce Protocol	범용 상거래 프로토콜
UI	User Interface	사용자 인터페이스
UN	United Nations	유엔 (국제연합)
UNDP	United Nations Development Programme	유엔개발계획
USCC	US-China Economic and Security Review Commission	미중경제안보검토위원회
UX	User Experience	사용자 경험
VLA	Vision-Language-Action	시각-언어-행동 (로봇 AI 모델)
VLM	Vision-Language Model	시각-언어 모델 (멀티모달 AI)
WASM	WebAssembly	웹어셈블리
WEF	World Economic Forum	세계경제포럼
XR	Extended Reality	확장현실



부록 E. 찾아보기

주요 용어가 비중 있게 다뤄지는 장을 부·장 번호로 안내합니다(예: "2부 10-1" = 제2부 10-1장). 장 제목은 목차에서 확인할 수 있습니다. 용어의 뜻풀이는 부록 C(용어집)와 부록 D(약어집)에 있습니다.

한글 용어

강화학습 — 2부 1

개인 에이전트 — 6부 1 · 7부 4 · 6부 2

검증 격차 — 5부 1-1

공공 API — 7부 8-2

기계적 해석 가능성 — 4부 9

데이터 주권 — 4부 9

독자 파운데이션 모델 — 5부 10 · 2부 10 · 2부 10-1

멀티모달 — 2부 5 · 6부 5

바이브 코딩 — 6부 10 · 6부 12

성공 작업당 원가 — 2부 10-1

소버린 AI — 2부 5 · 2부 11

스타게이트 — 2부 6 · 2부 7 · 2부 5 · 2부 9

에이전트 경제 — 6부 4 · 6부 4-2 · 6부 9

오케스트레이터 — 6부 11 · 6부 11-2

온디바이스 AI — 2부 1-1 · 1부 2 · 4부 11 · 7부 11

재귀적 자기개선(RSI) — 2부 12 · 2부 13 · 7부 3

정렬(제도적 정렬) — 2부 13

제로데이 — 6부 10-2

증거 점검(4단계 참고) — 3부 3 · 3부 4

지능 격차(디바이드) — 7부 2 · 7부 9

천하삼분지계 — 7부 10

초지능(ASI) — 2부 13 · 2부 12 · 6부 12-1 · 6부 12-2 · 7부 3

추론(인퍼런스) — 2부 10 · 2부 1

토큰 이코노미 — 2부 10

토큰 자급률 — 2부 10 · 2부 11

파인튜닝 — 5부 9

포스트 트레이닝 — 2부 1

플라이휠 — 4부 6

할루시네이션 — 7부 3

합성 데이터 — 4부 6

호모 프롬프트 — 4부 12 · 4부 13

휴머노이드 로봇 — 6부 6 · 5부 3 · 6부 5

영문·약어

A2A — 2부 3 · 4부 7 · 6부 4 · 6부 9 · 2부 4 · 6부 1

AI 2040 — 6부 12-2 · 2부 13 · 6부 12-1

AI 기본법 — 7부 10 · 4부 9 · 7부 3 · 7부 7

AI 기본사회 — 7부 8-1 · 7부 9

AI 제거 테스트 — 4부 2

AI 팩토리 — 2부 11 · 2부 5 · 2부 2 · 1부 2 · 1부 6 · 5부 3

AI 행동계획 — 7부 7 · 2부 9 · 5부 10

AIDC(AI 데이터센터) — 2부 9 · 2부 6

AIxCC — 6부 10

AX(AI 전환) — 3부 7 · 3부 3 · 3부 6-1 · 3부 6 · 5부 2 · 5부 8

Claude Code — 6부 10 · 6부 11-2 · 5부 1 · 6부 2-1

DeepSeek — 2부 10 · 1부 3 · 7부 5 · 7부 2 · 2부 1 · 2부 8-1

EU AI Act — 7부 3

Fable 5 — 6부 10-1 · 2부 10

HBM — 2부 9 · 7부 10 · 7부 7

Hermes — 6부 2

MCP — 2부 3 · 7부 8-2 · 3부 6-1 · 4부 7 · 6부 4-1 · 6부 3

METR — 1부 4 · 2부 12 · 5부 1

Mythos 사태 — 6부 10-1 · 1부 4 · 4부 6 · 5부 1

NVIDIA — 2부 5 · 6부 5 · 2부 2 · 2부 11 · 2부 7 · 2부 8

OpenClaw — 6부 2 · 6부 4-1

Physical AI — 6부 5 · 7부 7

RAG — 6부 8 · 5부 9

RentAHuman — 6부 4-2

SLM(소형 언어 모델) — 2부 1-1

SOUL.md — 6부 7 · 6부 9

인물

샘 올트먼 — 1부 6 · 6부 12-1 · 7부 6 · 2부 7

애디 오스마니 — 5부 1-1 · 6부 11-2

안드레 카파시 — 6부 12

젠슨 황 — 1부 6 · 2부 11 · 2부 5 · 2부 7 · 6부 5

AI 네이티브란 결국
지능 에너지 네이티브입니다.

전기가 처음 공장에 들어왔을 때, 증기 엔진 자리에 모터만 갈아 끼운 공장은 오래도록 제자리였습니다. 달라진 곳은 전기를 전제로 배치도를 처음부터 다시 그린 공장이었지요.

지금 AI 앞에서도 같은 갈림길이 놓여 있습니다. 이 책은 지능을 새로운 에너지로 보고, 그 발전소와 송전망과 가전제품이 어떻게 생겼는지, 회사와 나라가 배치도를 어떻게 다시 그려야 하는지 차근차근 이야기합니다.

이 책에서 만나는 장면

에이전트가 사람을 고용한다 AI에게 고용된 첫 번째 사람들

한 사람이 회사를 세운다 '1인 10억 달러 기업'이라는 기대

토큰값이 아니라 일값 AI 원가를 다시 계산하는 법

지능이 스스로를 고친다 초지능으로 가는 네 갈래 길

PDF·EPUB 최신판
v1.3 · 2026.10.23

전종홍 지음

지능을 전기처럼 쓰는 시대 · AI 네이티브 핸드북 2026

이 책은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스(CC BY-NC-ND 4.0)를 따릅니다.
creativecommons.org/licenses/by-nc-nd/4.0