AI GOVERNMENT TREND REPORT 2026-2

AI.GOV 해외동향

CONTENTS

NIA 한국지능정보원

일러두기 NOTE

「AI.GOV 해외동향」은 해외 주요국의 인공지능 정책 및 산업 동향을 파악하고 분석하기 위해 한국지능정보사회진흥원(NIA)에서 기획·발간하는 보고서입니다.

본 보고서의 내용은 생성형 AI를 활용하여 작성하였습니다.

내부 검수 과정을 거쳐 보완하였으나, 수록 정보의 정확성과 신뢰성 확보를 위해 참고문헌을 통한 원문 확인을 권고합니다.

본 보고서의 내용은 한국지능정보사회진흥원의 공식 견해와 다를 수 있으며, 문의 및 제안은 아래 연락처로 문의해주시기 바랍니다.

■ 발행처: 한국지능정보사회진흥원
■ 발행인: 김형철
■ 작성자: 한국지능정보사회진흥원 인공지능정부본부 AI정부기획팀
- 정부만 연구위원(cbm12@nia.or.kr)
- 이용건 수석(woojae@nia.or.kr)
- 박성하 선임(vin@nia.or.kr)
■ 보고서 온라인 서비스: www.nia.or.kr

목 차 CONTENTS

01

Mythos 쇼크와 AI 사이버보안의 새 임계점

- 데이터 유출 · Project Glasswing · 뉴로심볼릭 AI 부상과 국내 정책 함의

Reading Point

- 2026년 3월 유출로 존재가 확인된 Claude Mythos(코드명 Capybara)는 Anthropic 공개 모델 라인업 최상단을 넘어서는 별도 티어로, 사이버 보안계의 화두인 AI 모델
- Mythos의 고도화된 사이버보안 능력은 기존 LLM의 지도학습 한계를 뛰어넘어, 신경망과 기호 논리를 결합한 뉴로심볼릭 AI 기반의 고차원 추론 과정에서 생성된 역량으로 분석
- Anthropic은 파트너 12개사 제한 접근 체계(Project Glasswing)와 $1억 방어 이니셔티브로 대응하고 있으나 국내 기관은 초기 명단에 미포함된 상태로 선제적 정책 체계 구축 시급

Ⅰ Claude Mythos 개요: 유출에서 공식 발표까지

▸ 유출 사건과 공식 발표 경위

- 2026년 3월 26일, 모델 코드명 'Capybara'로 Anthropic의 Claude Code 소스 맵, 사이버 보안 영역 자율 작동 시나리오 등이 포함된 3,000여 건의 내부 기술 문서가 외부로 유출된 사고 발생[1)]
- Anthropic은 즉각 내부 조사에 착수하고 파트너사(12개社)에 'Project Glasswing'라는 비공개 협조체계를 요청하는 한편, 2주 후인 4월 7일 공식 성명 Claude Mythos Preview 발표[2)]

발표 타임라인

일자	사건	주요 내용
26.3.26.	데이터 유출	Claude Code 소스 맵·모델 정보 유출, Fortune 최초 보도
26.3.26.~4.6.	내부 조사 / 파트너 협의	정보 확산 차단, Glasswing 파트너 12개사 긴급 비공개 협의
26.4.7.	공식 발표	존재 인정, 제한 공개 방침 확정, 보안 능력 보고서 일부 공개
26.4.~	국내 대응	금융위원회 비공개 실무회의, AI 안전연구소 접근 시도

1) Fortune (2026.03.26.) Exclusive: Anthropic's powerful new AI model accidentally leaked early, https://fortune.com/2026/03/26/anthropic-claude-mythos-leaked/
2) Anthropic Red Team Blog (2026.04.07.) Claude Mythos Preview – Safety Report, https://red.anthropic.com/2026/mythos-preview/

- 별도의 보안 훈련 없이도 모든 주요 운영체제·브라우저에서 *제로데이(Zero-day) 취약점을 수천 건 자율 탐지·*익스플로잇(Exploit) 하는 능력을 보유한 것으로 확인

* 제로데이(Zero-day): 개발 바로 직후 보안패치가 존재하지 않는 무방비 상태
* 익스플로잇(Exploit): 취약점을 실제로 공격하는 코드 또는 행위

▸ Mythos가 무엇을 찾아냈는가[3)]

① OpenBSD TCP SACK – 27년 묵은 DoS 취약점

- 1998년 OpenBSD에 도입된 TCP SACK 구현부에 27년간 잠복해온 부호 정수 오버플로우 결함을 Mythos가 자율 발견하며 보안 평판 1위 OS의 신뢰성 타격

② FreeBSD NFS – 17년 묵은 원격 코드 실행

- 서버의 인증 우회·원격 코드 실행 취약점을 Mythos가 17년 만에 자율 발견하여 인증 없이 인터넷 어디서든 root 권한 탈취가 가능함을 약 4시간 컴퓨팅으로 입증
- 정교한 익스플로잇을 자율 작성했으며 Netflix CDN·WhatsApp·PlayStation OS 등 인터넷 기반 인프라 다수가 영향권에 포함되는 사안으로 확인

③ FFmpeg H.264 코덱 – 16년 묵은 메모리 변조 취약점

- FFmpeg H.264 코덱 내 sentinel 충돌로 인한 out-of-bounds write 결함을 Mythos가 16년 만에 자율 발견하며 거의 모든 미디어 플레이어·스트리밍·CCTV·화상회의 솔루션 기반 라이브러리의 잠재적 위협 노출 확인

▸ 자율형 사이버 행위 가능성의 시초 사례로 평가

- Mythos에서 관찰된 사이버보안 능력의 일부는 의도적으로 설계하지 않았음에도 모델 규모와 훈련 깊이가 임계점을 넘어서며 자발적으로 고도화 되어 출현
- 기존의 사이버 공격 체계(취약점 발견 - 공격Tool 설계 및 실행 - 보안 패치화)에서 '사람이 몰랐던 부분을 스스로 찾아내는 자율 판단 AI'로의 전환점으로 평가
- 취약점 발견부터 공격 실행까지 자율 수행하는 모습은 사람의 개입 없이 공격 주체가 될 수 있는 구조의 실증이자 AI의 사이버 행위 이론이 현실임을 방증
- 보안 평판이 가장 두터운 시스템들에서 16~27년간 누적된 안전, 신뢰성에 타격을 받으며 발견과 작동 익스플로잇 간 시차가 사실상 소멸하고 공격 비용이 $50~$20,000 수준으로 극단적 하향됨을 실증

3) Anthropic Red Team 공식 발표 (1차 자료): https://red.anthropic.com/2026/mythos-preview/

II Mythos의 사이버 능력과 대응 체제(Project Glasswing)

▸ Anthropic 전체 모델 라인업과 Mythos의 위치

Anthropic 전체 모델 비교표[4)]

모델명	출시	컨텍스트	주요 벤치마크	핵심 특징
Claude 1.0	2023.03	9K	MMLU 73%	최초 Constitutional AI
Claude 1.3	2023.04	100K	MMLU 77.0%	업계 최초 100K 컨텍스트
Claude 2.0	2023.07	100K	MMLU 78.5% HumanEval71.2%	claude.ai 베타 공개
Claude 2.1	2023.11	200K	MMLU 77.0%	환각 2배 감소, 시스템 프롬프트
Claude 3 Haiku	2024.03	200K	MMLU 75.2%	초고속·저비용 실시간 응답
Claude 3 Sonnet	2024.03	200K	MMLU 79.0%	기업용 워크로드 최적화
Claude 3 Opus	2024.03	200K	MMLU 86.8% HumanEval84.9%	복잡 추론 최고 지능
Claude 3.5 Sonnet v1	2024.06	200K	MMLU 88.7% HumanEval92.0%	Artifacts 기능, Opus 2배 속도
Claude 3.5 Sonnet v2	2024.10	200K	SWE-bench 49.0%	Computer Use(베타) 최초
Claude 3.5 Haiku	2024.11	200K	MMLU 80.6%	Opus급 성능을 Haiku 속도로
Claude 3.7 Sonnet	2025.02	200K	SWE-bench 63.7%	하이브리드 추론(추론시간 조절)
Claude 4.5 Sonnet	2025.10	200K	MMLU 89.2% HumanEval92.7%	지능 지수 대폭 향상
Claude 4.5 Haiku	2025.10	200K	-	Claude Code 최적화, 초저지연
Claude 4.6 Opus	2026.02	1M	MMLU 89.5%	100만 토큰 지원, 적응형 추론
Claude 4.7 Opus	2026.04	1M	SWE-bench 87.6%	자율 에이전트 코딩 특화
Mythos (Capybara)	**2026.04 (Preview)**	**1M+**	**CyberGym 전 문제 SWE-bench~100%**	**사이버보안 특화 Capybara 별도 티어 파트너12개사제한접근**

▸ 주목할 분기점

- Mythos (2026.4) – Opus를 넘는 별도 티어, 일반 공개 중단
- Opus 4.6 (2026.2) – 1M 토큰·14.5시간 자율 작업 도달
- Sonnet 4.5 (2025.9) – Sonnet이 처음으로 Opus 능가
- 3.5 Sonnet v2 (2024.10) – Computer Use 첫 등장
- Claude 3 (2024.3) – Opus/Sonnet/Haiku 3티어 체제 시작

4) Anthropic Newsroom 공식 모델별 발표문 종합, https://www.anthropic.com/news

▸ 사이버 보안 영역의 실질적 위협 vs 방어수단

- 수십년 간 보안 전문가 누구도 발견하지 못한 취약점을 단 수 시간 만에 식별하는 여러 사례들을 보아 인간 최고 수준의 보안 전문가가 수개월을 투입해도 찾지 못한 영역을 AI가 돌파했음을 의미
- 그 외에도 AI 기술 수준을 객관적으로 측정하는 주요 벤치마크에서 기존 최고 수준 모델을 크게 상회하는 성능을 기록
- 특히, Firefox 익스플로잇 생성 벤치마크에서 전체 장악 시도 중 84%의 성공률(181회)을 보여주면서 사이버 보안 분야에서 보유한 실질적 위협 수준을 가시화하는 한편 보안 방어 체계의 핵심 수단으로서의 효용성 또한 명확히 입증[5)]

생성형 AI 모델별 벤치마크 성적

벤치마크	평가내용	Claude Mythos	Claude Opus 4.6	Gemini 3.1 Pro	GPT -5.4
SWE-bench Verified	실제 소프트웨어 버그 수정	**93.9% ★**	80.8%	80.6%	미발표
SWE-bench Pro	실무급 고난도 코딩	**77.8% ★**	미발표	54.2%	57.7%
CyberGym	사이버 보안 문제해결	**83.1% ★**	66.6%	미발표	66.3%
Firefox 익스플로잇 생성	취약점 공격코드 설계 및 제어권 장악 (공격성공/제어권장악)	**181회/29회**	2회/0회	미평가	미평가

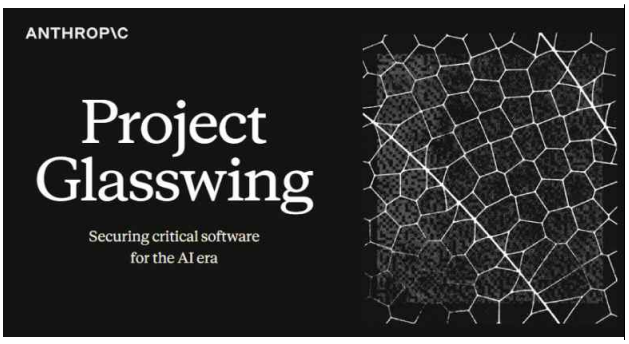
▸ Project Glasswing: 방어 목적 연구를 위한 민간 기업 협력체계 구성

- Anthropic은 Claude Mythos를 일반 공개 하지 않고 'Project Glasswing' 프레임워크 하에 파트너 12개사에만 제한 접근을 허용
- 미국·유럽 주요 기술 기업 및 사이버보안 전문 기업으로 구성된 파트너사와 함께 총 $1억 방어 이니셔티브를 구성하여 방어 연구 수행

Project Grasswing 주요 접근조건 및 용도

접근 조건	용도
① 방어 목적 취약점 탐지 연구 수행	① 방어 목적 취약점 탐지·패치 연구
② 발견 정보를 연합체 내에서만 공유	② AI 능력 임계점 평가를 위한 독립 검증 체계 구축
③ 공격적 활용 및 제3자 이전 금지	③ 접근 제한 AI 모델의 안전한 운용 기술 인프라 개발

5) 클로드 미토스 프리뷰 : 시스템 카드
https://www-cdn.anthropic.com/8b8380204f74670be75e81c820ca8dda846ab289.pdf



▸ 미국 정부 차원의 대응과 정보 공유 논의

- 트럼프 행정부 이래 AI Action Plan은 AI 기업 간 정보 공유에 긍정적 입장을 취하며 정보공유·분석센터(ISAC) 설립을 지속적으로 요구
- Mythos의 위협은 액션 플랜에서 강조한 “AI로 AI를 막는 선제적 방어 체계”를 실천할 수 있는 사례로 연방 차원에서 파트너사들에 대한 제한적 접근 허용
- 현재 Mythos의 실질적 위협 가시화와 맞물려 정보공유·분석센터를 통한 기업 간 보안 정보 공유를 강제하는 강력한 정책적 압박 명분으로 작용 중
- 그 외에도 *'증류(distillation)' 기술을 통한 비인가 복제 위험 및 중국 오픈웨이트 모델과의 결합 가능성도 주요 우려 사항에 포함
- 악용 방지를 위한 3사(Anthropic·OpenAI·Google) 공동 감지 알고리즘 개발 및 이상 API 요청 패턴 탐지 체계 구축 협의 중

* 증류 : 거대 AI의 지능을 핵심만 추출하여 가볍고 빠른 소형 AI로 이식하는 '모델 최적화 기술’

Ⅲ Mythos가 촉발한 차세대 방향 : 뉴로심볼릭 AI

▸ 뉴로심볼릭(Neuro-Symbolic) AI의 등장

- **(뉴로심볼릭 AI)** 10여 년간 학계 연구에 머물던 딥러닝 기반의 신경망(패턴 인식·언어 이해)과 규칙·논리 기반의 기호 추론(논리 추론·형식 검증)을 결합하는 AI 아키텍처 패러다임으로 추론의 신뢰성을 근본적으로 강화하는 강점을 보유
- 순수 신경만 구조에서 반복적 추론 결과를 복제하는 학습 방식만으로는 LLM의 성능 향상에 한계가 있다는 담론이 확산되며 기호 논리와의 결합 필요성이 본격 대두
- Mythos 자율 익스플로잇 정확도의 비약적 향상 원인은 기호 추론 엔진이 보안 취약점의 형식 검증(Formal Verification)에 직접 활용되었기 때문으로 분석

▸ 차세대 AI 경쟁의 핵심 방향 : 뉴로심볼릭 AI의 부상

- MIT Technology Review 2026년 10대 혁신 기술에 선정되며 AI의 효율적 추론과 지식 결합적인 접근이 AI 기술 경쟁의 핵심 지표로 부상
- 그 외에도 Google DeepMind, AlphaProof·IBM의 뉴로심볼릭 아키텍쳐 연구가 잇따르며 학술 담론을 넘어 차세대 AI 경쟁의 구조적 방향으로 공식화되는 추세

▸ 학술 담론에서 산업 현장으로 퍼지는 뉴로심볼릭 AI 주요 동향

- **(시장)** 23년 Hype Cycle에 뉴로심볼릭 AI를 첫 등재한 이후, 2025년에는 GenAI 환멸기 도래와 맞물려 'Composite AI' 핵심 구성요소로 주목
- **(산업)** Google DeepMind가 AlphaGeometry·AlphaProof를 통해 뉴로심볼릭 접근이 IMO 금메달 수준 수학 추론에 도달함을 입증하고, IBM Research가 MIT-IBM Watson AI Lab 중심으로 학술 허브를 운영하는 등 빅테크의 R&D 투자가 본격적으로 추진중

IV 주요 정책적 시사점

▸ Project Glasswing 접근 확보와 AI 안전연구소 역할 강화

- Mythos 수준의 AI가 공격에 활용될 경우 위협의 구체적 속성 파악 없이 방어해야 하는 구조로 접근 확보 뿐 아니라 발생 시 후속조치에 대한 대응 체계 구축

▸ 사이버보안 분야 '준비성'에서 '회복탄력성' 중심으로의 관점·체계 전환

- 사이버보안 성과 측정 기준을 인증·통과 중심에서 공격 탐지 속도·복구 수준 등 실전 지표로 전환하여 조직의 실전 대응 역량과 사이버 복원력 평가 관점 재설정 필요
- NIST CSF 2.0 Cyber AI Profile[6][7]·EU DORA(디지털 운영 복원력법)[8] 등의 글로벌 정책 분야에서는 사이버보안 회복탄력성 입증을 의무화하는 흐름
- 국내 보안 평가 체계의 결과 중심 KPI도 회복탄력성 중심으로의 개편·재설계 필요

▸ 전국 지자체 보안 격차 해소와 공동 AI 보안관제

- Mythos의 등장으로 지자체별 보안 격차는 국가망 전체의 제어권 상실로 직결가능, 망 분리 한계를 극복할 공동 AI 보안관제 체계를 구축하여 전 국가적 차원 방어체계 마련

6) NIST 공식: NIST Cybersecurity Framework (CSF) 2.0 https://www.nist.gov/cyberframework
7) KPMG 분석NIST AI 사이버보안 프레임워크: https://kpmg.com/us/en/articles/2026/cybersecurity-nist-draft-cybersecurity-framework-for-ai-reg-alert.html
8) EU DORA 공식: 디지털 운영 회복력법(Digital Operational Resilience Act) https://www.digital-operational-resilience-act.com/

02

에이전틱 AI 시대 작업 구조의 전환

- Vibe Coding에서 LLM Wiki 지식 체계까지

Reading Point

- AI 개발 패러다임은 Vibe Coding을 넘어 계획·구현·검증을 자율 수행하는 Agentic Engineering으로 이미 전환 중
- AI 에이전트의 고활용을 위해 AI가 조직 지식을 스스로 컴파일·정제·관리할 수 있는 AI 지식 관리 체계(LLM Wiki)를 제시, 기존 RAG의 맥락단절 및 확장성 한계 보완 가능
- Agentic Engineering과 LLM Wiki 결합 시 속도·정확성·자율 갱신이 동시에 실현가능

Ⅰ Agentic Engineering과 Software 3.0

▸ 기존의 기술 전환 주기를 압도하는 AI 분야 변화 속도

- 2025년 2월 Karpathy가 제안한 Vibe Coding이 Collins 2025년 올해의 단어로 선정되며 전 세계적으로 확산되었으나, 등장 1년 만에 그 스스로 구식으로 선언
- Vibe Coding의 탄생은 스타트업 창업 비용 급락, 1인 개발자 앱 출시 등 SW 개발 생태계에 새로운 변화를 촉진했으나, 품질·보안 측면에서의 한계 존재
- Karpathy는 이미 산업 흐름이 Vibe Coding을 넘어, 계획·구현·검증까지 자율 수행하는 Agentic Engineering 단계로 이동하고 있다고 평가[9)]
- Agentic Engineering을 통해 AI 에이전트 중심 작업 구조를 강조하며, 이후 *LLM Wiki를 통해 에이전트가 장기적으로 활용할 지식 체계 구조까지 제안

* LLM Wiki : AI Native 지식관리 구조 개념, AI가 조직 지식을 스스로 읽고 정리·연결할 수 있도록 설계된 방식

AI 패러다임 전환 타임라인

시점	키워드	핵심 내용	파급 효과
2025.02	Vibe Coding	자연어 대화만으로 소프트웨어 제작	Collins 2025 올해의 단어 신규 개발자 80% 첫 주 AI 활성화
2026.02	Agentic Engineering	"Vibe Coding은 구식" AI 중심 작업 구조로의 변화	AI 에이전트 위임 80% 역전 기업용 앱 80% 에이전트 내장
2026.04	LLM Wiki	AI가 조직 지식을 스스로 컴파일·관리	공개 2주 만에 1,700만 뷰 GitHub 13,000 Star

9) Karpathy, A. (2026.02) Sequoia AI Ascent 2026 기조연설 : https://www.youtube.com/watch?v=LCEmiRjPEtQ

▸ 소프트웨어 제작 방식 자체의 패러다임 전환 : Software 3단계 진화 프레임

- 2025.12 한 달 만에 에이전트 위임 비율이 20%에서 80%로 역전하면서 AI 에이전트 생성 코드에 대한 신뢰 수준이 구조적으로 상향되고 검증 과정이 안정화되면서 위임 비율이 급격히 확대
- 앞선 AI 산업 흐름의 변화 요인은 단순한 신규 개발 트렌드의 등장을 넘어, 소프트웨어를 설계·구현·운영하는 방식 자체가 AI 중심으로 재편되고 있음을 시사
- Karpathy는 소프트웨어 진화를 '코드 작성(1.0) → 신경망 훈련(2.0) → 프롬프트 의도 전달(3.0)'의 3단계로 정의하며, AI시대 인간의 역할 전환(실행→감독)을 시사

Software 3단계 진화 프레임

단계	작동 방식	인간의 역할	핵심 역량
Software 1.0	사람이 코드 한 행씩 직접 작성	실행자 (Implementer)	코드 작성 능력
Software 2.0	데이터셋으로 신경망 훈련	설계자 (Designer)	데이터 설계·레이블링
Software 3.0 (현재)	프롬프트로 의도 전달 컨텍스트 윈도우=레버 LLM=인터프리터	감독자·판단자 (Supervisor/Judge)	에이전트 오케스트레이션 맥락 설계·결과 검증

▸ AI 코딩 도구 시장: 패러다임이 바뀌는 속도에 맞춰 폭발적 성장

- **(규모)** 글로벌 AI 코드 도구 시장 2025년 79억 달러 → 2035년 910억 달러, CAGR 27.65% 전망[10)]
- **(경제)** Goldman Sachs는 경제 전반 AI 생산성 연계 미확인이나 코딩 등 2개 고영향 분야에서 중간값 30% 생산성 향상 보고[11)], McKinsey는 생성형 AI 연간 글로벌 경제 기여 최대 4.4조 달러로 추정[12)]
- **(인력)** GitHub Copilot 누적 사용자 2,000만 명 돌파, Fortune 100 기업 90% 채택[13)], 신규 개발자 80%가 가입 첫 주 내 AI 도구 활성화[14)]
- **(활용도)** Gartner는 2026년 전 세계 AI 지출 2.52조 달러(전년比 44%↑) 전망[15)], 별도 발표에서 2026년 말까지 기업용 앱의 40%가 태스크 특화 AI 에이전트 탑재 예측[16)]

10) Precedence Research, AI Code Tools Market (2026.03) : precedenceresearch.com
11) Fortune (Goldman Sachs 인용), 30% boost for 2 specific use cases (2026.03) : fortune.com
12) McKinsey, AI could increase corporate profits by $4.4 trillion a year (2023) : mckinsey.com
13) TechCrunch (Microsoft FY25 Q4 인용), GitHub Copilot crosses 20M users (2025.07) : techcrunch.com
14) GitHub Octoverse 2025 (2025) : github.blog
15) Gartner, Worldwide AI Spending $2.5 Trillion in 2026 (2026.01) : gartner.com
(가) 16) Gartner, 40% of Enterprise Apps with Task-Specific AI Agents by 2026

II AI Native 지식 체계의 등장: LLM Wiki

▸ 기존 방식(RAG)의 구조적 한계

- McKinsey 조사 결과 생성형 AI를 정기 활용하는 조직은 약 78%로 기존 RAG 방식의 맥락 단절·확장성 한계가 운영 진입 병목 요인 중 하나로 작용[17)]
- 문서를 조각 내 벡터 DB에 저장하는 방식은 문서 간 관계·맥락이 구조적으로 단절되며, 오래된 문서·불일치 정보가 혼재해도 자동 정제 불가

▸ 에이전틱 AI의 핵심병목 : 지식 관리의 비효율

- 지식 근로자는 업무 시간의 평균 19%(주 9.3시간)를 정보 탐색에 소비하며, 직원이 필요한 정보를 찾기 어렵다고 57%가 응답[18)]
- 부실한 지식 관리로 Fortune 500 기업이 연간 315억 달러 손실 추산
- Cloudera는 글로벌 서베이를 추진해 약 80%의 기업이 여전히 데이터 접근성 부족 문제로 인해 AI 및 데이터 이니셔티브가 제약을 받고 있다고 발표[19)]

▸ 조직 지식 체계의 새로운 패러다임 : LLM Wiki

- Karpathy가 공개한 LLM Wiki는 AI가 조직 지식을 스스로 컴파일·정제·관리하는 새로운 패러다임으로, 공개 2주 만에 1,700만 뷰·GitHub 13,000 Star를 달성하며 개발자·기업 커뮤니티에 즉각적 반향

LLM Wiki의 3단 폴더 구조와 작동원리

폴더/파일	역할	핵심 특징
raw/	원자료 수집 공간	보고서·논문·회의록 등 정제 전 원문 저장 정기 자동 수집 파이프라인 연계 가능
wiki/	AI 컴파일 지식 공간	AI가 원자료 읽고 요약·범주화·연결 백링크(Backlink)로 관계형 지식 구조 형성
index.md	전체 지식 지도	단일 컨텍스트 윈도우에 맞는 전체 목차 질의 시 AI가 먼저 확인 → 필요 문서 선택 호출

- LLM Wiki 구조 활용 시, 모든 정보가 특정 .md 파일로 추적 가능하고 에이전트가 필요 문서만 선택적 호출하여 불필요한 컨텍스트 로딩 최소화로 토큰 효율성 향상
- 다만, LLM Wiki와 RAG는 상호 대체 관계가 아닌 조직의 문서 규모와 활용 목적에 따라 선택적으로 설계해야 하는 지식 관리 아키텍처로 조직 특성별 맞춤 적용 필요

(2025.08) : gartner.com

17) McKinsey, The State of AI 2025: Agents, Innovation, and Transformation (2025.11) : mckinsey.com
18) WifiTalents (2026.02) 100+ Knowledge Management Statistics 2026, Verified https://wifitalents.com/knowledge-management-statistics/
19) https://www.epnc.co.kr/news/articleView.html?idxno=400432

III 일하는 방식과 기억하는 방식의 공진화

- Agentic Engineering(AI가 어떻게 일하느냐)과 LLM Wiki(AI가 무엇을 기억하느냐)는 "AI가 장기 자율 수행을 하기 위한 두 가지 전제 조건"으로 상호보완 관계
- 에이전트가 index.md를 먼저 확인 → 필요한 wiki 문서를 선택 호출 → 맥락을 확보한 상태에서 멀티스텝 작업 자율 수행 → 결과를 raw/에 저장하고 wiki/를 자동 업데이트하는 상호 보완 구조 시 기존의 병목 현상 완화 기대
- 싱가포르 IMDA에 따르면 기술 통제·인간 책임 소재·위험 평가·이용자 교육의 4대 축을 조직 설계 단계에서 내재화해야 성과 창출이 가능하다고 발표[20)]
- 즉, 모델 성능 뿐 아니라 내부 지식 구조 개선을 통한 맥락 공급이 성과 창출의 핵심

IV 주요 정책적 시사점

▸ 공공기관 내부 지식 자산의 AI 친화적 전환

- 공공데이터 100,000건 돌파(2025.2 기준)이나 대부분 PDF스캔본 혹은 비정형 상태 데이터로 사일로화된 시스템에 분산 존재하며 AI 도입 제한의 최대 요인[21)]
- 4단계(목록화→변환→정합→에이전트 연계) 순의 전환 방식 연구 필요

▸ AI 도입 구조 전략의 재검토

- 기관 내 가장 자주 참조되는 법령·지침·내규부터 AI 친화적 마크다운 형식으로 전환
- 4단계(목록화→변환→정합→에이전트 연계) 체계 제도화를 통해 도입 이후 AI 맥락 적용을 위한 지식 자산 구조 개편 대비
- 싱가포르 IMDA 4대 검증 축(위험 평가·책임 소재·기술 통제·이용자 교육)을 공공 SW 검수 기준 개정의 참조 모델로 활용하여 AI 산출물 품질 검수에 대한 체제 기반 마련 필요

▸ AI 시대 공무원 역량의 재정의

- 현재 AI 활용 공무원 역량 강화 교육 대부분 '도구 사용법' 중심으로, 에이전트에게 올바른 목표·맥락을 부여하고 결과를 비판적으로 검증하는 판단 역량으로의 전환 필요

---

20) 싱가포르 IMDA 에이전틱 AI 거버넌스 프레임워크 https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

21) https://www.segye.com/newsView/20250219509824?utm_source=chatgpt.com

03

하네스 엔지니어링, 자율형 AI 에이전트의 혁신

Reading Point

- 대규모 언어 모델(LLM)의 확산과 함께 '프롬프트 엔지니어링 → 콘텍스트 엔지니어링 → 하네스 엔지니어링'으로 AI 활용의 패러다임이 급속하게 진화 중
- 하네스 엔지니어링은 AI 모델 자체가 아닌, AI 모델이 실수없이 일을 잘 할 수 있는 작업 환경, 실행 규칙, 도구 연결, 피드백 구조 전체를 설계하는 개념
- 향후 하네스 엔지니어링은 공공부문에서 AI 모델의 책임감 있는 개발과 배포를 위한 정책 수립과 실행에 있어 매우 중요한 기반을 제공할 것으로 기대

Ⅰ 개요 : 프롬프트에서 하네스 엔지니어링으로 진화

▸ '하네스(Harness) 엔지니어링'이란?

- '하네스'는 원래 말을 제어하기 위한 마구를 의미하며, AI 분야에서는 AI 모델을 제어하고 안정적으로 작동하게 하는 기술과 프레임워크를 의미함[22)]
- 에이전트의 안전·통제된 작동을 보장하는 핵심 개념으로 급부상[23)]
- 2026년 2월, 실리콘밸리의 유명 개발자 미첼 하시모토(HashiCorp 공동창업자)가 이름을 붙이고, OpenAI가 공식화하면서 업계 최대 화두로 부상
- AI가 의도한 대로 작동하고 오류를 사전탐지 및 방지하여, 장기 실행에서 일관된 고품질 결과를 유지하도록 돕는 운영 구조·프레임워크를 설계하는 것을 의미
- OpenAI·Anthropic 등 선도 AI 기업들이 서로 다른 접근 방식을 통해 하네스 엔지니어링을 발전시키며, AI 책임 개발·배포 전략 수립의 핵심 기술로 부상
- AI의 책임감 있는 개발과 배포를 위한 정책 및 실행 전략 수립에 있어 중요한 시사점

22) https://www.mk.co.kr/mirakleai/newsletter/page/3145
23) https://www.mindstudio.ai/blog/what-is-harness-engineering-beyond-prompt-context-engineering

▸ 급격한 AI 진화로 'AI 블루(Blue)' 현상 확산

- 숙련 개발자 실력을 순식간에 압도하는 AI 작업 능력 앞에서 업계 전반의 무력감·불안감이 구조적으로 확산되는 현상으로 정의
- AI가 급속히 진화하는 핵심 원인은 모델 자체보다, 오류·실수를 차단하는 시스템 환경인 하네스 엔지니어링의 빠른 발전에 기인[24)]

▸ '말을 잘 거는 기술'에서 '시스템을 안전하게 설계하는 기술'로 전환[25)]

- AI 모델 자체의 성능 개선보다, 모델이 오류 없이 정확하게 작동할 수 있는 환경을 체계적으로 설계·구축하는 역량이 핵심 경쟁력으로 부상

엔지니어링 발전 단계별 변화

단계	시기	핵심기술	작동방식	한계
프롬프트 엔지니어링	2022~2024	단일 명령어 정교화	"잘해봐" (단순 지능 의존)	긴 작업 시 기억 손실·환각 빈발
콘텍스트 엔지니어링	2025	시스템 설계·피드백 루프	"이거 보고 잘해봐" (증강 지식 의존)	자율 실행·수정 역량 부족
하네스 엔지니어링	2026~	시스템 설계·피드백 루프	"안전한 환경에서 일해봐" (시스템 의존)	-

- 초기 프롬프트 엔지니어링(2022-2024) 단계에서는 단일 명령어의 정교화가 매우 중요하고 일회성 답변의 품질은 양호한 편이었으나, 길고 복잡한 업무 시 '기억 손실'과 '환각'의 문제점 등이 수시로 노출
- 다음 단계인 콘텍스트 엔지니어링(2025)에서는 이메일에 관련 참고 서류를 첨부하는 법을 알려주는 등 RAG(Retrieval-Augmented Generation)와 동적 컨텍스트 창 관리가 핵심 기술이지만 AI 모델이 스스로 실행하고 수정하는 '자율형 에이전틱(Agentic)' 역량은 부족
- 하네스 엔지니어링 단계부터는 AI 에이전트가 자율 실행할 수 있는 사무실·도구·규칙을 통째로 설계하는 것이 핵심
- 시스템적 제어와 피드백 루프를 통해 신뢰성·정확성을 확보하며 '자율형 AI 에이전트' 완성 단계로 급가속 중

24) https://madplay.github.io/post/harness-engineering#google_vignette
25) https://www.epsilla.com/blogs/harness-engineering-evolution-prompt-context-autonomous-agents

II 하네스 엔지니어링의 핵심 구성 요소

▸ **하네스 엔지니어링의 핵심 구성 요소**[26]

- OpenAI의 시니어 엔지니어인 라이언 로포폴로(Ryan Lopololo)에 따르면,

① Repository 문서화 : AI가 읽고 이해할 수 있는 정형화된 코드베이스

② AGENTS.md : 에이전트 전용 행동 강령 및 표준 운영 절차(SOP)

③ Linter : 구문 오류를 즉각 수정하게 하는 자동 피드백 루프

④ 구조 테스트 : 비즈니스 로직과 보안 준수 여부의 자동 검증

⑤ 리뷰 자동화 : 숙련 개발자를 대신하는 코드 품질 평가 시스템

⑥ 관측 가능성(Observability) : AI 모델의 사고 과정을 추적하는 로그 시스템

⑦ 워크트리(Worktree) : 안전한 실험을 위해 격리된 복제 작업 공간

▸ **하네스 엔지니어링은 AI의 반복적 실수 패턴을 해결하는 '시스템적 안전벨트' 이자 '환경적 가드레일'**[27]

- **(기억 손실)** AI 모델의 복잡·장기 작업 시 맥락을 잊고 엉뚱한 답변을 반환하는 반복적인 기억손실(Memory Loss)문제를 해결하기 위해 하네스 시스템 내 '상태 저장소' 구축 및 작업 이력 동기화 설계 필요[28]
- **(규칙 위반)** 권한 밖 실행을 원천 차단하기 위해 물리적 샌드박스와 API 접근 권한 제어(IAM, Identity and Access Management)[29] 체계를 하네스 시스템에 강제 적용 필요
- **(자기 착각)** AI가 잘못된 결과를 정답으로 확신하는 문제 해결을 위해 다른 모델 또는 확정적 코드가 결과를 재검토하는 '검증 단계(Evaluation Loop)' 설계 필요
- **(오류 복제)** 잘못된 패턴의 반복 답습 예방을 위해 중간 단계마다 '작업 분해'를 실시하여 오류가 다음 단계로 전이되지 않는 격리 장치 설계 필요

▸ **하네스 엔지니어링의 4대 핵심 구성 요소**[30]

- **(도구 연결)** AI가 외부 환경(DB·API·파일 시스템)에 접근할 수 있는 안전·표준화된 통로를

26) https://themiilk.com/articles/a8e11a610
27) https://mailchi.mp/e55b80f3fc4e/gpt-18254259?e=ca7cadb43a
28) http://www.itdaily.kr/news/articleView.html?idxno=239057
29) API 접근 권한 제어는 인증(Authentication)과 인가(Authorization)를 통해 민감한 데이터를 보호하고 보안을 강화하는 필수 절차. 주요 방식으로는 OAuth 2.0, API 키(API Key) 인증, 역할 기반 접근 제어(RBAC), IP 화이트리스트 및 리소스 기반 정책 등을 사용하여 적절한 사용자나 서비스에만 리소스 접근 권한을 부여하고 제어
30) https://nextsignalprediction.substack.com/p/decode-the-buzzword-why-harness-engineering

제공하며, MCP(Model Context Protocol) 등 표준을 통해 도구 인식·사용 방식을 정형화

- **(지식 저장소)** AI 전용 문서화·지식 저장소를 구축하여 "근거 없는 답변"을 원천 차단하며, "지식 저장소에 없는 것은 존재하지 않는 것"이라는 원칙 하에 확인 기반 실행 환경 구성
- **(작업 분해)** 거대 업무를 명확한 세부 작업 단위(Sub-tasks)로 분할하고 각 단계의 품질을 검증하는 DAG(유향 비순환 그래프) 기반 워크 플로우를 설계
- **(피드백 루프)** 실행 결과를 분석해 AI에 반복적 수정 기회를 부여하고, Linter·Unit Test 등 확정적 검증 도구와 디렉토리·명령 제한을 통해 시스템 신뢰성 확보

Ⅲ OpenAI와 Anthropic의 접근방식 비교

▸ **양사 공통적으로 도구 연결·문서화·컨텍스트 관리·작업 분해·피드백·평가를 강조**

- OpenAI는 '좋은 모델만으로 좋은 에이전트가 되지 않는다'는 사상 하에 저장소 구조·문서 체계·테스트 자동화·리뷰 흐름 등 개발 환경 설계가 성능을 결정한다는 관점에서 하네스 설계를 중시[31)]
- Anthropic은 하네스를 입력 처리·도구 호출 조율·결과 피드백을 아우르는 실행 시스템이자, AI 모델 발전에 따라 지속 업데이트가 필요한 '운영 설계'로 규정[32)]
- 양사 모두 하네스 엔지니어링을 AI 모델 성능이 아닌, 안전하고 신뢰성 있는 작동 환경을 설계하는 것이 AI 도입 성패를 좌우한다는 공통 인식에서 출발

▸ **하네스 설계에 있어 서로 다른 관점과 접근방식을 추구[33)]**

- OpenAI는 추론 능력(o-series 등) 극대화와 이를 지원하는 *범용 SDK 제공이 특징이며, 에이전트(Agents) SDK를 통해 개발자가 직접 환경을 구성하도록 유도

  * 범용 SDK : 특정 서비스·플랫폼에 한정되지 않고 다양한 환경과 시스템에서 공통적으로 활용할 수 있도록 만든 개발 도구 모음

- 오픈소스 성향이 강하며, 다양한 도구를 연결하여 AI가 넓은 범위의 작업을 처리하도록 유도함에 따라 AI 생태계의 유연성, 고성능 추론 모델과의 수직적 통합 가능
- Anthropic은 AI가 길고 복잡한 일을 끝까지 해내게 하는 '작업 구조'를 강조. 무엇보다 안전과 신뢰성을 중요하게 생각하며 AI가 복잡하고 긴 작업을 오류

31) https://openai.com/ko-KR/index/harness-engineering/
32) https://www.anthropic.com/engineering/harness-design-long-running-apps
33) https://stibee.com/api/v1.0/emails/share/xiVJMx4PPONdj8y84dYjlyf194PC014

없이 완수할 수 있는 관리형 실행 환경을 중시

- Claude Code와 전용 런타임을 통해 세션 기반 장기 작업(Long-running sessions)을 지원하며, 인프라 수준의 샌드박스로 보안성을 강화
- MCP 기반 표준화된 도구 생태계를 통해 안정적 작업 완수율을 중시[34)]

▸ **두 기업은 하네스 설계 철학에서 '인프라 중심'과 '세션 중심'이라는 뚜렷한 차이를 보이고 있음**

① OpenAI는 하네스를 지능적 에이전트를 위한 '보조 시스템'으로 인식

- 모델(o1, o3 등)의 추론 능력이 극대화되도록 주변 도구를 배치하는 데 주력
- AI를 '유능한 신입사원'으로 대우하며, 이들이 실수하지 않도록 풍부한 매뉴얼과 다양한 검증 툴을 제공해 주는 방식을 활용
- 고도의 자율성을 가진 단일 에이전트로 생산성 극대화를 도모

② Anthropic은 하네스를 장기 실행 앱을 위한 '세션(Session) 아키텍처'로 인식

- AI 모델이 일부 오류가 발생하거나 멈추어도 전체 시스템은 멈추지 않는 설계를 지향
- Harness(조율자): 전체 업무의 순서와 판단을 관리하는 두뇌 역할을 담당
- Session(일지): 모든 대화와 상태를 저장하여 모델이 교체되어도 작업 연속성을 보장
- Sandbox(작업장): 실제 코드가 실행되는 안전하고 독립적인 환경 조성

▸ **성공적인 AI 도입의 핵심 "어떤 AI 모델을 쓰느냐" → "어떤 하네스를 설계하느냐"**

- **(결정론적 울타리)** AI가 자유롭게 창의성을 발휘하되, 최종 결과물은 반드시 확정적 테스트(CI/CD 파이프라인 등)를 통과해야만 배포되도록 설계[35)]
- **(관측 가능성)** AI가 왜 그런 판단을 내렸는지 추적 가능한 로그와 실행 영속성(Execution Persistence) 기능을 하네스 시스템에 필수 내장
- **(인간 협업 구조)** DB 삭제·결제 실행 등 치명적 동작 직전 인간의 승인을 요구하는 Human-in-the-loop 체크포인트를 하네스 레벨에서 강제 설계
- **(도구 미니멀리즘)** 도구 수를 대폭 줄였을 때 작업 성공률이 비약적으로 상승하는 하네스 최적화(Harness Optimization) 원칙을 중요시[36)]
- 결국, 설계 철학의 방점은 AI 성능 극대화가 아닌, AI의 실수·이탈을 사전에 잡

34) https://www.anthropic.com/engineering/managed-agents
35) https://davletd.medium.com/deterministic-ai-building-reliability-around-intelligence-ada734c9234a
36) https://achan2013.medium.com/how-many-tools-functions-can-an-ai-agent-has-21e0a82b7847

아낼 수 있는 '통제 가능성(Controllability)' 확보에 있음

Ⅳ 주요 시사점 및 향후 전망

▸ **AI 모델의 위험·오류를 사전 차단하는 현실적 해결책이자 기술적 기반으로 자리매김 전망**

- AI 모델의 책임성 있는 개발과 배포를 위해서는 하네스 엔지니어링 관행을 포함하는 표준, 가이드라인, 규제 프레임워크를 수립하는 것이 중요
- AI 모델의 오용을 방지하고, 잠재적인 위험을 완화하며, 대중의 신뢰 구축에 직결
- 개발 초기 단계부터 하네스 엔지니어링을 고려한 안전성과 신뢰성 기반 설계 필요
- 'AI 블루' 시대의 무력감을 발전적으로 극복하려면, 조직은 물론 개인들도 AI 모델의 '사용자'를 넘어 '환경 설계자'가 될수 있는 역량 개발이 필수적
- 향후 하네스 엔지니어링 능력은 AI 시대의 필수 직무 역량으로 부상할 것으로 예상되며, 민관 협력 교육 프로그램을 통해 안전·통제 가능한 AI 시스템 구축 역량 적극 지원 필요

▸ **공공부문 하네스 엔지니어링의 주요 시사점 4가지**

- **(보안 환경)** 외부 API 호출보다 내부망에서 작동하는 '에어갭(Air-gap) 하네스' 설계를 우선 검토하고, 보안 중심의 프라이빗 하네스 환경 구축에 집중
- **(업무 특화)** 범용 AI 모델 도입보다 '재무 특화 하네스', '법률 특화 하네스' 등 도메인 지식과 제약 조건이 사전 세팅된 맞춤형 하네스 패키징 적용을 통해 생산성 혁신 필요
- **(인력 육성)** 프롬프트 작성 역량보다 AI가 실수 없이 작동할 수 있는 워크플로우를 설계하고 도구를 적시에 연결하는 '하네스 아키텍트' 육성 체계 마련이 시급
- 하네스 엔지니어링은 AI를 '비서'에서 '자율적 팀원'으로 격상시키는 핵심 기술, AI 경쟁력은 하네스 환경을 어떻게 설계·구축하느냐가 결정적 변수로 작용할 전망

▸ **AI 도입 성과는 모델 선택이 아닌 운용 환경 설계 역량에서 결정**

- 동일 모델 대비 하네스 최적화만으로 성능이 30위권→5위권으로 도약한 사례는 공공기관의 AI 도입 예산이 모델 구매보다 운용 체계 설계에 집중되어야 함을 시사
- AI 실수를 담당자 개인 책임이 아닌 시스템 설계 문제로 규정하는 조직 문화 전환과 AI 에이전트 도입 전 가드레일·데이터 거버넌스·피드백 순환 3대 축을 선행 설계가 선결 과제 대기 중으로 제도적 검수 기준 마련 시급

04

EU ‘AI 대륙 행동 계획’ 1년간의 주요 성과

Reading Point

- EU, AI 분야 글로벌 주도권 확보를 위한 'AI 대륙 행동 계획' 수립·발표(2025.4.9)
- 2025년 4월, EU는 미국·중국 주도의 AI 패권에 대응하여 유럽을 글로벌 AI 선도 대륙으로 도약시키기 위한 종합 정책 전략으로 'AI 대륙 행동 계획'을 공식 수립
- 인프라·데이터·혁신·인재·규제 5대 핵심 영역을 축으로 세부 과제를 체계화하였으며, 선언적 전략을 넘어 실행 가능한 구조로의 전환을 핵심 기조로 설정
- AI 팩토리·데이터 연합·디지털혁신허브·AI 역량 아카데미·AI협약 등 영역별 실행 수단을 동시 가동하여 산업·공공·연구 전 분야에 걸친 통합 생태계 조성을 목표로 추진

Ⅰ EU 'AI 대륙 행동 계획'(2025.4.9) 내용 요약

▸ EU가 AI분야의 글로벌 리더로 자리매김하고자 'AI 대륙 행동 계획 발표'

- 2025년 4월, EU는 글로벌 차원에서 AI분야의 선도 대륙으로 자리 매김하기 위한 정책 전략으로 'AI 대륙 행동 계획'을 수립

EU 'AI 대륙 행동 계획' 영역 및 세부과제

영역		세부 과제명
인프라	컴퓨팅 인프라 구축	▸ AI 기가팩토리 및 안테나의 구축, 지원 프레임워크 구축
데이터	AI 데이터 공유 증진	▸ 데이터 연합 전략, 데이터 실험실, 공유클라우드 ‘Simpl’, 대규모 언어 데이터 통합 프로젝트 툴(ALT-EDIC)
혁신	AI 혁신 촉진 및 도입 가속화	▸ 산업·공공 등 14대분야 사례 접근, 유럽 디지털 혁신 허브, 유럽 AI 연구개발 촉진
인재	AI 역량 및 인재강화	▸ AI 전문가 확대, 역량 및 재훈련 프로그램 개발
규제	AI 규제준수 및 절차 간소화	▸ AI법 이행 및 AI법 서비스 데스크, 규제간소화를 위한 AI 협약

- 특히 인프라, 데이터, 혁신, 인재, 규제 등의 5대 핵심 영역에서 세부과제 제시
- 단순 선언적 계획에 그치지 않고 AI 팩토리 19개 구축, 혁신허브 227개 운영, AI 협약 700여 개사 참여 등 실행 가능한 구조로의 전환을 1년 내 가시화

II EU ‘AI 대륙 행동 계획’ 1년의 주요 성과

▸ 컴퓨팅 인프라 구축 : AI 팩토리 및 팩토리 안테나, 지원 프레임워크 마련

- **(내용)** 고성능 슈퍼컴퓨팅 기반의 'AI 팩토리(AI Factory)' 네트워크 구축, 더 강력한 연산 능력을 갖춘 'AI 기가팩토리' 건설, 대규모 AI 컴퓨팅 생태계 지원을 위한 프레임워크인 통합 지원체계의 구축, 그리고 데이터 센터 용량을 3배로 늘리기 위한 '클라우드 및 AI 개발법' 제안 등이 계획에 포함

- **(추진성과)**

  ① AI 팩토리 19개 구축 및 운영 : AI 분야의 혁신, 협력 및 개발을 촉진하는 역동적 생태로써 16개국, 19개 팩토리를 통해 슈퍼컴퓨팅 파워, 데이터 및 인재를 한데 모아서 최첨단 AI 모델과 애플리케이션 개발 진행 중

- ② 지역 접근성을 높이기 위한 AI 팩토리 안테나 13개 구축 : EU 전역의 기술 격차 해소를 위해 자체 슈퍼컴퓨터가 없는 국가에 벨기에 등 7개 회원국과 영국 등 파트너국에서 국가의 AI 생태계를 지원

- ③ ’클라우드 및 AI 개발법(Cloud and AI Development Act, CADA)‘ : 산업계 및 이해관계자를 대상으로 증거 수집 및 공 컨설팅 완료(’25.4월~7월), 당초 2026년 1분기 제안을 목표로 했으나, 현재는 세부 조율 및 초기 준비 단계를 마쳤으며 곧 공식 법안 제출을 앞두고 있음

  ※ 센터 용량을 현재 대비 3배 확대, 센터 건설 허가 절차의 간소화, ‘공공부문 단일 클라우드 정책 수립’ 등 디지털 주권 강화

- ④ 지원 프레임워크 : AI 기가팩토리 선정을 위한 76건의 의향서 접수, 금융 지원 구조 마련, AI 팩토리 등의 운영 근거 규정의 개정 완료

- **(향후계획)** '26년까지 13개 이상 AI 팩토리 완전 가동, 약 34조원(200억 유로) 규모 InvestAI 펀드를 투입하여 보건· 제조 등 특정 산업 최적화 컴퓨팅 자원을 제공하는 AI 기가팩토리 최대 5개 구축 예정

▸ AI 데이터 공유 증진 : 데이터 연합·실험실·공유클라우드·언어 협력체로 범유럽 데이터 생태계 구축

- **(내용)** 고품질 데이터에 대한 접근성을 높이기 위해 '데이터 연합(Data Union) 전략' 수립, AI 팩토리 내에 데이터를 관리·정제하는 '데이터 연구소(Data Lab)' 설립, 오픈 소스 기반 공유 클라우드 플랫폼인 ‘Simple’ 제공, 유럽 언어 데이터 인프라 협력체(ALT-EDIC) 구성 등을 계획

- **(추진성과)** ① 데이터 연합 전략 : EU 단일 데이터 시장 구축을 위해 산업별 데이터 공간(Common European Data Spaces) 확장을 위한 기반 마련 및 데이터법(Data Act)과의 연계를 통해 데이터 재사용 권한의 명확화
- ② 데이터 실험실(Labs) : ’25년 말부터 주요 AI 팩토리 거점에 AI 전용 슈퍼컴퓨터와 연계하여 시범 데이터 실험실의 설치를 시작했으며, 개인정보보호 기술(PET)을 적용한 데이터 처리 샌드박스를 구축 중
- ③ 공유 클라우드 ‘Simpl’ : 25년 말, 'Simpl-Open'를 릴리스하여 유럽 내 공통 데이터 공간(Common European Data Spaces)에 배치되어 의료, 농업 등 10대 분야와 공공행정의 행정 데이터 공유체계에 도입·적용되어 사용 중
- ④ 유럽언어 데이터 인프라 협력체(ALT-EDIC) : 미국 주도의 LLM 한계 극복을 위해 프랑스, 폴란드 등 주요 회원국 주도로 5억 단어 이상의 다국어 말뭉치를 확보하고 통해 유럽 특화 모델인 '유로 링구아(EuroLingua)' 학습을 지원 중
- **(향후계획)** '27년까지 분야별 데이터 공간을 연결한 범유럽 통합 데이터 연합 완성, '26년까지 수조 토큰 단위의 데이터 공유 확대를 통해 국경을 넘는 EU 통합 데이터 시장 창출 예정

구분	주요 내용	추진성과
데이터 연합 전략 (Data Union)	의료, 제조 등 9개 핵심 산업분야에 공통 유럽 데이터 공간을 공유· 활용하여 상호 운용성을 확보하는 전략	의료 데이터 공간(EHDS) 등 주요 도메인별 데이터 표준화 완료 및 EU 회원국 간 데이터 교환 시스템을 가동 중
데이터 실험실 (Data Lab)	가치있는 원천 데이터를 AI 학습용 고품질 데이터로 가공·검증·테스트하는 민관 협력형 R&D 센터	제조 등의 분야를 중심으로 10여 개의 거점 실험실 구축, 데이터 정제 등 자동화 기술 보급
공유 클라우드 'Simpl'	데이터 공간을 연결하는 오픈소스 S/W, 클라우드 간 안전한 데이터 전송과 신뢰를 보장하는 인프라 구축	'Simpl-Live' 버전을 출시하여 기존 클라우드 프로젝트와 성공적으로 통합, 실시간 데이터 공유 환경 구축
유럽언어 데이터 협력체 (ALT-EDIC)	영어 편향성을 극복하기 위해 유럽 내 24개 공식 언어 데이터를 수집·공유하는 범유럽 협력 기구	유럽 내 각국 공공 데이터를 통합해 다국어 LLM(대형언어모델) 개발을 위한 대규모 말뭉치를 확보

▸ **AI 혁신 촉진 및 도입 가속화 : 산업 현장 밀착형 거점 네트워크로 실질적 AI 전환 견인**

- **(내용)** 전략 산업분야에서의 'AI 적용(AI Apply) 전략'과 과학 연구를 위한 'AI 과학(AI in Science) 전략'을 추진, 현장 밀착형 지원 거점인 유럽 디지털 혁신 허브 구축·운영, 유럽의 전략적 AI 연구개발(R&D) 프로젝트 등 계획 수립
- **(추진성과)** ① 산업 및 과학분야의 AI 가속화 : 분야별 등대 프로젝트 실시로 보건, 자동차, 제조 분야에서 100개 이상의 선도적 AI 적용 사례를 발굴하고,슈퍼컴퓨팅 공동사업(EuroHPC) 자원의 약 30%를 과학 연구개발용 AI 학습에 우선 배정 후 문헌분석과 가설 설정을 돕는 과학특화 AI 도구를 개발하여 배포 실시
- ② 유럽 디지털 혁신 허브(EDIH) : 중소기업 대상 AI 도입 전 테스트·기술·교육·네트워킹을 지원하는 유럽 디지털 혁신 허브(EDIH) 227개를 가동하여 5만 개 이상의 중소기업이 슈퍼컴퓨팅 자원 접근 등의 지원을 수혜
- ③ 유럽 AI의 연구개발 촉진 : 대규모 연구혁신 프로그램(Horizon Europe)을 통해 AI 연구에 누적 8.5조원(약 50억 유로) 이상을 투입하고 AI 우수 연구센터(CoE) 네트워크를 강화하여 세계적 수준의 논문·특허 성과 도출 시작
- **(향후계획)** '27년까지 범유럽 초거대 독자 LLM 모델 개발·AI 테스트베드 조성, '30년까지 EU 기업 75%의 AI 도입을 목표로 산업별 맞춤형 'AI 전환 패키지' 확대·보급 예정

구분	주요 내용	추진성과
산업 및 과학분야의 AI 가속화	제조 등 주요 산업분야에 'AI 적용 전략'을 결합하고, 과학적 발견의 속도를 높이기 위한 'AI 과학 전략' 프로젝트 지원	500개 이상 산업 현장에 AI 실증 프로젝트 완료, 100개 이상 선도적 적용 사례를 발굴
유럽 디지털 혁신 허브 (EDIH)	중소기업과 공공기관이 AI를 도입할 수 있도록 기술 테스트, 교육, 투자 자문을 제공하는 원스톱 지역 거점 네트워크 서비스 제공	230여 개의 혁신 허브(EDIH)가 활성화되어 운영 중이며, 3만 개 이상 중소기업이 AI 기술 도입 컨설팅 및 실무교육 혜택 수혜
유럽 AI 연구개발 촉진	전 세계 우수 인재를 유치하고, 유럽 AI 연구 위원회(ERC-AI)를 중심으로 회원국 간 연구 역량을 결합하는 대규모 R&D 전략	유럽 내 AI 우수 연구센터 네트워크를 단일 플랫폼으로 통합 및 대규모 연구개발 예산을 투입

▸ **AI 역량 및 인재 강화 : AI 전문가 확대, 역량 및 재훈련 프로그램 개발**

- **(내용)** 글로벌 AI 인재 유치를 위한 'AI 재능 비자' 제도 및 AI 학위과정 확대, 'AI 역량 아카데미(AI Skills Academy)'를 통해 전문 인력을 양성하는 계획을 수립
- **(추진성과)** ① AI 전문가 확대 :AI 펠로우십 프로그램을 통해 세계 우수 연구자 600명 이상을 유럽 내 대학·기업 연구소에 배치(목표: 1,000명 유치)하고, 범유럽 AI 석사 과정을 개설하여 졸업생 배출을 시작
- ② 역량 및 재훈련 프로그램 개발(Upskilling & Reskilling) : 자동차·제조 등 산업별 AI 맞춤 커리큘럼을 마련하고 유럽 디지털혁신 허브(EDIH)를 통해 현장 AI 활용 가이드 기반 교육을 실시하며, 중소기업 근로자 대상 AI 교육비 지원 바우처 제도로 정착 유도
- **(향후계획)** '30년까지 유럽 성인 80% 기본 디지털 숙련도 확보, 2천만 명 이상의 AI 전문가 양성을 목표로 비자 처리 30일 내 완료·평생 학습 플랫폼 고도화 등 추진 예정

구분	주요 내용	추진성과
AI 전문가 확대	글로벌 AI 인재 유치를 위한 'AI 재능 비자' 제도 및 유럽 내 AI 박사급 인력양성 네트워크 지원강화	'AI 탤런트 패키지' 및 'AI 펠로우 프로그램'을 통해 우수 연구자 유치 확대 및 범유럽 AI 석사 과정 개설로 졸업생 배출 시작
역량 및 재훈련 프로그램 개발	기존 노동자들이 AI를 활용할 수 있도록 돕는 'AI 스킬 아카데미'및 전 산업 종사자 대상 재교육(Reskilling)을 지원	유럽 내 150개의 'AI 직업 훈련 센터' 지정, 중소기업 임직원 및 공공부문 종사자 100만명에게 AI 기초 및 실무 교육 완료

▸ **AI 규제준수 및 절차 간소화 : AI법 이행 및 AI법 서비스 데스크, 규제 간소화**

- **(내용)** EU AI법(AI Act)의 원활한 이행을 지원하고, 스타트업 등 기업들이 AI법의 핵심 원칙을 자발적이고 선제으로 준수하도록 유도하는 민관 협력 프로그램인 'AI 협약(AI Pact)' 등 계획을 수립
- **(추진성과)** ① AI법 이행 : '수용 불가능한 위험' AI 관행 금지 조항 발효('25.2월)·범용 AI 모델 투명성 의무 적용('25.8월) 등 EU AI법을 단계적으로 이행하며, 유럽 AI 사무국을 중심으로 각 회원국별 감독 기관 지정 및 협력체계 구축 완료
- ② AI법 서비스 데스크 운영 : 법적·기술적 자산이 부족한 기업을 위해 5,000건 이상의 규제 상담을 처리하는 AI법 서비스 데스크를 운영하고, AI 위험 자가진단 도구를 배포('25. 하반기)하여 기업의 규제 대응 비용 절감 지원

- ③ 규제 간소화를 위한 AI 협약 : 구글·삼성(유럽법인) 등 글로벌 기업부터 유럽 내 700여 개 스타트업이 AI 협약에 서명하여 안전성 테스트 결과 공유·투명성 보고서 작성 등 자율 규제 문화 정착 중
- **(향후계획)** '26.8월 의료·교육·채용 등 고위험 AI 시스템 전면 적용, AI 협약 참여 기업 대상 인증 절차 간소화 '패스트 트랙' 혜택 및 인센티브 제도 도입 등 규제 이행과 혁신 지원의 병행 구조 강화 예정

구분	주요 내용	추진성과
AI법 이행	EU AI 오피스(AI Office)를 중심으로 AI법의 일관된 적용을 감독, 고위험 AI에 대한 안전 가이드라인 수립	AI 오피스가 공식 출범하여 범유럽 차원의 감독체계 구축, 기업이 참조할 수 있는 '분야별 고위험 AI 식별 표준' 배포
AI법 서비스 데스크	기업이 AI법을 준수하는 과정에서 겪는 법적·기술적 어려움을 해결해 주는 실시간 상담 및 지원 창구	온라인 통합 지원 플랫폼을 통해 5,000건 이상의 규제 상담 처리, 자동화된 'AI 위험 자가 진단 도구' 제공
규제 간소화용 AI 협약	기업이 자발적으로 안전 원칙을 준수하도록 유도하고, 규제 샌드박스를 통해 절차를 간소화하는 민관 협력체 운영	구글 등 글로벌 기업과 유럽 내 강소기업 100여 개가 참여하는 'AI 협약' 체결 및 기업의 자율규제 문화 정착 시도

Ⅲ 주요 분야별 정책적 시사점 및 한계점

▸ **(컴퓨팅 인프라 구축)**

- **인프라 공유 개방 및 생태계 보호** : 고가의 슈퍼컴퓨팅 자원을 스타트업에 무료 개방하여 기술 격차를 해소하고 빅테크 의존도를 낮춘 EU의 전략적 자율성 확보 모델을 국내 공공 AI 인프라 설계에 참조 필요
- **지역적 접근성 강화** : EU의 '안테나' 개념을 응용하여 수도권 외 지역·소규모 연구기관도 고성능 인프라를 원격으로 활용할 수 있는 네트워크 설계 방식의 국내 도입 검토 필요
- **지속 가능한 정책 프레임워크** : 클라우드 자생력을 높이기 위한 EU의 법안(Cloud and AI Development Act)을 참고하여 민간 투자의 불확실성 제거 필요

▸ **(AI 데이터 공유 증진)**

- **산업별 통합 거버넌스 강화** : 의료, 제조 등 산업별로 공공·민간 데이터를 상호

연결하여 실질적 비즈니스 창출로 이어주는 '데이터 연합체'와 통합 유통 체계 고려

- **즉시 활용 가능한 데이터 가공 서비스** : 단순 대량 개방을 넘어 기업이 AI 학습에 즉시 투입할 수 있도록 정제·라벨링·익명화 서비스를 제공하는 전문 거점을 확충하고 고품질 데이터셋 공급 체계 고도화 필요
- **기술 주권 및 상호 운용성 확보** : 국산 클라우드 경쟁력 강화와 병행하여, 특정 외산 클라우드 기업에 종속되지 않도록 플랫폼 간 데이터 이동을 자유롭게 하는 한국형 AI 상호운용성 기술 표준 확보 필요

▸ **(AI 혁신 촉진 및 도입 가속화)**

- **'산업별 특화 AI' 육성** : 우리나라의 제조 강국의 이점을 살려 '산업별 특화 AI'를 육성하고, 연구개발 예산을 AI 기반 연구 혁신에 집중 투자하는 방안을 고려
- **'AI 혁신 거점'을 산업 현장과 연결** : EU 혁신 허브(EDIH)가 공장·사무실 간 데이터를 즉시 전달하는 파이프라인 역할을 수행하듯, 국내 'AI 혁신 거점'도 실제 장비·연산 자원 테스트와 산업 현장 연결의 실질적 전진기지로 기능하도록 산학 연계 구조 강화 필요

▸ **(AI 역량 및 인재 강화)**

- **현장 중심의 '리스킬링' 가속화** : 고령화와 국내 산업구조 변화에 대응하기 위해 단순 소프트웨어 교육을 넘어, 유럽의 EDIH 연계 모델을 참조하여 기존 숙련공들이 AI를 도구로 쓸 수 있게 만드는 산업 밀착형 재교육 모델의 채택 고려
- **글로벌 인재 유치 경쟁력 제고** : 국내 대학 위주의 인재 양성뿐만 아니라, 해외 우수 인력이 국내 AI 생태계에 정착할 수 있도록 비자 제도 개선과 파격적인 연구개발 환경 제공 등 적극적인 '인재 인바운드' 전략 마련 필요
- **공공-민간 협력 기반의 아카데미** : EU의 'AI 역량 아카데미'와 같이, 우리 기업이 필요로 하는 AI 기술을 즉각 가르치는 민관 합동 교육 플랫폼을 마련하고, AI 관련 교육 수강자 등의 현장 재배치 강화 및 일자리 미스매치 해소에 기여 필요

▸ **(AI 규제준수 및 절차 간소화)**

- **지원 인프라 동시 구축** : 규제 법안만 제정하는 것이 아니라, '서비스 데스크'와 같은 행정 지원 도구를 동시에 마련하여 기업의 규제 대응 비용을 낮춰 주어야 함
- **자율 규제의 제도적 활용** : 'AI 협약' 모델을 참고하여, 법 집행 전까지 기업들이 자율적으로 안전 가이드라인을 지키도록 유도하고, 이에 참여한 기업에 실질적인 혜택(공공조달 가점 등)을 주는 방안이 필요

- **글로벌 정합성 확보** : EU AI법은 글로벌 표준(Brussels Effect)이 될 가능성이 높으므로, 우리나라 기업이 유럽 시장에 진출할 때 별도의 큰 비용 없이 대응할 수 있도록 국내 규제 체계를 EU와 상호 조화시키는 노력이 중요

▸ **주요 한계점**

- **투자 부족 우려** : 약 34조원(약 200억 유로) 규모 InvestAI 중 신규 자금은 10% 내외이며 나머지는 기존 기금 재활용으로 민간 투자 유치 및 예산 확보가 행동 계획 전체의 최대 리스크로 부상
- **규제의 복잡성** : 엄격한 AI법(AI Act)이 유럽 내 AI 스타트업의 혁신 속도를 저해하고 규제 불확실성을 증대시켜 EU 역내 기업이 오히려 해외로 이탈하는 역효과 우려가 상존
- **기술 격차 해소** : 슈퍼컴퓨터 인프라를 갖춰도 핵심 AI 파운데이션 모델 개발 역량 및 인재 측면에서 미국·중국 등 선도국과의 격차 해소에 상당한 시간이 소요될 것으로 예상

Ⅳ 종합 결론

▸ **EU 'AI 대륙 행동 계획'은 AI 주권 확립·인프라 자립·규제와 혁신의 균형이라는 3대 축에서 한국형 소버린 AI 전략의 실질적 벤치마킹 모델로 기능**

- **AI 주권 확립 노력** : 미국·중국 등에 대항하여 유럽을 글로벌 AI 기술 선도 대륙으로 도약시키기 위한 종합적인 산업 육성 및 도입 가속화를 위한 핵심 전략인 'AI 행동 계획'은 독자적인 한국형 소버린 AI를 마련하려는 우리나라에도 유용한 글로벌 전략의 본보기
- **인프라 자립화** : 슈퍼컴퓨팅 자원을 산업 생산 기지로 정의한 EU의 'AI 팩토리' 모델은 우리나라의 AI 컴퓨팅 인프라 확충 전략에도 유효한 정책으로 활용 가능
- **규제와 혁신의 균형** : EU가 법 제정 이후 '간소화 패키지'를 제시한 것을 참조, 우리나라도 규제 정립과 동시에 민간의 준수 부담을 줄여주는 지원책의 병행 필요
- **특화 산업의 집중 공략** : 제조 등의 전통 산업의 강점을 AI로 연결하는 EU의 '실제 적용 중심' 접근법은 국내의 제조업 기반 AI 경쟁력 강화 전략 마련에도 적절한 벤치마킹 사례로 활용 가능

AI.GOV 해외동향 지난 호 보기

2025-1호

- 해외 AI 에이전트 기술 및 서비스 개발 동향
- LLM을 보완하는 RAG 2.0 기술 현황
- 정부의 적극적 AI 활용 정책과 반향 : 미국 및 영국 사례
- 주요국 AI 인프라 투자 경쟁 가속화
- [OECD 최신 보고서] AI의 미래: 예견적 거버넌스 전략

2025-2호

- 미국, 연방기관의 AI 활용 및 조달에 관한 정책 발표
- Stanford HAI, 2025년 AI 인덱스 리포트 발간
- 최신 AI 모델 출시 동향
  (Gemini 2.5 Pro, Llama 4, Nova Act, Qwen2.5-Omni-7B)
- AI 에이전트 연결 기술의 표준 : MCP와 A2A
- EU, AI 대륙 행동 계획 발표
- 캐나다, 연방 공공 서비스를 위한 AI 전략 발표

2025-3호

- 미국, AI 실행계획(America's AI Action Plan) 발표
- EU, 범용 AI 실무규정(GPAI Code of Practice) 공개
- 국내외 최신 AI 모델 개발 동향
- OpenAI의 한국 법인 공식 설립과 주요 시사점

AI.GOV 해외동향 지난 호 보기

2025-4호

- 최신 AI 모델 개발 동향
- 중국 정부, AI+ 행동 심화 실시에 대한 의견 발표
- 일본, [Gennai 프로젝트] 정부 내 생성형 AI 활용 결과보고서 발표
- 일본 정부, 인공지능 전략본부 설치
- OECD ‘인공지능 시대의 정부 거버넌스’ 보고서 발표
- 영국, ‘디지털 더미’ 데이터 관리에 AI 도입 지침 발표

2026-1호

- 최신 AI 모델 개발 동향
- 일본 내각부, ‘인공지능 기본계획’ 발표
- OECD, ‘인공지능과 세계 생산성 격차’ 보고서 발표
- CES 2026 기조연설 및 핵심 트렌드 분석