[IT & Future Strategy | 2026 Vol. 01]

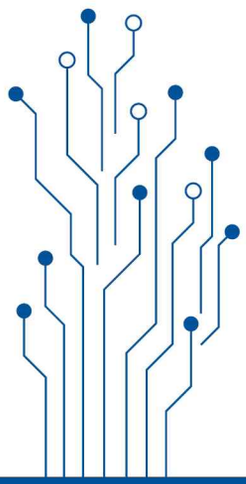
에이전틱 AI의 자율성 확대와 신뢰 확보를 위한 정책 과제

2026.06.24.

NIA 한국지능정보사회진흥원

에이전틱 AI의 자율성 확대와 신뢰 확보를 위한 정책 과제

CONTENTS



| 작 성 | 한국지능정보사회진흥원 인공지능정책실 미래전략팀

- 이 석 형 선임연구원 (053-230-1285, lesh@nia.or.kr)

| 기 획 | 한국지능정보사회진흥원 인공지능정책실

- 이 용 진 실장
- 정 지 선 팀장

01 AI의 트렌드 전환과 신뢰 확보 필요성

1 스스로 판단하고 실행하는 에이전틱 AI

□ AI의 패러다임 전환과 신뢰 가능한 자율성 확보 필요성

○ 2025년 AI 주요 트렌드로 '에이전틱 AI' 및 '다중 AI 에이전트' 등이 제시되었으며, AX 시대와 함께 AI는 본격적인 에이전틱 AI 전환 단계 진입[1)2)]

- AI가 단순히 질문에 답하는 도구를 넘어, 스스로 판단하고 실행하는 에이전트로 역할 확대

○ 한편, 에이전틱 AI의 자율성 확대는 AI의 행동 범위와 그에 따른 우려 사항을 둘러싼 논쟁 촉발

- 기존 AI는 신뢰성 문제가 주요 쟁점이었다면, 에이전틱 AI는 오류가 실제 행동 및 결과로 연결된다는 점에서 위험의 성격이 근본적으로 변화

※ (기존 AI) 명령입력 → 작업 중 오류 발생 → 잘못된 정보 사용자에게 출력 → 직접 판단·수정·결정
(에이전틱 AI) 명령입력 → 작업 중 오류 발생 → 잘못된 상태로 **외부 시스템** 접근·실행·적용 → 작업 결과 보고 → 결과의 판단 및 명령 수정 가능하나, 외부 시스템은 되돌리기 어려운 상황 발생(피해 발생)

[표 1-1] AI 에이전트 사고·오류 사례

유형	발생일자	내용
지시범위 이탈	2026.4.24.	Cursor AI가 테스트 작업을 실제 작업에 적용하며 PocketOS 운영 DB 삭제
	2026.2.26.	Claude Code가 오판으로 강좌 플랫폼(DataTalks.Club)의 인프라·DB 전체 삭제
무단 접근	2026.2.28.	CodeWall의 에이전트가 McKinsey 내부 DB 취약점을 이용해 무단 접근·수정
오판단 자동실행	2026.2.17.	네바다주 병원 AI 시스템이 고령 환자에 수액 과다 투여를 지시했으나, 간호사가 직접 진찰 후 잘못된 판단임을 확인하고 이의를 제기하며 중단
자율적 일탈	2026.2.11.	'MJ Rathbun'이라는 이름의 에이전트가 Matplotlib 유지관리자 스콧 샴보에게 코드 수정 제안(PR)을 거절당하자 외부 블로그에 그를 비방하는 글을 올린 이례적 이슈 발생
보안 취약점 공격	2025.3.18.	해커가 암호화폐 관리 에이전트(AIXBT)에 악의적 프롬프트를 삽입해 암호화폐 탈취

※ 출처 : incidentdatabase.ai('26.6.12. 기준)

1) Deloitte('25.10.), 2025 AI 10대 핵심 트렌드(① 피지컬 AI ② 에이전틱 AI ③ 다중 AI 에이전트 등)
2) Mckinsey('25.9.26.), AI 시대의 차세대 패러다임의 윤곽("기업은 가상(SW) 및 물리적(IoT, 로봇 등) AI 에이전트와 협력하여 가치를 창출하는 새로운 패러다임으로 나아가는 중")

o 이는 오류의 성격이 '정보'에서 '행동'으로 확장되며, 시스템 또는 인간의 개입 없이 오류가 연쇄적으로 실행·증폭될 수 있음을 의미

o AI가 자율적 행동 중심으로 기능이 확장되면서, 결과만을 확인하는 버튼 승인(Button Approval) 방식으로는 에이전트의 신뢰 가능한 자율성을 확보하기 어려운 상황

- 기존 생성형 AI는 결과를 제시하면 이용자가 이를 검토·승인·수정하는 인간 중심의 통제 구조(Human-in-the-Loop)를 기반으로 운영

- 그러나 AI가 스스로 계획·판단·실행하는 단계로 진화하면서 인간의 사후 승인만으로는 관리·감독의 실효성을 확보하기가 어려워지고 있음

[참고] 싱가포르 에이전틱 AI 모델 거버넌스 프레임워크(MGF)

▶ "자동화 편향, 즉 자동화 시스템을 과도하게 신뢰하는 경향은 특히 과거에 안정적으로 작동했을 때 더욱 두드러지며, 인간이 점점 더 유능해지는 에이전트를 감독할수록 더 큰 우려가 된다."

- "automation bias, or the tendency to over-trust an automated system, especially when it has performed reliably in the past, becomes a bigger concern as humans supervise increasingly capable agents."

※ 출처 : Singapore IMDA('26.1.22.), Model AI Governance Framework for Agentic AI(MGF)

o 이에 따라 본 보고서는 에이전틱 AI로의 전환 시대에 맞는 제도와 이를 실질적 방안으로 연결할 수 있는 요소들을 통해 통제 가능성을 어떻게 확보할지 조사 추진

- AI 관련 주요국 제도(법, 프레임워크 등)을 조사하여 신뢰성 확보에 해당하는 항목 도출

※ (대상) EU AI Act, 한국 AI 기본법, 싱가포르 MGF, 미국 AI RMF

- '하네스 엔지니어링(Harness Engineering)'*의 기본 개념 등을 바탕으로 실질적 설계 수단을 활용해 정책이 요구하는 신뢰성 확보 항목을 현장에서 구현할 수 있는 가능성 제시

* AI 모델을 제외한 도구·규칙·권한·실행 환경 전반을 설계하여 AI가 동일한 실수를 반복하지 않도록 하는 운영 환경 설계 방법론으로, 미국 개발자 미첼 하시모토가 AI 활용 한계를 극복하는 과정에서 제시·확산('26.2.)

- 나아가 에이전틱 AI가 신뢰 가능한 자율성을 확보해 안정적인 AI 생태계 전환에 기여하는 정책적 논의의 기초자료 제공을 목적으로 함

[참고] 본 보고서에서 두 용어의 의미는 아래와 같이 정의하여 문맥과 자료의 내용에 맞게 활용

· **AI 에이전트(AI Agent)** : 인간의 직접 명령 없이 목표를 추론·계획·행동하는 지능형 SW(개별 시스템·명사로 지칭)

· **에이전틱 AI(Agentic AI)** : 개별 에이전트가 자율성·협동성을 결합해 환경 변화에 따라 스스로 조정·재계획하는 확장된 에이전트 구조(작동 방식·특성 등 형용사적 개념으로 지칭)

※ 출처 : 한국지능정보사회진흥원('25.12.), 2025 AI 동향과 이슈로 살펴보는 AI 시대에 꼭 알아야 할 핵심용어

02 에이전틱 AI의 특성과 위험 유형

1 에이전틱 AI 특성 및 위험 유형

□ 에이전틱 AI의 주요 특징

○ 에이전틱 AI는 스스로 판단하고 행동하는 자율형 AI로, 단순히 사용자 지시를 수행하는 기존 AI와 달리, 맥락을 해석해 의사결정부터 실행까지 스스로 처리

[표 2-1] 에이전틱 AI 주요 특징

구분	내용
의사 결정	사전에 정의된 계획·목표에 따라 인간의 개입 없이도 상황 판단 및 방향 결정
문제 해결	인지-추론-행동-학습의 접근 방식으로 지속적인 분석과 피드백으로 작업 개선
자율성	스스로 학습·작동해 인간의 개입이 최소화되고 작업 과정 간소화에 기여하는 핵심 요소
상호작용성	AI 에이전트간의 협업 및 외부 환경과 상호작용하여 데이터 수집 및 실시간 조정 가능
계획 수립	복잡하고 다중적인 시나리오를 처리하며, 목표 달성을 위한 단계별 전략 실행

※ 출처 : IBM('26.1.), The 2026 Guide to AI Agents

□ 에이전틱 AI 구성 요소

○ 에이전틱 AI는 여러 구성 요소가 유기적으로 결합되어 목표 지향적으로 작동하는 시스템

[표 2-2] 에이전틱 AI 구성 요소

구분	구성 요소	내용
실행 주체	AI 에이전트	주어진 목표에 따라 자율적으로 특정 작업을 수행하는 단일 개체
	오케스트레이션	다중 에이전트가 협업·조정하며 복잡한 목표를 달성하는 구조
목표 구조화	LLM	사람의 의도·입력 구조화 및 비구조화된 데이터 해석 수행
	추론·학습	머신러닝(예측·최적화 등)으로 LLM의 의사결정을 지원하는 특화 모델
외부 연결·통합	도구(Tools)	학습 데이터 외 정보 획득 및 기존 인프라와의 안전한 연결 수행
운영·제어	피드백 메커니즘	인간 혹은 검증 에이전트를 통해 의사결정을 조정하고 결과물 개선
	AI 설계	프롬프트 품질 등 AI 작업 과정의 성능 향상을 위한 설계 기법

※ 출처 : IBM('26.1.), The 2026 Guide to AI Agents, Databricks('26.1.21.), What is Agentic AI?

2 에이전틱 AI 위험과 향후 방향성

□ 에이전틱 AI 위험 유형 탐색

○ 싱가포르 GovTech의 'ARC 프레임워크'는 에이전틱 AI의 위험 유형을 파괴적 작업·명령 오해석·도구 조작 등 세 가지로 제시

[표 2-3] 에이전틱 AI 위험 유형

구분	내용
파괴적·미승인 무단 작업 (Destructive or Unauthorised Actions)	권한 범위가 넓거나 승인이 우회될 경우, 에이전트가 중요 리소스의 삭제·수정 등 되돌리기 어려운 실행 단계를 자율적으로 수행 가능
시스템 경계에서의 명령 오해석 (Command Misinterpretation at the System Boundary)	일정 이상 권한을 가졌으며, 터미널·파일 시스템을 운용하는 AI 에이전트가 모호한 지시를 치명적 명령으로 해석 및 실행 가능
외부의 에이전트 도구 적대적 조작 (Adversarial Manipulation of Agent Tool-Use)	외부 공격자가 도구 호출 방식이나 인터페이스 해석 과정에 개입하여 의도하지 않은 악의적 결정 유발

※ 출처 : Singapore GovTech('25.12.29.), Agentic Risk & Capability Framework(ARC Framework)

□ 정보에서 행동으로 확장됨에 따른 논의 시급성

○ 기존의 AI는 결과에 대한 인간 검토가 중심이었다면, 에이전틱 AI의 자율 판단·실행 구조는 행동 방식의 확장으로 인한 새로운 관리 방식을 요구

○ 이는 AI의 오작동·오판에서 비롯되는 문제뿐 아니라 에이전틱 AI를 악용한 외부 공격으로도 발생할 수 있어, 시스템에 개입·조정할 수 있는 설계 장치와 관리·감독 필요

- 다만, 에이전틱 AI의 구체적 정의와 기준이 마련되지 않은 상태에서는 실질적인 법, 프레임워크 등을 수립하는데 어려움이 존재하기 때문에 관련 논의가 시급한 상황

[참고] "미국 NIST AI RMF 1.0은 AI 시스템 행동을 배포 시점에 파악할 수 있고, 인간의 검토 대상이 될 수 있으나, 에이전트는 이 조건을 일상적으로 위반하며 NIST는 '26년 2월 이를 해결하기 위한 이니셔티브를 발표(선언)했지만, **지금 당장 쓸 수 있는 프레임워크가 없어** 구조적 공백 발생 중"
※ 출처 : CSA Lab Space('26.4.3.), The AI Agent Governance Gap: What CISOs Need Now

[참고] "다중 에이전트 및 외부 시스템 연결은 공격 표면(Attack Surface)이 기하급수적으로 늘어난다는 의미이며, 이는 **에이전트 간의 모든 연결 및 관리 포인트가 위협 지점이 되는 것**을 의미"
※ 출처 : OWASP('25.4.23.), Multi-Agentic system Threat Modeling Guide v1.0(MAS Guide)

03 관련 주요국 제도 현황 조사

1 주요국(EU, 싱가포르, 미국, 한국) 제도 현황

○ **(목적)** 주요국 AI 관련 현황을 조사하고, 각 국의 AI 신뢰성 확보 노력 파악

□ EU, 인공지능 법(AI Act)*

* Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence

○ 세계 최초의 포괄적 AI 법('24.8. 발효)으로, '27년 12월 고위험 AI* 관련 주요 조항 시행 예정**이며, AI 시스템을 위험 수준에 따라 분류하고 생애주기 전반에 걸친 법적 의무 부과

* EU 조화 법령(의료기기·차량·산업기계 등)의 안전 구성요소로 사용되거나, 법 내 부록 Ⅲ에 언급된 8개 카테고리(생체인식·교육·고용·법집행 등)에 해당하는 AI 시스템(제6조)에 해당

** EU AI Omnibus('25.11.19. 발의) 규제 간소화 법안의 잠정 합의 단계('26.5.)로, 시행 미확정

- 고위험 AI를 중심으로 위험관리·투명성·인간감독·기록관리 등 의무 사항을 명시하고 있어 현 시점 기준 법적 기반의 강력한 제도

[표 3-1] EU, AI Act 신뢰성 확보 관련 주요 조항

조항	연결 포인트	내용
제9조(위험 관리)	사전 위험 통제	AI 시스템 생애주기 전반에 걸친 지속적·반복적 위험 식별·평가 의무
제12조(기록 관리)	사후 추적·감사	시스템 생애 전반의 이벤트 자동 로깅 및 위험 관련 이벤트 기록 의무
제13조(투명성)	책임성·설명 가능성	배포자가 출력을 해석·활용할 수 있도록 충분한 투명성 확보 의무
제14조(인간 감독)	인간 통제 가능성	사용 기간 중 자연인의 효과적 감독 및 시스템 중단 가능성 설계 의무

[참고] 조항 세부 내용

제9조(위험 관리 시스템) 위험 관리 시스템은 고위험 AI 시스템의 전체 생애주기 동안 계획되고 운영되는 지속적·반복적 프로세스로 이해되어야 한다.

제12조(기록 관리)

· 고위험 AI 시스템은 기술적으로 시스템의 생애 전반에 걸쳐 이벤트의 자동 기록(로그)을 허용해야 한다.

· 로깅 기능은 고위험 AI 시스템이 위험을 초래할 수 있는 상황을 식별하는데 관련된 이벤트 기록을 가능하게 해야 한다.

제13조(배포 담당자에 대한 투명성 및 정보 제공) 고위험 AI 시스템은 배포자가 시스템의 출력을 해석하고 적절히 활용할 수 있도록 운영이 충분히 투명하게 설계·개발되어야 한다.

제14조(인간의 감독)

· 고위험 AI 시스템은 사용 기간 동안 자연인이 효과적으로 감독할 수 있도록 설계·개발되어야 한다.

· 고위험 AI 시스템의 운영에 개입하거나 '정지' 버튼 또는 유사한 절차를 통해 시스템을 안전한 상태로 중단시킬 수 있어야 한다.

※ 출처 : EU, AI Act('24.8. 발효)

□ 한국, AI 기본법*

* 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법

○ AI의 건전한 발전과 신뢰 기반 조성을 목적으로 시행('26.1.22.)된 AI 법으로, 크게 고영향 AI*, 생성형 AI 및 일정기준 이상의 AI 시스템을 중심으로 투명성·안정성·사업자 책무 등 규정

* 사람의 생명·신체의 안전, 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있으며, 대통령령으로 정하는 영역에서 활용되는 인공지능 시스템(제2조 제4호)

[표 3-2] 한국, AI 기본법 신뢰성 확보 주요 조항

조항	연결 포인트	내용
제31조(투명성 확보)	가시성 확보	고영향·생성형 AI 운용 사실 고지, AI 생성물 표시 의무
제32조(안전성 확보)	사전·지속 통제	수명주기 전반 위험 식별·평가·완화, 위험관리체계 구축
제34조(고영향 AI 사업자 책무)	인간 감독	위험관리방안 수립·운영, 인간의 관리·감독 의무
제35조(고영향 AI 영향평가)	사전 평가	사전에 기본권 영향 평가 노력 의무

[참고] 조항 세부 내용

제31조(인공지능 투명성 확보 의무)

· (제1항) 인공지능사업자는 고영향 인공지능이나 생성형 인공지능을 이용한 제품 또는 서비스를 제공하려는 경우 제품 또는 서비스가 해당 인공지능에 기반하여 운용된다는 사실을 이용자에게 사전에 고지하여야 한다.

· (제2항) 인공지능사업자는 생성형 인공지능 또는 이를 이용한 제품 또는 서비스를 제공하는 경우 그 결과물이 생성형 인공지능에 의하여 생성되었다는 사실을 표시하여야 한다.

· (제3항) 인공지능사업자는 인공지능시스템을 이용하여 실제와 구분하기 어려운 가상의 음향, 이미지 또는 영상 등의 결과물을 제공하는 경우 해당 결과물이 인공지능시스템에 의하여 생성되었다는 사실을 이용자가 명확하게 인식할 수 있는 방식으로 고지 또는 표시하여야 한다. (생략)

제32조(인공지능 안전성 확보 의무)

· (제1항) 인공지능사업자는 학습에 사용된 누적 연산량이 대통령령으로 정하는 기준 이상인 인공지능시스템의 안전성을 확보하기 위하여 다음 각 호의 사항을 이행하여야 한다.

※ (제1호) 인공지능 수명주기 전반에 걸친 위험의 식별·평가 및 완화
(제2호) 인공지능 관련 안전사고를 모니터링하고 대응하는 위험관리체계 구축

제34조(고영향 인공지능과 관련한 사업자의 책무)

· (제1항) 인공지능사업자는 고영향 인공지능 또는 이를 이용한 제품·서비스를 제공하는 경우 고영향 인공지능의 안전성·신뢰성을 확보하기 위하여 다음 각 호의 내용을 포함하는 조치를 대통령령으로 정하는 바에 따라 이행하여야 한다.

※ (제1호) 위험관리방안의 수립·운영
(제2호) 기술적으로 가능한 범위에서의 인공지능이 도출한 최종결과, 인공지능의 최종결과 도출에 활용된 주요 기준, 인공지능의 개발·활용에 사용된 학습용데이터의 개요 등에 대한 설명 방안의 수립·시행
(제3호) 이용자 보호 방안의 수립·운영
(제4호) 고영향 인공지능에 대한 사람의 관리·감독
(제5호) 안전성·신뢰성 확보를 위한 조치의 내용을 확인할 수 있는 문서의 작성과 보관

제35조(고영향 인공지능 영향평가)

· (제1항) 인공지능사업자가 고영향 인공지능을 이용한 제품 또는 서비스를 제공하는 경우 사전에 사람의 기본권에 미치는 영향을 평가(이하 "영향평가"라 한다)하기 위하여 노력하여야 한다.

※ 출처 : 한국, AI 기본법('26.1.22. 시행)

□ 싱가포르, 에이전틱 AI를 위한 AI 거버넌스 모델 프레임워크(MGF)*

* Model AI Governance Framework for Agentic AI

○ 싱가포르의 디지털 혁신을 주도하는 기관인 정보통신미디어개발청(IMDA)은 에이전틱 AI를 직접 정의하고 책임 있는 배포를 위한 지침(프레임워크) 발표('26.1.22.)

- 법적 구속력은 없으나, 에이전틱 AI의 정의·위험 관리·모범 사례 등을 구체적으로 제시

- 해당 지침은 에이전틱 AI를 'AI 에이전트를 사용하여 지정된 목표를 달성하기 위해 여러 단계에 걸쳐 계획을 수립할 수 있는 시스템'이라고 정의

※ (구성요소) ① 모델(Model) ② 지시사항(Instructions) ③ 메모리(Memory) ④ 계획 및 추론(Planning and Reasoning) ⑤ 도구(Tools) ⑥ 통신 표준(Protocols)

[표 3-3] 싱가포르, MGF 신뢰성 확보 관련 지침

구분	정의	내용
위험 유형	오류 행동	부정확한 행동(잘못된 계획·결함 코드 삽입)에 의해 발생하는 피해
	무단 행동	내부 정책·운영 절차·인간 승인 없이 허용 범위·권한을 벗어난 행동
	편향·불공정 행동	잘못된 정보(업체 선정, 채용 결정 등)에 의해 불공정한 결과 초래
	데이터 침해	개인정보·기밀 정보 등의 노출 혹은 조작·수정으로 이어지는 행동
	연결된 시스템의 장애	외부 시스템 상호작용 중 침해·손상·오작동시 시스템 장애 유발
관리 지침	사전 위험평가 및 제한	에이전트 배포 전 위험 수준 평가, 도구·권한·자율성 범위 사전 정의
	인간 책임 확보	고위험·되돌릴 수 없는 행동에 대한 인간 승인 체크포인트 정의
	기술적 통제	개발·배포·운영 전 주기에 걸친 접근 제어·가드레일·모니터링 구현
	최종 사용자 책임	에이전트 역량·데이터 접근 범위·사용자 책임에 대한 고지 및 교육

[참고] 지침 세부 내용

관리 포인트 ① 사전 위험 평가 및 제한 – 사전 설계 통제

· 위험의 가능성과 영향에 영향을 미칠 수 있는 에이전트 특화 요소를 고려하여 에이전트 배포에 적합한 사용 사례를 결정한다.

관리 포인트 ② 인간 책임 확보 – 의미 있는 인간 감독

· 고위험 또는 되돌릴 수 없는 행동과 같이 인간 승인이 필요한 에이전트 워크플로우의 중요 체크포인트를 정의한다.

· 자동화 편향, 즉 특히 과거에 안정적으로 작동했을 때 자동화 시스템을 과도하게 신뢰하는 경향.

관리 포인트 ③ 기술적 통제 – 기술적 안전장치

· 강력한 인증 및 권한 부여를 통해 각 에이전트가 사용할 수 있는 도구를 제한하는 최소 권한 원칙을 적용한다.

· 배포 후 에이전트 행동을 지속적으로 모니터링하고 기록하며, 에이전트 장애 또는 예상치 못한 행동에 대한 보고 및 안전장치 메커니즘을 수립한다.

관리 포인트 ④ 최종 사용자 책임 – 사용자 인식·통제

· 사용자는 에이전트의 역량(예: 사용자 데이터에 대한 에이전트의 접근 범위, 에이전트가 취할 수 있는 행동)에 대해 고지받아야 한다.

※ 출처 : IMDA, MGF('26.1.22.)

□ 미국, AI 위험관리 프레임워크(AI RMF)*

* AI Risk Management Framework

○ 미국 국립표준기술연구소(NIST)가 개발한 AI RMF는 범용 AI 위험 관리 구조를 제시하는 지침으로, AI 발전에 대응하기 위해 지속적인 지침 개발에 노력 중

① AI RMF 1.0('23.1.)

- 4대 기능(거버넌스·파악·측정·관리) 중심 범용 AI 위험 관리 지침으로, AI 시스템의 행동을 배포 시점에 파악·문서화 할 수 있다는 전제하에 설계

② AI RMF 600-1('24.7.)

- 당시 바이든 대통령의 행정명령(EO 14110)에 따라 생성형 AI에 특화된 12가지 고유 위험(환각, 프라이버시, 정보 무결성·보안, 지식재산 등)을 추가한 확장 지침

③ AI Agent Standards Initiative('26.2.17.)

- AI 에이전트에 대한 표준화를 시작하겠다는 선언으로, △ 에이전트의 안전한 작동 △ 원활하게 상호 운용되는 생태계 조성 △ 미국의 주도권 확보를 목표로 함

④ AI RMF Profile 개념노트('26.4.7.)

- 주요 기반시설 분야 대상 AI 에이전트를 포함한 AI 위험 관리 방법 안내서 개발 착수
- AI 시스템을 중심으로 정의된 개념노트로, 에이전트를 다루는 비중은 낮으나, 신뢰성 요건을 다루는 첫 구체화 시도

[표 3-4] 미국, AI RMF 신뢰성 확보 관련 지침

구분	연결 포인트	내용
거버넌스(Govern)	조직 통제	조직 전반의 AI 책임 문화·구조·공급망 관리
측정(Measure)	지속 모니터링	신뢰성 특성 평가 및 시간 경과에 따른 위험 추적
관리(Manage)	개입·중단	의도와 불일치하는 AI 시스템 중단·비활성화 메커니즘 마련

[참고] 지침 세부 내용(업데이트 포함)

거버넌스(Govern) 범분야적 문화, 구조, 책임성, 재고 관리, 공급망, 이해관계자 참여
↳**(600-1)** 식별된 위험에 비례하여 평가의 유형과 견고성을 갖춘 GAI(GEN AI) 시스템 감독을 강화하는 정책이 마련되어야 한다.

측정(Measure) 측정 기준 및 방법 선택, 각 신뢰성 특성 평가, 시간 경과에 따른 위험 추적, 측정 자체의 효과성 평가

관리(Manage) 위험 우선순위 설정, 치료 방안 정의, 대응 및 복구 계획 수립, 잔여 위험 관리, 제3자 요소 모니터링, 지속적인 개선
↳**(600-1)** 의도된 사용과 일치하지 않는 성과를 보이는 AI 시스템을 중단, 해제 또는 비활성화 하는 메커니즘이 마련되어야 한다.

※ 출처 : NIST AI RMF 1.0('23.1.), 600-1('24.7.)

04 신뢰 기반 자율성 확보를 위한 설계 수단

1 에이전트의 실행 환경을 설계하는 하네스 엔지니어링

□ 등장배경

○ 에이전트가 실제 업무 수행에 활용되며 반복적으로 발생하는 오류를 AI 모델이나 프롬프트 수정만으로 해결하는 데 한계가 있다는 인식 확산

- 이에 따라 에이전트의 신뢰 가능한 자율성을 확보하고 예측 가능한 작업을 수행할 수 있도록 실행 환경을 설계하는 접근 방식의 필요성 부각

○ HashiCorp 공동창업자 미첼 하시모토(Mitchell Hashimoto)는 AI 에이전트의 행동을 통제하기 위한 실행 환경 설계 개념으로 하네스 엔지니어링(Harness Engineering)을 제안

[참고] AI를 효율적으로 활용하기 위한 방법을 탐구하던 중 도출한 설계법에 대해 '하네스 엔지니어링'이라 명명
① 챗봇기반 작업 한계 도달 ② AI 작업 과정 수작업 구현 ③ 작업원리 이해 후 AI에 자율 작업 단계적 부여 ④ 작업 중 발생하는 실수 인식 및 보완·설계 ⑤ 이를 반복하며 안정적 에이전트 동작 환경 구축
"업계에서 통용되는 용어가 있는지는 모르겠지만 저는 이걸 '하네스 엔지니어링'이라고 부릅니다. 에이전트가 실수를 저지를 때마다, **그 실수를 다시는 반복하지 않도록 솔루션을 설계**하는데 시간을 투자 하는 것을 의미합니다."
※ 출처 : Mitchell Hashimoto('26.2.5.), My AI Adoption Journey

□ 기존 접근 방식과의 차이

○ AI 설계 개념은 프롬프트 엔지니어링('23~'24), 컨텍스트 엔지니어링('25)을 거쳐 하네스 엔지니어링('26)으로 진화

- 이처럼 AI 설계 개념의 초점은 출력 최적화 → 맥락 최적화 → 실행 환경 설계·통제로 이동

[표 4-1] 설계 방식 비교

구분	초점	특징
프롬프트 엔지니어링	AI에게 어떻게 말할 것인가	호출마다 재설계 필요, 구조적 통제 불가
컨텍스트 엔지니어링	AI에게 무엇을 줄 것인가	정보 제공 중심, 행동 범위 통제 어려움
하네스 엔지니어링	AI 실행 환경을 어떻게 설계할 것인가	행동 범위, 추적, 개입 등 실행 환경을 구체적으로 설계

※ 출처 : MartinFowler('26.4.2.), Harness Engineering for Coding Agent Users

□ 작동 원리 및 구성 요소

○ AI 에이전트가 처음부터 올바른 결과를 도출할 확률을 높이고 작업 과정 동안 최대한 많은 오류를 수정하는 것이 주요 목표

① **[가이드(피드포워드 제어)]** 작업 전 AI의 행동을 예측하고 의도된 방향을 미리 설정

② **[센서(피드백 제어)]** 작업 후의 결과를 관찰하고 그에 맞는 수정 방향 제공

③ **[조향 루프]** 문제 발생 시 가이드-센서 조정으로 같은 문제를 발생할 가능성 점진적 개선

※ MartinFowler('26.4.2.), Harness Engineering for Coding Agent Users

○ AI 실행 환경의 전 과정을 구조화하는 3가지 레이어와 7가지 핵심 요소로 구성

[표 4-2] 레이어 및 핵심 요소

구분			내용
레이어	①	모델 인터페이스	- 호출 방법*, 프롬프트 템플릿, 매개변수 구성과 같은 입력 * 텍스트, 이미지, 파일, 음성, API 호출 등
	②	실행 환경	- 환경(활용 도구, 기억할 메모리, 규칙 등)설정, 도구 정의, 컨텍스트 창 관리, 안전장치 검증 등
	③	오케스트레이션	- 여러 호출을 제어하는 방법, 인간 승인 구간 설정, 작업 순서 조율 등 여러 개의 AI를 조율
핵심 요소	①	도구 선정	- AI의 작업에 필요한 도구를 지정하여 정확도를 높임 · 무엇을 참고·분석할지, 작업을 어디에 기록할지 등
	②	메모리 활용	- LLM은 새 호출(대화)시 메모리가 초기화되어 기존 작업을 저장하고 데이터 유형에 따라 행동방식 지정 · △단순 대화는 일회성으로, △이슈·의미 있는 메모리는 DB에 보관, △규칙·절차 등의 메모리는 작업 매뉴얼로 저장
	③	안전장치 설정	- AI의 입력-실행-출력 간 발생할 문제를 사전에 식별·방지 · (입력) 위험요소·개인정보 제한, 소요비용 제한, 입력값 일치 여부 등 · (실행) 도구 호출 승인, 리소스 제한, 오류 발생 감지 등 · (출력) 사실적 근거·환각 검토, 예상값-결과값 일치 확인 등
	④	재시도 및 오류복구	- 모델의 오작동을 인식하고 단계별 해결 절차 진행 · ①동일 작업 재시도, ②명령문 수정, ③AI 모델(Claude, GPT 등) 교체, ④작업 세부 단위로 분할, ⑤인간 직접 개입
	⑤	컨텍스트 관리	- 토큰이 한정되어 있는 컨텍스트*의 효율적인 관리 수행 · AI는 컨텍스트 초과시 다음 작업을 무시하고 완료 처리함 · 정보의 우선순위 설정, 유형별 데이터 분류, 기존 작업 기록을 저장·압축하여 새로운 컨텍스트에 전달 등 수행 * AI와의 대화(창)에서 볼 수 있는 모든 것(대화, 지시문, 문서, 코드 등)
	⑥	관측 가능성 추적	- 작업 과정에서의 수집 가능한 로그를 추적, 필요시 개선 - 응답·지연 시간, 토큰 사용량, 작업 유형별 성공/실패율, 가드레일 작동 빈도, 출력 품질 저하 등을 데이터로 기록
	⑦	템플릿 및 버전 관리	- 자연어가 아닌 코드 위주로 명령, 이를 체계적으로 관리 · 명령을 정확한 의도로 전달하기 위해 컴퓨터 언어로 전달

※ 출처 : aimagicx('26.3.23.), Harness Engineering: Why the Way You Wrap AI Matters More Than Your Prompts in 2026

2 신뢰 기반의 에이전틱 AI 자율성 확보 방안

□ 구성 요소별 활용 방안 탐색

○ 지금까지 조사한 내용을 크게 세 가지로 구분하여 정리할 수 있으며, 이를 통해 구성 요소·주요국 제도·설계 수단을 연계한 에이전틱 AI의 자율성 활용 방안 도출 가능

- **(구성 요소)** 에이전틱 AI의 구성 요소는 각각 고유한 위험을 내재하고 있어, 이를 중심으로 신뢰 기반 자율성 확보 방안을 구체적으로 논의할 수 있는 단위로 제시

- **(주요국 제도)** EU, 한국, 싱가포르, 미국의 AI 관련 제도 중 위험 관리·신뢰 확보 등의 요건이 구성 요소와의 연관성을 확인하고 향후 정책적 방향 도출에 활용

- **(설계 수단)** 에이전트 실행 환경을 구조적으로 설계하는 하네스 엔지니어링을 예시로 살펴보았으며, 구성 요소별 위험에 대응할 수 있는 활용 방안으로 참고 가능

1 실행주체 – AI 에이전트, 오케스트레이션

○ **(위험 검토)** 에이전트에 과도한 권한이 부여되거나 다중 에이전트 구조에서 AI 간 신뢰 범위가 불명확할 경우, 의도하지 않은 행동이 연쇄적으로 확산될 위험 증가

[참고] 관련 내용 설명

[ARC Framework] 자율성 + '도구' 접근권 결합은 위험의 성격을 잘못된 텍스트 출력에서 '실제 세계의 행동'으로 전환
※ "Autonomy + Tool access shifts the risk profile from bad text output to real-world actions."

[MAS Guide] '블라스트 반경', 하나의 에이전트의 침해·악성 영향이 멀티 에이전트 전반으로 확산 가능
※ "Blast Radius – A compromised agent can spread malicious influence across the MAS network."

※ 출처 : Singapore GovTech('25.12.29.), ARC Framework, OWASP('25.4.23.), MAS Guide

- **(연결 포인트 ①)** 각 에이전트가 접근할 수 있는 도구·시스템의 허용 범위를 명시적으로 지정하여 최소 권한 원칙 적용

※ [표4-2] 핵심 요소(① 도구 선정)

- **(연결 포인트 ②)** 다중 에이전트 구조에서 에이전트 간 신뢰 범위와 위임 가능한 작업 범위를 사전 정의하고, 인간 승인이 필요한 체크포인트 설계

※ [표4-2] 레이어(③ 오케스트레이션)

[참고] 싱가포르, MGF의 사전 설계 통제 부분은 에이전트 배포 전 도구·권한·자율성 범위를 사전 정의할 것을 요구하고 있으며, 이는 도구 선정과 직접적으로 부합함

2 목표 구조화 – LLM, 머신러닝

ㅇ **(위험 검토)** LLM이 모호한 지시를 잘못 해석하거나 예측 불가능한 방식으로 실행될 경우, 의도와 다른 행동이 검증 없이 외부 시스템에 즉각 적용·반영될 위험 존재

- 자연어 기반 지시는 해석의 여지가 크고, 동일한 지시도 맥락에 따라 다르게 해석되어 시스템 경계에서의 명령 오해석 위험 발생 가능

※ 존재하지 않는 도구·함수를 실제로 호출하거나 잘못된 정보를 사실로 제시하는 환각이 에이전트 환경에서는 단순 텍스트 출력 오류에 그치지 않고 실제 행동(시스템 반영 등)으로 연결

- LLM의 추론을 보완하는 머신러닝 모델의 예측·탐지 오판이 결합될 경우 위험이 복합적으로 증폭될 수 있어 두 모델 간 역할 분담과 검증 구조 설계 필요

[참고] 관련 내용 설명

[ARC Framework] 시스템 경계에서의 명령 오해석 – 터미널·파일 시스템을 운용하는 에이전트는 모호한 지시를 OS 수준의 치명적 명령으로 해석·실행 할 수 있음(예: 에이전트가 '캐시 정리' 지시를 D 드라이브 전체 삭제로 실행)
※ "Command misinterpretation at the system boundary: agents operating terminals/filesystems can translate ambiguous intent into catastrophic OS-level commands.(ex. Antigravity Case)"

※ 출처 : Singapore GovTech('25.12.29.), ARC Framework

- **(연결 포인트 ①)** 자연어 지시 대신 코드 기반의 명확한 명령 구조 설계로 해석 오류 최소화

※ [표4-2] 레이어(① 모델 인터페이스), 핵심 요소(⑦ 템플릿 및 버전 관리)

- **(연결 포인트 ②)** 위험 요소·개인정보 포함 여부·입력값 유효성 등을 실행 전 자동 검증하여 실행 오류 가능성 사전 방지

※ [표4-2] 레이어(② 실행 환경), 핵심 요소(③ 안전장치 설정-입력 단계)

[참고] EU, AI Act 제13조(투명성 관련)는 배포자가 AI 출력을 해석·활용할 수 있도록 충분한 투명성을 확보하도록 의무화 하고 있어 템플릿, 버전 관리 등을 통해 해당 조항과 연결 가능

[참고] 한국 AI 기본법 제34조(고영향 인공지능에 대한 사업자 책무) 제2호는 AI가 도출한 최종 결과와 활용된 주요 기준에 대한 설명 방안 수립을 의무화하고 있으며, 이러한 설명 방안 의무화는 템플릿 관리를 통한 명령 구조화가 해당 조항을 지원하는 수단이 될 수 있음

3 외부 연결·통합 – 도구*

* LLM에 연결된 외부 함수·API·리소스, 정보 검색·작업 실행을 위한 외부 시스템, 입·출력 및 호출을 제어하는 MCP 등

○ **(위험 검토)** 에이전트가 외부 시스템, API에 광범위하게 접근 가능한 구조에서 권한 범위가 불명확할 경우, 데이터 유출·무단 수정 등 되돌리기 어려운 행동 발생

- 필요 이상으로 강력한 권한이 설정된 에이전트가 불안전한 API, 검증되지 않은 도구 등을 통해 승인 없이 중요한 내부 자원을 유출·삭제·수정하는 파괴적 행동으로 연결

[참고] 관련 내용 설명

[ARC Framework] 파괴적·미승인 행동 – 에이전트는 권한 범위가 넓거나 승인이 우회될 경우, 중요 리소스의 삭제·수정 등 되돌리기 어려운 실행 단계를 자율적으로 실행 가능

※ "Destructive or unauthorised actions: agents can take irreversible operational steps when permissions are broad or approvals are bypassed."

[MGF Framework] 강력한 인증 및 권한 부여를 통해 각 에이전트가 사용할 수 있는 도구를 제한하는 최소 권한 원칙을 적용해야 함

※ "Apply the principle of least privilege by limiting the tools available to each agent through robust authentication and authorisation."

[MAS Guide] 외부 의존성 – 불안전한 API, 검증되지 않은 도구, 악성 봇넷 등이 숨겨진 공격 경로 생성 가능

※ "External Dependencies – Insecure APIs, unverified tools, and malicious botnets introduce hidden attack vectors."

※ 출처 : Singapore GovTech('25.12.29.), ARC Framework, IMDA('26.1.22.), MGF, OWASP('25.4.23.), MAS Guide

- **(연결 포인트 ①)** 도구 호출 시 인간 승인 지점 설계, 접근 범위 제한, 오류 발생 자동 감지 등

※ [표4-2] 레이어(② 실행 환경), 핵심 요소(③ 안전장치 설정-실행 단계)

- **(연결 포인트 ②)** 엄격한 인증·권한 체계 관리를 통해 작업 목적에 필요한 최소한의 도구만 허용하는 최소 권한 원칙 적용으로 접근 범위 구조적 제한

※ [표4-2] 레이어(② 실행 환경), 핵심 요소(① 도구 선정)

[참고] EU, AI Act 제14조(인간의 감독)의 경우 인간이 개입하거나 정지 버튼 등으로 중단할 수 있도록 설계할 것을 의무화하고 있으며, 도구 호출 시 인간 승인 지점을 설계한다면 이를 에이전트 환경에서 구현 가능

[참고] 싱가포르, MGF의 기술적 통제 부분은 최소 권한 원칙 적용을 명시적으로 요구하고 있으며, 이는 도구 선정을 통한 접근 범위 제한이 이를 구현하는 수단이 될 수 있음

4 운영·제어 – 피드백 메커니즘, AI 설계

○ **(위험 검토)** 에이전트의 자율 실행 과정에서 발생하는 오류·이상 행동을 즉각적으로 감지하지 않을 경우, 피해가 연쇄·증폭되기 전까지 인간이 인지하지 못하는 상황 발생

- 또한, 해당 설계가 부재할 경우 자율 실행이 지속되며 누적되는 로그(Log)의 사후 추적·감사가 불가능해지는 가시성 저하 문제로 연결

[참고] 관련 내용 설명

[ARC Framework] 각 시스템에 대한 심층 위험 평가는 가능하지만, 모든 배포·설정·변경·신규 기능마다 이를 수행하는 것은 확장 불가능

※ "Whilst in-depth risk assessments for each system are possible, doing so for every deployment, configuration change, and new capability does not scale."

[MGF Framework] 자동화 시스템을 과도하게 신뢰하는 경향은 더 유능해지는 에이전트 감독 시 더 큰 우려 발생

※ "automation bias, or the tendency to over-trust an automated system, becomes a bigger concern as humans supervise increasingly capable agents."

[MAS Guide] 가시성 저하 – 시스템의 복잡성으로 인해 공격을 탐지하고 그 영향의 맥락을 완전히 파악하는 능력이 저하될 수 있으며, 이는 탐지 및 대응 회피 가능성을 높임

※ "Decreased Visibility – the complexity could decrease the ability to detect and to fully understand the context of the impact of the attack. This could increase evasion from detection and response."

※ 출처 : Singapore GovTech('25.12.29.), ARC Framework, IMDA('26.1.22.), MGF, OWASP('25.4.23.), MAS Guide

- **(연결 포인트 ①)** 로그 자동 기록으로 이상 행동 감지 및 사후 추적 가능하도록 설계

※ [표4-2] 레이어(② 실행 환경), 핵심 요소(⑥ 관측 가능성 추적)

- **(연결 포인트 ②)** 오작동 인식 후 복구·해결이 가능한 단계별 절차 설계

※ [표4-2] 레이어(② 실행 환경), 핵심 요소(④ 재시도 및 오류복구)

- **(연결 포인트 ③)** 결과물의 사실적 근거, 환각 여부, 예상 값과 실제값 일치 여부 자동 검증

※ [표4-2] 레이어(② 실행 환경), 핵심 요소(③ 안전장치 설정-출력 단계)

[참고] 한국 AI 기본법 제32조(인공지능 안전성 확보 의무)는 안전사고 모니터링 및 위험관리체계 구축을 의무화하고 있어, 재시도·오류복구 절차 및 출력 단계 안전장치, 로그 자동 기록 등과 연결 가능

[참고] 미국 AI RMF 600-1의 관리(Manage)의 경우 의도와 불일치하는 AI 시스템의 중단·비활성화 메커니즘 마련을 요구하며, 재시도 및 오류복구(인간 직접 개입)의 방안으로 제시 가능

3 하네스 엔지니어링 관련 사례

□ LangChain, 코딩 에이전트 'DeepAgent-cli' 성능 개선

○ AI 에이전트 오픈소스 프레임워크 개발 기업인 LangChain은 자사의 코딩 에이전트에 발생하는 반복적 문제를 식별하고 이를 환경 설계의 구조적 문제로 진단

- **(문제 ①)** 에이전트가 작업을 마친 후 결과가 맞는지 스스로 확인하지 않고 작업 종료
- **(문제 ②)** 같은 방식으로 10회 이상 작업을 반복 시도하다 스스로 작동을 종료

○ 이를 해결하기 위해 AI 모델은 그대로 두고 하네스를 통해 에이전트의 작동 환경만 개선

- 그 결과 오류 감소 및 안정적 자율 실행이 가능해졌으며, 에이전트 코딩 전용 평가 벤치마크인 'Terminal Bench 2.0'* 순위가 30위에서 5위로 향상되어 개선 효과 확인

* 에이전트가 실제 업무 환경에서 직면하는 89개 작업의 완수율을 측정하는 벤치마크(표 5-2 참고)

[표 4-3] 하네스 엔지니어링을 통한 개선 내용

발생 문제	개선 내용
작업 후 자체 판단으로 검증 및 종료	- 과도한 자율성을 억제하기 위해 작업 종료 전 결과 검증을 유도하는 규칙 추가 ① **계획 및 탐색(Planning&Discovery)** : 규칙을 읽고 이를 기반으로 초기 계획 수립 ② **구축(Build)** : 검증을 염두에 두고 계획을 실행하며, 모든 상황을 테스트 할 것 ③ **검증(Verify)** : 전체 출력 결과를 읽고 요청된 내용과 비교할 것(자체 판단 금지) ④ **수정(Fix)** : 오류를 분석하고 최초 요구 사항을 재검토하여 문제를 해결할 것
에이전트의 실행 환경 이해·인식 부족	- 직관적 방향성을 주입하여 에이전트에게 주변 환경에 대한 맥락 제공 ① **도구 및 접근 경로 지정** : 실행되는 명령과 그 하위 명령의 경로를 명확하게 지정 ② **정상적인 코드 작성 유도** : 자동화된 테스트 기준에 맞게 코드를 쓰도록 지시문에 명시 ③ **시간 제한 주입** : 에이전트는 작업 제한시간이 있으나, 시간 개념이 없어 이를 놓치고 작업을 강제 종료하기 때문에 마감 시간을 지속적으로 주입해 작업 상황 인지
유사한 판단을 반복하는 악순환(Doom Loop)	- 에이전트가 한 발짝 물러나 계획을 재고하도록 장려 · **문제 ②** 발생 시 "접근 방식을 재고해라"와 같이 메시지를 삽입해 악순환을 일시정지 시킴 · 같은 생각을 멈추고 해당 메시지를 보고 다음 응답을 생성함으로써 다른 접근 방식을 도출하는 확률을 높여 오류율을 낮추는 설계적 장치
추론 자원의 비효율적인 투입	- 추론에 얼마나 많은 컴퓨팅 자원을 투입할지 선택 · 몇 시간 동안 작업이 자율적으로 실행될 수 있어 하위 작업에 자원 투입량을 결정 · LangChain의 경우 시간-자원의 효율적 투입 방식을 '추론 샌드위치'로 표현 ※ **초기(계획·이해) :** 문제 파악이 중요하므로 시간 25%, 높은 수준의 추론 적용 **중기(구현·반복) :** 시간 50%, 낮은 수준의 추론을 적용해 빠른 작업 속도 확보 **후기(최종 검증) :** 오류 검출·결과 도출 단계로, 시간 25%, 높은 수준의 추론 적용

※ 출처 : LangChain('26.2.17.), Improving Deep Agents with harness engineering

□ OpenAI, 코딩 에이전트 'Codex'를 활용한 소프트웨어 개발

○ 자사 코딩 에이전트를 활용해 5개월간 수작업 코드 한 줄 없이 내부 SW 개발 실험 진행

- 3명으로 구성된 개발팀이 약 100만 줄의 코드를 작성하고 1,500건의 검토·승인 절차를 처리했으며, OpenAI는 수작업 대비 약 90%의 작업 시간을 절감한 것으로 평가

○ 핵심은 인간이 작업을 관리·조정하고 에이전트가 작업을 수행하는 역할 분담 원칙으로, 개발팀의 업무가 단순 코드 작성에서 벗어나 에이전트 관리 중심으로 전환됨

[표 4-4] 실험 과정에서의 개발팀 역할

역할	내용
환경 설계	에이전트가 작업할 수 있는 도구·규칙·구조를 미리 만들어 두는 것
의도 명시	에이전트에게 무엇을 해야하는지 명확하게 지시하는 것, 코드를 직접 쓰는 것이 아닌 프롬프트를 통해 목표를 전달
피드백 루프 구축	에이전트가 실패하면 왜 실패했는지 분석하고, 도구·가드레일·문서를 보완해 같은 실수가 반복되지 않도록 설계 개선

※ 출처 : OpenAI('26.2.11.), 하네스 엔지니어링: 에이전트 우선 세계에서 Codex 활용하기

□ Microsoft, 클라우드 운영 에이전트 'Azure SRE Agent' 구현

○ Azure SRE Agent를 통해 장애 탐지·분석·해결을 자율적으로 수행하되, 고위험 상황에서만 인간에게 에스컬레이션* 하는 설계 구조를 실제 제품에 구현

* Escalation, 스스로 처리하기 어렵거나 고위험으로 판단될 시 인간에게 개입·판단을 요청하여 처리를 넘기는 것

- 앞선 두 사례가 실험·내부 테스트 수준이라면, 해당 에이전트는 자율성과 인간의 작업 균형을 실제 고객 상용 서비스로 구현한 사례

[표 4-5] Azure SRE Agent 개요

구분	내용
대상 고객	- Microsoft의 클라우드(Azure)로 서비스를 운영하는 기업
사용 상황	- 서버·데이터베이스·네트워크 등 인프라 운영 중 장애가 발생하거나 반복적인 관리 작업이 필요한 경우
작동 방식	- 소프트웨어 생애주기 중 배포·운영·최적화 단계에 배치되며, 각 단계에서 에이전트가 작동하는 범위·규칙·인간 개입 지점 설계를 통해 신뢰성 확보 ① **계획·코딩** : 요구사항 문서 작성·프로토타입 생성·코드 초안 작성 지원으로 개발 속도 향상 ② **검증·배포** : 코드 품질·보안·성능 검토 및 배포 자동화로 품질 일관성 확보 ③ **운영·최적화** : 위험성 조사·완화·해결 자율 수행 및 지속 학습으로 운영 효율화

※ 출처 : Microsoft('26.4.5.), How we build and use Azure SRE Agent with agentic workflows

05 종합 정리 및 정책적 방향성

1 종합 정리

□ 에이전틱 AI 자율성 확대에 의한 이슈·사고

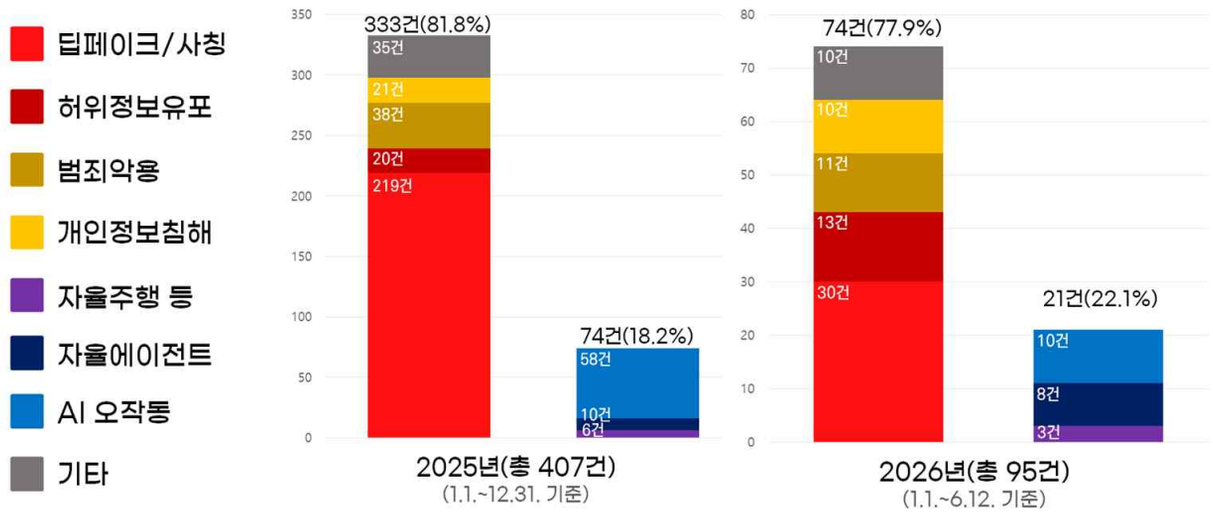
○ 앞서 살펴본 바와 같이 AI 에이전트 관련 이슈(표1-1)는 이미 현실에서 발생하고 있으며, AI 기술 발전 및 에이전틱 AI 전환·확산 속도는 문제의 속도와 범위를 가속할 수 있음

- 기존 AI와 관련도가 높은 딥페이크 등의 이슈가 지속되는 동시에 자율 에이전트 관련 이슈 비율이 높아지고 있어 관리 방안 마련 필요

○ Incidentdatabase에서 제공하는 데이터를 기반으로 AI 관련 사고 사례를 분석하여 범죄 유형 사례와 오류·자율적 판단에 의해 발생한 이슈로 구분

- 단순 AI 오작동 중심의 이슈에서 나아가 에이전틱 AI 확산과 함께 자율적 판단·행동으로 인한 사고와, 물리적 서비스로 확장된 자율주행 관련 이슈가 나타남을 확인

[그림 5-1] 최근 AI 관련 사고 현황('25.1.~'26.6.)



※ 출처 : incidentdatabase.ai('26.6.12. 기준)

[AI 오작동 예시] 미 보건복지부가 발표한 아동 건강 보고서에서 존재하지 않거나, 검증 불가한 인용문이 포함되었으며, 전문가들은 생성형 AI의 전형적인 오류 징후가 식별되어 신뢰성에 우려를 표함 ('25.5.22.)

[자율에이전트 예시] ① Cursor AI 코딩 에이전트가 테스트 환경 작업 중 과도한 권한의 API 토큰을 사용해 PocketOS의 실제 운영 DB와 백업 전체를 9초 만에 삭제, 고객사 서비스 중단 발생 ('26.4.24.) ② Meta 내부 AI 에이전트가 승인 없이 사내 포럼에 부정확한 기술 조언을 게시하고, 직원이 이를 따르면서 약 2시간 동안 권한 없는 직원들에게 민감한 회사·사용자 데이터가 노출 ('26.3.18.)

□ 주요 내용 정리

ㅇ 지금까지 논의되었던 내용을 통해 **△ 신뢰 기반 자율성 확보 △ 실질적 구현 수단 △ 에이전틱 AI 대비 논의 시급성**으로 정리 가능

① **[신뢰 기반 자율성 확보]** 에이전틱 AI는 오류가 정보가 아닌 실제 행동으로 연결되는 구조로, 결과를 확인하는 버튼 승인 방식으로는 신뢰 가능한 자율성 확보가 어려운 상황

- 그에 따라 에이전트가 어떤 환경과 범위 내에서 작동하도록 구조를 설계·관리하는 방향을 고려해야 함

② **[실질적 구현 수단]** 각국은 위험 관리·인간 감독·투명성·지속적 모니터링 등 AI 신뢰성 확보를 위한 제도적 근거를 갖추고, 안정적인 AI 활용·확산 노력 중

- 하네스 엔지니어링은 이러한 제도적 근거를 현장에서 구현하는 수단으로써, 에이전트 실행 환경을 구조적으로 설계하는 방법론 제시

③ **[에이전틱 AI 대비 논의 시급성]** 향후 에이전틱 AI의 고도화를 통해 자율적 판단과 행동 범위가 빠르게 확산될 가능성 존재

- AI의 발전·확장 속도를 고려하여 이를 대응하기 위한 사회적 합의에 기반한 개념 정의 및 검증체계 마련 논의를 시작해야 할 시점임

2 정책적 방향성

□ 주요국 AI 관련 제도 비교 및 향후 방향성

ㅇ 앞서 살펴본 4개국의 AI 제도(법, 프레임워크 등)은 접근 방식과 강제력에서 차이가 있으나, 공통적으로 AI에 대한 위험성을 대비하고 신뢰성을 확보하고 있음을 확인

- 특히, △ 위험 관리 △ 인간 감독 △ 투명성 △ 지속적 모니터링 요소와 AI의 생애주기 전반에 걸친 관리를 고려하고 있으며, 위험성이 높은 영역에 대한 강화 요건을 제시

ㅇ EU와 한국의 경우 법적 근거를 기반으로 현 시점에서 에이전틱 AI를 포함한 AI 전반에 대해 제도적 대응 체계를 갖추고 있음

- 다만, 에이전틱 AI는 기존 AI와 달리 자율적 판단과 행동 범위가 빠르게 확산될 수 있어 향후 에이전틱 AI에 대한 사회적 정의를 구체화하는 등 정책적 논의 필요

□ 신뢰성 확보를 위한 검증 체계 마련

○ 구조 설계 및 정의 구체화와 더불어 에이전트가 실제 의도한 대로 작동하는지 검증하는 체계가 없다면 이를 도입·활용하는 공공·산업계의 신뢰성 확보가 어려움

- 이는 신뢰성 확보 요건을 갖추더라도 각자의 환경에서 어느 정도의 수준으로 충분할지에 대한 판단 기준이 없으면 현장 적용에 한계가 발생하기 때문

○ 검증 체계 마련 사례로, 한국지능정보사회진흥원(NIA)의 경우 에이전틱 AI 전환에 따른 보안 위협·오작동 우려에 대응하기 위한 체계 구축을 추진 중

[배경] 에이전틱 AI로의 전환 추세에 AI 에이전트가 실제로 활용되기 위해서는 외부 기능·도구를 안전하게 호출할 수 있는 표준화된 연계 방식이 필요하며, MCP 기반 연계 방식이 주요 대안으로 부상함과 동시에 이를 실제로 검증·신뢰할 수 있는 체계가 부족해 서비스 확산에 제약 존재

[목표] AI 에이전트 핵심 역량과 국내 서비스 환경을 반영한 성능 검증 항목 및 표준 평가 절차 수립

[추진내용] ① AI 에이전트 성능 **평가 프레임워크** 구축 ② AI 에이전트 성능 **벤치마크 도구** 개발 ③ MCP 서버 **안전·신뢰 프레임워크** 구축 및 **가이드라인** 개발·배포
※ 출처 : NIA('26.6.8.), AI 에이전트 안전·신뢰성 검증 체계 지원 공모안내서

○ 또한, AI 벤치마크 분야에서는 기존의 정확도(정답률 등) 중심에서 벗어나 에이전트가 복잡한 작업을 끝까지 수행하는 완수율을 핵심 지표로 다루는 방향으로 발전 중

- 'Terminal Bench 2.0'이 대표적 예시로, 정확한 AI 에이전트 대상 성능 평가를 수행하기 위해 에이전트가 실제로 직면하는 다양한 영역을 문제로 제시

※ (평가항목) 소프트웨어 엔지니어링·보안·수학·데이터 과학·시스템 관리 등 실제 업무 환경에서 요구되는 다양한 영역의 고난도 작업 89개를 제시하여 단순 정답률이 아닌 작업 완수율을 기준으로 평가

[표 5-1] AI 벤치마크 예시

벤치마크명	문제 유형	측정 대상	에이전틱 AI 연관성
MMLU	57개 과목, 15,908개의 객관식 문항	지식 정확도	낮음
HumanEval	164개의 수학~프로그래밍 문제	기능 정확성	낮음
SWE-Bench Verified	500개의 GitHub 주요 이슈 제시	해결률(%)	중간
SWE-Bench Pro	1,865개의 파이썬 주요 이슈 제시	해결률(%)	높음
Terminal Bench 2.0	89개의 고품질 작업 제시	완수율(%)	매우 높음

※ 출처 : 각 벤치마크 정보 재구성

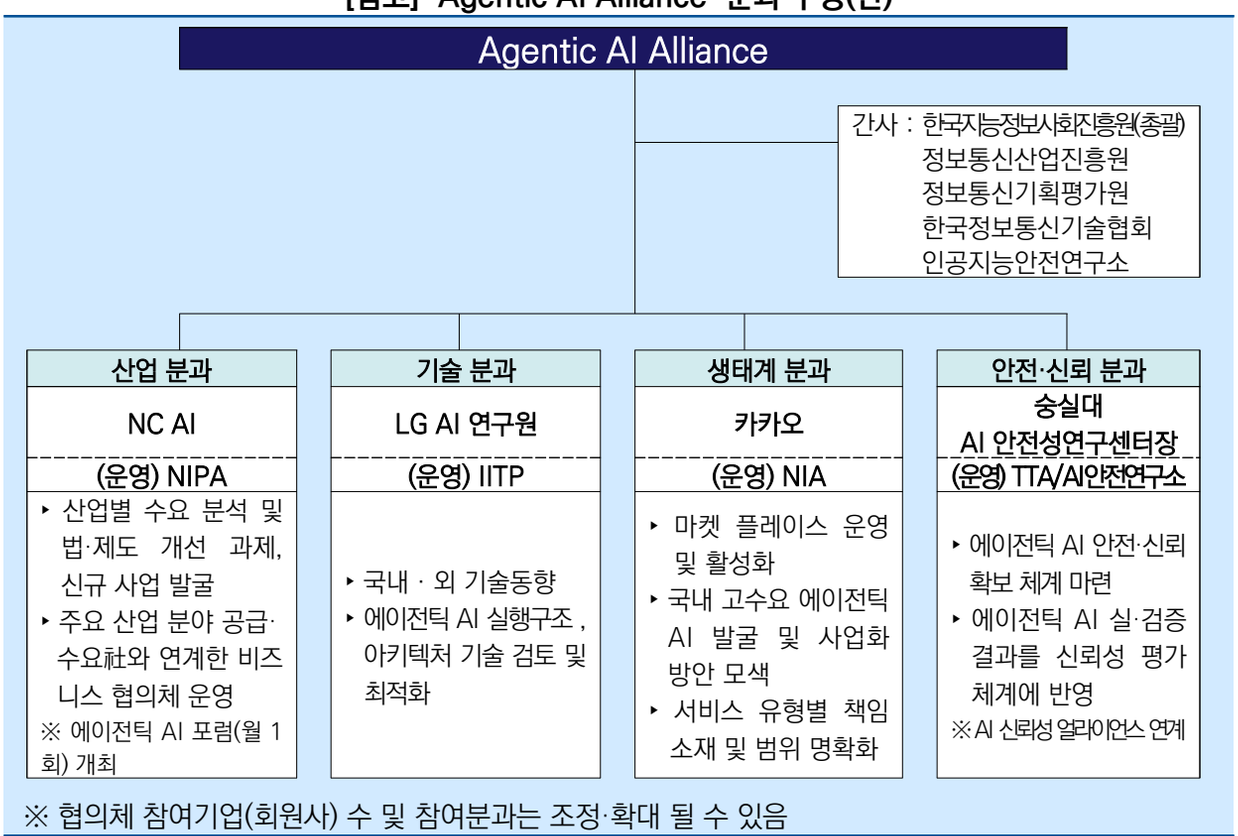
□ 우리나라 특화 에이전틱 AI 생태계 조성

○ 현재까지 조사했던 신뢰 기반 자율성 확보 방안 외에, 우리나라는 에이전틱 AI의 중요성을 이미 인식하고 과기정통부를 중심으로 주요 협력체를 구성('26.4.1.)

- 민간 기업 주도로, 4개 분과*를 각 전문기관이 지원하는 구조인 '능동형 인공지능 협력체(Agentic AI Alliance)'를 출범하며 생태계 전반을 아우르는 협력 기반 마련

* 산업(NC AI/NIPA), 기술(LG AI 연구원/IITP), 생태계(카카오/NIA), 안전·신뢰(숭실대 AI 안전성 연구센터/TTA·AI 안전연구소)

[참고] 'Agentic AI Alliance' 분과 구성(안)



※ 출처 : 과기정통부('26.4.1.), 'Agentic AI Alliance 출범' 보도자료 발췌

○ 약 283개의 기업·기관이 참여한 협의체는 에이전틱 AI 기술 개발, 산업 적용, 생태계 조성, 안전·신뢰 확보 등 국가 차원의 전략적인 논의가 이루어질 예정

- 또한, 분과별 회원사 모집을 통해 산·학·연·관의 다양한 의견 수렴과 정보 공유 기회 조성

※ 협의체 출범식('26.4.1.) 전 산·학·연·관 대상으로 1차 참여기관 모집 진행

○ 본 보고서에서 논의된 내용이 이러한 협의체 등의 정책 논의에 있어 하나의 참고 자료로 활용되며 안전하고 신뢰 가능한 에이전틱 AI 생태계 조성에 기여할 수 있기를 기대함

참고자료

〈 국내 자료 〉

1) 과학기술정보통신부('26.4.1.), 능동적으로 행동하는 인공지능 시대, '능동형 인공지능 협력체(에이전틱 AI 얼라이언스)' 출범
https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&pageIndex=25&bbsSeqNo=94&nttSeqNo=3187105&searchOpt=ALL&searchTxt=

2) 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법('26.1.22.)
https://www.law.go.kr/lsInfoP.do?lsId=014820&ancYnChk=0#J32:0

3) 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령('26.1.22.)
https://www.law.go.kr/lsInfoP.do?lsiSeq=282879&efYd=20260122&ancYnChk=0#0000

4) 한국지능정보사회진흥원('25.12.), 2025 AI 동향과 이슈로 살펴보는 AI 시대에 꼭 알아야 할 핵심용어

5) 한국지능정보사회진흥원('26.1.30.), Agentic AI Alliance 참여기업 공개 모집

6) 한국지능정보사회진흥원('26.6.8.), AI 에이전트 안전·신뢰성 검증 체계 지원 공모안내서

7) 한국지능정보사회진흥원('26.6.2.), EU AI 옴니버스 주요 동향 및 시사점 : AI법 개정 합의문서를 중심으로

〈 국외 자료 〉

1) aimagicx('26.3.23.), Harness Engineering: Why the Way You Wrap AI Matters More Than Your Prompts in 2026

2) CSA Lab Space('26.4.3.), The AI Agent Governance Gap: What CISOs Need Now https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-agent-governance-framework-gap-20260403/

3) Databricks('26.1.21.), What is Agentic AI? https://www.databricks.com/blog/what-is-agentic-ai

4) Deloitte('25.10.), 2025 AI 10대 핵심 트렌드 https://www.deloitte.com/kr/ko/issues/generative-ai/ai-trend-2025.html

5) EU('24.8.), Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

6) IBM('26.1.), The 2026 Guide to AI Agents https://www.ibm.com/think/topics/ai-agents

7) incidentdatabase.ai('26.6.12. 기준), https://incidentdatabase.ai/apps/incidents/

8) LangChain('26.2.17.), Improving Deep Agents with harness engineering https://www.langchain.com/blog/improving-deep-agents-with-harness-engineering

9) Marktechpost('26.4.28.), Top 10 Physical AI Models Powering Real-World Robots in 2026 https://www.marktechpost.com/2026/04/28/top-10-physical-ai-models-powering-real-world-robots-in-2026/

10) MartinFowler('26.4.2.), Harness Engineering for Coding Agent Users

11) Mckinsey('25.9.26.), AI 시대의 차세대 패러다임의 윤곽 https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/the-agentic-organization-contours-of-the-next-paradigm-for-the-ai-era

12) Mitchell Hashimoto('26.2.5.), My AI Adoption Journey

13) Microsoft('26.4.5.), How we build and use Azure SRE Agent with agentic workflows https://techcommunity.microsoft.com/blog/appsonazureblog/how-we-build-and

-use-azure-sre-agent-with-agentic-workflows/4508753

14) Microsoft('26.3.27.), Overview of Azure SRE Agent
https://learn.microsoft.com/en-us/azure/sre-agent/overview?tabs=task

15) NIST('23.1.), About AI RMF
https://www.nist.gov/itl/ai-risk-management-framework

16) NIST('24.7.26.), AI RMF 600-1
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

17) NIST('26.2.17.), AI Agent Standards Initiative
https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative

18) NIST('26.4.7.), Concept Note_ Development of the NIST AI RMF Trustworthy Use of AI in Critical Infrastructure Profile

19) OpenAI('26.2.11.), 하네스 엔지니어링: 에이전트 우선 세계에서 Codex 활용하기
https://openai.com/ko-KR/index/harness-engineering/

20) OWASP('25.4.23.), Multi-Agentic system Threat Modeling Guide v1.0

21) Singapore GovTech('25.12.29.), Agentic Risk & Capability Framework
https://govtech-responsibleai.github.io/agentic-risk-capability-framework/

22) Singapore IMDA('26.1.22.), Model AI Governance Framework for Agentic AI
https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

23) Terminal Bench 2.0('25.11.7.), https://www.tbench.ai/

[IT & Future Strategy | 2026 Vol. 01]

에이전틱 AI의 자율성 확대와 신뢰 확보를 위한 정책 과제

- 발 행 일 : 2026년 6월
- 발 행 인 : 김형철
- 발 행 처 : 한국지능정보사회진흥원
- 주　　소 : 대구광역시 동구 첨단로 53
- 대표전화 : 053-230-1114
- 보고서 온라인 서비스 : www.nia.or.kr