

SW 중심사회

ISSUE

AI 교육 혁신과 AI 융합인재 양성 전략: 해외 대학·교육기관 사례와 시사점

FOCUS

모든 산업을 알아야 하는 산업: AI 시대 금융의 변화와 정책 제언

신뢰할 수 있는 AI: 글로벌 현황과 향후 정책적 과제

월간 2026 September Vol.147

ISSN 2586-5021 | eISSN 3022-9537

09



Vol. 147

2026 September

09

Contents

ISSUE

AI 교육 혁신과 AI 융합인재 양성 전략:
해외 대학·교육기관 사례와 시사점

03

PHOTO ESSAY

사진산책 - 중간·이호준

30

FOCUS

1. 모든 산업을 알아야 하는 산업:
AI 시대 금융의 변화와 정책 제언

32

2. 신뢰할 수 있는 AI:
글로벌 현황과 향후 정책적 과제

42

발행인 박원재 소장

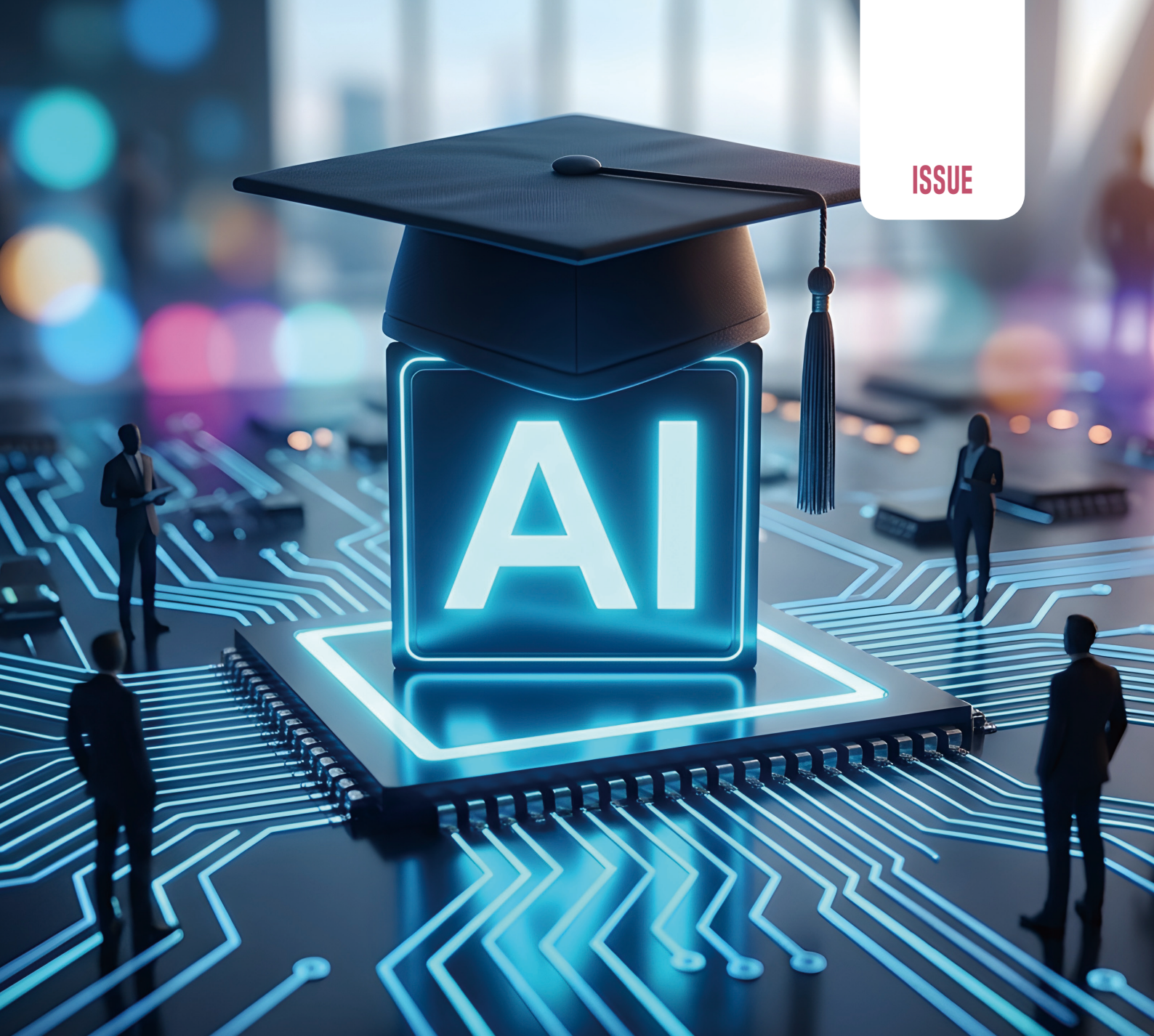
발행처 소프트웨어정책연구소

홈페이지 www.spri.kr

전화 031-739-7300

주소 경기도 성남시 분당구 대왕판교로 712번길 22 글로벌 R&D센터 연구동 A 4층, 소프트웨어정책연구소

디자인·제작 (주)늘품플러스 070-7090-1177



ISSUE

AI 교육 혁신과 AI 융합인재 양성 전략: 해외 대학·교육기관 사례와 시사점

Innovation in AI Education and Strategies for Developing
Interdisciplinary AI Talent: Global Practices and Policy Implications

- 봉강호 소프트웨어정책연구소 SW기반정책·인재연구실 선임연구원 | bk91@spri.kr
- 김항규 소프트웨어정책연구소 SW기반정책·인재연구실 책임연구원 | hkkim@spri.kr
- 윤보성 소프트웨어정책연구소 SW기반정책·인재연구실 책임연구원 | bsyun@spri.kr
- 진희승 소프트웨어정책연구소 SW기반정책·인재연구실 수석연구원 | hschin@spri.kr

EXECUTIVE SUMMARY

AI 인재가 군사력·에너지 자원과 같은 전통적인 국가 전략자산에 준하는 국가적 위상을 갖게 되면서, 주요국은 AI 인재 확보 및 양성, 그리고 교육 혁신에 대한 정책적 대응을 가속화하고 있다. 특히 생성형 AI와 범용 AI 도구가 확산되면서 AI 인재에 대한 수요는 단순한 양적 부족의 문제를 넘어 ‘어떤 역량을 갖춘 인재가 필요한가’라는 질적 변화의 문제로 확대되고 있다. 이에 따라 AI 인재상도 AI 모델·알고리즘을 직접 개발하는 전문 AI 인재를 넘어, 전공·산업 도메인 지식과 AI 역량을 결합한 융합형 인재, 그리고 AI 활용에 수반되는 사회적 책임을 이해하고 실천하는 책임 있는 AI 인재로 다층화되고 있다.

본고에서는 이러한 문제의식에서 출발하여 해외 주요 대학·교육기관의 AI 인재 양성 사례를 분석하고 국내 정책 및 대학 교육과정 혁신에 필요한 시사점을 도출하였다. 분석 대상은 미국(MIT·카네기멜런대), 중국(푸단대·칭화대), 영국(서리대·사우샘프턴대), 캐나다(토론토대), 스위스(ETH 취리히) 등 주요국 8개 대학·기관으로, 각 사례가 지향하는 인재상과 교육내용·방식, 학생에게 제공하는 학습경험 등을 중심으로 검토하였다.

이 사례들로부터 다음과 같은 공통점을 확인하였다. 첫째, AI 교육을 특정 전공의 전문교육이 아니라 전교적·전공횡단적 역량으로 재정의하고 있다. 둘째, AI+X 또는 X+AI 방식의 전공특화·도메인 결합형 교육이 핵심 모델로 부상하여, AI 기술 자체보다 ‘AI를 전공·산업·사회문제와 결합하는 능력’을 중시하고 있다. 셋째, 책임 있는 AI, 윤리, 그리고 AI 산출물을 검증·비판하는 역량이 교육과정의 필수 요소로 편입되고 있다. 넷째, 교육과정 설계뿐 아니라 지도교수 순환제·공동지도 등 거버넌스와 지도체계 자체를 융합형으로 재편하고 있다.

이를 바탕으로 본고는 국내 정책이 ‘전문인력 공급 확대’에서 ‘전공별 AI 역량의 체계적 확산’으로 목표를 확장해야 함을 제언한다. 구체적으로 공동교과·모듈형 교육 인프라 구축, AI+X·X+AI 복수전공·마이크로전공·융합트랙 제도화, 산업·공공문제 기반 프로젝트와 인턴십의 정규 교육과정 결합, 책임 있는 AI와 과정 중심 평가의 원칙 반영, 학제 간 공동지도·지도체계 개편 등 융합형 거버넌스의 단계적 도입 등이 필요하다. 다만 적용에는 단순 모방을 지양하고 대학 유형과 인재 양성 목표에 따른 맞춤형 설계가 요구된다. 결국 국내 AI 인재 양성 정책은 전문 인재의 양적 확대와 함께 전공별 AI 내재화, 경험 기반 학습, 책임 있는 AI 교육, 융합형 거버넌스를 포괄하는 방향으로 고도화되어야 한다.

AI talent has emerged as a core national strategic asset, and as global AI competition shifts toward securing and cultivating talent, major countries are accelerating their policy responses. With the spread of generative AI, the demand for AI talent has moved beyond a simple quantitative shortage toward a qualitative question of what competencies are needed. Accordingly, the notion of AI talent has diversified beyond specialized professionals who develop AI models and algorithms to include interdisciplinary talent—who combine AI competence with domain knowledge—and responsible AI talent.

This report analyzes how leading overseas universities and educational institutions cultivate AI talent and derives implications for domestic policy and curriculum reform. The analysis covers eight institutions across major countries focusing on the talent profiles each pursues, how curricula are designed, and what learning experiences and institutional support are provided to students. Our analysis identifies four common features: (1) the reorganization of governance structures, such as advisor-rotation and co-supervision systems, in an interdisciplinary direction; (2) redefining AI education as a campus-wide, cross-disciplinary competency; (3) the rise of discipline-specific, domain-integrated “AI+X” or “X+AI” education; and (4) the incorporation of responsible AI and the capacity to verify and critique AI outputs as essential curricular elements.

Building on these findings, this report argues that domestic policy should expand its objective from increasing the supply of specialized personnel to systematically diffusing discipline-specific AI competencies. Recommendations include building shared, modular educational infrastructure; institutionalizing AI+X and X+AI multi-majors, micro-majors, and convergence tracks; integrating industry- and public-problem-based projects and internships into the formal curriculum; and embedding responsible AI and process-oriented assessment as core principles. When applying these cases, simple imitation should be avoided in favor of tailored designs reflecting each institution’s type and talent objectives. Ultimately, domestic AI talent policy should be advanced to encompass not only the quantitative expansion of specialized professionals but also discipline-specific embedding of AI, experience-based learning, responsible AI education, and convergent governance.

I

배경 및 목적

○ AI 인재가 국가의 전략자산으로 부상

- AI 인재는 단순히 시장가치가 높은 고숙련 노동력이 아니라 군사력, 에너지 자원, 희토류, 핵심 산업 기반시설과 유사한 성격을 갖는 국가 전략자산으로 부상
 - 전략자산은 단순 가격이 높다는 이유만으로 규정되는 것이 아니라, ① 단기간에 시장을 통해 조달하기 어렵고, ② 공급이 특정 국가·기업·기관에 집중되며, ③ 부재 시 국가안보와 핵심 산업 기능에 직접적 제약을 초래하는 자원을 의미
 - AI 인재는 장기간의 고도 교육, 연구 경험, 대규모 데이터·컴퓨팅 인프라 접근, 글로벌 연구 네트워크를 통해 축적되기 때문에 단기간에 대체·확충하기 어려운 인적자원
 - 특히 최상위 AI 연구자와 엔지니어는 글로벌 빅테크, 선도 대학, 소수 연구기관에 집중되어 있어 일반 노동시장처럼 필요시 즉각 조달하기 어려운 구조적 희소성 내포
 - AI 인재의 부족은 단순한 인력난을 넘어, 첨단 제조·국방·사이버 보안·바이오·금융·공공서비스 등 국가 핵심 부문의 AI 전환 역량을 직접적으로 제약
- 이러한 점에서 AI 인재 확보는 개별 기업의 채용 경쟁을 넘어 국가 차원의 기술주권, 산업경쟁력, 안보 역량과 직결되는 전략 과제로 전환
 - 글로벌 빅테크 기업들이 우수한 AI 인재 영입을 위해 수백만~수천만 달러 수준의 보상 패키지를 제시하는 것은 AI 인재의 높은 시장가치를 보여주는 사례이나¹⁾, 보다 근본적으로는 최상위 인재의 공급이 제한적이라는 구조적 현실을 반영
 - 주요국이 파격적인 비자·이민 우대, 연구비 지원, 세제 혜택, 정주 여건 개선 등을 추진하는 것도 핵심 AI 역량의 해외 의존도를 줄이고 자국 내 전략자산을 축적하려는 대응으로 이해 가능(윤보성·진희승, 2025)
- 글로벌 AI 기술 경쟁의 핵심 축이 인재 확보와 양성에 집중되고, 주요국의 정책 대응이 가속화되는 양상
 - 주요국은 AI 인재 양성을 교육정책의 하위 과제가 아니라, 과학기술·산업·지역·노동·사회정책이 결합된 범부처 전략 의제로 설정

1 문상혁(2025), “경쟁사 직원에 1300억 제안도 … 빅테크 치열한 ‘인재 쟁탈전’”, 중앙일보, 2025.06.29.

[표 1] 주요국 AI 인재 양성 관련 국가 전략 및 정책 동향

국가	전략명/정책명	주요 내용
미국	AI 교육 추진 행정명령 (‘25.04)	- 백악관 과학기술정책실장 주재로 교육부·노동부·NSF 등이 참여하는 범부처 태스크포스 구성 - K-12 AI 교육 증진 및 노동부 도제 훈련의 AI 통합 추진
중국	AI+ 교육 실행계획 (‘26.04)	- 초·중·고등학교부터 박사에 이르기까지 전 교육과정에 AI와 교육을 결합 - AI 활용 확대-인프라 구축-제도 마련 등을 동시 추진
영국	AI 기회 행동계획 (‘25.01)	- AI 전공자 확대를 위한 고등교육기관 지원 - 대규모 AI 장학 프로그램 도입 - AI 인재풀의 다양성 제고 - Turing AI Fellowship 확대
일본	AI 기본계획 (‘25.12)	- 정부 차원 R&D·인프라·데이터·인재 양성·국제규범 종합 추진 2030년까지 AI 엔지니어·개발자 100만 명 양성 목표
한국	대한민국 AI 액션플랜 (‘25.09)	‘AI G3 도약’ 비전 - 초·중·고등학교 연속 AI 필수 교육 체계 구축 - 다층적 AI 인재 양성 - 부처 간 AI 인재 양성 사업 연계·효율화

○ AI 인재 수요의 확대 추세 및 질적 변화

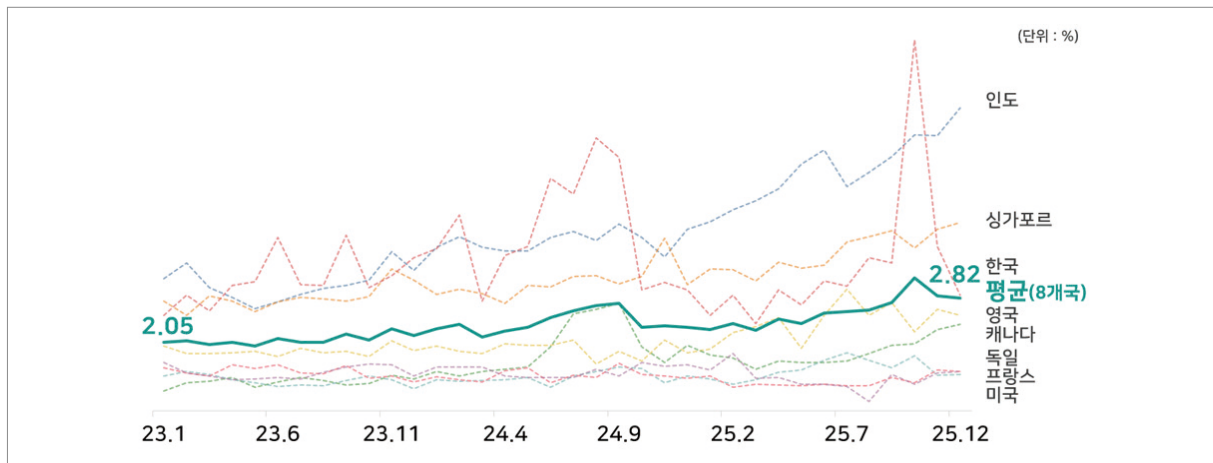
- AI 활용이 산업 전반에 확산되면서, 주요국 노동시장과 국내 기업 현장에서 AI 인재 수요가 빠르게 증가
 - 주요국의 전체 채용공고 중 AI 역량을 요구한 비중이 상승하는 추세
 - * 한국·미국·영국·인도·싱가포르·캐나다·독일·프랑스 등 8개국의 2023년 1월부터 2025년 12월까지 채용공고 추세를 보면 계절 변동 및 국가별 여건에 따른 등락은 존재하나, 전반적으로 AI 역량을 요구하는 채용 규모는 확대되는 흐름(<표 2> 및 [그림 1] 참고)
 - 한국은행 조사에 따르면, 국내에서 대기업뿐만 아니라 중견·중소기업까지 약 2/3가 향후 AI 인력 채용을 확대할 계획이 있다고 응답(박근용 외, 2025)
 - * AI 인력 채용 확대 계획 응답 비중: 대기업 69%, 중견기업 68.7%, 중소기업 56.2%

[표 2] 주요국 전체 채용공고 중 AI 인재 채용공고 비중의 연도별 최저치 비교

국가	인도	싱가포르	한국	영국	캐나다	독일	프랑스	미국
'23년 최저 (A)	2.62% (5월)	2.51% (2월)	2.50% (1월)	1.77% (6월)	1.22% (5월)	1.31% (12월)	1.36% (4월)	1.95% (5월)
'25년 최저 (B)	4.17% (1월)	3.12% (1월)	2.51% (1월)	1.90% (1월)	1.55% (3월)	1.22% (2월)	0.97% (8월)	2.24% (3월)
차이(B-A)	+1.55	+0.61	+0.01	+0.13	+0.33	-0.09	-0.39	+0.29

출처: oecd.ai(조회일: 2026.05.22.)

[그림 1] 주요국 전체 채용공고 중 AI 인재 채용공고 비중 추이

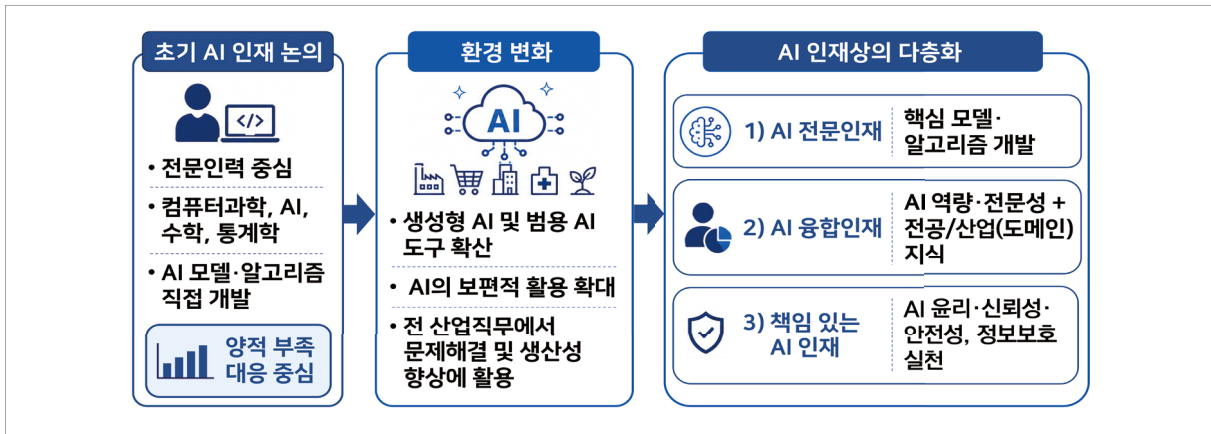


출처: oecd.ai(조회일: 2026.05.22.)

- 한편, AI 인재에 대한 수요는 단순한 양적 부족의 문제를 넘어, '어떤 역량을 갖춘 인재가 필요한가'라는 질적 변화의 문제로 확대
 - 초기 AI 인재 논의는 주로 AI, 컴퓨터과학, 수학, 통계학 등 AI 모델과 알고리즘을 직접 개발할 수 있는 전문인력 중심으로 전개
 - 그러나 생성형 AI와 범용 AI 도구의 확산으로 AI는 특정 전문가만 활용하는 기술이 아니라 모든 산업과 직무에서 문제해결과 생산성 향상에 활용되는 보편적 기술로 자리 잡는 양상
 - 이에 따라 AI 인재 수요는 AI 모델을 직접 개발하는 전문인력뿐 아니라 전공·산업 도메인 지식을 바탕으로 AI를 적절히 활용하고 결과를 검증하며, 현장 문제를 해결할 수 있는 인재로 확대
- AI 인재상 역시 전문 AI 연구자·개발자 중심에서 융합형 인재, 책임 있는 AI 인재 등으로 다층화
 - '전문 AI 인재'는 AI 핵심기술, 모델, 알고리즘, 시스템, 인프라 등을 개발·고도화할 수 있는 인재를 의미
 - 'AI 융합형 인재'는 AI 역량·전문성과 전공·산업 도메인 지식을 결합하여 현장의 문제를 정의하고 해결할 수 있는 인재를 의미
 - '책임 있는 AI 인재'는 AI 윤리, 신뢰성, 안전성, 편향·차별 방지, 개인정보 보호 등 AI 활용에 수반되는 사회적 책임을 이해하고 실천할 수 있는 인재
- 이러한 변화는 AI 인재 양성 정책의 방향이 'AI 전공자 확대'에만 머물러서는 충분하지 않음을 시사
 - AI 인재 양성은 전문 AI 인재의 고도화와 함께 비 AI 전공자의 AI 활용역량 강화, 전공별 AI 교육 내재화, 산업현장 기반 실습, 책임 있는 AI 교육을 포괄하는 방향으로 전환 필요
 - 특히 대학 단계에서는 전공과 AI를 분리된 교육영역으로 운영하기보다, 학생의 원전공 지식체계 안에서 AI를 이해·활용·검증할 수 있도록 교육과정을 재설계하는 것이 중요

- 이는 단순히 AI 관련 교과목 수를 늘리는 문제가 아니라, 학위과정, 전공체계, 평가방식, 산학협력, 지도 체계 등 대학 교육시스템 전반의 혁신을 요구하고 있으며, 세계 주요 대학은 이에 선도적으로 대응 중

[그림 2] AI 인재 수요 변화 양상



출처: 생성형 AI를 활용하여 제작한 이미지

○ 본고의 목적 및 내용

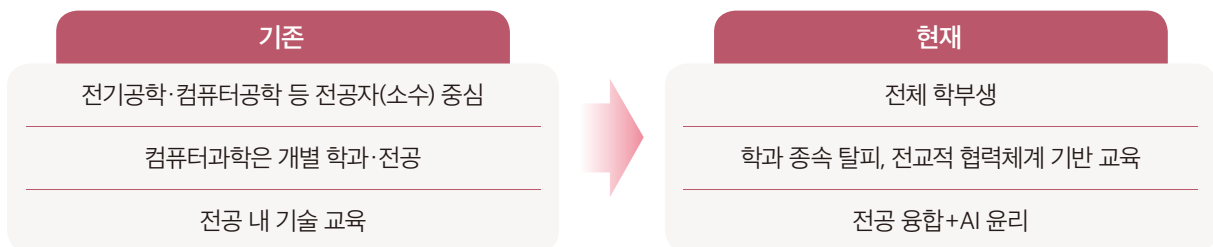
- **(목적)** 해외 대학·교육기관의 AI 인재 양성 사례를 분석하여 국내 AI 인재 양성 정책 및 대학 교육과정 혁신에 필요한 시사점을 도출
 - 해외 주요 대학·교육기관의 사례를 통해 AI 교육이 전공교육, 산업현장 경험, 책임 있는 활용, 융합연구 및 지도체계와 어떻게 결합되는지 검토
 - 이를 바탕으로 국내 AI 인재 양성 정책이 전문 인재 공급 확대를 넘어, 전공·산업·사회문제와 결합한 AI 역량을 체계적으로 함양하는 방향으로 발전하기 위한 정책적 시사점 도출
- **(내용)** 해외 대학·교육기관의 AI 인재 양성 사례를 중심으로 검토
 - 분석 대상은 AI 교육과정, 학위과정, 융합전공, 전공별 AI 내재화, 실습·프로젝트, 산업연계 프로그램, 공동지도 및 펠로우십 등을 운영하는 대학·교육기관 사례로 한정
 - * 미국(MIT·카네기멜런대), 중국(푸단대·칭화대), 영국(서리대·사우샘프턴대), 캐나다(토론토대), 스위스(ETH 취리히) 등 주요국의 8개 대학 및 기관
 - 각 사례가 어떠한 인재상을 지향하고, 어떻게 교육과정을 설계하며, 학생들에게 어떤 학습경험과 제도적 지원을 제공하는지를 중심으로 분석
 - 마지막으로 사례 간 공통점을 종합하고, 국내 AI 인재 양성 정책 및 대학 교육과정 혁신을 위한 시사점을 제시

II 해외 대학·교육기관 사례

○ (미국) MIT 슈바르츠만 컴퓨팅대학, 전공과 AI를 동등하게 구사하는 ‘컴퓨팅 이중언어자’ 양성 및 ‘Common Ground’ 공동교과 체계

- (개요) MIT는 2018년 10억 달러 규모의 투자를 발표하고 2019년 「슈바르츠만 컴퓨팅대학」을 기존 학과 구조에 종속되지 않는 캠퍼스 전체 학제 간 허브로 설립
 - 과거 MIT의 컴퓨팅·AI 교육은 전기공학·컴퓨터과학과(EECS) 중심의 전공자 대상 교육에 국한되어 타 전공 학생에게는 체계적인 컴퓨팅 학습 경로가 부족
 - 학부생 약 40%가 컴퓨팅에 몰리는 수급 불균형을 배경으로 2018년 투자 발표
 - 2019년 단과대 설립을 거쳐 ‘소수 전공자를 위한 분리된 컴퓨터과학 교육’에서 ‘Common Ground·혼합전공·윤리(SERC)를 통해 전 전공에 내재화된 융합교육’으로 전환
 - 핵심 목표는 컴퓨팅·AI와 타 전공 분야 양쪽에 능통한 ‘컴퓨팅 이중언어자(Computing bilinguals)’ 양성

[그림 3] 미국 MIT의 AI 교육 변화



- (Common Ground for Computing Education) 특정 학과에 종속되지 않고 여러 단과대 교수진이 참여하여 생물학·경제학·건축·인문학 등 학생의 원전공에 AI·컴퓨팅을 접목하는 ‘연계 교과(Bridge subjects)’ 방식의 모듈형 과목 운영²

² MIT 슈바르츠만 컴퓨팅대학 홈페이지: <https://computing.mit.edu/cross-cutting/common-ground-for-computing-education/common-ground-subjects/>

[표 3] 미국 MIT 슈바르츠만 컴퓨팅대학의 Common Ground 과목

분야	모듈명
머신러닝 및 최적화	- 머신러닝을 활용한 모델링: 알고리즘에서 응용까지 - 기후 행동을 위한 CO2 인공지능 - 지속 가능한 시스템 설계를 위한 계산 방법론 - 지구의 응용 원격 탐사
전산 과학 및 공학	- 전산 과학 및 공학 개론 - 학제 간 수치해석
디지털 인문학 및 사회과학	- 자기계발: 데이터 기반 학습 및 웰빙 - 대화형 데이터 시각화와 사회
컴퓨팅의 사회적·윤리적 책임	- 컴퓨팅 윤리 - AI와 합리성

- **(혼합 전공(Blended Majors))** 단순 AI 교과 추가가 아닌, 두 전공의 지식체계와 교육과정을 유기적으로 결합(혼합)한 학위과정 다수 운영³

[표 4] 미국 MIT 슈바르츠만 컴퓨팅대학의 전공 프로그램

핵심 전공(Majors)	혼합 전공(Blended Majors)	부전공(Minors)
- 컴퓨터과학 및 공학 - 인공지능 및 의사결정 - 컴퓨팅 기반 전자공학	- 컴퓨터과학 및 분자생물학 - 컴퓨터과학, 경제학, 데이터과학 - 도시과학·계획 및 컴퓨터과학 - 계산 및 인지	- 컴퓨터과학 - 통계학 및 데이터과학

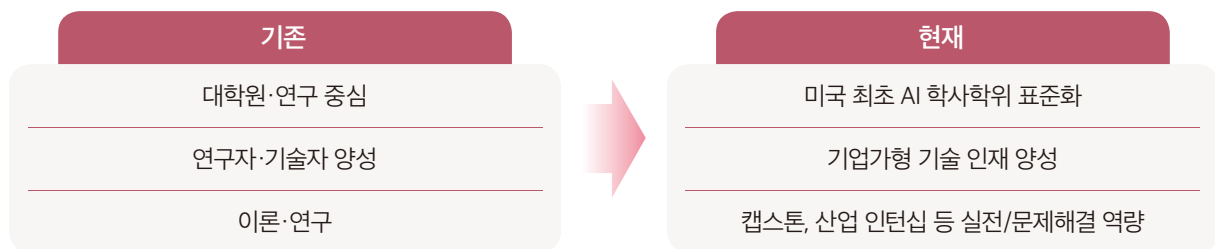
- **(공통 윤리교육)** 「사회·윤리적 책임 컴퓨팅(SERC, Social and Ethical Responsibilities of Computing)」을 통해 책임 있는 컴퓨팅 교육자료와 사례를 기존 컴퓨팅·AI 관련 교과에 통합·확산
 - 컴퓨터과학, 데이터과학, 인문학, 예술, 사회과학 등 다양한 분야의 구성원으로 이루어진 다학제 팀이 전 학위과정에 걸쳐 기존 교육에 접목할 수 있는 교육자료 개발
 - 생성형 AI 실패, 설명 가능한 AI, AI 동반자(Companion), 알고리즘 공정성 등의 다양한 사례를 수록한 출판물을 동료 평가를 거쳐 연 2회 발간

³ MIT 슈바르츠만 컴퓨팅대학 홈페이지: <https://computing.mit.edu/academics/undergraduate-programs/>

○ (미국) 카네기멜런대, AI 전문성과 기업가정신을 결합한 석사과정 운영

- **(개요)** 카네기멜런대는 AI·머신러닝 중심의 교육과정에 기업가정신 교육을 내재화한 「인공지능 및 혁신 석사과정(Master of Science in AI and Innovation, MSAIL)」 운영
 - 과거 카네기멜런대의 AI 교육은 1958년 미국 최초의 프로그래밍 강좌·세계 최초의 머신러닝학과 등 오랜 전통에도 불구하고 대학원·연구 중심이었으며, 학부 단계의 정식 AI 학위는 부재
 - 2018년 미국 최초의 AI 학사과정(B.S. in AI)을 신설해 학부 단계에서 AI를 독립 학위로 표준화
 - 기존 「생명공학·혁신·컴퓨팅 석사과정(MSBIC)」을 계승한 MSAIL을 AI·머신러닝 중심으로 재편하여, '연구 중심의 대학원 교육'에서 'AI 전문성과 기업가정신을 결합한 실전형 교육'으로 확장
 - 핵심은 AI로 시장기회를 발굴하고 외부 이해관계자와 협력해 제품·서비스로 구현하는 '기업가형 (Entrepreneurial) 인재' 양성

[그림 4] 미국 카네기멜런대의 AI 교육 변화



- **(교육과정)** AI·머신러닝·기업가정신 관련 기초 및 핵심과목 이수, 캡스톤 프로젝트 이수, 산업체 인턴십 수행 등으로 구성⁴
 - 기초 및 핵심과목을 통해 AI 관련 지식과 시장기회 발굴·평가 등의 기업가적 역량 함양
 - * 기초과목은 학사과정의 「컴퓨터 시스템 개론」이며, 이를 학위과정 첫 학기가 시작되기 전 또는 석사과정 기간 중 의무적으로 이수 필요
 - AI의 응용 가능성을 탐색하고 후속 개발이 가능한 제품으로 발전시키는 캡스톤 프로젝트 수행
 - 첫 번째 봄 학기를 마친 후 방학 동안 산업체 인턴십 참가 필수

⁴ 카네기멜런대 MSAIL 프로그램 홈페이지: <https://www.lti.cs.cmu.edu/academics/masters-programs/msail.html>

[표 5] 미국 카네기멜런대 MSAI 프로그램 교육과정 구성

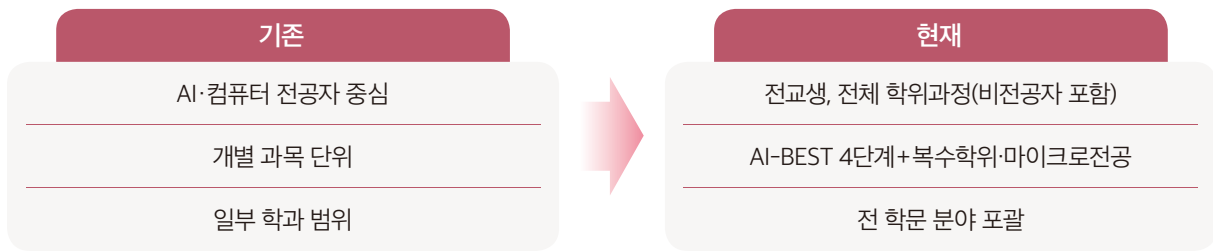
구분	주요 교과목(예시)	
핵심과목	• AI와 미래시장 • 컴퓨터 기술법 • AI 공학	• AI 혁신 • 캡스톤 프로젝트
필수과목	• 코딩 부트캠프 • 머신러닝 • 딥러닝 • 자연어처리	• 생성형 AI • 기타 AI, 자연어처리, 또는 머신러닝 분야의 과목 선택
인턴십	• 첫 번째 봄 학기를 마친 후 방학 동안 산업체 인턴십 참가	
선택과목 (택 3)	• 텍스트마이닝을 위한 머신러닝 • 자연어처리를 위한 신경망 • 고급 멀티모달 머신러닝 • 데이터베이스 시스템 • 고급 클라우드 컴퓨팅	• 컴퓨터 비전 • 웹 애플리케이션 개발 • 센싱 및 센서 • SW 개발 및 관리 • 스타트업을 위한 SW 엔지니어링

○ (중국) 푸단대, 교과목·복수학위·연합전공 3중 AI 융합교육 체계

- (개요) 푸단대는 2025년 발표한 「교육교학개혁 3.0판」의 일환으로 학사과정부터 박사과정까지 AI 기반 교육 혁신을 추진⁵
 - 과거 푸단대는 오랜 교양교육(통식교육) 개혁 전통에도 AI 교육이 컴퓨터·AI 전공 중심으로 제공되어, 비전공자의 체계적 학습 기회 부재
 - 2024년 「AI 대형강좌(AI大课)」를 설계 및 대폭 확대하여 전교생 대상 AI 교육 체계를 도입
 - 이를 토대로 'AI-BEST' 4단계, 'X+AI' 복수학위, 'AI+X' 마이크로전공으로 확장하며, '전공자 중심의 분절적 과목'에서 '전교생·전 학위과정의 단계적·제도적 융합교육'으로 전환
 - 핵심은 학사부터 박사까지 모든 학위과정에 학제 간 융합형 AI 교육을 내재화하는 것
 - * 모든 학위과정의 학생들이 학제 간 융합형 교육을 이수하도록 제도·커리큘럼 개편 및 복수학위·연합전공 프로그램 운영 추진

⁵ 赵天润(2026), "AI大课结项, 奔赴四个未来", 《复旦》校报, 2026.04.14., news.fudan.edu.cn/2026/0414/c3520a148722/page.htm

[그림 5] 중국 푸단대의 AI 교육 변화



- **(교과목)** 기초 과정부터 전문 과목까지 4단계로 구분해 ‘AI-BEST’ 체계를 구축하고, 약 116개의 AI 관련 강좌 개설
 - 「AI에 대한 이해」와 같은 기초적인 내용의 교과목부터 「LLM 이론 및 방법론」, 「자율주행 AI 원리와 실제」, 「영상의학에서의 AI 응용」 등 전문·응용 교과목까지 다양
 - 2025년까지 약 1.3만 명의 학생이 수강한 것으로 집계

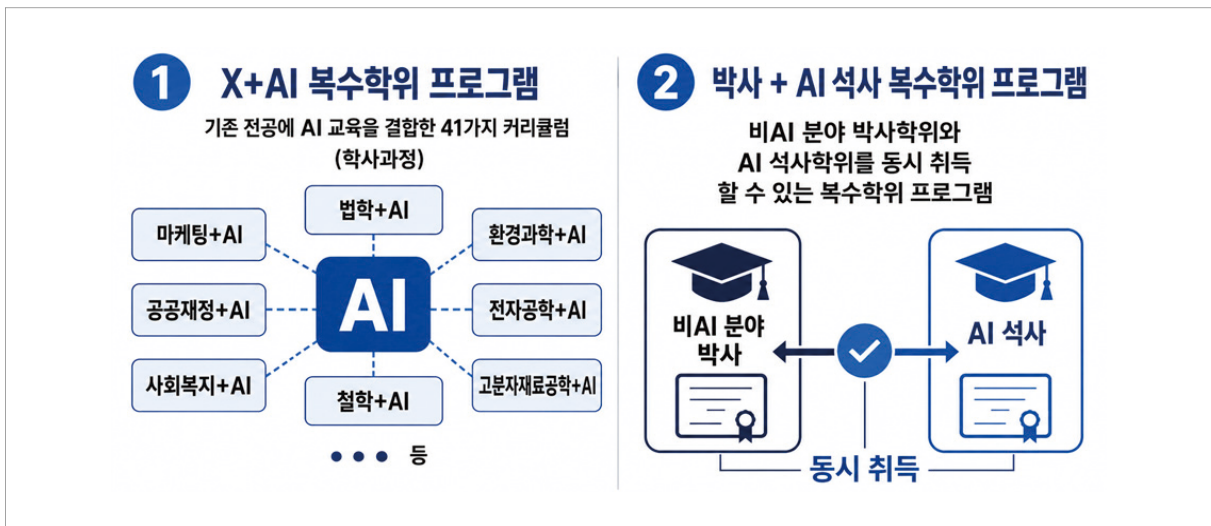
[표 6] 중국 푸단대 AI-BEST 교육 체계

단계	설명	주요 교과목(과목명)
AI-Basic (기초)	전교생 대상 AI 기초·소양 교육 (AI 비전공자와 AI 초심자도 수강 가능하도록 설계)	• 인공지능을 향한 첫걸음 • 소셜 컴퓨팅 개론
AI-Essential (핵심)	AI의 기본 원리, 공통 지식체계, 핵심기술에 관한 교육	• 패턴 인식과 머신러닝 • 인공지능 안전
AI-Subject (학제 간 융합)	인문학, 사회과학, 이·공학, 의학 등 기존 학문 분야에 AI를 접목하는 교육	• AI 기반 언어 분석·학습 • AI와 국가통치 • 대형모델 이론 및 방법론
AI-Thematic (응용)	산업 또는 과학연구에서의 AI 응용을 위한 교육	• 의료영상 분야에서의 AI 다중 시나리오 응용 • AI 고고학 • 뇌과학을 위한 AI

- **(연합전공)** 저장대·난징대·상하이교통대·중국과기대 등과 공동으로 ‘AI+X 마이크로 전공’ 프로그램 운영
 - AI 관련 전공(AI, 컴퓨터과학, SW 공학 등)이 아닌 학부생·대학원생 대상
 - AI 기초지식을 학습하고, 융합과목 및 산업현장 실습 과정을 통해 전공·산업 분야에 AI를 접목하는 역량 함양
 - 공동 운영 대학 간 학점 인정 및 ‘AI+X 마이크로 전공’ 수료증 발급

- **(복수학위)** ‘X+AI 복수학위 프로그램’, ‘박사+AI 석사 복수학위 프로그램’ 운영
 - ‘X+AI 복수학위 프로그램’은 기존 전공에 AI 교육을 결합한 커리큘럼(학사과정)으로 운영
 - * (예시) 「법학+AI」, 「철학+AI」, 「마케팅+AI」, 「공공재정+AI」, 「사회복지+AI」, 「환경과학+AI」, 「전자공학+AI」, 「고분자재료공학+AI」 등 41가지
 - 비 AI 분야 박사학위와 AI 석사학위를 동시 취득할 수 있는 복수학위 프로그램을 2025년 가을학기부터 시범적으로 개시

[그림 6] 중국 푸단대 복수학위 프로그램 개념도

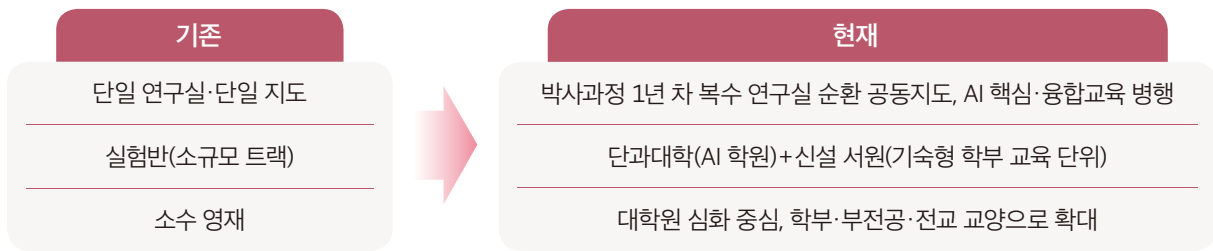


출처: 생성형 AI를 활용하여 제작한 이미지

○ (중국) 칭화대 AI 학원(College of AI), 융합적 시각 체득을 유도하는 ‘지도교수 그룹 순환제’ 운영

- **(개요)** 튜링상 수상자인 중국과학원 앤드류 야오 교수 주도로 2024년 4월 설립
 - 과거 칭화대의 AI 인재 양성은 2005년 설립된 ‘야오반(姚班, 컴퓨터과학 실험반)’을 통한 소수 영재 선발·집중 양성의 엘리트 모델 중심
 - ‘야오반’의 인재 양성 경험을 토대로 대학원 과정 중심의 AI 학원으로 출범
 - * 이와 함께 2025년 학부 과정(정원 150명)·AI 부전공·전교 AI 교양(117개 융합 교과목)·신설 서원을 더하여 양성 범위를 학부·전교로 확대

[그림 7] 중국 칭화대 AI 학원의 AI 교육 변화



- (교육과정) AI 핵심(AI Core) 트랙과 AI 융합(AI+) 트랙 운영

[표 7] 중국 칭화대 AI 학원 교육과정(트랙)

구분	교육과정명
AI 핵심(AI Core)	새로운 이론 및 효율적 알고리즘, 임베디드 AI 및 멀티모달 인식, 미래 컴퓨팅 모델, 과학적 지능 및 학제 간 통합, 기반 아키텍처 플랫폼
AI 융합(AI+)	AI 보안 거버넌스, 의료·헬스케어, 지능형 차량 및 제조, 에너지 및 환경

- (특징) 박사과정 학생이 1년 차에 3~6개월 단위로 2인 이상의 교수 연구실을 순환하도록 하는 ‘지도교수 그룹 순환제’ 도입⁶ 운영
 - 특정 분야에 매몰되지 않고 다양한 분야와 연구문화를 경험하며 융합적 시각을 체득하도록 유도하는 것이 목적

[그림 8] 칭화대 AI 학원 박사과정 순환제 운영 예시

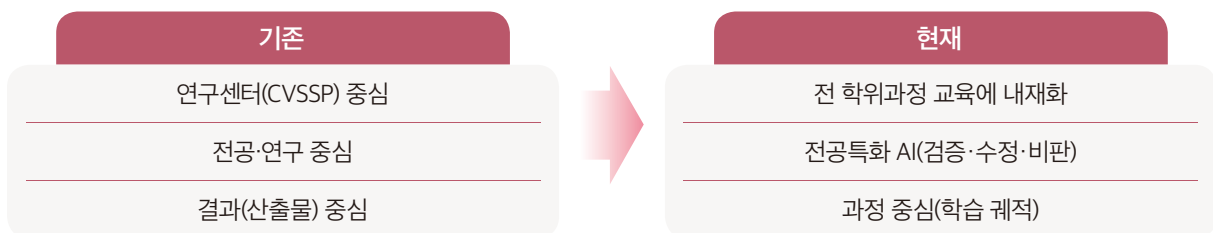
월	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
박사과정 A	A교수 지도			➡	B교수 지도				➡	C교수 지도		
박사과정 B	A교수 지도					➡	B교수 지도					
박사과정 C	A교수 지도				➡	B교수 지도			➡	C교수 지도		

⁶ 칭화대 AI 학원의 대학원 프로그램 홈페이지: https://collegeai.tsinghua.edu.cn/en/Graduate/Graduate_Programs.htm

○ (영국) 서리대(University of Surrey), 2026년 9월부터 전 학위과정에 전공특화 AI 교육 내재화 — ‘과정 중심 평가’ 전환 병행

- **(개요)** 서리대는 2026년 9월부터 모든 학위과정(학부·대학원)에 전공특화 방식의 AI 교육을 내재화하는 전면적 커리큘럼 재설계 시행⁷
 - 과거 서리대의 AI는 30년 이상의 역사를 지닌 「영상·음성·신호처리 연구센터(CVSSP)」를 중심으로 한 연구 강점에 기반했으나, 교육은 전공·연구 중심
 - 2021년 「사람 중심 AI 연구소」를 설립해 연구역량을 ‘교육과정 내재화’로 전환하는 목표를 제시
 - 2026년 9월부터 모든 학위과정에 전공특화 AI를 내재화하고 평가를 ‘과정 중심’으로 전환하며, ‘연구실 중심의 연구역량’에서 ‘전 학위과정 교육에 내재화된 전공특화 AI’로 전환
 - 핵심은 단순 리터러시 추가가 아니라, 모든 학위과정의 커리큘럼을 체계적으로 재설계함으로써 전공 지식으로 AI 산출물을 검증·수정·비판하는 역량을 함양하는 것

[그림 9] 영국 서리대의 AI 교육 변화



- **(전공특화 AI 교육)** 학생들이 전공 분야에서 AI를 사용할 수 있을 뿐 아니라, 전공 지식으로 AI 산출물을 검증·수정하고 책임지는 학습활동으로 설계

[표 8] 영국 서리대의 전공특화 AI 교육내용 사례

전공	전공특화 AI 교육내용(예시)
정치학	생성형 AI로 선거 투표행동 분석 → 정치행동 이론, 학술 데이터, 선행연구 등과 대조하여 부정확하거나 모호한 부분을 식별
토목공학	AI로 호텔 초기 설계안 생성·검토 → 수작업 계산 및 유한요소 모델링으로 AI가 생성한 결과물을 검증
경영학	AI로 실제 기업의 시장분석 및 마케팅 계획 초안 작성 → AI의 오류·일반화 한계를 비판한 후 전공 지식으로 재구성

⁷ 서리대 보도자료(2026.04.27.): <https://www.surrey.ac.uk/news/ai-be-embedded-discipline-specific-ways-every-university-surrey-degree-september-2026-training>

- **(평가방식 개편)** AI 교육을 내재화하는 동시에, 성취도 평가방식을 ‘과정 중심’으로 전환
 - 학습과 인지능력 변화의 증거 포착, 오류와 수정의 가치 인정, 투명성과 학습자 주동성, 완성된 산출물이 아닌 학습 궤적 평가 등을 원칙으로 제시
 - 이러한 원칙에 따라 AI가 생성할 수 있는 산출물을 평가하는 대신, 학생의 실제 이해와 판단력을 확인하기 위해 검증·수정·비판 과정을 평가
 - * (예시) 영문학: 에세이 제출 시 주석이 달린 발췌본·초안·동료 피드백·수정 메모 포함 의무화
 - 연극학: 리허설 일지·폐기된 아이디어·실패 실험·주석 영상 클립 포함 포트폴리오 제출

○ (영국) 사우샘프턴대(University of Southampton), 영국 최초로 전체 학사과정 AI 역량 필수 이수제 도입

- **(개요)** 사우샘프턴대는 2026년부터 모든 학사과정에서 AI 역량 교육을 필수 이수하는 제도 도입⁸
 - 과거 사우샘프턴대는 오랜 전통·연구 우수성을 갖춘 전자컴퓨터공학부(ECS)와 웹 사이언스(Web Science)*를 중심으로 전공생 대상 교육 및 연구
 - * 웹 사이언스: 컴퓨터과학·AI뿐 아니라 사회학·법학·경제학·심리학 등을 결합해, 웹(Web)이 사회에 미치는 영향과 사회가 웹을 형성하는 방식을 분석·연구하는 학제 간 학문 분야
 - 2026년 영국 최초로 전 학부생 AI 역량 필수 이수제를 도입하며, ‘전공자·연구 중심’에서 ‘전 학부생 대상 보편적 AI 역량(AI fluency) 교육’으로 전환
 - * 학사과정의 모든 학생이 AI 스킬·윤리·창의성을 다루는 교육과정을 이수하게 되며, 단순 사용법을 넘어 AI 도구를 효과적이고 책임 있게 활용하는 역량(AI fluency)을 함양하도록 설계

[그림 10] 영국 사우샘프턴대의 AI 교육 변화



- 핵심은 모든 학부생이 AI 스킬·윤리·창의성을 이수하고 단순 사용법을 넘어 책임 있게 활용하는 역량을 함양하는 것

⁸ 사우샘프턴대 보도자료(2026.02.24.): <https://www.southampton.ac.uk/news/2026/02/southampton-leads-the-way-with-introduction-of-artificial-intelligence-skills-for-all-students.page>

[표 9] 영국 사우샘프턴대 AI 역량 교육의 주요 내용

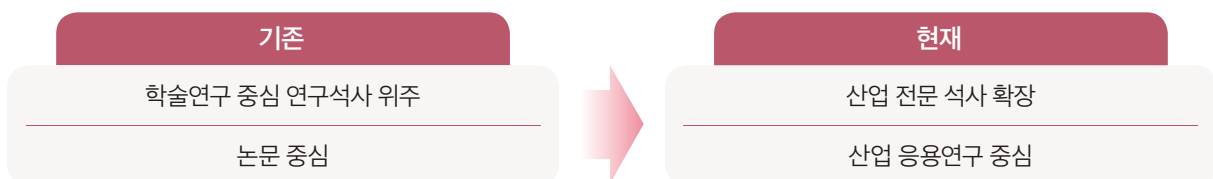
분야	교육내용
AI 스킬(AI Skill)	AI 도구 활용, AI 기반 문제해결, AI 기술 이해 등
AI 윤리(AI Ethics)	책임 있는 AI 활용, 학업·실무 맥락에서의 윤리적 판단 등
AI 창의성(AI Creativity)	AI를 활용한 창의적 작업, 아이디어 생성, 프로토타이핑 등

- **(전공별 AI 응용 학습)** 기본적 AI 역량을 확보한 이후, 각 전공 분야별 AI 활용에 대해 학습하도록 설계
- 전공 분야(공학·의학·예술·인문학) 맥락에서의 문제해결 방식, 윤리기준, 직업적 책임 등을 학습
- **(실습·프로젝트 연계)** AI를 직접 적용·구현하는 경험 기회 제공을 위해 AI 해커톤, 프로토타입 워크숍, 산업체 주도 프로젝트 등을 운영

○ (캐나다) 토론토대, 「MScAC-AI」— 8개월 교육 이수+8개월 응용연구 인턴십 ‘일체형’ AI 전문 석사

- **(개요)** 토론토대 「응용컴퓨팅 이학석사(MScAC)」의 AI 집중과정은 교과목 이수+응용연구 인턴십(Applied research internship)으로 구성⁹
- 과거 토론토대의 대학원 AI 교육은 전통적으로 학술연구 지향 연구석사 중심으로 이루어져, 산업 현장과의 연계는 제한적
- 2010년 MScAC를 출범해 학위논문 대신 8개월 산업 응용연구 인턴십을 결합한 ‘일체형’ 모델을 도입
- 이후 AI 등 7개 집중분야로 확대(2017년 설립된 벡터연구소 등 지역 생태계와 연계)하며, ‘논문 중심의 연구 인재 양성’에서 ‘산업 응용연구를 결합한 현장 문제해결형 인재 양성’까지 확장
- 핵심은 교과 8개월+산업 인턴십 8개월(총 16개월) 구조에서 산업·학술 공동지도로 현장 문제해결 역량을 기르는 것

[그림 11] 캐나다 토론토대의 AI 교육 변화



⁹ 토론토대 홈페이지 MScAC AI 집중과정 관련 설명: <https://mscac.utoronto.ca/concentrations/ai/>

- **(교과목 이수 과정)** 총 6개의 대학원 과목을 이수해야 하며, 이 중 절반(3과목)은 AI 영역 과목 이수 필수¹⁰
 - AI 영역에서는 AI 핵심(AI Core)에 해당하는 과목 중 2개, 그리고 컴퓨터공학과 커리큘럼 내 [그룹 2]에 포함되지만, AI 핵심에는 포함되지 않는 과목 중 1개를 선택(AI 영역 총 3과목)
 - AI 영역 외 과목으로서 전공 필수과목 2개*, 그리고 컴퓨터공학과 커리큘럼 내 [그룹 1], [그룹 3], [그룹 4]에 포함되는 과목 중 1개 선택하여 이수
- * '컴퓨터 과학자를 위한 커뮤니케이션(Communication for Computer Scientists)'과 '기술 기업가정신(Technical Entrepreneurship)'은 필수과목으로 반드시 이수 필요

[표 10] 캐나다 토론토대 MScAC-AI의 교과목 이수 조건

구분		요건	주요 교과목(과목명)	
AI	AI 핵심과목 (AI Core)	2개 과목 이수	- 머신러닝 입문 - 자연어 컴퓨팅 - 지식표현과 추론 - 신경망과 딥러닝	- 컴퓨터 비전 기초 - 로보틱스 제어 - 디지털 영상처리와 응용 - 계산언어학
	추가 AI 과목 (Additional AI course)	AI 핵심과목에 포함되지 않는 [그룹 2]의 1개 과목 이수	- 문장 구조의 이산수학적 모델 - 머신러닝 심화: 신약개발을 위한 AI	- 의미변화 계산모델 - 컴퓨터 비전 심화: 시각 컴퓨팅을 위한 물리기반 신경 표현
비 AI	기타 컴퓨터공학과 커리큘럼 과목	1개 과목 이수 (그룹 1, 3, 4 범위 내)	- 분산 컴퓨팅 이론 - 컴퓨터 네트워크 - 데이터베이스 시스템 기술	- 프로그램 설계 방법론 - 인간-컴퓨팅 상호작용 - AI의 윤리적 측면 - SW 비즈니스
	전공 필수	모든 과목(2개) 이수	컴퓨터 과학자를 위한 커뮤니케이션	기술 기업가정신

- **(응용연구 인턴십)** 사전 검토 및 승인을 거친 파트너 기관이 제안한 프로젝트를 수행하는 응용연구 인턴십 이수 필수
 - 파트너십이 체결된 산업체·연구기관·공공기관·비영리기관에서 제안한 프로젝트를 학생에게 공개하기 전 사전 검토를 진행하고 승인된 프로젝트에 한해 홍보
 - * 프로젝트는 단순한 현장실습 성격이 아닌, 현장의 실제 문제를 해결하는 응용연구 중심의 활동으로 설계되어야 한다는 가이드라인을 제시
 - 학생은 매칭된 파트너 기관에서 산업 지도교수(Industrial supervisor)와 학술 지도교수(Academic supervisor)의 지도하에 8개월간 프로젝트를 수행

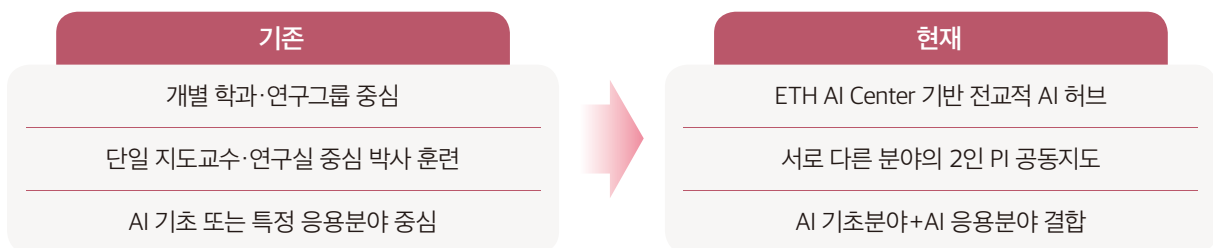
¹⁰ 토론토대 컴퓨터공학과 대학원 과정 설명: <https://web.cs.toronto.edu/graduate/course-descriptions>

- 최종 프로젝트 보고서, 공개용 프로젝트 개요, 프로토타입 발표(ARIA 쇼케이스 참가) 등을 산출해야 하며, 최종 통과(Pass)/탈락(Fail) 여부로 평가

○ (스위스) ETH 취리히, AI 센터 설립 및 학제 간 융합 AI 박사 펠로우십 운영

- **(개요)** ETH 취리히는 2020년 10월 대학 차원의 AI 연구 허브인 「ETH AI Center」를 설립하고, AI 기초·응용·사회적 함의를 포괄하는 학제 간 연구체계 구축
 - ETH AI Center 설립 전 ETH 취리히의 AI 연구·교육은 컴퓨터과학·공학·자연과학·보건·로보틱스 등 개별 학과·연구그룹 중심
 - ETH AI Center 설립 이후에는 개별 학과·연구그룹에 분산되어 있던 AI 역량을 대학 차원의 허브로 연결하고, AI 기초연구, AI 응용연구, AI의 사회적 함의를 함께 다루는 학제 간 플랫폼으로 재편
 - * AI 교과목 확대보다는 대학 차원의 AI 허브, 공동지도, 공동연구 공간, AI 기초+응용 주제 설계를 통해 박사급 AI 인재 양성체계를 재구성
 - 2026년 기준 16개 학과 전반에 걸친 120명 이상의 교수진과 외부 연구그룹이 참여하는 대규모 학제 간 AI 연구 플랫폼으로 확대

[그림 12] 스위스 ETH 취리히의 AI 교육 변화



- **(박사 펠로우십)** 학제 간 AI 연구를 수행할 박사과정생을 대상으로 「AI 센터 박사 펠로우십(AI Center Doctoral Fellowship)」을 운영¹¹⁾
 - 매년 약 5~8인의 박사 펠로우를 선발하며, 펠로우십 기간은 기본 3년
 - 지원자는 컴퓨터과학, 머신러닝, 통계학, 물리학, 공학, 응용수학 또는 관련 분야 석사학위를 보유하거나 취득 예정이어야 하며, ETH 취리히 박사과정 입학요건을 충족해야 함

¹¹⁾ ETH 취리히 홈페이지의 AI 센터 박사 펠로우십 설명: <https://ai.ethz.ch/research/phd-and-postdoc-programs/phd-fellowships.html>

- **(공동지도 제도)** 서로 다른 분야 또는 학과에 속한 2인의 PI가 동등한 비중으로 박사과정생을 공동 지도하는 구조를 제도화
 - 박사과정생이 단일 지도교수와 단일 연구실에 소속되는 방식이 아니며, 서로 다른 전문성을 가진 두 명의 지도교수가 공동으로 지도
 - 이를 통해 AI 기초기술과 응용 도메인, 기술적 문제와 사회적 함의, 방법론 연구와 실제 문제해결 간 연결을 연구 초기부터 경험하도록 설계
 - 펠로우는 ETH AI Center를 주된 연구 공간으로 사용하며 다양한 연구팀·교수진과 교류 및 공동연구 수행
 - * 모든 박사 펠로우가 AI Center 공간에 함께 배치되며, 이는 물리적 공간과 소속감을 통해 학제 간 협력을 구조적으로 유도하기 위한 목적
- **(연구분야 선택)** 모든 펠로우는 AI 기초분야와 함께 AI 응용분야를 반드시 선택해야 하는 구조
 - 펠로우는 AI 이론·기초·방법론·시스템에 해당하는 관심 분야와 AI 응용분야를 함께 선택하는 것이 필수
 - * AI 기초분야만 선택하는 것은 불가
 - 고급 AI 인재를 단순 알고리즘 개발자가 아니라, AI 기반 학제 간 문제해결자로 양성하는 모델로 이행 가능

[표 11] 스위스 ETH 취리히 AI 센터 박사 펠로우십의 주요 연구분야

AI 기초분야	AI 응용분야
• 머신러닝, AI 및 데이터과학의 수학적·통계적 기초 • AI 시스템 설계, 엔지니어링 및 HW • 에이전트 및 인간-AI 협업 • 컴퓨터 비전 및 기계 인식 • LLM과 자연어 이해 • 강화학습 및 데이터 기반 제어 • 파운데이션 모델 • 보안 및 개인정보 보호 • 신뢰할 수 있는 AI 시스템 • 안전성, 신뢰성 및 강건성	• 양자컴퓨팅을 위한 AI • 의료·의학 AI • 교육 AI • 에너지시스템, 기후 모델링 및 환경을 위한 AI • 금융 및 경제 AI • 산업 및 제조를 위한 AI • 과학·공학 AI • 우주 연구 및 탐사를 위한 AI • 로봇공학 및 자율시스템을 위한 AI • 데이터과학 기초에 대한 학제 간 협력

* 주: 분야 선택 시, 상기 분야에 국한되지는 않음

III 종합 및 시사점

1. 종합: 사례의 특징

[표 12] 사례별 주요 내용 및 AI 교육 변화 요약

사례	과거	현재	핵심 변화
MIT (미국)	유관 전공자(소수) 중심, 컴퓨터과학은 개별 학과·전공	Common Ground로 전 전공에 내재화, '컴퓨팅 이종언어자' 양성	분리된 CS → 전공 융합
카네기멜런대 (미국)	대학원·연구 중심, 학부 정식 AI 학위 부재	미국 최초 학부 AI 학위(2018) +기업가정신 결합(MSAII)	대학원/연구 → 학부 표준화·실전
푸단대 (중국)	유관 전공자 중심의 개별 과목	전교생 대상 AI-BEST 4단계 +복수학위·마이크로전공	개별 과목 → 전교 제도화
칭화대 (중국)	소수 영재 집중 양성	대학원 융합연구(순환제·AI 핵심/융합) 중심+학부·전교 확산	엘리트 소수 → 대학원 융합 심화·다층 확산
서리대 (영국)	연구센터 중심의 AI 연구역량	전 학위과정 전공특화 AI 내재화+과정 중심 평가	연구 → 교육 내재화
사우샘프턴대 (영국)	유관 전공·연구 중심	전 학부생 AI 역량 필수 (영국 최초)	전공 선택 → 전원 필수
토론토대 (캐나다)	논문 중심 연구석사, 산학 연계 제한적	산업 응용연구 결합 '일체형' 전문 석사로 확장	연구 → 연구+산업 응용연구
ETH 취리히 (스위스)	학과별 단일 지도교수· 연구실 중심 박사 양성	전교 횡단 학제 간 공동지도	학과 → 학제 간 융합
종합	• 전공자·소수 → 전교생 • 특정 학과·연구실 → 전공에 내재화 • 기술·연구 중심 → 응용·윤리·학제 간 문제해결 중심		

○ 교육과정 설계뿐 아니라 거버넌스·지도체계 자체를 융합형으로 재편

- 칭화대는 박사과정 학생이 1년 차에 여러 연구실을 순환하도록 하는 지도교수 그룹 순환제 운영
- ETH 취리히는 서로 다른 분야의 2인 PI가 박사과정생을 공동지도하고, 모든 펠로우가 AI 기초와 AI 응용 분야를 함께 선택하도록 설계
- MIT와 푸단대 역시 특정 학과 단위의 폐쇄적 교육이 아니라, 전교적·학제 간 협력체계를 기반으로 AI 교육을 확산

○ AI 교육을 특정 전공의 전문교육이 아니라 전교적·전공횡단적 역량으로 재정의

- MIT는 슈바르츠만 컴퓨팅대학을 캠퍼스 전체의 학제 간 허브로 설립하고, 학생의 원전공에 AI·컴퓨팅을 접목하는 Common Ground 과목 및 혼합전공 운영
- 푸단대는 AI-BEST 체계를 통해 기초·핵심·학제 간 융합·응용 단계의 AI 교육을 전 학위과정에 확산
- 서리대와 사우샘프턴대는 모든 학위과정 또는 모든 학부생을 대상으로 AI 교육을 내재화·필수화하는 방향을 제시

○ AI+X 또는 X+AI 방식의 전공특화·도메인 결합형 교육이 핵심 모델로 부상

- MIT의 혼합전공, 푸단대의 X+AI 복수학위 및 AI+X 마이크로전공, 칭화대의 AI+ 트랙, ETH 취리히의 AI 기초분야+AI 응용분야 결합 구조는 모두 AI 자체보다 'AI를 전공·산업·사회문제와 결합하는 능력'을 중시 - 즉, AI 기술 자체를 이해하는 능력과 함께, 법학·의학·경영학·공학·인문사회·예술 등 각 전공 분야의 문제를 AI 관점에서 재정의하고 해결하는 능력을 갖춘 AI 인재 양성을 목표로 교육과정 및 구조 설계
- 카네기멜런대는 MSAIL 과정에서 AI·머신러닝 전문성에 기업가정신을 결합하여, AI를 시장기회 발굴 및 제품·서비스 개발과 연결하는 융합형 인재 양성을 지향

○ 책임 있는 AI, 윤리, 검증·비판 역량이 교육과정의 필수 요소로 편입

- MIT는 사회·윤리적 책임 컴퓨팅을 기존 컴퓨팅·AI 교과에 통합
- 푸단대와 사우샘프턴대도 AI 안전, 윤리, 책임 있는 활용을 교육내용에 포함
- 서리대는 AI 산출물을 그대로 평가하는 방식이 아니라, 학생이 AI 결과를 검증·수정·비판하는 과정을 평가 대상으로 설정
- 이는 생성형 AI 확산 이후 대학 교육의 핵심 과제가 'AI 사용 여부를 제한하는 것'이 아니라 AI를 책임 있게 사용하고 그 결과를 검증할 수 있는 역량을 길러내는 것으로 이동하고 있음을 의미
 - AI 활용역량은 단순한 프롬프트 작성이나 도구 사용 능력에 그치지 않으며, AI가 생성한 결과의 정확성·편향·한계·사회적 영향·전공 지식과의 정합성을 판단하는 비판적 검증 능력이 핵심 역량으로 부상 중

[표 13] 해외 대학·교육기관 주요 사례 특성

특성	핵심 취지	주요 사례
지도체계 융합	교과목뿐 아니라 지도방식도 학제 간 결합	- 칭화대: 지도교수 그룹 순환제 - ETH 취리히: 타 분야 2인 PI가 박사과정생 공동지도 - MIT·푸단대: 전교적·학제 간 협력체계 기반 AI 교육
전교적 교육 체계 개편	AI 교육을 특정 전공의 전문교육이 아닌 전공횡단적 역량 교육으로 재정의	- MIT 슈바르츠만 컴퓨팅대학: 학제 간 허브로 설립, 학생의 원전공에 AI·컴퓨팅을 접목 - 푸단대 AI-BEST 체계: 기초·핵심·학제 간 융합·응용 단계의 AI 교육을 전 학위과정에 확산 - 서리대, 사우샘프턴대: 모든 학위과정 학생 대상 AI 교육 내재화·필수화 방향성 제시
AI 융합	전공 분야 문제를 AI 관점에서 재정의·해결하는 능력 함양	- MIT 혼합전공 - 푸단대의 X+AI 복수학위 및 AI+X 마이크로전공 - 칭화대의 AI+ 트랙 - ETH 취리히의 AI 기초분야+AI 응용분야 결합 - 카네기멜런대 MSAIL의 AI-머신러닝 전문성과 기업가정신 결합
AI 책임성	책임 있는 AI, 윤리, 검증·비판 역량 교육 필수화	- MIT: AI 교과에 사회·윤리적 책임 컴퓨팅 결합 - 푸단대·사우샘프턴대: AI 안전, 윤리, 책임 있는 활용 등을 교육내용에 포함 - 서리대: 학생이 AI 산출물을 검증·수정·비판하는 과정을 평가 대상으로 설정

2. 시사점

○ AI 인재 양성 정책의 목표를 '전문인력 공급 확대'에서 '전공별 AI 역량의 체계적 확산'으로 확장 필요

- AI가 범용기술로 확산되면서 비 AI 전공자와 일반 전공 학생에게도 AI를 이해·활용·검증할 수 있는 역량이 요구되는 동시에, AI 전문인재 역시 전공·산업 도메인, 기업가정신, 커뮤니케이션, 윤리·책임성, 현장 문제해결 경험이 요구되는 상황
- 따라서 국내 정책은 AI 전공자 확대와 함께 모든 전공의 학생이 자신의 전공 분야에서 AI를 활용할 수 있도록 하는 '전공별 AI 내재화' 전략 병행 필요
 - 단기적으로는 대학별 AI 공통교과, AI 리터러시 교과, 전공별 AI 응용 모듈 개발을 지원
 - 중장기적으로는 전공 교육과정 자체를 AI 시대에 맞게 재설계하고, 학과·단과대학 단위의 교육과정 개편을 유도

○ 대학 전반의 AI 교육 내재화를 위한 공동교과·모듈형 교육 인프라 구축 필요

- MIT의 Common Ground, 푸단대의 AI-BEST, 사우샘프턴대의 보편적 AI 역량(AI fluency) 교육은 AI 교육 확산을 위해 대학 전체가 공유할 수 있는 교육 인프라가 중요하다는 점을 보여주는 사례
- 국내 대학에서도 개별 교수자의 자율적 교과목 개설에만 의존하기보다, 대학 차원의 공통 AI 교육 체계 구축 필요
 - AI 기초소양, 데이터 이해, 생성형 AI 활용, AI 윤리·안전, 전공별 AI 응용, AI 기반 프로젝트 등을 수준별·모듈형으로 구성
 - 각 전공은 공통 모듈을 기반으로 전공 특성에 맞는 사례, 데이터, 과제, 평가방식을 결합
 - 이를 통해 교육의 중복을 줄이고, 비 AI 전공 학생에게도 일정 수준 이상의 AI 역량을 안정적으로 제공할 수 있도록
- 특히, 공통 AI 교육이 단순한 온라인 강좌 또는 기초교육으로 축소되는 것을 지양
 - AI 도구 사용법 중심의 단기 교육을 넘어 전공 지식과 AI 결과물을 연결해 판단하는 비판적 사고, 책임 있는 활용, 협업 기반 문제해결 포함 필요

○ AI+X, X+AI 방식의 복수전공·마이크로전공·융합트랙 제도화

- 해외 사례에서 확인되는 중요한 흐름은 AI와 전공 도메인을 결합한 학위·전공제도의 다양화
 - MIT는 컴퓨터과학과 분자생물학·경제학·도시계획 등 다양한 혼합전공을 운영
 - 푸단대는 X+AI 복수학위와 AI+X 마이크로전공을 통해 비 AI 전공 학생의 AI 융합역량을 제도적으로 인정
 - ETH 취리히는 박사 펠로우십에서 AI 기초분야와 AI 응용분야를 함께 선택하도록 설계
- 국내 대학에서도 기존의 AI 부전공·복수전공을 넘어, 전공별 특성과 학생 수요에 부합하도록 유연한 학위·이수체계 확대 필요
 - 예컨대, 법학+AI, 의학+AI, 경영학+AI, 제조+AI, 에너지+AI, 문화예술+AI 등 전공별 AI 융합트랙 설계
 - 마이크로전공, 나노디그리, 인증과정 등 단기·중기 이수체계를 통해 비 AI 전공 학생의 진입장벽 완화
 - 복수전공·연계전공·융합전공 이수 결과가 졸업요건, 취업, 대학원 진학, 산학 프로젝트 참여와 연계될 수 있도록 제도적 인정체계를 마련

○ 산업현장·공공문제 기반 프로젝트와 인턴십을 정규 교육과정에 결합

- 카네기멜런대, 토론토대, 사우샘프턴대는 AI 교육에서 프로젝트·인턴십·해커톤·캡스톤이 보조 활동이 아니라 핵심 학습경험으로 제도화되고 있음을 보여주는 사례
- 국내에서도 산업현장 기반 실습을 단기 체험이나 비교과 활동이 아닌, 정규 교육과정의 일부로 편입하는 AI 교육 설계 확산 필요
 - 산업체·공공기관·연구기관이 실제 문제를 제안하고, 대학은 이를 교육과정 내 프로젝트로 전환 (기획·도입)
 - 학생은 학술 지도교수와 산업 지도교수의 공동지도를 받으며 문제정의, 데이터 분석, 모델 설계, 성능 검증, 결과 해석, 이해관계자 피드백 반영 등을 수행
 - 이러한 방식은 학생에게 실제 문제해결 경험을 제공하는 동시에, 대학과 산업계 간 AI 인재 양성 협력을 실질화하는 수단이 될 것으로 기대

○ 책임 있는 AI와 과정 중심 평가를 AI 교육의 핵심 원칙으로 반영

- 생성형 AI 확산 이후, 대학 교육에서 활용을 금지하거나 제한하는 방식만으로는 학습성과를 보장하기 어려운 상황
- 해외 사례는 AI를 교육과정에 적극적으로 내재화하되, 학생이 AI 결과를 검증·수정·비판하는 과정을 평가하는 방향으로 전환 중
 - 서리대의 과정 중심 평가 방식은 AI 산출물 그 자체보다 학생의 사고 과정과 판단력, 오류 인식, 수정 노력, 전공 지식에 기반해 검증
 - MIT, 푸단대, 사우샘프턴대 사례 역시 AI 윤리, 안전, 책임 있는 활용을 교육과정의 필수 요소로 포함
- 국내 대학도 AI 활용을 학업 윤리 문제로만 접근하기보다, AI 시대 학습성과에 대해 새로운 정의 필요
 - 학생이 AI를 어떻게 활용했는지, 어떤 한계를 발견했는지, 어떤 근거로 수정했는지, 최종 판단에 어떤 전공 지식을 적용했는지를 평가
 - 이를 위해 초안, 프롬프트 기록, 수정 메모, 검증 자료, 동료 피드백, 실패한 시도 등을 포함하는 포트폴리오형 평가 도입 고려
 - 책임 있는 AI 교육은 정보보호, 편향·차별, 저작권, 설명 가능성, 안전성, 사회적 영향 등을 전공별 맥락에서 다루도록 설계

○ 학제 간 공동지도·지도체계 개편 등 융합형 거버넌스의 단계적 도입

- 칭화대의 지도교수 그룹 순환제, ETH 취리히의 서로 다른 분야 2인 PI 공동지도, 토론토대의 산업·학술 공동지도는 교육과정뿐 아니라 지도체계·거버넌스 자체를 학제 간 융합형으로 재편한 사례
- 국내 대학·대학원에서도 단일 지도교수·단일 연구실 중심의 양성 구조를 넘어, 학제 간 공동지도와 연구실 간 협업을 제도적으로 지원 필요
 - 서로 다른 전공·연구실의 교수가 공동지도하는 학제 간 공동지도제, 박사과정 초기 복수 연구실 경험(순환), 산업·학술 공동지도 등을 시범 도입
 - 학과·단과대학 간 공동 교육·연구 운영, 겸임·교차 임용, 학제 간 학위·연구 트랙 등 조직 칸막이를 낮추는 행정·재정 기반 마련
 - 다만 공동지도가 책임 소재가 모호해지지 않도록, 지도교수 간 역할 분담과 학생 지원·평가 책임 구조를 함께 설계

○ 해외 사례의 시사점을 국내 적용할 때, '단순 모방'보다 대학 유형과 인재 양성 목표에 따른 맞춤형 설계 필요

- 해외 사례의 공통점은 국내 정책 설계에 중요한 참고가 되지만, 모든 대학에 동일한 모델을 적용하는 것은 부적절
- 정책적으로는 대학 유형별·전공 분야별로 차별화된 AI 인재 양성모델을 인정하고, 획일적 성과지표보다 교육과정 혁신의 질적 성과 평가 필요
 - 단순 수강자 수, 교과목 수, 전공 신설 수 중심의 지표만으로는 AI 교육 내재화의 성과를 측정하는 데 한계
 - 전공별 AI 교육과정 개편 수준, 산업·공공문제해결 경험, 책임 있는 AI 평가체계 도입 여부, 학과 간 공동 운영 등을 함께 고려 필요
- 결론적으로, 국내 AI 인재 양성 정책은 전문 AI 인재의 양적 확대와 전공별 AI 내재화, 경험 기반 학습, 책임 있는 AI 교육, 융합형 거버넌스를 포괄하는 방향으로 고도화해야 할 것
 - 이는 AI 인재 양성을 단순한 인력 공급정책에서 대학 교육시스템 혁신정책으로 확장하는 것을 의미
 - 해외 사례는 이러한 전환이 이미 주요 대학·교육기관에서 구체적 제도와 교육과정으로 구현되고 있음을 보여주며, 국내에서도 이를 참고한 체계적 정책 설계 필요

참고문헌

1. 국내문헌

- 박근용·서동현·오삼일·한진수(2025), AI 전문인력 현황과 수급 불균형: 규모, 임금, 이동성 분석. BOK 이슈노트 제2025-36호, 서울: 한국은행
- 윤보성·진회승(2025), 주요국 AI 인재 양성 및 유치 정책: 현황 및 시사점. 이슈리포트 IS-203, 성남: 소프트웨어정책연구소

2. 해외문헌

- 赵天润(2026), “AI大课结项，奔赴四个未来”，《复旦》校报, 2026.04.14., <https://news.fudan.edu.cn/2026/0414/c3520a148722/page.htm>

3. 웹사이트

- ETH 취리히 AI 센터 홈페이지, <https://ai.ethz.ch>
- MIT 슈바르츠만 컴퓨팅대학 홈페이지, <https://computing.mit.edu>
- OECD AI 홈페이지, <https://oecd.ai>
- 사우샘프턴대 홈페이지, <https://www.southampton.ac.uk>
- 서리대 홈페이지, <https://www.surrey.ac.uk>
- 칭화대 AI 학원 홈페이지, <https://collegeai.tsinghua.edu.cn>
- 카네기멜런대 홈페이지, <https://www.cmu.edu>
- 토론토대 컴퓨터과학과 홈페이지, <https://web.cs.toronto.edu>

PHOTO ESSAY

사진산책

중간

디지털과 아날로그의 중간. 일상과 쉼의 중간 같은 사진.

특정 주제나 내용에 치우치지 않는 누구나 공감할 수 있는 사진

이호준 Lee, Ho-Joon | 성공회대 강사·사진가

대한항공 여행사진 공모전에서 2회 수상하고, 여덟 차례의 개인전과 아홉 번의 단체전을 개최했다. 포토에세이 <직조>와 <걸으면 보이는>을 출간했으며, 여러 월간지와 주간신문에 사진이야기를 연재했다. 다수의 인쇄, 방송 매체에 소개되었고 지자체, 공공기관, 사진모임 초청으로 사진 강연을 한다. 대학에서 사진학을 강의하고 초등학교 미디어리터러시 강사로도 활동하고 있다.



본문: 이호준 작가의 사진

문학평론가 김현 선생의 에세이를 읽다가 한 문장이 눈에 들어와 몇 번을 반복해 읽고 형광펜으로 강조 표시를 했다. “책을 읽고 쓴다는 것은, 세계를 새롭게 볼 수 있게 만든다. 세계를 새롭게 볼 수 있다는 것은, 일상 속에 파묻혀 일상인의 눈으로 세계를 보지 않게 되었음을 뜻

한다.” 내가 평소 사진에 대해 갖고 있던 생각을 문장으로 정확히 옮겨놓은 듯했다. 익숙한 풍경을 낯설게 바라보는 일, 무심히 지나치던 장면에서 숨겨진 무언가를 발견하는 일. 사진을 찍는 행위와 글을 읽고 쓰는 행위가 크게 다르지 않다는 것을 말해주고 있었다.



모든 예술은 장르를 뛰어넘어 같은 목적을 공유하는 게 아닐까. '세계를 새롭게 보게 하는 것' 말이다. 그렇기에 예술가는 자기 장르의 울타리 안에만 머물지 않도록 해야 한다. 당장 사진과 문학만 해도 서로에게 필요한 존재이지 않은가. 문학은 사진에 사유의 깊이를 더해주고, 사진은 문학이 말하려는 것을 한 컷의 이미지로 보여줄

수 있다. 사진가에게 가장 필요한 것은 자기만의 시선을 갖는 것인데 그것은 타고난 재능만으로 완성되지 않는다. 이웃한 예술 장르와 교류하며 자연스럽게 영향을 주고받는 가운데 조금씩 자라나는 것이다. 그래서 오늘도 촬영길에 나서며 카메라 옆에 책도 함께 꽂는다. 좋은 문장을 만난 날은 사진도 잘 찍히는 느낌이 든다.

모든 산업을 알아야 하는 산업: AI 시대 금융의 변화와 정책 제언

· 서병기 | UNIST 경영과학부 교수 | bkseo@unist.ac.kr



세 장면, 하나의 질문

<1썸> 앤트로픽의 미토스 소동

지난 4월, 앤트로픽(Anthropic)이 개발한 새로운 인공지능 모델인 ‘미토스(Mythos)’가 발표되자마자 작지 않은 소동이 일어났다. 미토스는 아직 아무도 모르는 보안 취약점, 이른바 제로데이(zero-day)를 스스로 찾아내는 녀석이다. 사람 손과 자동화 검사를 20년 넘게 빠져나간 것들이 걸려 나왔다. 가장 오래된 것은 보안으로 이름난 OpenBSD의 27년 묵은 버그였다. 회사 발표에 따르면, 한 달 남짓의 테스트에서 찾아낸 것만 수천 개다. 그런데 정작 흥미로운 것은 발표 다음에 벌어진 일이다. 24시간이 채 지나기 전에 미 연준 의장과 재무장관이 재무부 청사로 대형 은행 CEO들을 긴급 소집했다. 국방부도 정보기관도 아니고, 왜 하필 은행이었을까? 그 자리에서 내려진 주문이 또 의미심장하다. 미토스를 막으라는 것이 아니라, 미토스를 직접 써서 각자의 약점부터 찾으라는 것이었다. 파장은 곧 한국으로도 전해졌다. 금융위원회는 지난 5월, 2013년 이후 십수 년간 금융보안의 대원칙으로 지켜져 왔던 망분리 규제를 완화하겠다고 밝혔다. “AI 공격은 AI로 방어”한다는 것이 당국의 표현이었다. 막아야 할 대상이 곧 막을 수단이 된 셈인데, 그렇다면 규제는 대체 무엇을 규제해야 하는 것일까?

<2썸> 레오폴드의 몰락

지난 7월의 뉴욕. 시추에이셔널 어웨어니스(Situational Awareness)라는 헤지펀드가 마진콜을 맞았다. 2024년 여름 혜성처럼 등장해 2억 달러 남짓을 2년 만에 450억 달러까지 불린, 스물다섯 살의 천재 레오폴드 아셴브레너(Leopold Aschenbrenner)가 운용하던 펀드였다. AGI(Artificial General Intelligence)가 머지않아 도래한다는 전망 하나로 데이터센터와 반도체, 전력 인프라를 네 배 레버리지까지 사들인 결과였다. 그리고 한 달 만에 운용자산의 3분의 2가 사라졌다. 상장주식은 통째로 경쟁 헤지펀드에 10% 가까이 싼 값에 넘어갔고, 그 안에는 SK하이닉스도 있었다. 그 주 국내 증시가 이렇다 할 이유 없이 급락했던 것이 사실은 이 때문이었다는 분석도 있다. 그렇다고 이 펀드의 전망이 틀렸다고 판명된 것도 아니다. 미래를 옳게 보고도 무너질 수 있다면, 무너진 것은 대체 무엇인가?



<3썬> 금융 특화 파운데이션 모델

얼마 전 산업 특화 파운데이션 모델을 개발하는 공모 사업에 금융 주제로 참여해 보려고 계획서를 준비한 적이 있다. 그런데 금융에 특화된 파운데이션 모델이란 대체 어떤 물건이어야 할까? 공리를 거듭하다 보니 아무래도 증권사 리서치센터 정도는 되어야겠다는 데 생각이 미쳤다. 리서치센터가 어떤 곳인가? 반도체와 조선, 유통과 제약과 건설을 각각 맡은 애널리스트들이 거의 모든 산업 섹터를 나누어 커버하는 조직이다. 거기까지 생각이 닿는 순간 계획서를 쓰던 손이 멈췄다. 아니, 이건 결국 범용 모델이 아닌가.

생각해 보면 당연한 귀결이었다. 금융은 스스로 무엇을 생산하지 않는다. 다른 산업이 벌이는 일을 자금과 가격의 형태로 되비추는 것이 금융이 하는 일이다. 조선사에 대출을 내주려면 조선업을 알아야 하고, 정유사의 회사채를 평가하려면 정유업을 알아야 한다. 금융을 잘하려면 모든 산업을 알아야 한다는 결론은 처음부터 정해져 있던 셈이다. 본 기고의 제목은 여기서 나왔다.

위의 세 장면은 각각 시스템과 인프라, 투자와 자산 가격, 그리고 고객을 마주하는 서비스의 문제를 건드리고 있다. 세 영역에서 공통으로 드러나는 문제는 AI가 판단의 속도와 범위를 빠르게 넓히는 동안, 그 판단을 확인하고 책임지는 체계는 아직 그 속도를 따라가지 못하고 있다는 점이다. 이 글에서는 세 영역에서 지금 무엇이 벌어지고 있는지를 가장 앞서 있는 곳들의 사례로 살펴보고자 한다. 우리가 앞으로 마주할 문제 대부분은 이미 그곳에서 한 번씩 모습을 드러냈다.

[표 1] 금융 시 적용 영역

적용 영역	주요 내용	핵심 리스크	대응 방향
시스템	보안 관제, 이상 거래 탐지, 결제·정산 자동화, 감독 업무(SupTech)	공급자 집중, 사이버 위협의 고도화, 인프라 설계 전제의 붕괴	제3자 리스크 관리, 결제 인프라 재설계, 표준 형성 과정 참여
투자	로보어드바이저, 알고리즘 트레이딩, AI 리서치·리포트 생성	모델 동조화에 따른 쓸림, 레버리지 확대, 생성 정보의 신뢰성	자동주문 식별, 레버리지 익스포저 관측, 산출물 검증
서비스	신용평가 고도화, 상담·판매 채널, 언더라이팅, 내부 업무 자동화	설명책임 공백, 편향, 책임 소재의 분산	설명·이의제기 절차, 사람의 개입 통제, 책임 소재의 정리

1. 시스템 - 방어의 도구가 곧 위협의 도구일 때

미토스 사태가 드러낸 것은 새로운 해킹 기법이 아니다. 공격과 방어 사이의 비대칭이 사라졌다는 사실이다. 금융보안은 그동안 시간차 위에 서 있었다. 취약점을 찾는 시간, 그것을 악용하는 시간, 그 사이에 패치를 배포하는 시간. 망분리도 결국 이 시간을 벌기 위한 장치였다. 그런데 취약점이 알려진 뒤 그것을 뚫는

코드가 나오기까지 걸리던 며칠에서 몇 주가 몇 시간으로 줄면, 패치를 배포할 시간차 자체가 사라진다. 방어하는 쪽이 같은 모델을 쓰지 않으면 애초에 경기가 성립하지 않는 것이다.

진짜 문제는 그다음이다. 방어를 위해 고성능 AI를 쓴다는 것은, 그 모델을 만드는 소수의 회사와 그것을 엮어 돌리는 소수의 클라우드 사업자에게 금융 시스템의 안전을 맡긴다는 뜻이기도 하다. 금융안정위원회(FSB)가 AI 관련 취약성 가운데 가장 앞에 놓은 것도 이 제3자 의존과 공급자 집중 문제였다. 종래의 금융 리스크는 대체로 금융회사 안에 있었다. 자본과 유동성이 충분한지를 보면 됐다. 그런데 AI가 매개하는 리스크는 금융회사 밖에, 그것도 금융감독의 관할 밖에 놓인다. 감독의 대상은 은행인데, 정작 그 은행의 판단을 좌우하는 것은 감독 대상도 아닌 어느 기술회사의 모델 업데이트인 상황이 오는 것이다.

한편, 더 근본적인 변화는 결제와 정산 인프라에서 이미 시작되고 있다. 오늘의 금융 인프라는 예외 없이 사람을 전제로 설계되어 있다. 영업시간이 있고, 건마다 본인 확인을 하고, 하루 단위로 모아 정산한다. 그런데 사람이 아니라 AI 에이전트가 거래의 주체가 되면 어떻게 될까? 기계는 밤낮을 가리지 않고, 건당 금액은 몇 원 수준이지만 빈도는 초 단위다. 사람이 매 건을 승인한다는 것은 애초에 성립하지 않는 이야기다.

그 답으로 등장한 것이 x402라는 결제 규약이다. 이름의 유래가 재미있다. 웹의 통신 규약인 HTTP에는 402번 'Payment Required(결제 필요)'라는 상태 코드가 있다. 웹 초창기 설계자들이 훗날 값을 치르는 일이 생길 것을 내다보고 '장래를 위해 예약만 해둔 코드인데, 카드 결제의 최소 단가 탓에 소액 결제가 성립하지 않아 30년 가까이 이렇다 할 쓰임을 얻지 못했다. 2026년 7월, 리눅스 재단이 이 코드를 되살린 x402 재단을 출범시켰다. 회원사 명단이 불만하다. 비자, 마스터카드, 아메리칸익스프레스, 스트라이프, 아마존웹서비스. 카드망과 스테이블코인 진영이 같은 테이블에 앉은 것이다.

규모도 이미 실험 단계를 넘어섰다. 비자의 스테이블코인 결제는 2026년 4월 기준 연 70억 달러 수준이고, 마스터카드는 같은 해 3월 스테이블코인 결제 플랫폼을 최대 18억 달러에 인수하기로 했다. 아마존은 자사 에이전트 서비스에 x402를 넣어 건당 약 200밀리초, 1센트 미만의 비용으로 결제가 끝나게 했다. 사람이 물건을 사는 결제가 아니다. 에이전트가 연산과 데이터와 API 호출의 값을, 카드로는 감당이 안 되는 빈도와 단가로 치르는 것이다.



여기서 눈여겨봐야 할 것은 규칙이 정해지는 자리다. 기계가 돈을 주고받는 방식의 표준이 어느 나라 감독 당국도 아니고 국제기구도 아닌, 카드망과 클라우드와 암호자산 사업자가 함께 앉은 민간 기구에서 만들어 지고 있다. 한국 기업도 카카오페이를 비롯해 세 곳이 이름을 올리기는 했지만, 규약의 방향을 의결하는 프리미어 등급 17개 자리에는 포함되지 않았다. 표준은 한번 굳으면 되돌리기 어렵다. 아직 논의가 열려 있는 지금, 참여의 수준을 한 단계 올려둘 필요가 있다.

규칙의 문제에서 한 걸음 더 들어가면 분배의 문제와 만난다. AI가 만들어내는 가치는 데이터를 제공한 쪽, 모델을 만든 쪽, 업무에 적용한 쪽에 걸쳐 발생하는데, 현재의 정산 구조는 이 사슬을 반영하지 못한다. 기여도에 따라 토큰으로 가치를 배분하자는 토크노믹스도, 자동화에 과세하자는 로봇세도, 결국 같은 질문에 대한 다른 답안이다. 누가 만든 가치를 누가 가져가는가? 이 질문에 답하지 않은 채 속도만 높이면, 그 이익이 어디로 흘러갔는지는 한참 뒤에야 확인하게 될 것이다.

2. 투자 - 옳은 전망과 감당할 수 있는 전망

다시 7월의 헤지펀드로 돌아가 보자. 이 펀드의 몰락을 두고 AI 낙관론의 파산이라고들 하지만, 정확한 진단은 아니다. AGI가 언제 오는지에 대한 이 펀드의 판단은 지금도 틀렸다고 확인된 바 없다. 무너진 것은 전망이 아니라 그 전망을 담은 그릇이었다. 간단한 산수를 해보자. 네 배 레버리지로 산 자산은 가격이 25%만 빠져도 자기자본이 전부 사라진다. 전망이 옳더라도 실현되기까지의 경로에서 그런 조정을 한 번이라도 겪으면 결과는 같다. 옳은 전망과 감당할 수 있는 전망은 전혀 다른 것이다.

이 대목에서 조금 뜻밖의 사실이 있다. 복수의 대형 투자은행이 신용을 공여했지만, 이번에는 손실을 본 곳이 없었다. 증거금을 올려 받고, 포지션을 통째로 다른 헤지펀드에 넘기는 것으로 정리가 끝났다. 2021년 아케고스(Archegos) 사태 이후 프라임브로커들이 다듬어 온 위험관리가 제 몫을 한 셈이다. 아케고스는 TRS (Total Return Swap)를 써서 포지션 자체를 감쌌지만, 이번 펀드의 상장주식 보유는 공시에 드러나 있었다.

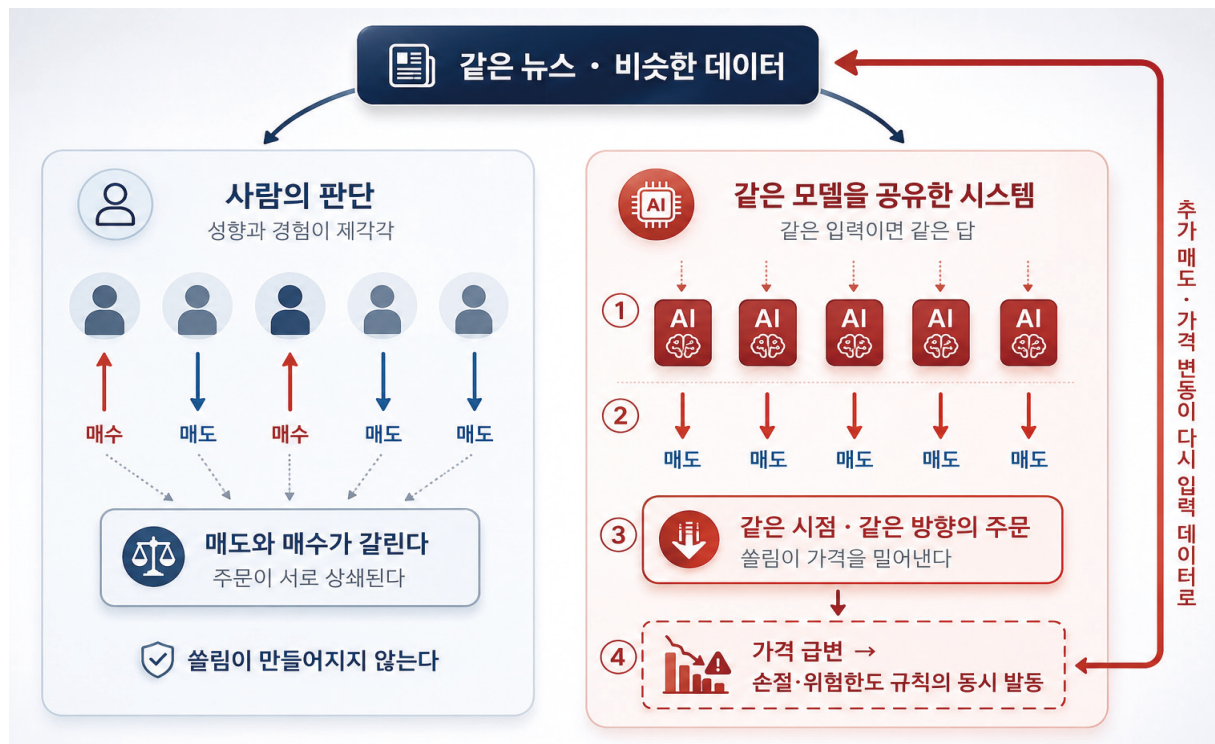
은행은 무사했는데 시장은 흔들렸다. 강제 매각이 진행되는 동안 국내 증시가 이유 없이 밀린 것이 그 증거다. 그리고 정리가 끝난 뒤에도 이 펀드에는 100억 달러 규모의 자산이 남았는데, 대부분이 값을 매기기 어려운 비상장 지분이다. 개별 금융회사의 장부가 안전해지는 것과 시장 전체가 안전해지는 것은 다른 문제다.

같은 시기, 투자은행 안에서 인공지능은 도구를 넘어 인력이 되어가고 있다. JP모건은 내부 생성형 AI 도구를 20만 명이 넘는 직원에게 배포했고, 이 회사의 최고 분석 책임자는 “장시간 자율 에이전트의 시대에 들어섰다”라며 에이전트가 한두 시간씩 스스로 일하고 앞으로는 며칠, 몇 주 단위로 갈 것이라고 했다.

골드만삭스는 자체 AI 어시스턴트를 4만 6천여 명 전 직원에게 열었고, 모건스탠리는 한 발 더 나가, 기업 고객의 AI 에이전트가 자사 주식 보상 관리 플랫폼에 직접 접속해 데이터를 가져가도록 여는 중이다. 사람 보라고 만든 화면을 거치지 않고 기계가 곧장 들어오는 구조다.

그런데 골드만삭스의 그 어시스턴트가 무엇으로 만들어졌는지를 들여다보면 이야기가 달라진다. 오픈AI와 구글, 엔트로픽의 모델 위에 골드만 자체의 보안 계층을 얹은 구조다. 세계에서 가장 앞선 투자은행조차 자기 모델을 갖고 있지 않은 것이다. 남들과 같은 소수의 파운데이션 모델을 쓰면서, 그 위에 자기 규율을 얹고 있을 뿐이다.

[그림 1] 모델 동조화에 따른 시장 쓸림



여기서 조용히 자라나는 문제가 모델 동조화다. 같은 모델에 비슷한 데이터를 넣으면 비슷한 답이 나온다. 평시에는 이것이 오히려 장점처럼 보인다. 판단의 품질이 고르게 유지되니까. 문제는 시장이 한쪽으로 기울기 시작할 때다. 사람의 판단이라면 성향과 경험이 제각각이어서 같은 뉴스에도 매도와 매수가 갈리지만, 같은 모델을 공유하는 시스템들은 같은 시점에 같은 방향으로 움직인다. 금융안정위원회가 시장 상관성과 쓸림의 증폭을 AI의 시스템 리스크로 지목한 것도 이 때문인데, 정작 감독 당국은 관측되는 쓸림 가운데 어디까지가 AI 탓인지 식별할 방법이 마땅치 않다고 토로한다.

이 문제는 우리에게 더 무겁게 다가온다. 골드만조차 남의 모델을 쓴다면, 자체 파운데이션 모델을 감당할 여력이 있는 국내 금융회사는 사실상 없다고 봐야 한다. 후발 시장일수록 같은 모델에 더 깊이 기대게 되고, 그만큼 판단이 얕아갈 여지도 크다. 동조화는 프런티어보다 오히려 우리 쪽에서 먼저 나타날지도 모른다.

여기에 하나가 더 겹친다. AI가 만들어낸 분석이 시장 가격에 반영되는 경로다. 종래의 애널리스트 리포트는 작성자의 이름과 소속, 평판이라는 담보를 달고 시장에 나왔다. 생성형 모델이 만들어낸 요약과 전망에는 그런 담보가 없다. 근거가 주장을 실제로 뒷받침하는지 확인할 길이 없는 채로 정보가 유통되고, 매매로 이어진다. 가격이란 결국 정보를 반영하는 장치인데, 그 정보의 출처를 검증할 수 없다면 가격이 반영하고 있는 것은 대체 무엇일까?

마지막으로 하나만 짚어두자. 요즘 투자라는 말은 두 방향을 동시에 가리킨다. AI로 투자하는 것과, AI에 투자하는 것. 그리고 후자의 규모는 이미 전자를 압도한다. 데이터센터와 반도체, 전력 설비로 흘러드는 자금은 상당 부분 차입에 의존하고 있으며, 국제결제은행(BIS)은 AI 관련 자산의 가치 평가가 실제 기업 이익보다 앞서 나가고 있다고 경고한 바 있다. 앞서 본 헤지펀드는 이 두 방향이 만나는 지점에서 무너졌다. AI를 활용해 AI에 베팅했는데, 그 베팅을 담아낼 그릇이 없었던 것이다.

투자 영역에서 지금 검토해 볼 만한 것은 레버리지와 자동 주문에 대한 통합적인 관측이다. 해외 펀드의 강제 청산이 국내 종목으로 옮겨붙는 경로가 어딘가에서는 관측되어야 하고, AI가 발주한 주문에는 그것을 식별할 표시가 있어야 한다. 그래야 시장이 밀릴 때 그것이 국내 사정 때문인지 밖에서 옮겨온 것인지, 그리고 그 가운데 AI에 기인한 부분이 얼마인지 사후적으로나마 분해해 볼 수 있다.

3. 서비스 - 특화될 수 없는 산업

고객을 마주하는 자리에서 인공지능이 가장 깊이 들어간 영역은 신용평가다. 종래의 신용평가는 금융 거래 이력이라는 좁은 창으로 사람을 들여다보는 일이었다. 대출을 받아본 적도, 카드를 써본 적도 없는 사람은 위험한 사람이 아니라 알 수 없는 사람이다. 그런데 알 수 없다는 이유로 위험한 사람과 같은 취급을 받아왔다. 이른바 싹파일러(thin filer) 문제다.

미국의 온라인 대출 플랫폼 업스타트(Upstart)는 이 창을 넓히는 데 가장 멀리 간 사례로 꼽힌다. 회사 발표에 따르면 2,500개가 넘는 변수를 쓰고, 대출의 91%가 업스타트 쪽 사람 손을 거치지 않고 처리된다. 자사 모형을 가상의 전통적 심사 모형과 비교했더니 같은 손실률에서 44% 더 많은 신청자를 승인할 수 있었다는 것이 이 회사가 내세우는 수치다. 후불결제 회사 클라르나(Klarna)도 1억 명 이상의 고객을 상대로 심사를 1초도 안 되는 사이에 끝낸다. 보이지 않던 사람들이 보이기 시작한 것이다.

그런데 창이 넓어지면 설명해야 할 것도 함께 늘어난다. 2,500개의 변수를 비선형으로 결합한 모형을 붙들고 “이 사람이 왜 거절되었는가”를 물으면 어떤 답이 나올까? 정확도가 높아질수록 설명은 어려워진다는 오래된 긴장이 이제는 규제 형태로 되돌아오고 있고, 각국의 대응은 갈린다. 미국은 신용 거절에 이유를 알릴 의무를 개별법에 두고, 감독 당국이 개별 사업자와 협의해 길을 터주는 방식을 병행해 왔다. 업스타트가 2017년 소비자금융보호국(CFPB)으로부터 받은 노액션 레터(No-Action Letter)가 그 예다. 반면 유럽은 신용평가를 고위험으로 분류해, 시장에 들어가기 전에 요건을 채우게 하는 쪽을 택했다.

[표 2] 신용평가에 대한 각국의 접근

구분	미국	EU	한국
근거	ECOA·FCRA 등 개별법의 거절 사유 고지 의무	AI Act, GDPR 제22조 (자동화된 결정)	인공지능 기본법 및 시행령
신용평가의 지위	별도 분류 없이 기존 소비자보호법제 적용	고위험(high-risk)으로 분류	고영향 영역으로 분류
규제 방식	사후 설명 중심, 감독 당국과의 개별 협의(노액션 레터) 병행	사전 규율, 사람의 개입과 절차적 권리 요구	사전 규율에 가까움, 위험·영향평가와 문서화 의무
대표 사례	업스타트(CFPB 노액션 레터, 2017년 발급·2022년 종료)	레볼루트 (리투아니아 중앙은행 감독 리뷰)	시행 초기 단계

유럽의 방식이 실무에서 어떤 모습이 되는지는 레볼루트(Revolut) 사례가 잘 보여준다. 리투아니아 중앙은행은 이 회사의 신용 모형을 검토하면서 설명가능성과 사람의 개입 통제를 강화하라고 요구했고, 회사는 거절 사유를 변수 기여도로 환산해 제시하고 일정 기준을 넘는 거절 건은 사람이 다시 들여다보는 대기열을 두는 것으로 답했다. 규제가 문서 한 장으로 끝나지 않고 업무 절차의 재설계로 이어진 것이다. 우리의 인공지능 기본법도 대출 심사를 고영향 영역으로 분류해 두었으니, 국내 금융회사들도 머지않아 같은 종류의 재설계를 요구받게 된다.

보험 쪽에서는 성격이 다른 문제가 자라고 있다. 인공지능이 언더라이팅과 보험금 심사를 돕는 것은 이제 새로운 이야기가 아니다. 정작 어려운 것은 AI 자체를 담보로 삼는 상품이다. 자율주행의 판단 오류, 의료 영상 판독의 실수, 투자 자문 알고리즘의 손실. 누가 어디까지 책임지는가? 위험을 인수하려면 그것이 얼마나 자주, 어느 크기로 발생하는지 알아야 하는데, 그럴 데이터가 아직 세상에 없다. 손해율을 모르는 위험에는 값을 매길 수 없고, 값을 매길 수 없는 위험은 거래되지 않는다.

앞서 모건스탠리가 외부 에이전트에 플랫폼을 열고 있다고 했는데, 사실 이것이 서비스 영역에서 일어날 가장 큰 변화의 예고편이다. 고객이 자신의 AI 비서에게 “조건 괜찮은 대출 좀 알아봐 줘”라고 말하고, 그 비서가 여러 금융회사를 비교해 신청까지 마치는 구조가 자리 잡으면 어떻게 될까? 금융회사가 마주하는

상대는 더 이상 사람이 아니라 다른 기계다. 상품 설명 의무는 누구에게 이행해야 하는가? 적합성 원칙은 누구를 기준으로 판단하는가?

이 대목에서 미리 정리해 둘 것이 책임의 소재다. 모델을 만든 회사, 그것을 업무에 쓴 금융회사, 고객을 대신해 거래를 실행하는 에이전트 사업자. 이 셋 사이에서 손해가 났을 때 누가 어디까지 부담하는지 정해지지 않으면, 보험은 값을 매기지 못하고 금융회사는 도입을 주저하게 된다.

돌고 돌아 이야기는 다시 한곳으로 모인다. 조선사의 대출을 심사하려면 수주 잔고를, 정유사의 회사채를 평가하려면 정제 마진을, 한 사람의 신용을 보려면 그가 속한 업종과 지역의 사정까지 읽어야 한다. 금융의 서비스는 예외 없이, 다른 산업과 다른 삶에 대한 이해 위에서만 성립한다.

그렇다면 금융회사의 경쟁력은 어디에서 나오는가? 적어도 모델 자체에서 나오지는 않는다는 것이 필자의 잠정적인 답이다. 골드만이 그랬듯, 남들과 같은 모델을 쓰면서도 그 산출물을 검증하고 관리하는 능력에서 나온다. 어떤 데이터를 붙였는가, 그 판단을 어떻게 확인하는가, 틀렸을 때 무엇이 작동하는가. 특화는 모델에 있지 않고, 그 위에 얹는 검증 체계에 있다.

도입의 속도, 검증의 능력

세 영역을 지나오며 마주친 문제들은 결국 하나로 모인다. 판단은 빨라지는데, 그 판단을 확인할 방법이 따라오지 못하고 있다는 것이다. 영역마다 하나씩 짚어둔 것 외에, 두 가지만 덧붙이고 싶다.

하나는 공통의 기준이다. AI가 생성한 금융정보를 무엇으로 믿을 만하다고 판단할 것인가? 지금의 논의는 대체로 모델의 성능과 개발 과정을 향해 있다. 그러나 시장에 실제로 영향을 미치는 것은 모델이 아니라 그 결과로 나온 산출물이다. 근거를 추적할 수 있는가? 그 근거가 주장을 실제로 뒷받침하는가? 사실과 다른 내용이 어느 정도 섞여 있는가? 이런 항목에 공통의 척도와 제3자 검증 절차가 마련된다면, 리포트든 상품 설명이든 비로소 신뢰의 근거를 갖게 될 것이다.

다른 하나는 검증을 위한 공동의 인프라다. 개별 회사가 각자 검증 체계를 갖추는 것은 대형사에게도 부담이고, 중소형사에게는 사실상 불가능한 일이다. 대학과 금융회사, 감독기관이 함께 쓰는 시험 환경을 두고, 실제 데이터를 대신할 합성 금융 데이터의 활용 기준까지 함께 정리한다면, 개인정보 제약과 데이터 부족이라는 두 병목을 한 번에 다뤄볼 수 있다. 이때 감독기관은 최소한의 공통 기준과 시험 환경을 마련하고, 실제 검증은 금융회사와 대학·전문기관 등 독립된 주체가 함께 수행하는 구조가 현실적일 것이다. 모든 AI 활용을 일률적으로 규율하기보다 소비자나 시장에 미치는 영향이 큰 영역부터 검증 수준을 높여가는 방식도 생각해 볼 수 있다.

오해를 피하기 위해 덧붙이자면, 위의 이야기들은 규제를 하나 더 얹자는 것이 아니다. 오히려 그 반대인가깝다. 확인할 수 있는 것이 늘어날수록 풀어줄 수 있는 여지도 함께 넓어지기 때문이다. 무엇이 일어나고 있는지 모르는 상태에서는 규제를 강화하자는 쪽도 완화하자는 쪽도 근거를 대기 어렵다.

다시 첫 장면으로 돌아가 보자. 십수 년을 버텨온 규제가 모델 하나에 흔들렸을 때, 당국이 내린 결론은 그 모델을 막는 것이 아니라 쓰는 것이었다. 원칙의 후퇴라고 볼 수도 있겠지만, 필자는 오히려 정직한 판단이었다고 본다. 막을 수 없는 것을 막겠다고 버티기보다는, 쓰되 확인할 수 있는 체계를 만드는 편이 낫다. 문제는 쓰기로 한 속도만큼 확인하는 능력이 함께 자라고 있는가에 있다.

금융은 특화될 수 없는 산업이다. 다른 산업이 벌이는 모든 일이 자금과 가격이라는 형태로 이 산업에 모여들기 때문이다. 그래서 금융에 인공지능을 들이는 일은 신기술 하나를 도입하는 일이 아니라, 세상 전체를 읽어내는 장치를 사회의 자금 배분 기구에 엮는 일에 가깝다. 그 장치가 무엇을 근거로 판단했는지 물을 수 없다면, 우리는 이유도 모르는 채로 자금이 흘러가는 사회를 갖게 된다.

경쟁은 이미 도입의 속도에서 검증의 능력으로 옮겨가고 있다. 남들보다 먼저 쓰는 회사가 아니라, 남들과 같은 것을 쓰면서도 그 결과를 확인할 수 있는 회사가 결국 앞선다. 이번 국면이 우리 금융이 그 능력을 키워가는 계기가 되기를 바란다.

참고문헌

- [1] Anthropic Frontier Red Team, "Assessing Claude Mythos Preview's cybersecurity capabilities," 2026. 4. 7. <https://www.anthropic.com/research/mythos-preview>
- [2] AI Security Institute(UK), "Our evaluation of Claude Mythos Preview's cyber capabilities," 2026.
- [3] Centre for Emerging Technology and Security(CETaS), "Claude Mythos: What Does Anthropic's New Model Mean for the Future of Cybersecurity?," 2026.
- [4] 금융위원회, 망분리 규제 완화 및 금융보안 대응 관련 보도자료, 2026. 5.
- [5] 미 재무장관·연준 의장의 대형은행 CEO 긴급 소집 및 Project Glasswing 관련 보도(CNBC, Bloomberg 등), 2026. 4.
- [6] 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 및 동법 시행령, 2026. 1. 시행.
- [7] Financial Stability Board, 인공지능과 금융안정 관련 보고서, 2024; 책임 있는 AI 도입을 위한 모범관행 협의보고서, 2026. 6.
- [8] Bank for International Settlements, Annual Economic Report 중 인공지능 관련 장, 2026.
- [9] Linux Foundation, x402 재단 출범 관련 발표 및 관련 보도, 2026. 7.
- [10] Visa·Mastercard·AWS의 스테이블코인 결제 및 에이전트 결제 관련 발표와 보도, 2026.
- [11] JPMorgan Chase의 AI 에이전트 도입 관련 보도(CNBC 등), 2026. 6.
- [12] Goldman Sachs GS AI Assistant 관련 보도, 2025~2026; CNBC, Morgan Stanley의 외부 AI 에이전트 접속 개방 보도, 2026. 6. 3.
- [13] Upstart 및 Klarna의 신용평가 모형 관련 공시·발표 자료; CFPB No-Action Letter(2017).
- [14] Bank of Lithuania의 Revolut 신용모형 감독 검토 관련 자료, 2024.
- [15] 시추에이션얼 어웨어니스 마진콜 관련 보도, 2026. 7.

신뢰할 수 있는 AI: 글로벌 현황과 향후 정책적 과제

- 김종원 | 슈어소프트테크 AIV Lab 팀장 | bk3206@suresofttech.com
- 김영현 | 슈어소프트테크 AX 센터 팀장 | yhkim1@suresofttech.com
- 정재룡 | 슈어소프트테크 AX 센터 실장 | jjjung@suresofttech.com

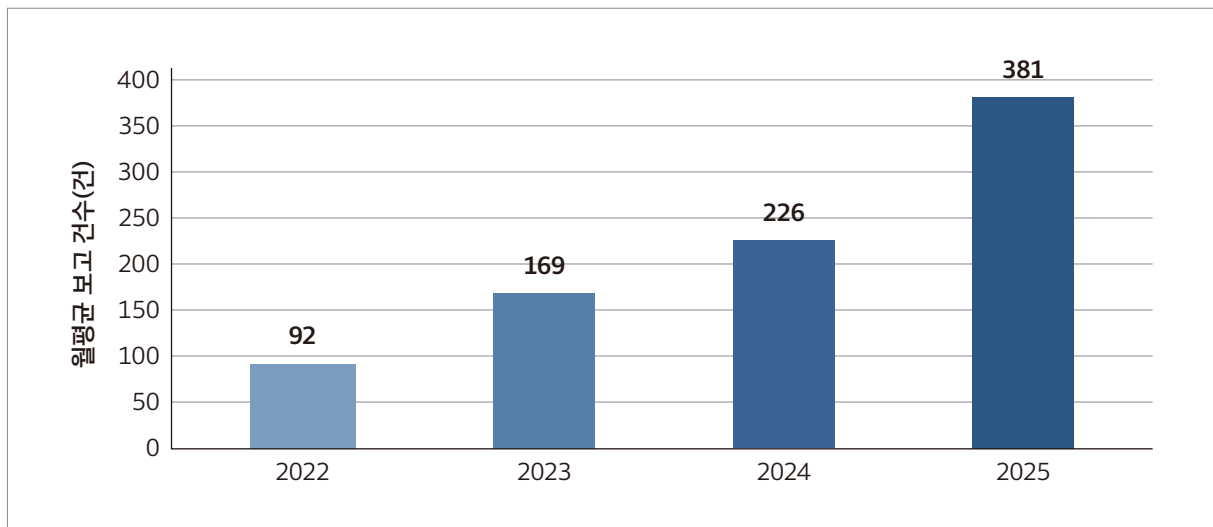


서론: “AI 중심 사회가 마주한 새로운 도전: 신뢰할 수 있는 AI”

인공지능(Artificial Intelligence, AI)은 단순한 기술적 혁신을 넘어 현대 사회의 기반 인프라로 안착하고 있다. OpenAI가 개발한 ChatGPT의 주간 활성 사용자 수(WAU)가 9억 명을 돌파한 사례에서 보듯^[1], AI는 이제 개인의 생산성 도구나 기업의 사무 보조를 넘어 의료기기, 자율주행, 로봇틱스 등 물리 시스템과 국가 핵심 인프라에도 깊숙이 이식되고 있다. 이는 마치 전력망이나 통신 인프라처럼 사회 시스템 전체가 AI의 정상 작동을 전제로 작동하는 시대에 도달했음을 의미한다.

그러나 이러한 기술적 확산 속도가 AI의 ‘안전성’을 담보하는 것은 아니다. OECD AI 사고 모니터(AIM)¹의 집계에 의하면, AI 관련 사고는 2022년 월평균 92건에서 2025년 381건으로 3년 만에 약 4.1배 증가했다.^[2]

[그림 1] OECD AI 사고 모니터의 최근 4년간 월평균 AI 관련 사고 건수



AI 시스템의 불확실성은 기업과 사회에 실질적이고 막대한 손실을 야기할 수 있다. 2023년 구글의 챗봇 Bard의 오답 논란으로² 모기업 알파벳의 시가총액 1,000억 달러가 하루 만에 증발했고^[3], 2024년에는

¹ OECD AI 사고 모니터(AIM, AI Incidents and Hazards Monitor)는 전 세계 언론 보도를 자동으로 수집·분석해 AI 관련 사안을 집계하는 OECD의 모니터링 체계이다.

² 2023년 2월 구글이 공개한 Bard 시연에서 챗봇이 “제임스 웹 우주망원경이 태양계 밖 행성을 최초로 촬영했다”고 답했으나, 외계행성 최초 촬영은 2004년 유럽남방천문대의 초대형망원경(VLT)이 수행한 것이었다. 이 여파로 알파벳 주가는 하루 만에 7.6% 하락했다.

딥페이크 기술을 악용한 화상회의 사기로³ 2,500만 달러가 유출됐다.^[4] 물리 시스템 영역인 자율주행에서는 인명 피해까지 발생했는데, 테슬라의 오토파일럿과 자율주행(FSD, Full Self-Driving) 기능 관련 사망자는 2016년 첫 사망사고 이후 2025년 10월까지 65명에 이르고 있다.^{4[5][6]}

이러한 사고들은 일시적/우발적 결함이 아닌, AI 기술 특성에 기인한 구조적 위험이라는 사실에 주목할 필요가 있다. 데이터 기반의 확률적 추론에 의존하는 AI 알고리즘의 특성상, 특정 출력에 대한 원인을 명확히 설명하기 어려울 뿐 아니라 오류를 수정했다고 해서 오류가 완벽히 제거되었는지 증명하는 것 역시 까다롭다.

이런 한계를 극복하기 위해 국제 표준화 기구는 '신뢰할 수 있는 AI(Trustworthy AI)'라는 개념을 제시했다. ISO/IEC TR 24028:2020에서는 신뢰가능성(Trustworthiness)을 “이해관계자의 기대를 검증 가능한 방식으로 충족하는 능력”으로 정의한다.^[7] EU AI Act와 한국 AI 기본법 등 주요 국가의 법제는 고위험·고영향 AI에 대해 안전성 확보를 법적 의무로 명시하여 강력한 증명을 요구하고 있다.^{[8][9]}

그러나, 법적·제도적 규제 요구와 이를 입증할 객관적 검증 수단 사이에는 여전히 격차가 존재한다. 법적 당위성은 마련되었으나, 과학적 검증 도구가 부재한 이 간극이 오늘날 AI 중심 사회가 마주한 본질적 도전이다.

글로벌 AI 신뢰성 및 안전 기술 동향

앞서 언급한 ISO/IEC TR 24028 외에도, 미국 NIST의 AI Risk Management Framework, EU의 Ethics Guidelines for Trustworthy AI 등은 안전성, 강건성, 보안성, 투명성, 책임성 등 AI 신뢰성을 구성하는 다양한 요소를 제시하고 있다.^{[10][11]} 이러한 요소를 검증의 관점에서 재구성하면 기능성 검증, 강건성 검증, 보안성 검증, 운영·관리 영역으로 구분할 수 있는데, 운영·관리 영역은 시스템의 구성·배포·모니터링 등 전통적인 소프트웨어 검증에서도 다뤄지는 공통 영역이므로, 본문에서는 AI 모델 자체의 동작과 성능을 직접 시험하는 기능성·강건성·보안성 검증(EU AI Act 제15조 대응 영역)⁵에 초점을 맞춘다.^[8]

³ 2024년 1월 영국 엔지니어링 기업 아룹(Arup)의 홍콩 지사 직원이 본사 최고재무책임자(CFO)와 동료들로 위장한 화상회의에 참여한 뒤 홍콩 소재 5개 계좌로 15차례에 걸쳐 송금한 사건이다. 회의에 등장한 얼굴과 음성은 모두 공개된 임원 영상·음성으로 만든 딥페이크였다.

⁴ Tesla Deaths는 언론에 보도된 테슬라 관련 사망사고를 집계하는 민간 데이터베이스로, 2025년 10월 기준 전체 772건 가운데 오토파일럿이 관여한 사망자를 65명으로 집계했다. 이 중 2건은 미국 도로교통안전국(NHTSA) 결함조사국이 2022년 이후 FSD 작동 중 발생한 것으로 판정한 사례다.

⁵ EU 인공지능법 제15조의 표제는 “Accuracy, robustness and cybersecurity”로, 고위험 AI 시스템이 정확성·강건성·사이버보안을 적절한 수준으로 달성하고 생애주기 전반에 걸쳐 일관되게 작동하도록 설계·개발할 것을 요구한다. 이 세 요건이 본문의 기능성·강건성·보안성 검증에 각각 대응한다.

1) 기능성 검증

기능성 검증은 AI 모델이 의도한 기능을 명확히 수행하는지 확인하기 위해 시험 요구사항을 충족하는 데이터셋을 구성하는 것이 핵심이다. 과거에는 평균/표준편차 등 통계적 특성을 활용하거나 차원 축소 및 클러스터링 기반으로 분류하여, 데이터셋 전체의 통합 결과를 확인하는 방식이 주로 사용되었다.^[12] 이러한 방식은 분류 기준은 정량적일지라도 각 데이터 조건이 실제 어떤 기능적 요구사항에 대응하는지 추적 및 해석하기 어렵다는 한계가 있었다.

최근에는 입력 데이터로부터 사람이 이해할 수 있는 정보를 추출하여 시험 조건을 세분화한다(예: 자율주행 도메인 기준 운전 방향, 도로 유형, 날씨, 계절 등).^[13] 이를 조합해 테스트 케이스를 구성함으로써 모델의 전체 성능뿐 아니라 각 요구사항별 준수 여부까지 확인할 수 있다. 즉 기능성 검증은 암묵적 기준에 의존해 통합 성능만 측정하던 수준에서 벗어나, 운용 조건별로 판정 기준을 명시하고 요구사항 단위로 충족 여부를 추적하는, 보다 정밀하고 정량적인 검증으로 발전하고 있는 것이다.

2) 강건성 검증

강건성 검증은 AI 모델이 어떠한 악의적 혹은 비정상적 입력에 대해서도 의도한 결과를 정확히 낼 수 있는지를 검증하는 것이다. 이를 위해 AI 모델의 입력 데이터셋에 노이즈나 현실적인 변형이 가해진 상황에서도 AI 모델의 출력이 원본 입력의 결과와 일관성이 유지되는지, 성능 요구사항이 허용 기준 이하로 저하되지 않는지를 확인하는 검증을 수행한다. 강건성 검증은 무작위 변형을 반복 적용하는 랜덤 테스트 기술을 시작으로, 오류 유형의 데이터를 변형·조합하는 유전 알고리즘을 거쳐, 모델 내부의 기울기(Gradient) 정보를 활용하는 기울기 기반(Gradient-based) 테스트로 발전하면서 AI 모델의 강건함을 효율적으로 확인할 수 있게 되었다.^{[14][15][16]} 그러나, 이러한 테스트들을 통해 발견된 오류가 없다고 하더라도 모델이 강건하다고 단정하기는 어렵다.

이를 보완하기 위해 정해진 입력의 변형 범위 내에서 오류가 발생하지 않음을 수학적으로 증명하는 형식 검증⁶ 연구가 활발하게 진행되고 있다.^{[17][18][19]} 즉, 강건성 검증은 단순한 오류 사례의 효율적 탐색을 넘어, AI가 어느 범위까지 정상 동작을 보증할 수 있는지 정량적으로 제시하는 방향으로 확장되고 있다.

⁶ 형식 검증(Formal Verification)은 정해진 입력 변형 범위 전체에 대해 오류가 발생하지 않음을 수학적으로 증명하는 방법이다. 통계적 분석이 성능 유지를 '보여주는' 데 그치는 반면 형식적 방법은 이를 '증명'할 수 있으나, 입력 차원과 신경망 규모가 커질수록 계산량이 급증하는 확장성 한계가 있다.

3) 보안성 검증

보안성 검증은 공격자가 AI의 추론 메커니즘과 입력 특성을 악용하더라도 AI가 안전하게 동작하는지를 확인한다. 초기에는 보안 전문가의 경험과 직관에 의존한 레드티밍⁷이 중심이었으나, 전문가의 역량과 선택된 공격방식에 따라 결과가 달라지는 한계가 있었다.

최근에는 NIST의 적대적 공격 분류와 MITRE ATLAS 같은 공통 위험 프레임워크를 적용하여 평가 대상을 체계화하고^{[20][21]}, 적대적 공격 입력을 자동 생성해 대규모로 반복 시험하는 기술이 도입되고 있다.^{[22][23]} 아울러 공격 성공률(Attack Success Rate)⁸ 및 안전성 등급 등으로 결과를 수치화하려는 시도도 확대되고 있다.^{[24][25]} 즉 보안성 검증도 정성적 취약점 점검에서 시험 범위와 실패 수준을 명시적으로 측정하는 구조로 전환되고 있다.

3가지 검증의 방법과 대상은 다르지만 발전의 지향점은 동일하다. AI 검증은 정성적·단편적 검증에서 벗어나 시험 조건과 판정 기준을 명확히 하고, 검증 결과를 수치 데이터와 Pass/Fail 기준으로 제시하는 정량적·객관적 검증으로 진화하고 있다. 한편, 산업계에서는 Cisco의 Cisco AI Defense, Patronus AI의 Patronus Platform, 슈어소프트테크의 VERIFAI 시리즈 등 AI 모델과 시스템을 검증하는 도구들이 속속 출시되면서 AI 테스트를 자동화하여 효율적인 검증이 가능해지고 있다.^{[26][27][28]}

그러나 여전히 근본적인 질문은 남아 있다. “어떤 조건까지 시험해야 충분한지”, “어느 수준의 강건성과 보안성을 확보해야 하는지”에 대한 공통된 합격 기준은 아직 확립되지 않았다. 기술은 “측정할 수 있는 단계”로 발전했지만, “어느 정도면 충분한가”에는 정답에 다다르지 못한 것이다. 다음 장에서는 이 미완의 과제가 주요 국가의 정책과 국내 제도에서 어떻게 논의되고 있는지 살펴본다.



⁷ 레드티밍(Red-teaming)은 공격자의 관점에서 시스템의 결함과 취약점을 의도적으로 찾아내는 구조화된 시험 활동을 말한다. AI 분야에서는 모델이 내놓지 말아야 할 답변을 하도록 유도하는 입력을 탐색하는 활동을 가리킨다.

⁸ 공격 성공률(ASR, Attack Success Rate)은 시도한 공격 가운데 방어를 우회해 목표한 결과를 얻어낸 비율이다.

글로벌 AI 신뢰성 및 안전 규제 동향

1) 글로벌 AI 규제화

앞서 살펴본 바와 같이 AI 검증 기술은 정량적/객관적 단계로 진화하고 있으나, '어느 수준을 안전한 수준으로 볼 것인가'에 대한 글로벌 공통 기준은 아직 확립되지 않았다. 그러나 AI의 적용 범위가 물리 시스템과 핵심 인프라로 급속히 확대되면서, 각국 정부는 검증 기술의 성숙을 기다리지 않고, 규제 체계를 먼저 마련하기 시작했다.

이 흐름이 국제사회에서 본격적으로 공론화된 계기가 2023년 '블레츨리 선언(Bletchley Declaration)'이다.⁹[29] 기존 AI 모델보다 훨씬 강력한 성능을 보유한 AI의 위험에 공동 대응하자는 선언에서 출발한 AI 안전 논의는 이후 각국의 법률, 행정명령, 가이드라인 등으로 구체화되었다. 다만, 구체적 방식은 각국의 제도적 여건과 규제 전략에 따라 서로 다른 형태로 발전하고 있다.

2) 주요국 규제 현황

EU, 미국, 중국, 일본 등 주요 국가들의 규제와 주요 내용, 이행 수단을 정리하면 다음 표와 같다.

[표 1] 글로벌 AI 신뢰성 관련 규제

국가	규제 및 법령	주요 내용	이행 수단
EU	인공지능법(AI Act) ^[8]	위험도에 따라 4등급으로 분류, 고위험 AI 시스템은 출시 전 적합성 평가 및 CE 마킹 의무화	위반 시 글로벌 매출의 최대 7% 과징금 부과
중국	생성형 AI 서비스 잠정관리방법 ^[30]	여론 형성과 관련된 AI 출시 전 안전성 평가 및 알고리즘 신고	신고 미이행 시 서비스 출시 불가
인도	정보기술 규칙 개정안 ^[31]	AI로 생성한 정보에 라벨 및 출처에 대한 메타데이터 포함 의무화	플랫폼(중개자)에게도 책임 부과
영국	AI 안전연구소의 자발적 협약 ^[32]	최첨단 AI 모델의 위험성에 대해 배포 전 평가 수행	AI 안전연구소와 협업을 통한 AI 모델 평가 진행
미국	AI 표준혁신센터(CAISI) 및 행정명령 ^[33]	주요 AI 개발사와 모델 접근 및 배포 전 안전성 평가 협약 체결	AI 표준혁신센터의 표준 프레임워크에 대해 실증 및 평가
일본	AI 추진법 및 공공조달 가이드라인 ^{[34][35]}	AI 시스템의 안전성/신뢰성 확보를 공공조달 입찰 조건으로 연계	공공 조달 지침을 통한 품질 확보 유도
싱가포르	AI 어슈어런스 파일럿 ^{[36][52]}	제3자 시험기관에 의한 Application 검증	시장에 의해서 AI 신뢰성 확보 유도

⁹ 2023년 11월 1~2일 영국 블레츨리 파크에서 열린 제1차 AI 안전성 정상회의에서 한국·미국·중국을 포함한 28개국과 EU가 서명했다. 광범위한 작업을 수행할 수 있는 고성능 범용 AI를 '프론티어 AI'로 규정하고 그 위험에 국제적으로 공동 대응할 것을 합의했다.

국가별 제도화 과정은 ‘측정 기술의 성숙도’와 ‘규제 현실’ 간의 밀접한 관계가 있음을 보여준다. EU는 AI Act를 구체화한 표준과 AI 적합성 평가 체계의 준비가 지연되어 고위험 AI에 대한 의무 적용 시점을 2027년 12월 이후로 연기한 바 있다.¹⁰[37][38] 인도도 정보기술 규칙을 개정하는 과정에서 AI가 생성한 ‘시각 콘텐츠 표시 면적 10% 이상’이라는 정량적 표시 기준을 제시했으나, 최종안에서는 해당 수치 기준을 삭제하고 “명확하고 두드러진 라벨링”이라는 원칙 중심의 기준으로 조정했다.¹¹[31] 반면, 영국 AI 안전연구소는 최첨단 AI 모델을 측정하고 평가하는 평가 프레임워크를 빠르게 수립하여 공개함으로써^[39] AI 개발사들은 협약에 맞춰 배포 전에 평가하고 있다.

3) AI 규제 유형과 4대 공통 특징

각국의 대응은 규제 대상(위험등급 기반 vs 행위 기반)과 이행 수단(제재형 vs 참여 유도형)에 따라 다음 표처럼 4가지 유형으로 분류된다.

[표 2] AI 규제 유형

	위험등급 기반	행위 기반
제재형	등급규제형 — EU - AI 위험 등급을 판정해 의무 부과 - 위반 시 과징금 및 과태료 부과	행위규제형 — 중국, 인도 - 신고·표시·삭제 등 특정 행위를 직접 명령 - 위반 시 출시 금지, 법적 책임 부과
참여 유도형	역량평가형 — 영국, 미국 - 국가 연구소 중심의 AI 모델 평가 프레임워크를 공개 - 기업들은 이를 통해 자발적 평가 및 배포	절차이행형 — 일본, 싱가포르 - AI 개발/검증 절차 준수를 요구 - 공공 조달 및 시장에 의한 평가를 유도

이처럼 각국은 규제 대상과 이행 수단을 달리하고 있지만, 그 가운데에서도 다음과 같은 공통점이 나타난다.

- ① 사전 검증 체계로의 전환: AI 신뢰성을 확인하는 절차가 시장에 출시되기 전에 진행되고 있다. 적합성 평가, 사전 신고, 표시 의무, 배포 전 시험은 모두 시장 출시에 앞서 안전을 확인하는 절차다.

¹⁰ 유럽위원회는 2025년 11월 ‘Digital Omnibus on AI’를 통해 고위험 AI 관련 의무의 적용 시점을 조화표준(harmonised standards)의 준비 상황에 연동해 늦추는 개정안을 제안했고, 이사회와 유럽의회는 2026년 이에 합의했다.

¹¹ 인도 전자정보기술부(MeitY)가 회람한 개정 초안은 AI 생성 시각물에 표시 면적의 10% 이상을 차지하는 라벨을, 음성물에는 앞부분 10% 구간에 걸친 고지를 요구했다. 최종 규칙에서는 이 정량 기준이 삭제되고 ‘명확하고 두드러지게(clearly and prominently) 표시’라는 원칙 기준으로 대체되었다.

- ② 측정 가능성과 규제 정밀성: 규제 기준을 구체화하기 위해서는 이를 객관적으로 측정할 수 있는 기술적 기반과 산업 현장에서의 적용 가능성이 함께 뒷받침되어야 한다. EU와 인도의 사례는 정량적이고 세부적인 규제 기준을 마련하기 위해 측정 기술의 성숙도뿐만 아니라 실제 이행 가능성까지 고려해야 함을 보여준다.
- ③ 정부의 평가 기준·프레임워크 제시: 정부는 AI 검증 및 측정 역량을 확보하고, 이를 민간이 활용할 수 있는 평가 체계로 제공하고 있다. 영국과 미국은 국가 연구기관을 중심으로 평가 프레임워크를 구축·공개하고, 이를 기반으로 민간의 자발적인 평가와 협력을 유도하고 있다.
- ④ AI 검증 체계의 산업화: 싱가포르의 사례처럼 시험 방법의 표준화, 시험기관 인정 체계 구축 등 AI 검증 자체를 하나의 신산업 영역으로 육성하려는 정책적 시도가 구체화되고 있다.

국내 AI 신뢰성 및 안전 규제의 현황 및 글로벌 규제와의 차이점

1) 국내 AI 기본법 체계의 현황과 특성

2026년 1월 22일 시행된 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 AI 기본법)은 고영향 AI라는 위험 등급 체계¹²⁾를 도입하여 판정하고 이에 따라 법적 의무를 부과하며, 의무 위반이나 시정명령 불이행 시 과태료를 부과하는 구조를 취하고 있다.^{[9][40]} 이는 앞서 3장에서 분류한 유형 중 EU와 동일한 ‘등급규제형’에 해당한다. AI 기본법은 생성형 AI 결과물의 고지·표시 의무, 고영향 AI의 위험관리와 설명 가능성 확보, 일정 연산량 이상의 고성능 AI¹³⁾에 대한 안전성 확보 체계 구축 등을 기업의 법적 책임으로 명시하고 있다. 정부도 시행에 맞춰 투명성·안전성·고영향 판단·사업자 책무·영향평가 등 5대 가이드 라인을 제시하고^{[43][44][45][46][47]} 최소 1년의 계도 기간을 부여했다.^{[41][42]}

이로써 한국은 AI 안전성 확보를 위한 법적 체계를 마련했으나, AI 산업 현장에서 이를 이행하기 위한 구체적인 기준은 아직 충분치 마련되지 않았다. 고영향 AI의 위험등급을 객관적으로 판정할 기술과 AI 평가 프레임워크, 정량적 기준 등이 구체화되지 않은 상태에서 법적 의무가 먼저 발효됐기 때문이다.

¹²⁾ AI 기본법상 고영향 인공지능은 사람의 생명·신체 안전 및 기본권에 중대한 영향을 미칠 우려가 있는 영역에 사용되는 AI를 말하며, 에너지·보건의료·원자력·범죄수사·채용·대출심사·교통·공공서비스·교육 등이 해당 영역으로 열거되어 있다.

¹³⁾ 학습에 사용된 누적 연산량이 10의 26승 부동소수점연산(FLOP) 이상인 인공지능을 가리킨다. 해당 사업자는 인공지능 수명주기 전반의 위험을 식별·평가·완화하고 위험관리체계 구축 결과를 과학기술정보통신부장관에게 제출해야 한다.

2) 글로벌 기준과 국내 현황 간 차이점

3장에서 도출한 글로벌 공통점과 국내 AI 기본법을 비교/분석하면 다음 표와 같은 차이가 있다.

[표 3] 글로벌 AI 규제 공통점과 국내 AI 규제 현황 간 비교 분석

글로벌 AI 규제 공통점	국내 AI 규제 현황	구조적 한계 및 차이점 분석
사전 검증 체계로의 전환	AI 기본법을 제정하여 고영향 AI 판별과 안전성 확보를 법적 의무로 규정	산업 현장 적용이 가능한 수준의 세부 검증 표준 및 평가 가이드라인 미비
측정 가능성과 규제의 정밀성	AI 등급 판정 요건은 가이드라인에 제시되고 실질적 요건은 고시에 위임	AI 등급, AI 안전도 등 객관화, 수치화할 수 있는 정량 지표 부재
정부의 평가 기준·프레임워크 제시	AI 안전연구소 중심으로 42종의 AI 모델 평가 및 한국어 벤치마크를 구축했으나, 평가 도구 및 세부 방법론은 미공개 ^{[48][49]}	민간 기업이 참조할 공개 레퍼런스·프로토콜 미흡
AI 검증 체계의 산업화	CAT, AI-MASTER 등 민간 검증체계 운영 중 ^{[50][51]} * CAT: TTA에서 진행하는 AI 신뢰성 인증 제도 * AI-MASTER: 한국인공지능산업협회, 슈어소프트테크에서 주도하는 AI 신뢰성 인증 제도	국제 상호인정 제도와 AI 특화 시험기관 인정 체계는 부재

종합하면, 한국은 AI 기본법 제정을 통해 사전 검증 체계 도입 및 정부 중심의 평가 인프라 구축 등 글로벌 규제 트렌드의 제도적 틀을 발 빠르게 확보했으나, 이를 산업 현장에 실효성 있게 적용하는 부분에 있어서는 ‘기술적·구조적 이행 격차(Implementation Gap)’가 존재한다. 또한, 고영향 AI 판정 및 안전성 수준을 객관적으로 입증할 정량적 측정 지표와 세부 검증 표준이 부재하여 기업의 법적 불확실성을 높이고 있다. 국가 차원에서 구축한 평가 도구와 방법론이 민간 기업의 개발 실무에 즉시 활용될 수 있도록 배포되고 있지 않다. 아울러 CAT, AI-MASTER 등 민간 중심의 신뢰성 인증제도가 운영 중임에도 불구하고, 글로벌 상호인정 제도 및 AI 특화 시험기관 인정 체계의 미비로 인해 국내 검증 결과가 국제적으로 통용되고, 검증 산업이 확대되는 데 한계를 나타내고 있다.

신뢰할 수 있는 AI를 위한 3대 정책 제언

앞서 확인한 글로벌 기준과 국내 상황을 고려했을 때, AI 신뢰성 구축을 위해 다음 세 가지 정책 방향이 필요하다.

- ① AI 신뢰성에 대해 산업 현장 중심의 정량적 세부 검증 표준 및 정밀 측정 지표를 신속하게 확립해야 한다. AI 기본법 제정으로 안전성 확보를 위한 법적 의무는 마련되었으나, 이를 산업 현장에서 이행하기 위한 정량적 판정 기준과 세부 검증 방법은 충분히 구체화되지 않았다. 정부는 주요 분야별(의료, 자율주행, 금융, 제조 등) 운용 특성을 반영한 정량적 세부 검증 기준(적대적 공격 성공률 허용치, 워터마크 존속성 등)을 하위 고시에 명확히 규정해야 한다. 또한 무엇을, 어떤 절차로, 어떤 판정 기준에 따라 측정할 것인지를 규정한 표준 시험 방법을 제시하여 산업 현장에 명확한 이행 기준을 제공해야 한다. 다만, 이러한 기준이 중소/스타트업에게 과도한 진입장벽이 되지 않도록 위험등급과 기업 규모에 따라 검증 수준을 차등화하고, 공용평가 데이터셋과 검증 바우처 등 완충 장치를 함께 마련해야 한다.
- ② 정부는 직접 AI 평가 도구와 데이터셋, 방법론을 개발해 보급하는 공공 주도 방식보다 시장이 준수해야 할 명확한 안전 기준과 도메인별 평가 가이드라인을 제시하는 선제적 역할에 집중해야 한다. 한편, 해당 기준을 입증할 실질적인 검증 도구, 자동화 평가 솔루션(레드티밍, 형식 검증 등), 그리고 자율주행·로보틱스 등 물리 시스템을 위한 시뮬레이션과 실증 환경의 개발 및 확산은 민간의 전문성과 기술 생태계가 주도하도록 유도해야 한다. 정부는 민간 중심의 검증 기술 생태계가 자율적으로 형성되고 활성화될 수 있도록 연구개발(R&D) 지원 및 인프라 연계 등 시장 친화적 환경을 조성하는 데 중점을 두어야 한다. 특히, 개별기업이 자체적으로 구축하기 어려운 공개 벤치마크와 평가 데이터셋, 물리 시스템 실증을 위한 공동 테스트베드는 공공이 구축하되, 운영과 고도화는 시험기관 및 산업계가 참여하는 개방형 생태계에 맡길 수 있다.
- ③ AI 인증/검증 제도가 국제적으로 통용될 수 있게 해야 하며, 공공조달 연계를 통해 'AI 검증 산업'을 육성할 필요가 있다. AI에 특화된 시험기관의 자격과 검증 품질을 관리하는 인정 체계를 구축하고, 국내 시험 결과가 해외 적합성 평가에서도 인정되도록 상호인정 제도를 확보해야 한다. 동시에 공공조달과 국가 AI 사업 평가에 안전성 검증을 결합해 AI 검증에 대한 시장 수요를 만들어야 한다. 이를 위한 단계적 이행 과제와 추진 주체를 정리하면 다음 표와 같다.

[표 4] AI 검증 인프라 구축 단계별 로드맵

단계	주요 과제	주관·협력
1단계 (계도 기간 내)	분야별 정량 검증 기준 및 표준 시험 방법 하위 고시 제정 AI 특화 시험기관 인정 기준 마련	과학기술정보통신부 국가기술표준원
2단계 (제도 정착기)	인정 체계 운영 및 시험기관 지정 공공조달 입찰 요건에 안전성 검증 결과 반영	국가기술표준원·KOLAS 조달청, TTA·한국인공지능산업협회 등 인증기관
3단계 (국제 연계기)	국내 시험 결과의 해외 적합성 평가 인정 확보 EU·싱가포르 등과 상호인정 협력 추진	과학기술정보통신부, KOLAS 한국AI안전연구소

법적 의무와 기술적 기준 사이의 격차는 상호 조정할 필요가 있다. EU가 AI 의무의 적용 시점을 표준 준비 상황에 연동한 것처럼 법적 의무 이행의 시기를 조절하거나, 규제 샌드박스를 운영하여 부분적으로 적용하는 방법 등 산업계와 정부가 서로 소통해야 한다.

마지막으로 위의 세 가지 정책 제언은 서로 밀접한 관계가 있다. 기준이 없으면 무엇을 측정할지 정할 수 없고, 측정 수단이 없으면 기준은 문서에 머물며, 국제적으로 통용되지 않으면, 결과는 국내로 한정된다. 지금 필요한 것은 규제의 강도를 높이는 일이 아니라 안전을 판정하고 그 판정을 통용 가능한 증명으로 만드는 기반을 갖추는 일이다. 그 기반이 갖춰질 때 AI 안전은 혁신을 늦추는 제약이 아니라, 우리 AI가 더 빠르고 더 넓은 시장으로 나아갈 수 있게 하는 조건이 될 것이다.

참고문헌

- [1] TechCrunch. (2026). *ChatGPT reaches 900 million weekly active users*. <https://techcrunch.com/2026/02/27/chatgpt-reaches-900m-weekly-active-users/>
- [2] Organisation for Economic Co-operation and Development. (2026). *OECD AI Incidents and Hazards Monitor (AIM)*. <https://oecd.ai/en/incidents>
- [3] CNN Business. (2023). *Google shares lose \$100 billion after company's AI chatbot makes an error during demo*. <https://www.cnn.com/2023/02/08/tech/google-ai-bard-demo-error>
- [4] CFO Dive. (2024). *Scammers siphon \$25M from engineering firm Arup via AI deepfake 'CFO'*. <https://www.cfodive.com/news/scammers-siphon-25m-engineering-firm-arup-deepfake-cfo-ai/716501/>
- [5] Tesla Deaths. (2025). *Digital record of Tesla crashes resulting in death*. <https://www.tesladeaths.com/>
- [6] National Highway Traffic Safety Administration. (2024). *Preliminary evaluation PE24031: Tesla Full Self-Driving (FSD)*. U.S. Department of Transportation.
- [7] International Organization for Standardization & International Electrotechnical Commission. (2020). *ISO/IEC TR 24028: 2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*.
- [8] European Parliament & Council. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. Official Journal of the European Union.
- [9] *인공지능 발전과 신뢰 기반 조성 등에 관한 기본법*, 법률 제20676호 (2025).
- [10] National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce.
- [11] High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019). *Ethics guidelines for trustworthy AI*. European Commission.
- [12] Eyuboglu, S., Varma, M., Saab, K., et al. (2022). *Domino: Discovering systematic errors with cross-modal embeddings*. International Conference on Learning Representations (ICLR).

- [13] Chen, M., et al. (2025). *HiBug2: Efficient and interpretable error slice discovery for comprehensive model debugging*. arXiv: 2501.16751.
- [14] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. International Conference on Learning Representations (ICLR).
- [15] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. International Conference on Learning Representations (ICLR).
- [16] International Organization for Standardization & International Electrotechnical Commission. (2021). *ISO/IEC TR 24029-1:2021 Artificial intelligence — Assessment of the robustness of neural networks — Part 1: Overview*.
- [17] International Organization for Standardization & International Electrotechnical Commission. (2023). *ISO/IEC 24029-2:2023 Artificial intelligence — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods*.
- [18] Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). *Certified adversarial robustness via randomized smoothing*. International Conference on Machine Learning (ICML).
- [19] Kaulen, T., et al. (2025). *The 6th International Verification of Neural Networks Competition (VNN-COMP 2025): Summary and results*.
- [20] National Institute of Standards and Technology. (2025). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2)*. U.S. Department of Commerce.
- [21] MITRE Corporation. (2025). *ATLAS: Adversarial threat landscape for artificial-intelligence systems*.
- [22] Mazeika, M., Phan, L., Yin, X., et al. (2024). *HarmBench: A standardized evaluation framework for automated red teaming and robust refusal*. International Conference on Machine Learning (ICML).
- [23] Chao, P., DeBenedetti, E., Robey, A., et al. (2024). *JailbreakBench: An open robustness benchmark for jailbreaking large language models*. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track.
- [24] Souly, A., Lu, Q., Bowen, D., et al. (2024). *A StrongREJECT for empty jailbreaks*. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track.
- [25] Ghosh, S., et al. (2025). *ALuminate: Introducing v1.0 of the AI risk and reliability benchmark from MLCommons*. arXiv: 2503.05731.
- [26] Cisco. (2026). *AI model and application validation (Cisco AI Defense) [Product overview]*. <https://www.cisco.com/site/us/en/products/security/ai-defense/ai-model-application-validation/index.html>
- [27] Patronus AI. (2026). *Patronus platform [Product overview]*. <https://www.patronus.ai/>
- [28] 시타임스. (2024). *슈어소프트테크, AI 모델 검증 도구 '베리파이엠(VERIFAI-M)' 최초 공개*.
- [29] UK AI Safety Summit. (2023). *The Bletchley Declaration*.
- [30] Cyberspace Administration of China. (2023). *生成式人工智能服务暂行管理办法 [생성형 AI 서비스 잠정관리방법]*.
- [31] Ministry of Electronics and Information Technology. (2026). *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026*.
- [32] UK AI Security Institute. (2025). *Frontier AI trends report*.
- [33] National Institute of Standards and Technology. (2024). *U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI*.

- [34] Cabinet Office of Japan. (2025). 人工知能関連技術の研究開発及び活用の推進に関する法律 [AI 관련 기술 연구개발 및 활용 추진법], 법률 제53호.
- [35] Digital Agency of Japan. (2025). 行政の進化と革新のための生成AIの調達・利活用に係るガイドライン (DS-920) [행정의 진화와 혁신을 위한 생성형 AI 조달·활용 지침].
- [36] Infocomm Media Development Authority & AI Verify Foundation. (2025). *Global AI assurance pilot: Findings report*.
- [37] European Commission. (2025). *Digital omnibus: Proposal amending Regulation (EU) 2024/1689*.
- [38] Council of the EU & European Parliament. (2026). *Artificial intelligence: Council and Parliament agree to simplify and streamline rules [Press release]*.
- [39] UK AI Safety Institute. (2024). *Inspect: An open-source framework for large language model evaluations*. <https://inspect.aisi.org.uk/>
- [40] 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령, 대통령령 (2026).
- [41] 과학기술정보통신부. (2026). 인공지능기본법 22일 시행 [보도자료].
- [42] 과학기술정보통신부. (2026). 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령 개정안 국무회의 통과 [보도자료].
- [43] 과학기술정보통신부. (2026). 인공지능 투명성 확보 가이드라인.
- [44] 과학기술정보통신부. (2026). 인공지능 안전성 확보 가이드라인.
- [45] 과학기술정보통신부. (2026). 고영향 인공지능 판단 가이드라인.
- [46] 과학기술정보통신부. (2026). 고영향 인공지능 사업자 책무 가이드라인.
- [47] 과학기술정보통신부. (2026). 인공지능 영향평가 가이드라인.
- [48] 한국AI안전연구소. (2026). 국내외 AI 모델 42종 안전성 평가 수행 실적.
- [49] 한국AI안전연구소. (2026). LLM 안전성 및 신뢰성 평가 데이터셋 구축 사업.
- [50] 한국정보통신기술협회. (2026). 인공지능 신뢰성 인증(CAT) 안내.
- [51] 한국인공지능산업협회. (2026). AI-MASTER 인증 제도 안내.
- [52] Infocomm Media Development Authority. (2025). *Singapore launches new tools to help businesses protect data and deploy AI in a trusted ecosystem*.



이호준 Lee, Ho-Joon

성공회대 강사·사진가

대한항공 여행사진 공모전에서 2회 수상하고, 여덟 차례의 개인전과 아홉번의 단체전을 개최했다. 포토에세이 <직조>와 <걸으면 보이는>을 출간했으며, 여러 월간지와 주간신문에 사진 이야기를 연재했다. 다수의 인쇄, 방송 매체에 소개되었고 지자체, 공공기관, 사진모임 초청으로 사진 강연을 한다. 대학에서 사진학을 강의하고 초등학교 미디어리터러시 강사로도 활동하고 있다.

