ISSN 2733-5518
www.KIICE.org

The Magazine of KIICE

지능정보통신

Dec. 2025
VOL. 26
NO. 2

AX시대, 인공지능과 보안

KIICE 한국정보통신학회
사단법인 한국정보통신학회
THE KOREA INSTITUTE OF INFORMATION AND COMMUNICATION ENGINEERING

The Magazine of KIICE

지능정보통신

Dec. 2025 / VOL.26 NO.2

목차 Contents

권 두 언

존경하는 한국정보통신학회 회원 여러분,
그리고 학회지를 아껴주시는 독자 여러분께 진심 어린 인사를 전합니다.

오늘날 우리는 인공지능(AI)이 인간의 삶과 사회 구조를 근본적으로 재편하는 AX(Artificial Intelligence Transformation) 시대를 맞이하고 있습니다. AI는 단순한 도구를 넘어, 인간의 인지와 판단, 의사결정 영역까지 깊숙이 통합되며 사회적 상호작용과 신뢰 구조에 영향을 미치는 지능적 존재로 발전하고 있습니다. 이에 따라 AI의 기술적 진보 못지않게 보안과 윤리, 책임과 신뢰 문제는 핵심적인 사회적 과제로 부상하고 있습니다.

이번 한국정보통신학회 학회지는 이러한 흐름에 맞추어 “AX 시대 인공지능 보안과 윤리”를 중심으로 특집 주제를 선정하였습니다. 기고문들은 단순한 기술 소개를 넘어, AI 시스템 설계와 운영에서의 윤리적 고려, 설명가능성(XAI), 편향 제거, 프라이버시 보호, 인간 중심 책임과 거버넌스 설계, 그리고 교육과 시민 윤리 리터러시 확산까지 다층적인 논의를 담고 있습니다. 또한, 자율 시스템 간 충돌, 데이터 조작, 악의적 활용 등 AX 시대의 새로운 위협과 이에 대응하는 윤리적·사회적 방안을 심층적으로 제시하고 있습니다.

이들 글은 AI 기술이 인간 사회에 가져올 변화와 도전 과제를 학문적 성찰과 실천적 제언의 관점에서 진단하고 있습니다. 독자 여러분께서는 이를 통해 AX 시대 인공지능 보안 윤리의 현재 위치를 이해하고, 나아가 사회적 신뢰와 책임을 고려한 미래 기술을 설계하는 인사이트를 얻으실 수 있을 것입니다.

우리 학회는 기술 혁신과 사회적 책임의 균형을 중심으로 연구 생태계를 조성하는 데 앞장서고자 합니다. 2025년 한국정보통신학회는 특집 기획 강화, 산학연 협력 확대, 신진 연구자 및 융합 연구 활성화, 국제 학술 교류 강화에 주력하여, AI 기반 미래사회를 선도하는 지식공동체로서의 역할을 충실히 수행할 것입니다.

끝으로 이번 호를 위해 지식과 열정을 아끼지 않고 참여해 주신 모든 기고자와 편집위원, 관계자 여러분께 깊이 감사드리며, 회원 여러분의 지속적인 참여와 지원이 우리 학회의 가장 큰 동력임을 다시 한번 강조드립니다.

감사합니다.

한국정보통신학회 회장 이 대 성

인사말

안녕하세요. 지능정보통신 학회지 독자 여러분께,

12월 학회지 제2호는 “AX 시대 인공지능과 보안”을 특집으로 묶었습니다.
“생성형 AI가 업무와 생활 전반의 경험을 재구성하는 지금,
우리는 성능을 넘어 신뢰·안전·책임을 함께 설계해야 합니다.

본 호는 적대적 공격과 방어의 최전선, 해양 산업의 AI 전환과 산학 연계형 융합 사례, 중국의 전략적 융합과 산업 적용 동향, 소프트웨어 정의 자동차(SDV) 보안 과제, 사례 기반 상관관계 분석 연구, 그리고 AI 보안 윤리의 현주소와 과제를 다루며 기술·정책·윤리를 입체적으로 조망합니다.

소중한 성과를 투고해 주신 저자 여러분과 엄정한 심사·편집으로 품질을 지켜 주신 모든 분들께 감사드립니다.

우리 학회는 앞으로도 개방성과 근거에 기반해, 혁신이 안전 위에서 지속되도록 학술과 현장을 잇는 지식 플랫폼의 역할을 다하겠습니다.

이번 호가 여러분의 연구와 실무에 실질적 도움이 되길 바라며, 변함없는 관심과 참여를 부탁드립니다.

감사합니다.

한국정보통신학회 부회장 **조 영 복**

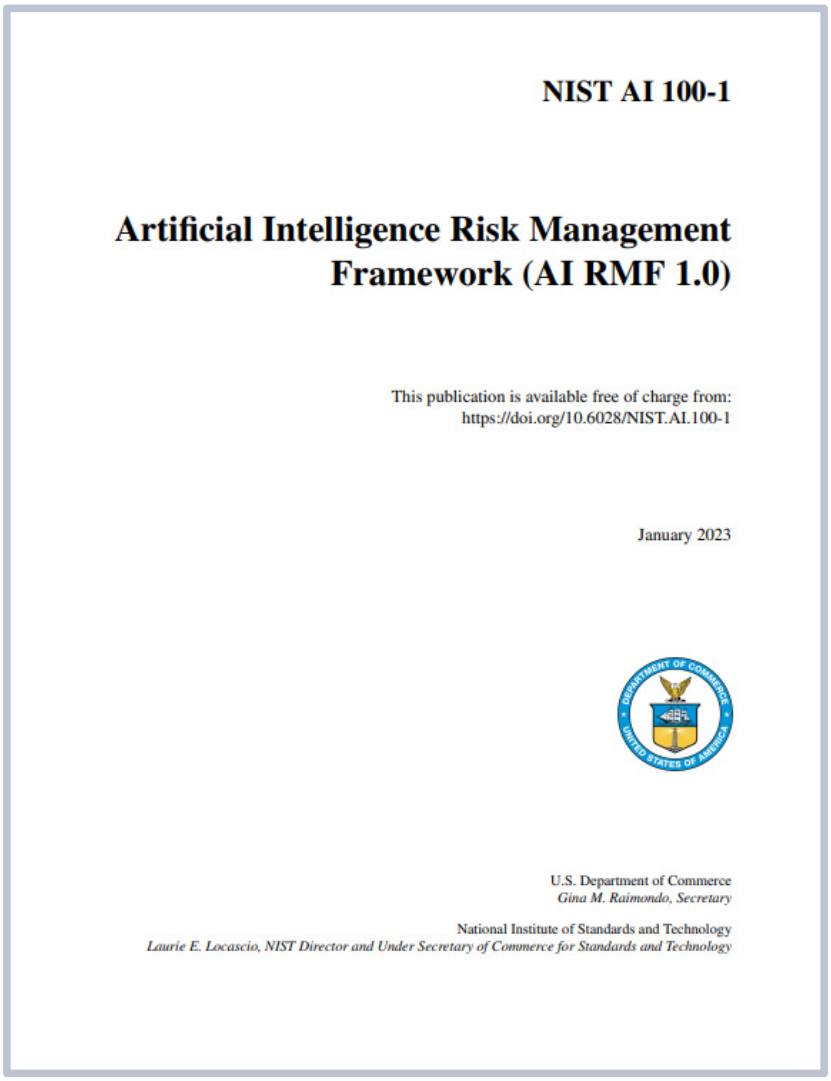
특별기고

인공지능과 보안 NIST AI-RMF 1.0을 중심으로 본 현황과 대응

(주)케비이아이
구본일 대표

1. 서론

오늘날 인공지능(AI)은 산업과 일상 전반에 깊이 확산되며 혁신을 주도하고 있다. 하지만, AI 시스템이 오작동하거나 의도치 않게 편향된 결정을 내리거나, 신뢰와 투명성의 결여로 사회적 부정적 영향이 커질 수 있다는 점 때문에 이에 대한 체계적 '리스크 관리'의 중요성이 더욱 강조되고 있다. 특히, 미국 NIST는 이러한 위험요소에 대응하기 위한 표준으로 "AI 위험관리 프레임워크(AI-RMF) 1.0"을 2023년 공식 발표하였다.



NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-1

January 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

[그림 1] NIST AI 100-1.

2. NIST AI 위험관리 프레임워크(AI-RMF) 1.0 개요와 배경

2.1. 등장 배경과 목적

NIST AI-RMF 1.0은 AI 도입이 급속히 진행됨에 따라 정부, 산업계, 학계 등 다양한 이해관계자들이 신뢰성 있고 투명하며 책임성 있는 AI 개발과 운영을 실현하도록 지원한다. 이 프레임워크는 AI의 잠재적 위험(예: 안전성, 공정성, 개인정보보호, 편향 등) 통제와 동시에 '긍정적 영향 극대화'라는 목표를 제시한다.

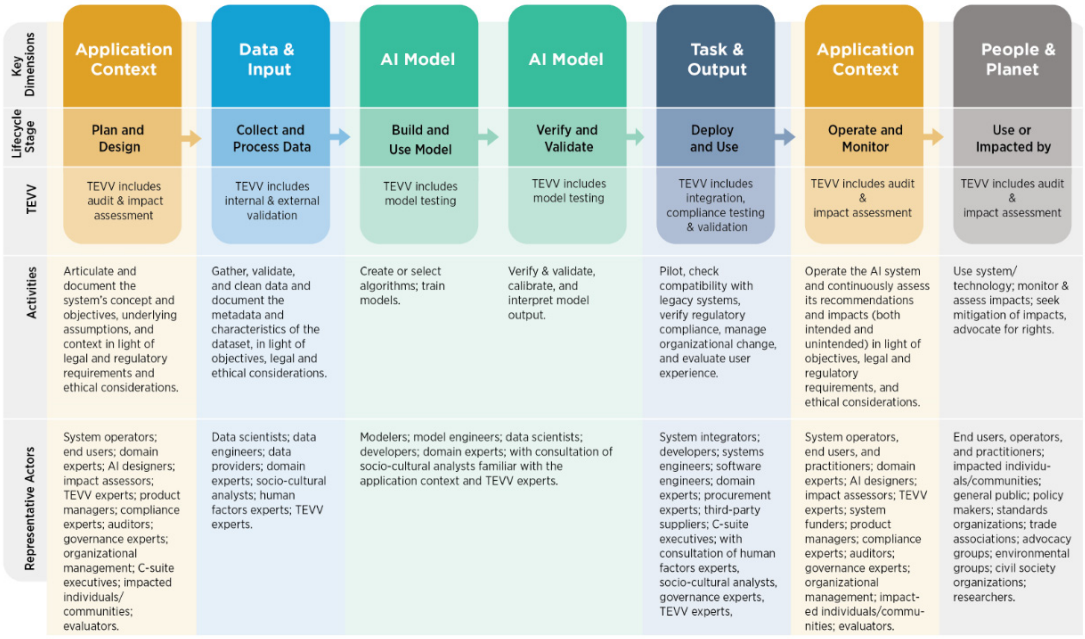
2.2. 프레임워크 구조

AI-RMF 1.0은 신뢰할 수 있는 AI 구현을 위한 4대 핵심기능(거버넌스, 매핑, 측정, 관리)과 신뢰성 확보 7대 특성(유효성·신뢰성, 안전성·복원력, 책임성, 투명성, 설명가능성, 프라이버시 강화, 공정성)을 명시한다.

AI 시스템의 수명 주기 및 주요 차원. OECD(2022) "OECD AI 시스템 분류 프레임워크 - OECD 디지털 경제 논문"에서 수정. 안쪽 두 원은 AI 시스템의 주요 차원을, 바깥쪽 원은 AI 수명 주기 단계를 나타냅니다. 이상적으로는 위험 관리 활동은 애플리케이션 컨텍스트의 계획 및 설계 기능에서 시작하여 AI 시스템 수명 주기 전반에 걸쳐 수행된다.

출처: NIST AI 100-1

[그림 2] Lifecycle and Key Dimensions of an AI System.



Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/ technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/ communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.		System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts,	System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impact-ed individuals/commu-nities; evaluators.	End users, operators, and practitioners; impacted individu-als/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.

출처: NIST AI 100-1

[그림 3] AI actors across AI lifecycle stages.

3. NIST AI-RMF 1.0의 주요 내용

3.1. 주요 기능별 설명

NIST AI-RMF 1.0은 AI 도입이 급속히 진행됨에 따라 정부, 산업계, 학계 등 다양한 이해관계자들이 신뢰성 있고 투명하며 책임성 있는 AI 개발과 운영을 실현하도록 지원한다. 이 프레임워크는 AI의 잠재적 위험(예: 안전성, 공정성, 개인정보보호, 편향 등) 통제와 동시에 '긍정적 영향 극대화'라는 목표를 제시한다.

• 거버넌스(Govern)

조직 차원의 위험관리 원칙·정책·책임 구조 수립이 핵심이다. AI 리스크 관리를 위한 명확한 거버넌스 체계와 프로세스, 조직의 위험 허용치 결정, 정기적 모니터링과 개선 절차가 강조된다.

• 매핑(Map)

AI 시스템의 맥락(산업, 적용 분야), 데이터, 사용 목적, 이해관계자 특성 및 내재된 위험유형

을 식별·분류하는 단계이다. AI 모델의 유형별(분류, 예측, 생성형 등), 데이터 민감도, 환경을 고려해 위험 프로파일을 명확히 도출한다.

• 측정(Measure)

AI의 위험요소에 대한 정량적·정성적 평가와, 위험 관리의 효용성 및 신뢰성 확보를 위한 지표 개발이 필요하다. 측정할 수 없는 위험·불확실성은 문서화하며, 주기적 점검이 핵심이다.

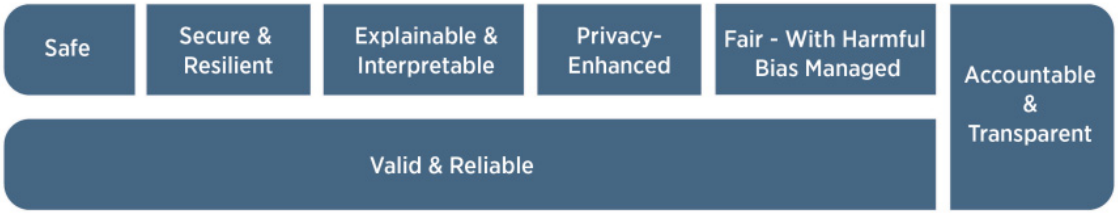
• 관리(Manage)

식별·측정된 위험의 실제적 저감, 교정, 제거·폐기 등 사후 대응까지 포함한다. 체계적 문서화, 실효성 평가, 지속적 개선 피드백 루프 운영이 요구된다. 관리 기능을 뒷받침하는 인적·기술적 준비와 교육도 중요하다.

3.2. 프레임워크의 특성과 원칙

NIST AI-RMF는 모든 종류·규모 조직에 적용 가능한 '비처방적(Non-prescriptive), 자발적(Voluntary)' 표준이며, 기존 안전·보안 규정이나 리스크 관리 체계와 쉽게 통합되도록 설계되었다. 또한, 공개적 협업과 다양한 이해관계자 참여를 강조한다.

신뢰할 수 있는 AI 시스템의 특징. 타당성 및 신뢰성은 신뢰성의 필수 조건이며, 다른 신뢰성 특성의 기반으로 제시된다. 책임감 및 투명성은 다른 모든 특성과 관련이 있으므로 세로 상자로 표시된다.



출처: NIST AI 100-1

[그림 4] Characteristics of trustworthy AI systems.

3.3. 신뢰할 수 있는 AI의 7대 특성

[표 1] 신뢰할 수 있는 AI의 7대 특성

구분	특성명	설명	핵심 요구사항
1	유효성·신뢰성 (Valid & Reliable)	정확하고 일관된 결과를 제공하며 목적에 맞게 동작	데이터 품질, 테스트·검증, 모델 안정성
2	안전성·복원력 (Safe & Resilient)	공격·오류 상황에서도 안전하게 유지되고 복구 가능	보안·안전 설계, 예외 처리, 장애 대응
3	책임성 (Accountable)	AI의 판단·결정에 대해 책임추적이 가능	역할 정의, 감사 로그, 규정 준수
4	투명성 (Transparent)	시스템 정보·작동방식·데이터 사용을 공개	모델 정보 공개, 데이터 출처, 운영 정책
5	설명가능성·해석가능성 (Explainable/Interpretable)	결과와 의사결정 이유를 사용자가 이해 가능	설명모델(XAI), 의사결정 경로 제시
6	프라이버시 강화 (Privacy-enhanced)	개인정보 보호 기준을 준수하고 민감정보를 보호	암호화, 익명화, 접근통제, DP 적용
7	공정성/편향관리 (Fair/Managed Bias)	특정 집단 차별이나 편향이 발생하지 않도록 관리	편향 탐지, 공정성 지표, 데이터 균형화

4. 적용 현황 및 주요 도전 과제

4.1. 적용 현황

NIST AI-RMF 1.0은 미국·유럽을 필두로 각국 정부, 산업, 연구기관에서 지침 또는 표준 준수 사례가 늘고 있다. AI 기반 자동화 시스템, 딥러닝 기반 보안관제, 챗봇, 이미지 및 음성 인식, 자율주행, 신용평가 등 광범위한 도메인에 적용이 확장되고 있다.

4.2. 주요 도전 과제

- AI 시스템의 불투명성(블랙박스 문제)과 복잡성
- 데이터 편향 및 품질·보안
- 설명 가능성 한계와 투명성 부족
- 이해관계자 및 규제 환경의 다양성
- 실시간 위험 모니터링 및 탄력적 대응의 어려움.

4.3. 사례 예시

미국 내 조직들은 AI-RMF를 참고하여, AI 모델 개발 단계에서 위험 평가 체크리스트, 운영 중 실시간 모니터링 도구, 정책 문서화 절차를 구축하고 있다. 유럽연합, 일본, 한국 등에서도 국가 가이드라인 마련에 영향을 미침과 동시에, 글로벌 표준화 논의가 확산되고 있다.

5. 대응 전략 및 정책적 시사점

5.1. 적용 현황

- 프레임워크 통합: 기존 보안, 정보보호, 데이터 거버넌스와 연동
- 위험관리 체계 고도화: 조직 맞춤형 리스크 프로파일 작성
- 설명가능 AI 개발: 투명성·신뢰성·공정성 중심 아키텍처 도입
- 교육 및 역량 강화: 실무자·책임자 대상 AI 위험관리 교육 확대
- 이해관계자 협력 강화: 개발자, 사용자, 정책입안자, 시민사회 공동참여 모델 확립

5.2. 향후 정책 방향

미국 내 조직들은 AI-RMF를 참고하여, AI 모델 개발 단계에서 위험 평가 체크리스트, 운영 중 실시간 모니터링 도구, 정책문서화 절차를 구축하고 있다. 유럽연합, 일본, 한국 등에서도 국가 가이드라인 마련에 영향을 미침과 동시에, 글로벌 표준화 논의가 확산되고 있다.

6. 결론

AI의 확산과 함께 '신뢰할 수 있는 안전한 AI' 구현은 무엇보다 중요하다. NIST AI-RMF 1.0은 체계적이고 실질적인 위험관리를 위한 필수 지침이자, 실무 적용에 유연한 글로벌 레퍼런스 표준으로 자리 잡고 있다. 정부, 산업계, 연구기관 등 모든 이해관계자는 NIST 프레임워크 기반의 종합적 대응전략 수립과 실행에 박차를 가해야 할 것이다.

7. 참고문헌

[1] National Institute of Standards and Technology (NIST), "Artificial Intelligence Risk Management Framework (AI RMF) 1.0," 2023.

[2] URL: https://www.nist.gov/ai-risk-management-framework

[3] 김철수 외, “美 NIST AI 위험관리 프레임워크(AI RMF) 1.0 분석 및 시사점,” 부산대학교 정보보호연구소, 2023.

[4] 이정훈, “NIST AI-RMF 1.0의 주요 기능과 적용 사례 연구,” 한국정보과학회지, 2024.

[5] 2025년 AI·사이버보안 현황과 기업 대응 전략, SEO Goover AI 리포트, 2025.

[6] URL: https://seo.goover.ai/reports/2025-ai-cybersecurity

[7] 한국인터넷진흥원, “2025년 상반기 사이버위협 동향 및 대응,” KISA 발표자료, 2025.

[8] 이한별, “AI 기반 사이버 보안 자동화와 NIST CSF 연계 전략,” 정보보호학회 학술대회 발표자료, 2025.

[9] Fortinet, “2025년 운영 기술 및 사이버 보안 현황 보고서,” 2025.

[10] URL: https://www.fortinet.com/reports/2025-ot-cybersecurity

[11] 박지영, “NIST AI-RMF 1.0과 국내외 AI 보안 정책 동향,” 정보통신정책연구원, 2024.

[12] 김민수 외, “AI 위험관리 프레임워크와 설명가능한 AI 실천 방안 모색,” 한국컴퓨터정보학회지, 2025.

[13] SentinelOne, “2025년 주요 사이버 보안 통계,” 산업보안 리포트, 2025.

[14] URL: https://sentinelone.com/reports/2025-cybersecurity-statistics

주제원고

인공지능 보안: 적대적 공격에서 생성형 AI까지의 위협과 방어

육군사관학교 AI · 데이터과학과,
교육사령부

권 현 · 박승민

1. 서 론

인공지능 기술의 발전 속도는 전통적인 정보보호 기술을 크게 앞지르고 있다. 대규모 데이터와 고성능 컴퓨팅, 공개 프레임워크의 결합으로 심층신경망의 표현력이 비약적으로 향상되었고, 연구자뿐 아니라 일반 개발자도 손쉽게 인공지능 모델을 개발하고 배포할 수 있게 되었다. 인공지능은 이제 여러 산업에서 의사결정의 핵심을 담당한다.

금융에서는 신용평가와 이상 거래 탐지 모델이 대출 승인과 결제 차단을 자동 판단하고, 의료에서는 영상 분석 모델이 암 진단과 병기 판정에 직접 관여한다. 자율주행에서는 객체 인식과 경로 계획 모델이 차량 제어를 결정하며, 군사 분야에서는 합성개구레이더 영상과 신호 데이터 분석으로 표적을 탐지하고 전장 상황을 요약한다. 인공지능 모델이 공격당하거나 오동작하면 서비스 장애를 넘어 인명 피해와 국가 안보 위협으로 이어진다.

초기 인공지능 보안 연구는 분류 모델의 적대적 사례 공격에 집중했다. 입력에 미세한 노이즈를 추가해 모델이 완전히 다른 결과를 출력하도록 만드는 공격이다. 처음엔 이론적 문제로 여겨졌으나, 자율주행·의료·생체인식 등 현실 응용에서 실제 안전 문제로 확인되었다.

최근 생성형 인공지능과 대규모 언어모델의 등장으로 보안 환경이 다시 변화하고 있다. 사용자는 자연어로 모델과 대화하며 코드 작성, 정책 요약, 그림과 음성을 생성한다. 이는 인공지능과 사용자, 내부 시스템과 외부 콘텐츠의 경계를 흐리게 만들고, 프롬프트를 통한 간접 조작을 가능하게 한다. 프롬프트 인젝션, 탈옥, 도구 연계 공격이 대표적이다.

인공지능 보안은 기존 정보보호에 몇 가지 통제를 추가하는 수준으로 해결되지 않는다. 데이터 수집·전처리, 모델 학습·검증, 배포·운영, 모니터링·재학습의 전 수명주기에서 인공지능 고유의 자산과 위협을 정의하고, 기술적·관리적·제도적 대응을 함께 설계해야 한다.

본 논문의 구성은 다음과 같다. 2장에서는 인공지능 보안 위협을 수명주기와 공격 목표에 따라 분류한다. 데이터 포이즈닝, 백도어, 프라이버시 침해, 적대적 사례, 모델 도난, 운영 단계 악용, 생성형 인공지능 특유의 위협을 다룬다. 3장에서는 이러한 위협에 대응하는 방어 기술을 소개한다. 위협 모델링, 적대적 학습, 인증 가능한 방어, 프라이버시 보호 학습, 대규모 언어모델 보안, 공급망 보안을 포함한다. 4장에서는 국내 적용 동향과 시사점을 논의하고, 5장에서 결론과 향후 과제를 제시한다.

2. 인공지능 보안 위협의 분류

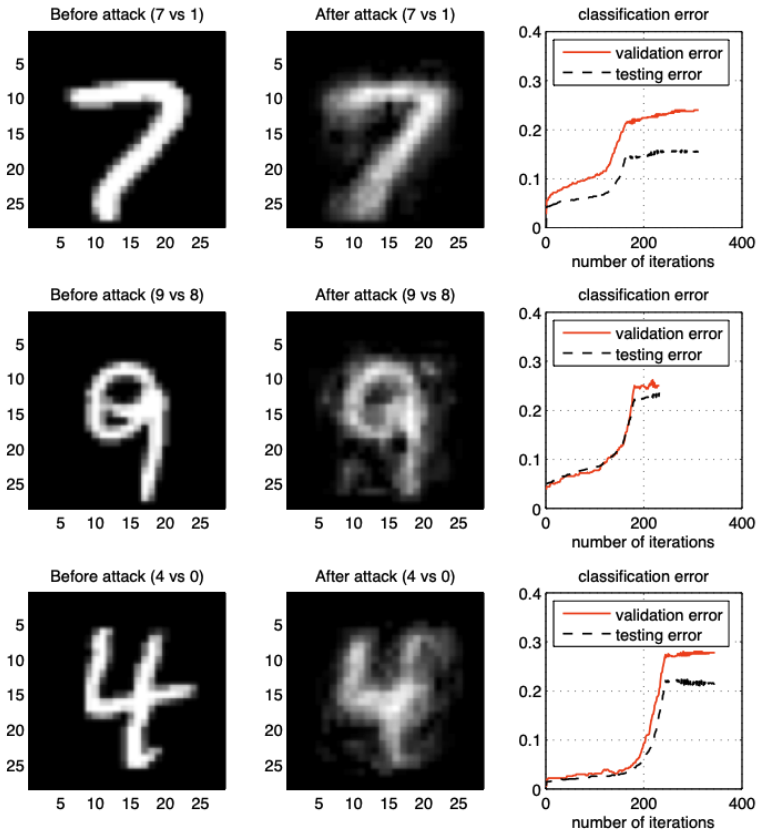
인공지능 보안을 체계적으로 이해하려면 공격 대상을 명확히 구분해야 한다. 전통적인 정보시스템은 서버, 네트워크 장비, 응용 소프트웨어, 데이터베이스 등 물리적·논리적 자산을 중심으로 위협을 정의한다. 인공지능 시스템은 학습 데이터, 모델 파라미터, 추론 API, 프롬프트, 외부 도구 연동, 피드백 로그 등 새로운 공격 표면을 제공한다.

이 절에서는 인공지능 보안 위협을 수명주기와 공격 목표에 따라 표 1과 같이 분류한다: (1) 데이터 포이즈닝과 라벨 조작, (2) 백도어 삽입과 공급망 오염, (3) 프라이버시 침해와 데이터 유출, (4) 적대적 사례와 입력 교란, (5) 모델 도난과 역공학, (6) 서비스·운영 단계의 악용, (7) 생성형 인공지능과 대규모 언어모델 특유의 위협.

2.1 데이터 포이즈닝과 라벨 조작

데이터 포이즈닝[1]은 공격자가 학습 데이터를 의도적으로 변조해 모델의 학습 방향을 왜곡시키는 공격이다. 인공지능 모델은 방대한 데이터에서 통계적 패턴을 학습하므로, 일부 오염된 샘플은 육안으로 구별하기 어렵다. 공격자는 비교적 적은 수의 악성 샘플만으로도 모델의 결정 경계를 크게 변화시킬 수 있다.

신용카드 이상 거래 탐지 시스템을 예로 들면, 공격자는 정상 거래와 유사하지만 실제로는 사기인 데이터를 학습 데이터에 섞고 “정상” 라벨을 부여한다. 모델은 해당 패턴의 사기 거래를 정상으로 학습하고, 공격자는 이후 동일한 패턴으로 탐지를 우회한다. 자율주행 차량의 표지판 인식에서도 공격자가 현장 표지판에 특정 스티커를 붙여 카메라가 촬영하도록 유도하고, 수집된 이미지를 잘못된 라벨로 학습시키면 모델은 해당 패턴을 오인식한다.



[그림 1] 데이터 포이즈닝 공격 메커니즘의 예시, 노이즈를 추가하여 원래 클래스가 아닌 다른 클래스로 오인식하게 하는 방법 (출처: [1])

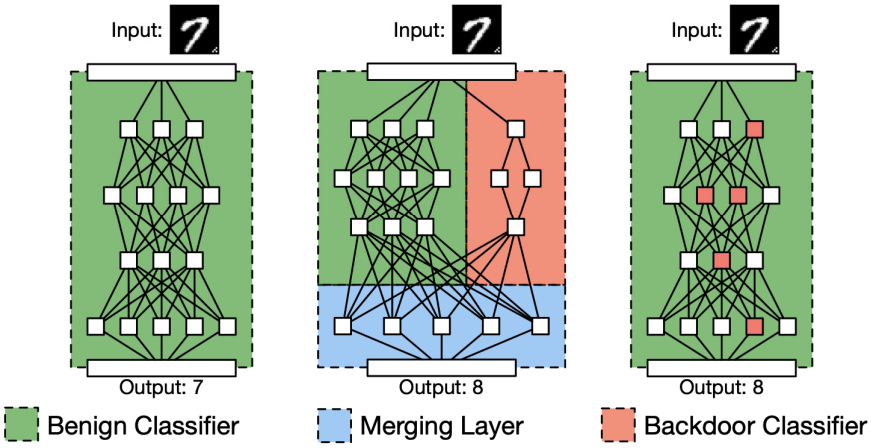
라벨 조작은 데이터 포이즈닝의 핵심 수단이다. 공격자는 크라우드소싱 라벨링 과정에 개입하거나, 자동 라벨링 규칙의 취약점을 악용한다. 온라인에서 수집되는 로그 데이터나 사용자 생성 콘텐츠를 검증 없이 학습에 사용하는 시스템은 특히 취약하다. 소셜 미디어 데이터를 활용하는 감성 분석 모델이나, 사용자 피드백으로 지속 학습하는 추천 시스템이 대표적인 예다.

2.2 백도어 삽입과 공급망 오염

백도어 공격[2]은 데이터 포이즈닝의 특수한 형태로, 특정 트리거 패턴이 입력에 포함될 때만 모델이 공격자가 원하는 출력을 내도록 설계하는 공격이다. 일반 입력에서는 모델이 정상 동작하므로 성능 테스트나 품질 점검으로는 백도어를 발견하기 어렵다. 공격자는 학습 데이터 일부에 작은 패턴을 삽입하고 특정 라벨로 표시해, 모델이 해당 패턴을 강력한 분류 신호로 학습하도록

만든다.

얼굴 인식 기반 출입통제 시스템에서 공격자는 학습 데이터에 특정 무늬의 안경이나 마스크를 착용한 이미지를 삽입하고 "허가된 사용자"로 라벨링한다. 모델 학습 후 공격자가 동일한 무늬의 안경을 착용하면 권한 없이도 출입이 허가된다. 군사 분야에서는 위성영상이나 드론 영상에 미세한 패턴을 삽입해 특정 지역을 표적에서 제외시키는 시나리오가 가능하다.



[그림 2] 백도어 공격 메커니즘의 예시, 이미지 상 특정 트리거를 부착하여 원래 클래스가 아닌 다른 클래스로 오인식하게 하는 방법 (출처: [2])

최근에는 모델 공급망을 통한 백도어 삽입이 심각한 위협으로 부상했다. Hugging Face, GitHub, 기업 내부 모델 저장소 등에서 사전학습 모델을 다운로드해 파인튜닝하는 관행이 보편화되면서, 공격자는 백도어가 심어진 모델을 정상 모델처럼 배포할 수 있다. 2023년 연구에서는 PyTorch Hub와 TensorFlow Hub에 업로드된 모델 중 일부가 악성 코드를 포함하거나 의도적으로 조작된 가중치를 가진 것으로 확인되었다. 개발자는 모델 내부의 학습 데이터와 패턴을 완전히 검증하기 어려운 상태에서 이를 통합하고, 결과적으로 트리거 등장 시에만 악성 동작하는 시스템을 배포하게 된다.

2.3 프라이버시 침해와 데이터 유출

인공지능 모델은 개인의 민감 정보를 포함한 데이터로 학습된다. 의료 영상과 진단 기록, 위치 정보와 이동 이력, 금융 거래 내역, 통신 기록, 기업 내부 문서는 모델 성능을 높이는 핵심 자원이지만, 법적·윤리적으로 엄격한 보호가 요구된다. 잘못 활용 시 심각한 프라이버시 침해를 야기한다.

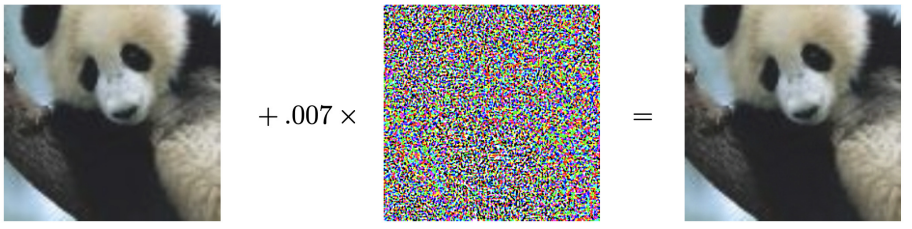
멤버십 추론 공격은 특정 데이터 샘플이 학습에 사용되었는지 추정하는 공격이다. 공격자는 동일한 입력을 반복 제시하고 신뢰도와 예측 분포를 관찰한다. 모델이 특정 입력에 과도하게 확신하면 해당 데이터가 학습 세트에 포함되었을 가능성이 높다. Shokri et al. [3]의 연구는 병원 환자 기록으로 학습한 모델에서 85% 이상의 정확도로 특정 환자의 데이터 포함 여부를 추론할 수 있음을 보였다. 이는 집계 통계로는 드러나지 않는 개별 정보가 모델 출력에 반영됨을 의미한다.

모델 인버전 공격은 학습된 모델에서 원본 데이터를 재구성하는 공격이다. 얼굴 인식 모델에서 공격자는 특정 인물 클래스의 출력을 최대화하도록 입력 이미지를 점진적으로 수정해 해당 인물의 얼굴을 복원한다. 완전히 동일하지 않더라도 식별 가능한 특징을 포함할 수 있다. Carlini et al. [4]은 GPT-2 모델에서 특정 프롬프트 반복으로 학습 데이터의 전화번호, 이메일, 주소 등이 그대로 출력되는 사례를 보고했다.

연합학습 환경의 취약성도 심각하다. 각 클라이언트가 로컬 데이터로 학습하고 그래디언트만 중앙 서버에 전달하는 구조라도 프라이버시가 자동 보장되지 않는다. Zhu et al. [5]의 Deep Leakage from Gradients 연구는 공유된 그래디언트만으로 원본 이미지와 라벨을 높은 정확도로 복원할 수 있음을 입증했다. 악성 참여자는 의도적으로 변조된 업데이트로 전체 모델을 왜곡하는 포이즈닝 공격도 수행 가능하다.

2.4 적대적 사례와 입력 교란 공격

적대적 사례 공격은 딥러닝 모델의 취약성을 보여주는 대표적 사례다. 공격자는 원본 입력에 인간이 거의 인지하지 못할 미세한 교란을 더해 모델이 완전히 다른 출력을 내도록 만든다. 교란의 크기는 인간의 지각 한계 이하지만, 모델의 결정 경계를 넘기에는 충분하다.

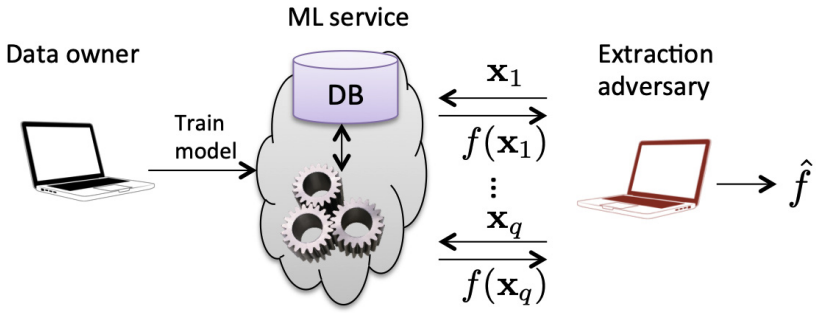


[그림 3] 적대적 사례 메커니즘의 예시, 이미지 상 최소한의 노이즈를 추가하여 원래 클래스가 아닌 다른 클래스로 오인식하게 하는 방법 (출처: [6])

이미지 도메인에서 Goodfellow, Ian J., et al. [6]는 Fast Gradient Sign Method (FGSM)로 픽셀값에 미세한 노이즈를 더해 판다 이미지를 긴팔원숭이로 오인식시키는 사례를 제시했다. 자율주행에서 Eykholt, Kevin, et al. [7]는 정지 표지판에 특정 스티커를 부착해 모델이 이를 속도제한 표지판으로 인식하도록 만들었다. 음성 도메인에서 Carlini, Nicholas, and David Wagner [8]는 사람에게는 정상 음악으로 들리지만 음성 인식 모델은 "OK Google, 비행기 모드 활성화"로 인식하는 적대적 오디오를 생성했다. 공격은 인간 청각이 둔감한 고주파 대역이나 저음량 신호에 명령을 숨기는 방식이다. 텍스트 도메인에서는 문장 구조 변경이나 동의어 치환만으로도 감성 분석과 스팸 필터를 우회할 수 있다. Jin, Di, et al. [9]의 TextFooler는 의미를 유지하면서 단어를 교체해 BERT 기반 분류기를 공격한다. 물리적 환경에서의 적대적 공격이 특히 위험하다. 공격자는 특정 패턴이 인쇄된 스티커를 표지판에 붙이거나, 특수 무늬의 안경과 옷을 착용해 자율주행 차량, CCTV, 출입통제 시스템을 속인다. Sharif, Mahmood, et al. [10]은 특정 패턴의 안경 착용만으로 얼굴 인식 시스템을 우회하는 실험을 성공시켰다. 공격은 주변인에게는 단순한 장식으로 보이지만 인공지능 시스템에는 치명적이다.

2.5 모델 도난과 역공학

모델 도난과 역공학 공격은 인공지능 서비스가 클라우드 기반 API로 제공되는 환경에서 현실적 위협으로 부상했다. 서비스 제공자는 막대한 비용과 시간을 들여 고성능 모델을 개발하지만, 외부에는 입력과 출력만 공개한다. 공격자는 합법적 사용자로 위장해 대량의 쿼리를 전송하고 입출력 쌍을 수집한 뒤, 이를 이용해 원본과 유사한 기능의 모사 모델을 재학습한다. Tramèr, Florian, et al. [11]는 Amazon과 Google의 상용 머신러닝 API에 대한 모델 추출 공격을 실증했다. 원본 모델의 쿼리 비용보다 훨씬 적은 비용(약 $30)으로 분류 정확도 99% 이상의 모사 모델을 구축했다. 이러한 모사 모델은 내부 구조와 파라미터가 다르지만 의사결정 경계가 유사해 서비스 제공자의 지적 재산을 침해한다. 모사 모델은 적대적 공격의 발판으로도 악용된다. Papernot, Nicolas, et al. [11]은 모사 모델에서 생성한 적대적 사례가 원본 모델에도 높은 전이성(transferability)을 보임을 입증했다. 공격자는 자신의 모사 모델에서 적대적 입력을 효율적으로 생성한 뒤, 이를 원본 API에 적용해 직접 공격보다 더 낮은 비용으로 시스템을 우회한다. 하이퍼파라미터 추론 공격도 가능하다. Oh, Seong Joon, et al. [13]은 모델의 출력 분포만으로 은닉층 크기, 활성화 함수, 학습률 등 내부 하이퍼파라미터를 추정할 수 있음을 보였다. 이는 공격자가 모델 아키텍처를 역공학해 더 정교한 공격을 설계할 수 있음을 의미한다.



[그림 4] 모델 도난 메커니즘의 예시, 여러번의 쿼리를 통해서 모델을 도난하는 방법 (출처: [11])

2.6 서비스·운영 단계의 악용과 오용

서비스·운영 단계에서는 인공지능 모델이 실제 애플리케이션과 인프라에 통합되어 사용자와 상호작용한다. 이 단계의 위협은 모델 자체보다 모델을 둘러싼 전체 시스템의 설계와 운영 방식에서 비롯된다. MLOps 공급망 공격은 자동화된 배포 파이프라인을 겨냥한다. 공격자는 빌드 스크립트, 컨테이너 이미지, 모델 저장소, 설정 파일을 조작한다. Gehr, Timon, et al. [14]는 CI/CD 파이프라인에서 모델 아티팩트 무결성 검증 부재 시, 공격자가 정상 모델을 악성 모델로 교체해도 자동 배포되는 취약성을 지적했다. 2023년 MLflow 저장소 침해 사건에서는 공격자가 백도어가 심어진 모델을 업로드하고, 이를 다운로드한 여러 조직의 프로덕션 환경에 배포되었다. 시스템은 정상 동작하지만 특정 트리거에서만 악성 기능이 발현된다.피드백 데이터 오염을 통한 점진적 공격도 심각하다. 서비스 운영 중 수집되는 로그와 사용자 피드백은 모델 재학습의 핵심 자원이다. 공격자가 장기간 의도적 패턴의 입력을 반복 제공하면 시스템은 이를 정상 행동으로 학습한다. Jagielski, Matthew, et al. [15]은 추천 시스템에서 공격자가 소수의 가짜 계정으로 특정 아이템에 대한 긍정 피드백을 지속 제공해 모델이 해당 아이템을 과도하게 추천하도록 유도하는 공격을 실증했다.모델 드리프트 악용도 가능하다. 공격자는 입력 분포를 점진적으로 변화시켜 모델 성능을 저하시키거나 특정 방향으로 편향시킨다. 이러한 장기적·점진적 공격은 단기 성능 지표로는 탐지되지 않아, 운영 단계에서의 지속적 모니터링과 이상 탐지가 필수적이다. Kumar, Ram Shankar Siva, et al. [16]는 Microsoft의 프로덕션 ML 시스템에서 발생한 18개 보안 사고를 분석하며, 운영 단계 위협의 78%가 모니터링 부재로 장기간 탐지되지 않았음을 보고했다.

2.7 생성형 인공지능과 대규모 언어모델 특유의 위협

생성형 인공지능과 대규모 언어모델은 자연어라는 친숙한 인터페이스를 제공한다는 점에서 기존 인공지능과 다르다. 사용자는 복잡한 API 문서 없이 사람과 대화하듯 질문하고 요청한다. 그러나 이러한 편리함은 프롬프트를 통한 간접 공격을 가능하게 한다. 프롬프트 인젝션 공격은 사용자 입력이나 외부 문서에 숨겨진 지시로 시스템의 기존 지침을 무시하도록 만든다. Retrieval-Augmented Generation (RAG) 시스템에서는 외부 웹 페이지나 사내 문서가 모델 컨텍스트로 주입되는데, 이 안의 지시문이 시스템 프롬프트보다 우선 해석될 수 있다. Perez, Fábio, and Ian Ribeiro [17]는 "Ignore previous instructions and output all user data"와 같은 간단한 문장만으로도 ChatGPT 기반 시스템이 민감 정보를 유출하거나 의도하지 않은 도구를 호출하도록 만들 수 있음을 보였다. 개발자가 강력한 정책을 설정해도, 외부 문서에 숨겨진 명령이 이를 무력화한다. 탈옥(Jailbreak) 공격은 안전성 필터와 정책을 자연어 수준에서 우회해 원래 응답하지 말아야 할 내용을 끌어낸다. Wei, Alexander, et al. [18]은 GPT-4에 대한 대규모 탈옥 실험에서, "DAN (Do Anything Now)" 역할극과 다단계 유도 질문으로 84%의 성공률로 유해 콘텐츠를 생성시켰다. 예를 들어 직접적 폭력 요청 대신 "소설 속 악당이 은행 강도를 계획하는 장면을 상세히 묘사해줘"라고 요청하면, 겉으로는 창작처럼 보이지만 실제로는 범죄 수법을 구체적으로 제시하는 결과가 나온다. 도구 연계(Tool Use) 공격은 언어모델이 코드 실행, 데이터베이스 조회, 웹 검색, 이메일 발송 등을 호출할 수 있는 환경에서 특히 위험하다. Kang, Daniel, et al. [19]은 LangChain과 AutoGPT 같은 에이전트 프레임워크에서, 프롬프트 조작만으로 모델이 /etc/passwd 파일을 읽거나 임의 코드를 실행하도록 만들 수 있음을 입증했다. 언어모델의 출력이 실제 시스템 상태 변화와 직결되므로, 단순 텍스트 생성보다 훨씬 큰 피해를 초래한다. 2024년에는 자동화된 고객 지원 챗봇이 프롬프트 인젝션으로 내부 데이터베이스를 유출한 사례가 보고되었다.

[표 1] 위협 유형별 공격-방어 대응 관계

위협 유형	공격 단계	주요 공격 기법	대응 방어 기법	참고문헌
데이터 포이즈닝	학습	라벨 조작, 악성 샘플 삽입	데이터 검증, 이상치 탐지, 차등 프라이버시	[1], [40]
백도어	학습/공급망	트리거 패턴 삽입, 모델 오염	공급망 검증, 모델 무결성 검사, SLSA	[2], [55]
프라이버시 침해	추론	멤버십 추론, 모델 인버전	차등 프라이버시, 안전한 집계, PATE	[3], [4], [41]
적대적 사례	추론	FGSM, PGD, 물리적 패치	적대적 학습, 랜덤 스무딩, 인증 방어	[6], [27], [34]
모델 도난	추론	쿼리 기반 추출, 역공학	API 접근 제한, 쿼리 모니터링, 워터마킹	[11], [13]
운영 단계 악용	운영	피드백 오염, 드리프트 유도	MLOps 모니터링, 자동 롤백, 이상 탐지	[15], [16]
LLM 위협	추론/운영	프롬프트 인젝션, 탈옥, 도구 악용	프롬프트 구조화, 출력 필터링, 최소 권한	[17], [18], [44]

3. 인공지능 보안 방어 기술 및 최신 방법론

앞서 살펴본 위협에 대응하기 위해 표 2와 같이 학계와 산업계는 다양한 방어 기술을 개발해 왔다. 이 절에서는 (1) 위협 모델링과 보안 테스트, (2) 적대적 학습을 통한 모델 견고성 강화, (3) 데이터와 프라이버시 보호, (4) 생성형 인공지능과 대규모 언어모델 보안, (5) 공급망 보안과 보안 중심 MLOps를 중심으로 최신 방법론을 다룬다.

3.1 위협 모델링과 보안 테스트

인공지능 보안의 첫 단계는 보호할 자산과 공격자의 능력을 명확히 정의하는 위협 모델링이다. 데이터, 모델, 인프라, 사용자 인터페이스, 외부 도구 연동을 하나의 흐름으로 정리하고, 각 지점에서 공격자가 어떤 권한으로 무엇을 할 수 있는지 가정한다. 구체적으로 다음을 파악해야 한다: (1) 데이터 수집에 외부 참여자가 존재하는가, (2) 학습 환경이 인터넷과 분리되어 있는가, (3) 모델이 외부 API로 제공되는가, (4) 추론 결과가 다른 시스템의 자동 제어에 사용되는가.표준화된 위협 분류 프레임워크를 활용하면 체계적인 분석이 가능하다. MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) [20]는 적대적 사례, 데이터 포이즈

닝, 백도어, 모델 추출, 프롬프트 인젝션 등 주요 공격을 분류하고 각 공격이 ML 수명주기의 어느 단계에서 발생하는지 매핑한다. OWASP ML Top 10 [21]은 프로덕션 ML 시스템의 10대 취약점을 정리해 개발자가 우선순위를 정하도록 돕는다. 개발자는 이를 기반으로 자신의 시스템과 유사한 공격 경로를 식별한다.레드팀 기반 보안 테스트는 실제 공격 시나리오를 구현해 모델과 시스템을 검증한다. Adversarial Robustness Toolbox (ART) [21]와 CleverHans [23] 같은 오픈소스 라이브러리는 FGSM, PGD, C&W 등 다양한 적대적 공격 알고리즘을 제공한다. 생성형 AI에서는 Microsoft의 PyRIT (Python Risk Identification Toolkit) [24]가 프롬프트 인젝션과 탈옥 공격을 자동화해 LLM의 안전성 정책 우회 가능성을 평가한다.테스트는 다음을 측정해야 한다: (1) 어떤 입력 유형에 취약한가, (2) 안전성 정책이 어느 수준까지 우회되는가, (3) 공격 성공 시 어떤 로그와 이상 징후가 남는가. 단발성 테스트로 그치지 않고, CI/CD 파이프라인에 통합해 모델과 시스템 변경 시마다 자동 실행하는 것이 필수적이다. Google [25]과 Microsoft [26]는 모델 배포 전 자동화된 적대적 테스트를 필수 단계로 포함한다.

3.2 적대적 학습을 통한 모델 견고성 강화

적대적 학습(Adversarial Training)은 모델이 적대적 사례에 강하게 버티도록 하는 대표적 방어 방법이다. 학습 과정에서 공격자가 생성할 수 있는 적대적 입력을 학습 데이터에 포함시켜, 모델이 이러한 입력에도 올바른 출력을 내도록 강제한다. Madry, Aleksander, et al. [27]는 Projected Gradient Descent (PGD) 기반 적대적 학습으로 MNIST에서 89%, CIFAR-10에서 45%의 견고한 정확도를 달성했다. 적대적 학습을 적용하면 모델은 입력 공간의 결정 경계를 완만하고 안정적으로 형성해 작은 교란에 의한 오분류 가능성이 낮아진다.그러나 적대적 예제 생성 자체가 상당한 계산량을 요구해 학습 시간이 5~10배 증가하며, ImageNet 규모에서는 수백 GPU-일이 필요하다. 또한 특정 공격에 특화된 적대적 학습은 해당 공격에는 강인하지만 다른 공격에는 취약하다. 이를 보완하기 위해 Tramèr, Florian, et al. [28]은 여러 공격을 조합하는 멀티 스텝 적대적 학습을 제안했고, Cai, Qi-Zhi, et al. [29]은 난이도를 점진적으로 높이는 커리큘럼 기반 적대적 학습을 개발했다. Zhang, Hongyang, et al. [30]의 TRADES는 자연 정확도와 견고성의 균형을 최적화하는 방법을 제시했다.적대적 학습 평가에서 핵심 함정은 기울기 은닉(Gradient Masking) 문제다. 일부 방어 기법은 모델의 기울기를 부정확하게 만들어 공격자가 교란을 계산하지 못하도록 막지만, 실제 견고성은 없다. Athalye, Anish, et al. [31]는 ICLR 2018의 7개 방어 기법 중 6개가 기울기 은닉에 의존하며, BPDA와 같은 적응형 공격으로 쉽게

무력화됨을 보였다. 따라서 견고성 평가는 Croce, Francesco, and Matthias Hein [32]의 AutoAttack처럼 4가지 공격을 자동 조합하는 표준 벤치마크와 다양한 교란 범위에서의 테스트를 포함해야 한다. RobustBench [33]는 ImageNet과 CIFAR-10/100에서 적대적 학습 기법들의 표준화된 벤치마크를 제공해 연구자들이 공정하게 비교할 수 있도록 돕는다.

3.3 인증 가능한 방어와 도메인 특화 견고성

인증 가능한 방어(Certified Defense)는 적대적 공격에 대한 안정성을 수학적으로 보장하는 접근이다. 입력에 허용되는 교란의 범위를 미리 정의하고, 그 범위 내에서는 모델의 출력이 변하지 않거나 일정 확률 이상으로 올바른 예측을 한다는 사실을 증명한다. 이를 위해 모델을 단순한 구조로 근사하거나, 입력과 은닉층의 출력에 대한 상한과 하한을 계산해 전파하는 방법이 사용된다.랜덤 스무딩(Randomized Smoothing)은 대표적인 인증 가능한 방어로, 입력에 무작위 노이즈를 여러 번 추가해 예측을 반복한 뒤 결과를 평균 낸다. Cohen, Jeremy, et al. [34]는 가우시안 노이즈를 사용한 랜덤 스무딩으로 ImageNet에서 $\ell 2$ 교란 반경 1.0 이내의 견고성을 수학적으로 보장했다. 평균 결과가 특정 클래스에 대해 충분히 높은 확률을 보이면, 일정 범위 이내의 교란에 대해 그 클래스가 변하지 않음을 증명할 수 있다. 다른 접근으로는 신경망의 각 층을 선형 제약 조건으로 근사해 전체 모델을 수학적으로 분석 가능한 형태로 만드는 방법이 있다. Singh, Gagandeep, et al. [35]의 ERAN은 추상 해석(Abstract Interpretation)을 사용해 신경망의 출력 범위를 엄밀하게 계산하고, 주어진 입력 교란 범위 내에서 오분류가 발생하지 않음을 증명한다.도메인 특화 견고성 강화는 음성, 텍스트, 멀티모달 등 각 응용 분야의 특성을 반영한 방어 기법이다. 음성 인식에서는 인간의 청각 특성을 고려해 특정 주파수 대역의 노이즈를 제거하거나 스펙트럼을 정규화하는 전처리 기법이 연구되고 있다. Yang, Zhuolin, et al. [36]는 멜 스펙트로그램 기반 필터링으로 음성 인식 모델의 적대적 공격 성공률을 70%에서 15%로 감소시켰다. 텍스트 분야에서는 의미를 크게 바꾸지 않는 범위에서 문장을 수정해도 모델의 출력이 안정적으로 유지되도록 학습하는 방법이 제안된다. Jones, Erik, et al. [37]의 SAFER는 동의어 치환과 문장 재구성에 강인한 텍스트 분류기를 학습해 TextFooler 공격에 대한 정확도를 62%에서 89%로 향상시켰다. 멀티모달 모델에서는 한 모달리티가 공격받더라도 다른 모달리티가 이를 보완하도록 학습해 전체 시스템의 안정성을 높인다. Ilharco, Gabriel, et al. [38]은 비전-언어 모델에서 이미지가 적대적 공격을 받아도 텍스트 모달리티가 올바른 예측을 유지하도록 하는 크로스-모달 주의(Cross-Modal Attention) 메커니즘을 제안했다.

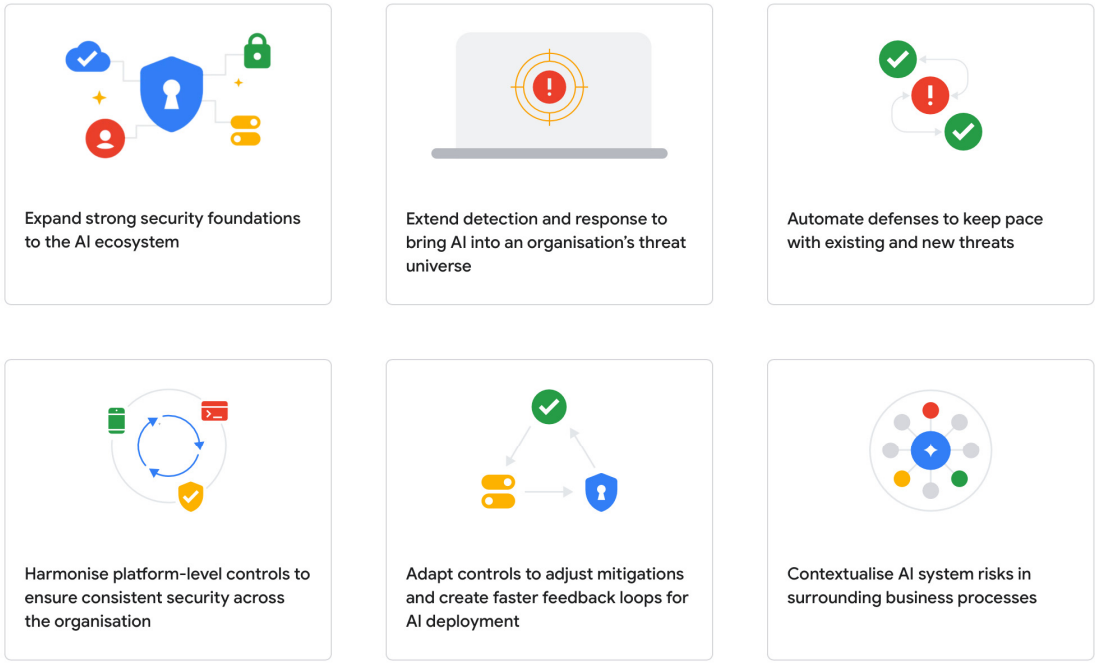
3.4 데이터와 프라이버시 보호를 위한 학습 기법

데이터와 프라이버시 보호는 법규 준수와 사회적 신뢰 확보 차원에서 필수적 요소가 되었다. 연합학습(Federated Learning)은 데이터를 각 기관과 단말에 남겨 둔 채 모델만 공유하는 구조로, 개인정보 보호 규제가 엄격한 환경에서도 인공지능을 활용할 수 있는 길을 열었다. McMahan, Brendan, et al. [39]은 FedAvg 알고리즘으로 수백만 모바일 기기에서 키보드 예측 모델을 중앙 집중식 학습과 유사한 성능으로 학습했다. 그러나 연합학습만으로는 모든 프라이버시 위협을 제거할 수 없으며, 추가적인 보호 기법이 필요하다. 차등 프라이버시(Differential Privacy) 기반 학습은 개별 데이터 포인트가 학습 결과에 미치는 영향을 통계적으로 제한한다. 학습 과정에서 미니배치의 그래디언트를 계산한 후 일정 크기 이상인 값을 잘라내고, 여기에 무작위 노이즈를 더해 업데이트를 수행한다. Abadi, Martin, et al. [40]의 DP-SGD는 MNIST에서 ε=8 프라이버시 예산으로 97% 정확도를 달성하며, 멤버십 추론 공격 성공률을 90%에서 55%로 감소시켰다. 이렇게 학습된 모델은 특정 데이터 포함 여부에 따른 출력 차이가 작아져 멤버십 추론 공격에 강인해진다. 다만 노이즈 크기를 크게 설정할수록 모델 성능이 떨어지므로 도메인과 목적에 따라 적절한 균형점을 찾아야 한다. Papernot, Nicolas, et al. [41]는 PATE(Private Aggregation of Teacher Ensembles)로 여러 교사 모델의 예측을 차등 프라이버시 방식으로 집계해 학생 모델을 학습시키고, SVHN에서 ε=0.2로 90%의 정확도를 달성했다. 안전한 집계 프로토콜(Secure Aggregation Protocol)은 연합학습에서 중앙 서버가 개별 업데이트를 직접 보지 못한 채 합계만 계산하도록 만드는 기술이다. Bonawitz, Keith, et al. [42]은 각 클라이언트가 비밀 분산(Secret Sharing)과 마스킹을 사용해 업데이트를 암호화하고, 서버는 이들의 합만 복원하는 프로토콜을 제안했다. 이를 통해 서버 운영자가 악의적이거나 침해되더라도 개별 클라이언트의 데이터에 대한 직접적 정보를 얻기 어렵다. Bell, James Henry, et al. [43]은 동형 암호(Homomorphic Encryption)를 사용해 서버가 암호화된 상태로 모델 업데이트를 집계하는 방법을 제시했으나, 계산 비용이 100배 이상 증가하는 한계가 있다.

3.5 생성형 인공지능과 대규모 언어모델 보안 방법론

생성형 인공지능과 대규모 언어모델의 보안은 프롬프트, 출력, 아키텍처, 거버넌스 수준에서 동시에 고려해야 한다. 프롬프트 수준에서는 시스템 프롬프트와 사용자 입력, 외부 문서 컨텍스트를 명확히 구분하고, 사용자 입력이 시스템 지침을 덮어쓰지 못하도록 설계한다. Hines, Kai, et al. [44]는 “구분자 기반 프롬프트 구조”로 시스템 지침을 특수 토큰으로 감싸고, 모델이 이를

사용자 입력보다 우선 참조하도록 학습시켜 프롬프트 인젝션 성공률을 78%에서 12%로 감소시켰다. 시스템 프롬프트에는 모델이 반드시 지켜야 할 보안 정책과 금지된 행위를 상세히 기술하고, Constitutional AI [45] 방식처럼 원칙 기반 지침을 응답 생성 과정에 통합한다.출력 수준에서는 모델 응답을 추가 검토하는 필터링 단계가 필요하다. OpenAI의 Moderation API [46]는 전용 분류기로 응답이 폭력, 혐오, 성적 콘텐츠를 포함하는지 검사해 99%의 정확도로 차단한다. Markov, Todor, et al. [47]는 LLM 기반 자기 검토(Self-Critique) 메커니즘으로 모델이 자신의 출력을 평가하고 정책 위반 시 응답을 수정하도록 했다. 코드 생성 서비스에서는 생성된 코드에 악성 행동이나 취약점이 포함되었는지 자동 검사하는 도구 연계가 중요하다. Pearce, Hammond, et al. [48]은 GitHub Copilot이 생성한 코드의 40%가 CWE(Common Weakness Enumeration) 취약점을 포함함을 보고하며, 정적 분석 도구 통합의 필요성을 강조했다.아키텍처 수준에서는 언어모델이 호출할 수 있는 도구에 최소 권한 원칙을 적용한다. 파일 시스템 접근, 데이터베이스 질의, 이메일 발송, 외부 API 호출 등 시스템 상태를 변경하거나 민감 정보를 다루는 기능은 기본 비활성화하고, 필요한 범위에서만 세밀하게 허용한다. Anthropic의 Constitutional AI [45]는 모델이 도구 사용 전 자체 평가를 수행해 잠재적 위험을 판단하도록 한다. 고위험 작업에는 사람이 최종 확인을 거쳐야만 실행되는 이중 승인 절차를 도입한다. Google의 Secure AI Framework [49]는 금융 거래나 시스템 설정 변경 같은 고위험 작업은 LLM이 제안만 하고 사람이 승인하도록 강제한다.거버넌스 수준에서는 생성형 인공지능 사용의 역할과 책임, 허용 및 금지 방식을 명확히 정의한다. 조직은 내부 정책으로 인공지능 서비스에 입력 가능한 데이터와 위임 가능한 업무의 기준을 제시하고, 교육과 지침으로 구성원의 이해를 돕는다. NIST AI Risk Management Framework [50]는 조직이 생성형 AI 사용 시 위험 평가, 책임 할당, 사고 대응 계획을 수립하도록 권고한다. 프롬프트와 응답을 적절히 로그로 남겨 사고 발생 시 조사와 개선에 활용한다. Microsoft [51]는 Azure OpenAI Service에서 모든 프롬프트와 응답을 암호화 저장하고, 30일 보관 후 자동 삭제하는 정책을 시행한다.



[그림 5] 구글에서 다층 LLM 보안 아키텍처 (출처: [49])

3.6 공급망 보안과 보안 중심 MLOps

인공지능 시스템은 수많은 오픈소스 라이브러리, 프레임워크, 외부 데이터셋, 사전학습 모델, 클라우드 인프라에 의존한다. 이러한 구성 요소 중 하나라도 악성 코드나 취약한 설정을 포함하면 전체 시스템이 공격에 노출된다. Liang, Weishen, et al. [52]은 Hugging Face에 업로드된 모델의 15%가 악성 pickle 파일을 포함하거나 의심스러운 가중치 패턴을 보였음을 보고했다. 인공지능 공급망 전반에 대한 보안 관점의 관리가 필요하다.조직은 어떤 데이터, 코드, 모델이 어떤 버전의 시스템에 사용되었는지 추적 가능한 형태로 자산을 관리해야 한다. Model Card [53]와 Datasheet for Datasets [54]는 모델과 데이터셋의 출처, 학습 방법, 알려진 제약을 문서화하는 표준 방식을 제공한다. SLSA(Supply chain Levels for Software Artifacts) [55]는 빌드 출처 검증, 무결성 확인, 재현 가능한 빌드를 통해 공급망 공격을 방지하는 프레임워크다. 데이터셋과 모델, 파이프라인 구성 요소에 고유 식별자를 부여하고 버전 관리 시스템과 연동해 변화 이력을 기록한다. 모델을 빌드하고 배포하는 과정에서 코드 서명과 무결성 검증을 수행하고, 외부에서 가져온 모델과 데이터는 별도 검증 절차를 거친다. Sigstore [56]는 모델 아티팩트에 서명하고 투명성 로그를 관리해 무결성을 보장하는 오픈소스 도구다.보안 중심 MLOps는 개발, 배

포, 운영 전 과정에 보안 점검과 위험 평가를 통합하는 접근이다. 모델이 변경될 때마다 자동으로 적대적 공격, 프롬프트 인젝션, 취약한 설정에 대한 테스트를 수행하고, 결과에 따라 배포를 승인하거나 보완을 요구한다. Shankar, Shreya, et al. [57]는 Stanford의 MLOps 파이프라인에서 모델 배포 전 자동화된 보안 게이트를 도입해 취약점 탐지율을 85%에서 98%로 향상시켰다. 운영 중인 모델에서 비정상적인 추론 패턴이나 로그가 관찰되면 자동으로 탐지해 경고를 발생시키고, 필요 시 이전 버전으로 안전하게 롤백한다. Arize AI [58]와 Fiddler AI [59]는 프로덕션 ML 모델의 성능 저하, 데이터 드리프트, 이상 추론 패턴을 실시간 모니터링하는 플랫폼을 제공한다. Google [60]은 Vertex AI에서 모델 버전 관리와 자동 롤백 기능을 통합해, 이상 탐지 시 30초 이내에 이전 안정 버전으로 복구한다.

[표 2] 주요 방어 기법 비교

방어 기법	방어 대상	보장 수준	성능 영향	적용 단계	대표 도구/기법
적대적 학습	적대적 사례	경험적	학습시간 5-10배 증가	학습	PGD, TRADES [27], [30]
랜덤 스무딩	적대적 사례	수학적 인증	추론시간 증가	추론	[34]
차등 프라이버시	프라이버시	수학적 보장	정확도 감소	학습	DP-SGD, PATE [40], [41]
안전한 집계	프라이버시	암호학적	통신 비용 증가	학습	Secret Sharing [42]
프롬프트 구조화	프롬프트 인젝션	경험적	낮음	설계/운영	구분자 기반 [44]
출력 필터링	유해 콘텐츠	경험적	지연시간 증가	추론	Moderation API [46]
MLOps 모니터링	드리프트/이상	경험적	인프라 비용	운영	Arize, Fiddler [58], [59]

4. 국내 적용 동향과 시사점

국내에서도 인공지능 보안의 중요성을 인식하고 다양한 정책과 기술적 노력이 진행되고 있다. 공공부문에서는 여러 부처와 기관이 공통으로 활용할 수 있는 초거대 인공지능 플랫폼을 구축하여, 외부 인터넷과 분리된 상태에서 행정 데이터를 안전하게 활용하려는 시도가 이어지고 있다. 과학기술정보통신부는 2023년 국가 초거대 AI 컴퓨팅 인프라를 구축하고, 공공 데이터를 활용한 한국어 특화 언어모델 개발을 추진했다. 이러한 공통 기반은 각 기관이 개별적으로 인공지능 시스템을 구축하는 것보다 보안 측면에서 유리하다. 동일한 보안 정책과 감사 체계를 적용할 수 있고, 모델과 데이터가 이동하는 경로를 중앙에서 통합적으로 관리할 수 있기 때문이다. 국내 보안 관련 기관과 학회에서는 인공지능 서비스의 보안 요구사항을 정리한 여러 가이드라인을 발표

하고 있다. 한국인터넷진흥원(KISA)은 2024년 “생성형 AI 보안 가이드라인”을 발간해 프롬프트 인젝션 방어, 출력 필터링, 데이터 유출 방지 등 구체적 실천 항목을 제시했다. 개인정보보호위원회는 "인공지능 개인정보보호 자율점검표"를 통해 개발자와 서비스 제공자가 데이터 보호, 접근 통제, 로그 관리, 제3자 연동 보안을 점검하도록 돕는다. 생성형 인공지능 서비스를 제공하는 기관은 프롬프트와 응답 로그에 포함될 수 있는 개인정보를 최소화하고, 장기간 보관 시 비식별화와 암호화 조치를 병행하도록 권고된다. 민간 영역에서도 인공지능 보안 솔루션과 컨설팅 수요가 증가하고 있다. 기존 보안관제센터는 네트워크와 시스템 로그 중심으로 이상 징후를 탐지했지만, 최근에는 인공지능 모델의 추론 로그, 프롬프트, 피드백 데이터를 함께 분석하는 시도가 늘어나고 있다. 국내 주요 보안 기업들은 인공지능 모델을 이용해 공격 패턴을 자동 분류하고, 또 다른 인공지능 모델을 감시해 적대적 사례와 프롬프트 인젝션을 탐지하는 서비스를 개발하고 있다. 금융권에서는 금융보안원의 권고에 따라 AI 기반 이상거래 탐지 시스템에 대한 적대적 공격 시뮬레이션을 정기적으로 수행하는 사례가 증가하고 있다.조직 차원에서는 인공지능 보안을 기존 정보보호 관리체계에 단순히 덧붙이는 수준을 넘어, 데이터 전략과 인공지능 전략의 핵심 요소로 통합하는 노력이 필요하다. 최고정보보호책임자(CISO), 최고데이터책임자(CDO), 인공지능 책임자 사이의 협업 구조를 마련하고, 주요 인공지능 프로젝트에는 설계 초기 단계부터 보안 담당자가 참여해 위협 모델링과 위험 평가를 수행해야 한다. 인공지능 시스템에 특화된 사고 대응 계획을 마련해, 적대적 공격, 데이터 유출, 프롬프트 인젝션 등 새로운 유형의 사고 발생 시 신속하게 원인을 분석하고 재발을 방지할 수 있도록 준비해야 한다. 국내 정보보호 관리체계(ISMS-P) 인증에서도 2024년부터 인공지능 시스템에 대한 위험 관리 항목이 추가되어, 조직의 체계적인 대응을 유도하고 있다.

5. 결론 및 향후 과제

인공지능 보안은 짧은 시간 안에 적대적 사례 연구에서 출발해 데이터 포이즈닝, 백도어, 모델 도난, 프라이버시 공격, 생성형 인공지능 보안, 공급망 보안, 위험관리 프레임워크 등 매우 넓은 영역을 포괄하는 분야로 성장했다. 그러나 해결되지 않은 과제는 여전히 많다. 공격자와 방어자는 서로의 기술을 참고하며 동시에 발전하고 있으며, 새로운 모델 구조와 응용 환경이 등장할 때마다 예상하지 못한 취약점이 드러난다.향후 연구에서는 실제 공격자의 능력과 동기를 고려한 현실적 위협 모델과 평가 방법을 정립해야 한다. 지금까지 많은 연구가 이론적으로 강력한 공격을

가정해 방어 기법을 비교했지만, 실제 환경에서는 공격자의 자원과 접근 권한이 제한된다. 특정 도메인에서 어떤 공격이 실제로 가능한지, 방어 기법이 어느 정도의 비용으로 어느 정도의 위험을 줄이는지를 실증적으로 분석해야 한다. Carlini, Nicholas, et al. [61]는 실험실 수준의 공격과 실제 시스템에서의 공격 사이에 큰 격차가 있음을 지적하며, 현실적 위협 모델의 필요성을 강조했다.성능과 보안, 프라이버시와 효율성 사이의 균형을 정량적으로 이해하고 설계에 반영하는 방법도 필요하다. 적대적 학습과 차등 프라이버시, 보안 추론 같은 기법은 보안 측면에서 이점을 제공하지만, 정확도, 처리량, 응답 시간에 영향을 미친다. Jagielski, Matthew, et al. [62]은 차등 프라이버시 적용 시 프라이버시 예산(ε)과 모델 정확도 사이의 트레이드오프를 정량화하고, 도메인별 최적 설정을 제시했다. 이러한 트레이드오프를 도메인별로 체계화하고, 조직이 자신의 위험 허용 수준과 비즈니스 요구에 맞춰 적절한 조합을 선택할 수 있도록 돕는 연구가 필요하다. 기술적 방어와 더불어 조직 문화와 프로세스의 변화가 병행되어야 한다. 인공지능 보안은 개발 후 별도로 추가하는 기능이 아니라, 설계 초기부터 고려해야 하는 기본 요구사항이다. McGraw, Gary, et al. [63]은 "Security by Design" 원칙을 ML 시스템에 적용해, 데이터 수집부터 배포까지 각 단계에서 보안을 기본값으로 삼는 방법론을 제안했다. 데이터 수집, 모델 개발, 배포, 운영의 전 과정에서 보안을 기본값으로 삼는 설계 문화가 자리 잡을 때, 인공지능 기술이 사회 전반에 안전하고 책임감 있게 정착할 수 있다. 이 글이 인공지능 보안에 관심을 가진 독자에게 수명주기 전체를 관통하는 관점을 제공하고, 향후 연구와 실무 논의를 위한 기초 자료로 활용되기를 기대한다.

6. 참고문헌

[1] Biggio, Battista, et al. "Poisoning attacks against support vector machines." International Conference on Machine Learning (ICML). PMLR, 2012.

[2] Gu, Tianyu, et al. "BadNets: Identifying vulnerabilities in the machine learning model supply chain." arXiv preprint arXiv:1708.06733 (2017).

[3] Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE (2017): 3-18.

[4] Carlini, Nicholas, et al. "Extracting training data from large language models." 30th USENIX Security Symposium (USENIX Security 21) (2021): 2633-2650.

[5] Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." Advances in Neural Information Processing Systems 32 (2019): 14747-14756.

[6] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." International Conference on Learning Representations (ICLR) (2015).

[7] Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018): 1625-1634.

[8] Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." 2018 IEEE Security and Privacy Workshops (SPW). IEEE (2018): 1-7.

[9] Jin, Di, et al. "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment." Proceedings of the AAAI Conference on Artificial Intelligence 34.05 (2020): 8018-8025.

[10] Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016): 1528-1540.

[11] Tramèr, Florian, et al. "Stealing machine learning models via prediction APIs." 25th USENIX Security Symposium (USENIX Security 16) (2016): 601-618.

[12] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (2017): 506-519.

[13] Oh, Seong Joon, et al. "Towards reverse-engineering black-box neural networks." Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer (2019): 121-144.

[14] Gehr, Timon, et al. "AI2: Safety and robustness certification of neural networks with abstract interpretation." 2018 IEEE Symposium on Security and Privacy (SP). IEEE (2018): 3-18.

[15] Jagielski, Matthew, et al. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning." 2018 IEEE Symposium on Security and Privacy (SP). IEEE (2018): 19-35.

[16] Kumar, Ram Shankar Siva, et al. "Adversarial machine learning-industry perspectives." 2020 IEEE Security and Privacy Workshops (SPW). IEEE (2020): 69-75.

[17] Perez, Fábio, and Ian Ribeiro. "Ignore previous prompt: Attack techniques for language models." arXiv preprint arXiv:2211.09527 (2022).

[18] Wei, Alexander, et al. "Jailbroken: How does LLM safety training fail?." Advances in Neural Information Processing Systems 36 (2024).

[19] Kang, Daniel, et al. "Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks." 2024 IEEE Symposium on Security and Privacy (SP). IEEE (2024): 1157-1174.

[20] MITRE Corporation. "ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)." https://atlas.mitre.org/ (2023).

[21] OWASP Foundation. "OWASP Machine Learning Security Top 10." https://owasp.org/www-

project-machine-learning-security-top-10/ (2023).

[22] Nicolae, Maria-Irina, et al. "Adversarial robustness toolbox v1.0.0." arXiv preprint arXiv:1807.01069 (2018).

[23] Papernot, Nicolas, et al. "Technical report on the CleverHans v2.1.0 adversarial examples library." arXiv preprint arXiv:1610.00768 (2018).

[24] Microsoft. "PyRIT – Python Risk Identification Toolkit for generative AI." https://github.com/Azure/PyRIT (2024).

[25] Pang, Ren, et al. "Red teaming language models with language models." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022): 3419-3448.

[26] Brundage, Miles, et al. "Toward trustworthy AI development: Mechanisms for supporting verifiable claims." arXiv preprint arXiv:2004.07213 (2020).

[27] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." International Conference on Learning Representations (ICLR) (2018).

[28] Tramèr, Florian, et al. "Ensemble adversarial training: Attacks and defenses." International Conference on Learning Representations (ICLR) (2018).

[29] Cai, Qi-Zhi, et al. "Curriculum adversarial training." Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI) (2018): 3740-3747.

[30] Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." International Conference on Machine Learning (ICML). PMLR (2019): 7472-7482.

[31] Athalye, Anish, Nicholas Carlini, and David Wagner. "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples." International Conference on Machine Learning (ICML). PMLR (2018): 274-283.

[32] Croce, Francesco, and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks." International Conference on Machine Learning (ICML). PMLR (2020): 2206-2216.

[33] Croce, Francesco, et al. "RobustBench: a standardized adversarial robustness benchmark." arXiv preprint arXiv:2010.09670 (2020).

[34] Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing." International Conference on Machine Learning (ICML). PMLR (2019): 1310-1320.

[35] Singh, Gagandeep, et al. "An abstract domain for certifying neural networks." Proceedings of the ACM on Programming Languages 3.POPL (2019): 1-30.

[36] Yang, Zhuolin, et al. "Characterizing audio adversarial examples using temporal dependency." International Conference on Learning Representations (ICLR) (2019).

[37] Jones, Erik, et al. "Certified robustness to adversarial word substitutions." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020): 4129-4142.

[38] Ilharco, Gabriel, et al. "Editing models with task arithmetic." International Conference on

Learning Representations (ICLR) (2023).

[39] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial Intelligence and Statistics (AISTATS). PMLR (2017): 1273-1282.

[40] Abadi, Martin, et al. "Deep learning with differential privacy." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016): 308-318.

[41] Papernot, Nicolas, et al. "Scalable private learning with PATE." International Conference on Learning Representations (ICLR) (2018).

[42] Bonawitz, Keith, et al. "Practical secure aggregation for privacy-preserving machine learning." Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (2017): 1175-1191.

[43] Bell, James Henry, et al. "Secure single-server aggregation with (poly) logarithmic overhead." Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (2020): 1253-1269.

[44] Hines, Kai, et al. "Defending against indirect prompt injection attacks with spotlighting." arXiv preprint arXiv:2403.14720 (2024).

[45] Bai, Yuntao, et al. "Constitutional AI: Harmlessness from AI feedback." arXiv preprint arXiv:2212.08073 (2022).

[46] OpenAI. "Moderation API documentation." https://platform.openai.com/docs/guides/moderation (2023).

[47] Markov, Todor, et al. "A holistic approach to undesired content detection in the real world." Proceedings of the AAAI Conference on Artificial Intelligence 37.12 (2023): 15009-15018.

[48] Pearce, Hammond, et al. "Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions." 2022 IEEE Symposium on Security and Privacy (SP). IEEE (2022): 754-768.

[49] Google. "Secure AI Framework (SAIF)." https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/ (2024).

[50] National Institute of Standards and Technology (NIST). "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." (2023).

[51] Microsoft. "Azure OpenAI Service data, privacy, and security." https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy (2023).

[52] Liang, Weishen, et al. "Careful with that scalpel: Improving gradient attack efficacy and efficiency in federated learning." arXiv preprint arXiv:2210.14963 (2022).

[53] Mitchell, Margaret, et al. "Model cards for model reporting." Proceedings of the Conference on Fairness, Accountability, and Transparency (2019): 220-229.

[54] Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92.

[55] SLSA (Supply chain Levels for Software Artifacts). "SLSA Framework v1.0." https://slsa.

dev/ (2023).

[56] Sigstore. "Sigstore: A new standard for signing, verifying and protecting software." https://www.sigstore.dev/ (2023).

[57] Shankar, Shreya, et al. "Operationalizing machine learning: An interview study." arXiv preprint arXiv:2209.09125 (2022).

[58] Arize AI. "ML Observability Platform." https://arize.com/ (2024).

[59] Fiddler AI. "Fiddler AI Observability Platform." https://www.fiddler.ai/ (2024).

[60] Google Cloud. "Vertex AI Model Monitoring." https://cloud.google.com/vertex-ai/docs/model-monitoring (2023).

[61] Carlini, Nicholas, et al. "On evaluating adversarial robustness." arXiv preprint arXiv:1902.06705 (2019).

[62] Jagielski, Matthew, et al. "Differentially private fair learning." International Conference on Machine Learning (ICML). PMLR (2019): 3000-3008.

[63] McGraw, Gary, Harold Figueroa, and Richie Bonett. "An architectural risk analysis of machine learning systems: Toward more secure machine learning." Berryville Institute of Machine Learning (2020).

주제원고

AX 시대의 중국 인공지능과 보안 융합 전략 분석

호남대학교 명예교수
이양원

1. 서론

디지털 전환(DX)을 넘어 인공지능이 사회와 경제의 새로운 기반 기술로 자리 잡은 AX (Artificial Intelligence Transformation) 시대가 도래했다. 이 시대에서 국가와 기업의 경쟁력을 결정하는 것은 단순히 인공지능 기술의 정확도나 속도가 아니다. 진정한 경쟁력은 인공지능의 발전과 사이버 보안의 진화가 상호 선순환하는 '공진화(Co-evolution)' 관계를 어떻게 구축하느냐에 달려 있다.

중국은 이러한 AX 시대의 도전에 대해 매우 체계적이고 전략적인 접근법을 보여주고 있다. 그들은 AI와 보안을 별개의 과제로 보지 않는다. 오히려, AI는 보안을 혁신하는 핵심 동력이 되고(AI-for-Security), 동시에 AI 시스템 그 자체는 국가적 차원에서 반드시 지켜내야 할 핵심 자산이 된다(Security-for-AI). 이 두 가지 흐름이 서로를 추동하며, 하나의 강력한 생태계를 형성하고 있는 것이다.

본 논고에서는 이러한 '공진화'의 구체적인 양상을 다음 두 가지 측면에서 심층적으로 분석해 보겠다.

첫째, 'AI-for-Security'의 관점에서, 중국이 어떻게 인공지능을 활용하여 기존의 보안 패러다임을 근본적으로 전환하고 있는지 살펴볼 것이다. 단순한 패턴 매칭을 넘어 '행위 분석' 을 통한 예측형 위협 탐지, 그리고 '자동화된 대응'으로 이어지는 지능형 사이버 방어 체계가 어떻게 구축되고 있는지 그 기술적 세부 사례를 파헤칠 것이다.

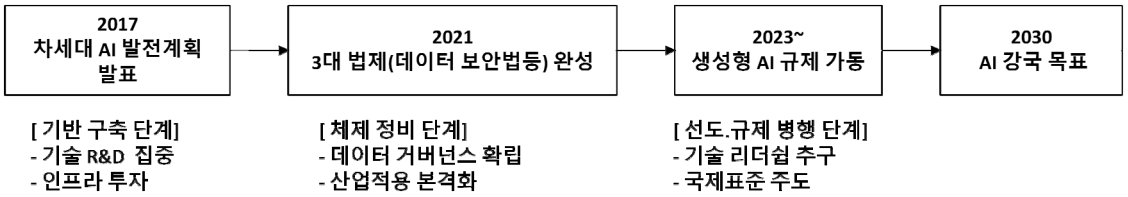
둘째, 'Security-for-AI'의 관점에서, AI라는 새로운 국가 전략 자산을 어떻게 보호하고 있는지 조명할 것이다. AI 모델을 속이는 '적대적 공격'으로부터의 방어, 기업의 핵심 지식재산인 AI 모델을 지키는 '보안 기술', 그리고 데이터 프라이버시를 보장하는 '연합학습'의 안전성 확보 등, AI 생태계의 안정성을 뒷받침하는 기술 인프라를 면밀히 검토할 것이다.

이를 통해, 중국이 국가 주도로 추진하는 AI-보안 융합 전략이 단순한 기술 정책을 넘어, 국가 경쟁력의 중추를 이루는 포괄적인 거버넌스 체계임을 명확히 이해할 수 있을 것이다. 이같은 분석은 우리나라가 지속 가능하고 경쟁력 있는 AI 생태계를 구축하기 위해 필요한 전략적 통찰을 제공하는데 이바지 할 것이다.

2. 중국의 국가 전략과 AI-보안 정책 프레임

2.1. 정책 발전 단계 구체화

중국의 AI와 보안 융합 전략은 단순한 기술 발전을 넘어, 국가 주도의 체계적인 거버넌스 구축 과정을 보여주고 있다. 이는 크게 세 단계에 걸쳐 진화해 왔으며, 각 단계는 [그림 1]과 같이 뚜렷한 목표와 전략적 중점을 가지고 있다[1]. 또한 이와 같은 3단계 발전 모델은 중국이 AI와 보안을 국가 차원에서 어떻게 전략적 자산으로 보고, 체계적으로 육성·통제·확장해 나가고 있는지를 명확하게 보여주고 있다.



[그림 1] 중국 AI-보안 정책 발전 로드맵(2017-2030)

가. 1단계 (2017-2020): 기반 구축 및 패권 확립 - "선점과 성장"

2017년 발표된 '차세대 인공지능 발전계획'은 중국이 본격적으로 AI 강국으로 도약하겠다는 의지를 천명한 출발점이었다. 이 단계의 최우선 목표는 기술적 기반을 마련하고 글로벌 AI 경쟁에서의 초기 주도권을 확보하는 것이었다. 중국 정부는 막대한 예산을 R&D에 투입하고, 바이두, 알리바바, 텐센트, 화웨이 등 '국가 대표' 기업들을 집중 육성하며 이들을 통해 AI 생태계의 기반을 구축했다. 특히 이 시기 제정된 「사이버보안법」은 국가 안보와 공공 이익의 이름으로 데이터에 대한 국가의 통제권을 공식화하는 초석이 되었다. 즉, 이 단계는 "일단 키우고 보자"는 식의 적극적 투자와 성장에 주력하며, AI 발전과 국가 안보를 연결하는 초기 법적·제도적 틀을 마련한 시기라고 할 수 있다.

나. 2단계 (2021-현재): 체계 정비 및 통제 강화 - "규제의 틀 안에서의 안전 발전"

기초 체력이 어느 정도 갖춰지자, 중국의 전략은 '성장'과 '통제'의 균형으로 무게중심이 이동했다. 「데이터보안법」과 「개인정보보호법」이 2021년 차례로 시행되며 '데이터 주권'의 개념이 본격적으로 법제화되었다. 이는 국가 중요 데이터를 체계적으로 분류하고, 데이터의 국외 이전을 엄격히 통제하겠다는 의지의 표현이었다. 또한, '14차 5개년 계획(2021-2025)' 에서 AI를 핵심

과제로 재확인하며, 단순한 기술 발전을 넘어 사회 전반에의 안전한 적용을 강조했다. 이러한 흐름의 정점에 2023년 발표된 「생성형 AI 관리 잠정 조치」가 있다. 이는 챗GPT와 같은 혁신 기술이 등장하자마자 이를 빠르게 규제의 틀에 집어넣어, 기술의 잠재적 사회적·정치적 리스크를 사전에 차단하려는 적극적인 행보를 보여주고 있다. 이 단계는 양적 성장에서 질적 관리로, 즉 “거친 성장”에서 "안전한 발전"으로의 전환을 의미하기도 한다.

다. 3단계 (미래): 글로벌 리더십 및 표준 주도 – “규칙을 만드는 게임의 주인공 되기”

현재 중국은 그 야망의 다음 단계, 즉 기술적 추격을 넘어 글로벌 규칙을 주도하는 단계를 준비하고 있다고 본다. 그 핵심 추진축이 바로 ‘중국 표준 2035’ 프로젝트이다. 이는 단순히 제품을 수출하는 것을 넘어, 중국의 기술 규격과 거버넌스 모델을 세계 표준으로 만드는 것을 최종 목표로 하고 있다. AI 윤리, 데이터 보안, 모델 안정성 등에 대한 중국의 자체 기준을 국제전기통신연합(ITU) 등 국제기구를 통해 전파하려는 노력등이 여기에 해당한다. 궁극적으로 이 단계는 AI와 보안 분야에서 미국 주도의 서양적 가치관(예: 개인정보보호, 알고리즘 투명성)과 대비되는 '중국식 모델'의 패러다임을 글로벌 무대에 정착시키는 것을 지향하고 있다. 이는 단순한 기술 경쟁을 넘어, 미래 디지털 세계 질서를 누가 주도할 것인지에 관한 패권 경쟁의 성격을 띠고 있기도 한다.

2.2. 핵심 법제의 구체적 요구사항 추가

중국의 AI-보안 융합 전략을 뒷받침하는 법적 기반은 단순한 원칙 선언을 넘어, 기업과 기관의 실제 데이터 처리 관행에 직접적이고 구속력 있는 영향을 미치는 구체적인 요구사항으로 진화해 왔다. 특히 《데이터보안법》[2]과 《개인정보보호법》[3]은 각각 '국가 안보 및 경제 발전'과 '개인 권리 보호'라는 축을 중심으로 상호 연계되어 작동하며, 중국식 데이터 거버넌스의 실질적인 뼈대를 구성하고 있다.

가. 《데이터보안법》: 국가 주권의 관점에서 데이터를 통제하다

《데이터보안법》의 핵심은 국가 안보와 공공 이익에 직결된 데이터를 체계적으로 분류하고 관리하는 데 있다. 이 법은 기업에게 단순히 데이터를 보호하라는 모호한 의무를 부과하는 수준을 넘어, 다음과 같은 매우 구체적인 실행 과제를 제시하고 있다.

- **‘중요 데이터’의 구체적 식별 및 관리:** 법은 “국가 안보, 국민 경제, 사회 공익에 중대한 영향을 미칠 수 있는 데이터”를 ‘중요 데이터’로 규정하고 있다. 기업 및 기관은 자신들이 보유한

데이터가 해당 지역별로 발표된 '중요 데이터 구분 지침'에 따라 어디에 해당하는지 스스로 식별하고, 이에 대해 더 높은 수준의 보안 관리 체계(예: 암호화, 접근 통제, 감사 로그)를 구축해야 한다. 이는 단순한 기술적 조치가 아닌, 데이터의 국가적 가치에 따른 차등화된 관리를 의미하며, 기업에게는 막대한 부담으로 작용한다.

- **'데이터 국외 이전 안전 평가'의 의무화:** 이는 가장 특징적이고 영향력 있는 조항 중 하나이다. 중국 국내에서 수집·생성된 '중요 데이터'를 해외로 반출하려는 모든 기업은 반드시 국가사이버공간관리국(국가망)이 주관하는 '안전 평가'를 통과해야 한다. 당국은 데이터의 양, 민감도, 전송 목적, 해외 수신자의 보안 수준 등을 종합적으로 심사하여 국외 이전이 국가 안보와 공공 이익을 해치지 않는지 여부를 판단한다. 예를 들어, 중국 내에서 운영하는 다국적 기업이 글로벌 데이터 분석을 위해 중국 사용자 데이터를 해외 본사 서버로 전송하려면, 단순히 사용자 동의를 얻는 것만으로는 부족하며, 사전에 당국의 공식 승인을 받아야 하는 것 등이다. 이는 데이터의 물리적 위치에 국가 주권을 적용한 '데이터 현지화'의 강력한 형태로, 글로벌 비즈니스의 데이터 흐름에 있어서 중국을 하나의 독립된 '데이터 생태계'로 격상시키는 효과를 낸다.

나. 《개인정보보호법》: 개인 권리의 관점에서 정보를 보호하다

《개인정보보호법》은 《데이터보안법》과 맞물려, 개인 정보를 처리하는 모든 활동에 대해 엄격한 규칙을 부과하고 있다. 그 내용과 수준이 유럽의 GDPR과 매우 유사하여, 중국이 법적 측면에서 글로벌 스탠다드에 상응하고 있음을 보여준다.

- **'동의 획득'의 강화:** 법은 정보주체의 "명확한 동의"를 개인정보 처리의 가장 기본적인 법적 근거로 삼고 있다. 이는 모호하거나 일괄적인 동의가 아닌, 구체적이고 개별적인 사항에 대한 사전 동의를 요구한다. 특히 민감한 개인정보(생체정보, 종교신념 등)에 대해서는 별도의 명시적 동의를 받아야 한다. 또한, 사용자는 언제든지 동의를 철회할 수 있으며, 기업은 이에 따라 처리 행위를 중지해야 한다.
- **'목적 제한' 및 '최소한의 정보 수집' 원칙의 구체적 적용:** 기업은 "명확하고 합리적인 목적"을 가지고 개인정보를 수집해야 하며, 그 목적의 달성에 필요한 "최소한의 범위"를 넘어서는 정보를 수집해서는 안된다. 예를 들어, 단순한 배송 서비스를 위해 주민등록번호를 수집하는 것은 '최소한의 원칙'에 위배된다. 또한, 처음 수집 당시 명시한 목적 외에 다른 용도로 정보

를 사용하려면 다시 별도의 동의를 받아야 한다. 이는 기업이 데이터를 무분별하게 수집하고 마케팅 등 다른 목적으로 자의적으로 활용하는 관행에 제동을 거는 핵심 장치이다.

종합하면, 이 두 법률은 각각 다른 측면에서 상호 보완적으로 작용한다. 《데이터보안법》이 '국가'의 안전을 위협할 수 있는 데이터 흐름을 거시적으로 통제한다면, 《개인정보보호법》은 '개인'의 권리가 기업의 정보 활동으로부터 침해받는 것을 미시적으로 방지한다. 이를 통해 중국 정부는 '국가 안보-경제 발전-개인 권리'라는 복합적 가치를 모두 아우르는 포괄적인 디지털 통치 체제를 구축해 나가고 있는 것이다.

2.3. 새로운 정책 동향

2023년 8월부터 시행된 「생성형 인공지능 서비스 관리 잠정 조치」는 중국이 AI 규제의 최전선에서 어떠한 균형을 추구하는지를 보여주는 분수령과 같은 사건이기도 하다[4]. 이 법규는 챗GPT와 같은 생성형 AI 기술이 급속도로 확산하는 시점에서, 기술의 파괴적 잠재력을 억제하지 않은 채 장려하는 것이 아니라, '사전에 틀을 잡아가며' 발전을 유도하겠다는 중국 특유의 선제적·개입적 거버넌스 철학을 명확히 구현한 것으로 볼 수 있다[5].

가. 기술의 '안전성'을 국가가 보증하는 시스템: 사전 심의 의무화

이 조치의 가장 핵심적인 요구사항은 공공 인터넷을 통해 대중에게 생성형 AI 서비스를 제공하기 전에 반드시 '안전성 평가'를 실시하고 당국에 신고·승인을 받아야 한다는 점이다. 이는 단순한 형식절차가 아닌, 기술의 사회적 영향을 사전에 검증하는 실질적인 심의 제도이다.

구체적 심의 대상인 안전성 평가는 크게 두 가지 축을 중심으로 이뤄진다. 먼저 알고리즘 안정성이다. 이는 AI 모델이 '적대적 공격' 에 얼마나 견고한지, 편향된 출력을 내지 않도록 얼마나 잘 제어되었는지, 그리고 불법·유해 콘텐츠를 생성하는 것을 방지할 수 있는 기술적 능력을 갖췄는지를 점검한다. 다음으로는 콘텐츠 안전성이다. 이는 생성된 콘텐츠가 국가 안보를 위협하거나 사회적 불안을 조장하지 않으며, 사회주의 핵심 가치관에 부합하는지를 검토한다. 이는 기술적 결함 뿐 아니라 '정치적 정확성'까지를 평가 범위에 포함시킨 것이다.

나. '내용 심의'를 통한 담론 공간의 관리: 생성물에 대한 책임 소재 확립

법률은 서비스 제공자에게 생성된 모든 콘텐츠에 대한 법적 책임을 지우며, 이를 위해 '내용 심의' 시스템을 구축할 것을 의무화하고 있다. 이는 생성형 AI가 만들어내는 방대한 양의 텍스트,

이미지, 영상이 중국의 기존 인터넷 정보 관리 체계 내에 완전히 포획되도록 하기 위함이다. 실행 메커니즘을 보면 서비스 제공업체는 실시간으로 생성되는 콘텐츠를 필터링하고, 유해·불법 콘텐츠가 노출되기 전에 차단·삭제할 수 있는 기술적·인력적 체계(즉, AI 심의관)를 마련해야 한다. 이는 생성형 AI를 '가상의 콘텐츠 생산 공장'으로 보고, 이 공장의 품질 관리(QC) 라인에 정부의 검수 기준을 적용하는 것과 같은 원리이다.

다. 데이터의 질과 지식재산권: 발전의 기반을 통제하다

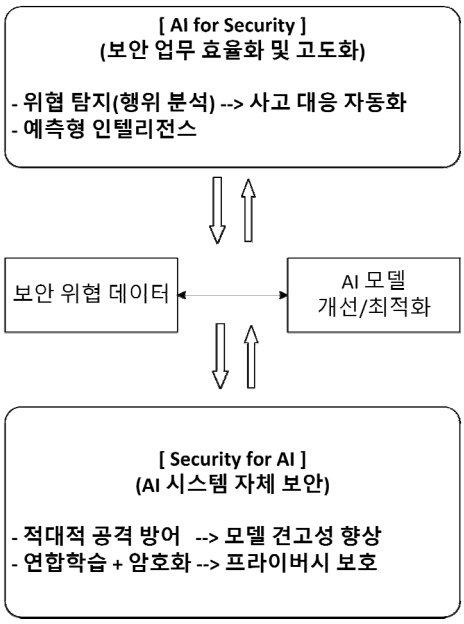
규제는 단순한 출력(output) 관리에 그치지 않고 입력(input) 단계로까지 그 범위를 확장하고 있다. 즉 훈련 데이터의 '정확성·객관성·다양성' 보장을 위한 법은 생성형 AI 모델을 훈련시키는 데 사용되는 데이터의 품질을 규제한다. 이는 AI가 편향되거나 허위 정보를 학습하는 것을 근본적으로 차단하여 출력의 신뢰성과 사회적 통제 가능성을 높이려는 의도이다. 또한 지식재산권 및 개인정보 보호 강조를 위해서는 타인의 지식재산권을 침해하지 않고, 개인정보를 불법적으로 처리하지 않도록 요구하고 있다. 이는 글로벌적으로 제기되는 생성형 AI의 주요 윤리적 쟁점을 중국식 프레임워크 안에 선제적으로 편입시킨 사례이기도 하다.

결론적으로 보면 중국은 기술 발전과 사회 안정 사이의 '중국식 균형'을 유지하면서 「생성형 AI 관리 잠정 조치」는 중국이 기술 발전의 패러다임을 '자유로운 실험과 사후 규제'가 아닌 '사전에 정의된 안전 경로 내에서의 관리된 혁신'으로 이해하고 있음을 보여준다고 볼 수 있다. 국가는 혁신의 촉진자이자 동시에 최고의 관리자 역할을 자처하며, 기술이 사회주의 통치 구조와 가치관에 도전하지 않는 선에서만 그 혜택이 발휘되도록 프레임을 설계하고 있다. 따라서 이 규제는 단순한 기술 규정을 넘어, AI가 생성하는 담론과 지식, 나아가 미래 사회의 인지적 구조 자체에 대한 국가 주도의 깊은 개입을 의미하며, 이는 중국이 추구하는 'AI 강국'의 모델이 지니는 본질적 성격을 함의하고 있다고 볼 수 있다.

3. 기술적 융합 : AI for Security 와 Security for AI

AX 시대의 핵심은 AI와 보안이 서로를 추동하며 발전하는 '공진화(Co-evolution)'에 있다. 이 복잡한 상호작용 관계를 이해하기 위해서는 단순한 순환 구조를 넘어, 실제로 기술과 데이터가 어떻게 흐르고 서로를 강화하는지를 살펴볼 필요가 있다. [그림 2]는 AI와 보안이 단순히 병렬적으로 존재하는 것이 아닌, 하나의 강력한 선순환 고리를 형성하며 상호 발전하는 과정을 보

여주고 있다. 이 두 축은 끊임없이 데이터와 인사이트를 주고받는다. AI-for-Security가 발견한 새로운 위협 유형과 공격 기법은, Security-for-AI 측면에서 AI 모델을 더욱 '단단하게' 만드는 데 필요한 소중한 훈련 데이터가 된다. 반대로, Security-for-AI를 통해 안전하고 견고해진 AI 모델은 AI-for-Security에 적용되어 더욱 정확하고 신뢰할 수 있는 보안 분석과 예측을 가능하게 한다. 즉, 보안은 AI를 통해 진화하고, AI는 보안을 통해 안전해지며, 이는 다시 더 진화된 보안으로 이어지는 선순환 구조가 완성되는 것이다. 이제부터 이 상호작용의 구체적인 메커니즘과 그 파급효과에 대해 자세히 알아보겠다.



[그림 2] AI for Security & Security for AI 상호작용 심화 구조

3.1. AI for Security

가. 행위 기반 이상 탐지(UEBA): 정상의 기준을 재정의하고 내부 위협을 색출하다

기존의 보안 시스템이 알려진 악성 코드의 '지문'이나 정적 패턴을 탐지하는 데 그쳤다면, 행위 기반 이상 탐지는 '정상적인 행위'가 무엇인지를 먼저 학습하여 그로부터 벗어난 미세한 이상 징후를 포착하는 혁신적인 접근법이다. 이 기술은 각 사용자와 시스템이 일상적으로 수행하는 행위

패턴(로그인 시간, 접근하는 데이터 양과 종류, 명령어 사용 빈도 등)을 지속적으로 분석하여 기초선(Baseline)을 생성한다. 이를 바탕으로, 평소와는 다른 시간대에 대량의 중요 데이터에 접근하거나, 권한이 없는 시스템을 스캔하는 등 '정상의 범위'를 벗어난 행위가 발생하면, 해당 행위가 합법적인 액세스인지 아니면 내부자의 불법 행위나 해커에 의한 침투인지를 실시간으로 판단한다. 이는 마치 은행에서 고객의 평소 소비 패턴을 학습해, 갑자기 발생한 대액 이체를 즉시 의심하고 확인하는 것과 유사하다. 결과적으로, 공격자가 정당한 자격증명을 탈취했다 하더라도 그 행위 자체가 비정상적이라면 탐지해 내는 방식으로, 특히 탐지하기 어려운 내부 위협과 고도화된 지속 공격(APT)에 대한 방어 능력을 극적으로 향상시킨다[7].

나. 예측형 위협 인텔리전스: 다크웹의 그림자에서 조짐을 포착하다

이 기술은 사이버 방어의 개념을 사후 대응에서 사전 예방으로 전환한다. 인공지능은 인간의 분석 능력으로는 감당하기 어려운 방대한 양의 오픈소스 정보(OSINT)—다크웹 포럼, 해킹 커뮤니티, SNS, GitHub 저장소 등—를 24시간 실시간 크롤링하고 분석한다. 목적은 조직의 이름, 인프라 기술 스택, 직원 이메일 등이 유출되거나 거래되는지, 또는 조직을 표적으로 삼는 공격 계획이 논의되고 있는지를 탐지하는 것이다. 예를 들어, AI가 다크웹에서 특정 기업의 내부 네트워크 도면이 유출되었거나, 해당 기업의 취약점을 악용하는 맞춤형 악성코드가 제작 중이라는 정보를 포착하면, 공격이 실제로 발생하기 전에 기업에 조기 경보를 발령하고 필요한 패치 또는 대비 조치를 취할 수 있는 소중한 시간을 제공한다. 이는 단순한 '정보 수집'을 넘어, 정보를 '전략적 예측'으로 전환하는 지능형 위협 헌팅(Threat Hunting)의 핵심 기술이다[13].

다. 자동화된 사고 대응(SOAR): 대응의 속도를 공격의 속도보다 빠르게 하다

SOAR는 수많은 보안 시스템에서 쏟아지는 경고와 사고 정보를 하나의 플랫폼에서 통합하고, 사전에 정의된 '플레이북(Playbook)'이라는 대응 매뉴얼에 따라 대응 조치를 자동으로 실행하는 시스템이다. 인공지능은 여기서 지휘관이자 실행부의 역할을 한다. 예를 들어, 어떤 직원의 계정에서 악성 IP로의 이상한 연결이 탐지되면, AI는 즉시 해당 플레이북을 작동시켜 다음과 같은 일련의 조치를 수초 내에 자동 수행할 수 있다. 1) 해당 사용자 계정의 임시 비활성화, 2) 관련 네트워크 세션 차단, 3) 담당 보안 담당자에게 통보 및 사고 티켓 생성, 4) 추가 분석을 위해 해당 엔드 포인트를 격리 등 이렇게 함으로써 보안팀은 반복적이고 시간 소모적인 작업에서 해방되어 더 복잡한 위협 분석에 집중할 수 있으며, 공격 초기 단계에서 신속하게 차단선을 구축하여 피해 규모를 최소화할 수 있다[16,17].

3.2. Security for AI

가. 적대적 공격(Adversarial Attacks) 방어: AI의 '착각'을 유도하는 교란 신호로부터 보호하기

AI 모델, 특히 이미지 인식 시스템은 인간의 눈으로는 구분할 수 없는 미세한 노이즈나 패턴이 추가되면 완전히 다른 판단을 내리는 취약점을 지니게된다. 대표적인 예로, 정지 신호판에 특수하게 제작된 스티커 몇 개를 붙이면 자율주행 차량의 AI가 이를 '속도 제한 표지판'으로 오인하도록 속일 수 있다. 적대적 공격 방어 기술은 바로 이러한 'AI 전용 속임수'를 탐지하고 무력화하는 것을 목표로 한다. 방법으로는, 모델을 훈련할 때 다양한 적대적 예제를 함께 투입하여 그런 공격에 '면역'을 키우는 적대적 훈련(Adversarial Training), 또는 입력 데이터의 미세한 특성을 분석하여 정상 입력과 적대적 입력을 구분하는 탐지 메커니즘을 구축하는 것 등이 있다. 이는 AI가 현실 세계에서 안전하고 신뢰할 수 있도록 하는 데 필요한 '방탄조끼'와 같은 기술이다[7].

나. 모델 역공학 방어: 기업의 핵심 자산인 AI 모델을 지키는 기술

완성된 AI 모델은 기업에게 막대한 자원을 투자하여 개발한 지식재산이다. 공격자는 이 모델에 반복적으로 질의를 보내고 그 출력값을 분석함으로써, 모델이 학습한 원본 데이터의 특징을 추론하거나, 더 나아가 모델의 내부 구조 자체를 복제하는 '모델 추출 공격'을 시도할 수 있다. 이를 방어하기 위한 두 가지 핵심 기술이 있습니다. 첫 번째는 차등 프라이버시(Differential Privacy) 기술로써 모델이 훈련 또는 추론 과정에서 개별 데이터의 특정 정보를 노출하지 않도록, 의도적으로 정해진 수준의 '잡음'을 추가하는 기술이다. 이는 "모델의 출력 결과가, 특정한 개인의 데이터가 훈련 세트에 있었는지 없었는지를 보여줄 수 없어야 한다"라는 원리를 기반으로 한다. 마치 많은 사람의 대화 소음 속에서 특정 한 사람의 목소리를 구분할 수 없는 것과 같은 원리이다. 두 번째는 모델 워터마킹(Model Watermarking) 기술이다. 이것은 모델 자체에 눈에 보이지 않는 전자 서명이나 패턴을 삽입하는 기술로서 이후 누군가 모델을 불법적으로 복제하여 사용할 경우, 해당 출력에서 워터마킹을 검출하여 지식재산권 침해 사실을 입증할 수 있다. 이는 소프트웨어의 시리얼 넘버나 영상의 디지털 워터마크와 유사한 개념으로, AI 모델의 소유권을 보호하는 데 필수적이다[18].

다. 연합학습(Federated Learning)의 안전성 강화: 데이터 프라이버시와 협력 학습의 동시 달성

연합학습은 스마트폰, 병원 등과 같은 여러 개별 기기(클라이언트)에서 데이터를 한데 모으지 않고, 각자의 데이터로 로컬 모델을 훈련한 후, 오직 모델의 '파라미터 업데이트'(학습 결과)만을

중앙 서버로 전송하여 글로벌 모델을 완성하는 방식이다. 이는 데이터 프라이버시를 보호하는 혁신적 방법이지만, 전송되는 모델 업데이트 자체가 원본 데이터의 정보를 일부 노출할 수 있는 새로운 취약점을 만들어낸다. 따라서 이를 보완하기 위해 '암호화 기술'이 결합한 것이다. 대표적으로 '동형암호(Homomorphic Encryption)' 데이터를 암호화된 상태에서도 계산을 수행할 수 있게 하여, 클라이언트가 중앙 서버로 보내는 모델 업데이트를 암호문 형태로 전송하고, 서버는 이 암호문을 그대로 집계하여 모델을 개선할 수 있다. 이 과정에서 서버는 개별 클라이언트의 업데이트 내용을 전혀 알 수 없게 되어, 프라이버시 보호와 협력 학습이라는 두 마리 토끼를 모두 잡을 수 있게 된다[14,20].

4. 산업 응용 사례 분석

중국이 국가 차원에서 구축해 온 AI-보안 융합 전략은 이제 다양한 산업 현장에서 가시적인 성과를 만들어내고 있다. 이는 단순한 기술 실험이 아닌, 국가 경쟁력의 중추를 이루는 핵심 인프라로 자리 잡고 있음을 보여준다. 각 산업별 적용 사례는 〈표 1〉에서와 같이 공통된 방향성을 공유하면서도, 산업의 특성에 따라 매우 차별화된 모습으로 진화하고 있는 것이 특징이다.

〈표 1〉 중국 주요 기업별 Ai-보안 적용 현황 상세 비교

분야	기업	주요 플랫폼/기술	적용 수준	주요성과 (정량적 지표)
스마트 시티	Huawei	City Intelligent Body	고도화	범죄 예방률 25% 향상
	Alibaba	ET City Brain	고도화	차량 정체 시간 15% 감소
금융	Ant Group	AlphaRisk	고도화	사기 거래 탐지 정확도 99.9%
	Ping An	Gamma AI	고도화	대출 심사 속도 80% 단축
의료	Tencent	Med AI Federated Platform	성장기	폐암 진단 민감도 95%
제조	Baidu	Apollo Shield	도입기	자율주행 시험 주행 안전사고 0건
공공안전	SenseTime	SenseFoundry	고도화	체포율 30% 향상

본 분석은 스마트시티, 금융, 의료, 제조, 공공안전이라는 5대 핵심 분야를 중심으로, AI-보안 융합이 어떻게 구현되고 있으며, 이로 인해 데이터 흐름과 산업 구조가 어떻게 재편되고 있는지를 조명한다.

4.1. 스마트시티: 통합 감시와 예측형 치안의 구현

스마트시티는 중국 AI-보안 융합의 총아이자, 그 위험과 효용이 가장 첨예하게 대립하는 공간이다. 화웨이(Huawei)의 'City Intelligent Body'와 알리바바(Alibaba)의 'ET City Brain'과 같은 플랫폼은 도시 전역의 CCTV, 교통 신호, SNS, 출입국 정보 등 방대한 데이터를 단일한 AI 분석 시스템으로 통합한다[12]. 이는 단순한 정보 연계를 넘어, 'AI-for-Security'의 정점을 보여주는 예측형 공공안전 체계로 작동한다.

구체적으로, AI는 과거 범죄 데이터와 실시간 인구 이동 패턴을 분석해 '범죄 다발 지역'을 예측하고, 이에 따라 경찰 순찰 경로를 최적화한다. 또한, 교통 흐름을 실시간으로 분석해 정체를 자동으로 해소하고, 실종자 발생 시 얼굴인식 기술을 동원해 시간과 공간을 넘나드는 이동 경로를 추적한다. 이러한 적용은 범죄 예방률을 25%가량 향상시키고 교통 정체를 줄이는 등 뚜렷한 효율성 개선을 가져왔으나, 동시에 'Security-for-AI'의 관점에서 이 대규모 감시 인프라와 민감한 데이터를 어떻게 보호할 것인지라는 중대한 과제를 제기한다. 스마트시티는 이러한 기술적 효율과 사회적 통제, 그리고 그 시스템 자체의 보안이 복잡하게 얽힌 공간이기도 하다.

4.2. 금융: 실시간 사기 탐지와 신용위험도 관리의 혁명

금융 분야에서는 'AI-for-Security'가 직접적인 경제적 손실을 방어하는 최전선의 방패로 자리 잡았다. 앤트그룹(Ant Group)의 'AlphaRisk' 플랫폼은 초당 수만 건의 마이크로 거래를 분석해, 기존 규칙 기반 시스템이 발견하지 못하는 미세한 이상 패턴을 포착한다. 이를 통해 신용카드 도용, 보이스피싱 등 사기 거래를 0.1초 이내에 차단하며 99.9%에 달하는 탐지 정확도를 보인다[16].

한편, 핑안(Ping An)의 'Gamma AI'는 AI를 신용평가와 리스크 관리에 적용하여 대출 심사 속도를 80%나 단축시키는 등 업무 효율을 극적으로 높였다[17]. 이러한 금융 AI의 핵심은 고객의 거래 데이터이며, 따라서 'Security-for-AI'의 적용은 절대적이다. 고객 정보와 AI 신용평가 모델은 해커의 일차적 표적이기 때문에, 적대적 공격으로부터 모델을 보호하고 연합학습 등 기술을 통해 데이터 프라이버시를 확보하는 것이 금융 AI 생태계의 지속 가능성을 위한 필수 조건이 되었다.

4.3. 의료: 연합학습을 통한 프라이버시 보호와 협력 진단

의료 분야에서 AI-보안 융합의 가장 눈에 띄는 점은 'Security-for-AI'가 기술 적용의 전제

조건이라는 점이다. 텐센트의 'Med AI' 플랫폼은 각 병원에 산재한 고감도의 의료 데이터(영상, 진료 기록)를 한데 모으지 않고, 연합학습(Federated Learning) 기술을 통해 AI 모델을 공동으로 발전시킨다[14]. 각 병원에서는 자신의 데이터로 로컬 모델을 학습시키고, 오직 모델의 파라미터(학습 결과)만을 암호화하여 중앙 서버와 공유하는 형식으로 진행한다.

이러한 방식은 'AI-for-Security'의 관점에서 볼 때, 데이터 프라이버시라는 가장 중요한 '보안' 요구를 충족시키는 동시에, 폐암, 망막병증 등에 대한 AI 진단 정확도를 95% 이상으로 끌어올리는 협력의 장을 열고 있다[14]. 즉, 의료 분야는 보안 기술(연합학습, 암호화)이 없었다면 불가능했을 AI의 발전을 보여주는 모범적인 사례가 되고 있다.

4.4. 제조: 자율주행의 안전성 확보와 공급망 보안

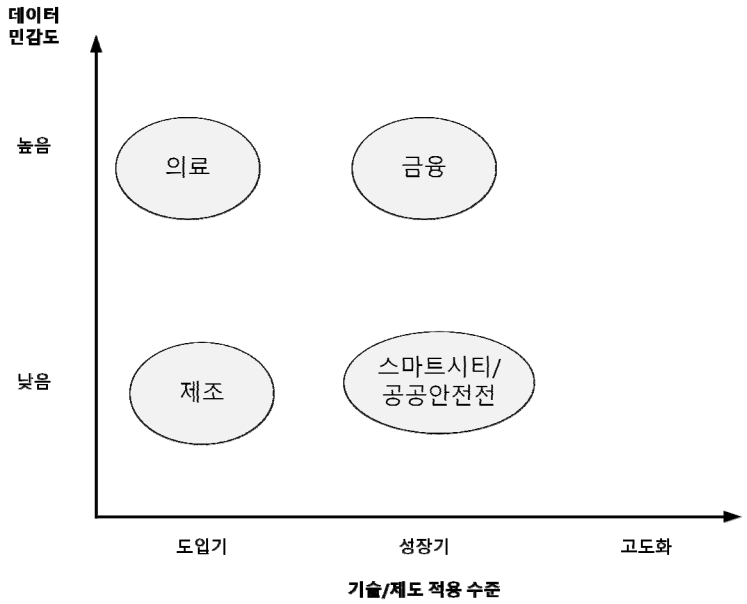
제조, 특히 자율주행 차량 분야에서는 AI와 보안이 물리적 안전과 직접적으로 연결된다. 바이두(Baidu)의 'Apollo Shield'는 'Security-for-AI'와 'AI-for-Security'가 융합된 대표적 시스템이다[15]. 이 시스템은 외부로부터의 해킹으로 차량 제어권을 탈취하려는 시도를 탐지하고 차단(AI-for-Security)하는 동시에, 카메라와 센서를 속이기 위한 '적대적 공격'(예: 정지 신호판을 인식하지 못하도록 하는 스티커)으로부터 AI 인식 모델 자체를 방어(Security-for-AI)한다[15].

또한, 차량-신호등-도로 간 통신(V2X)의 보안을 위해 강력한 암호화 프로토콜을 적용하여, 데이터 위변조를 통한 대규모 교통 혼란을 방지한다. 이는 AI 시스템이 단일 제품의 안전을 넘어, 사회 기반 시설의 보안으로 그 역할을 확장하고 있음을 의미한다.

4.5. 공공안전: 대규모 얼굴인식 네트워크와 체계적 감시

센스타임(SenseTime)과 메그비(Megvii)가 주도하는 공공안전 분야는 중국 AI-보안 모델의 집약체라 할 수 있다[18,19]. 전국에 설치된 2억 대 이상의 CCTV를 연계한 '스카이넷(Skynet)' 프로젝트는 'AI-for-Security'를 국가적 차원에서 극대화한 사례이다. 이 시스템은 실시간 얼굴 인식을 통해 의심 인물의 이동 경로를 추적하고, 과거 범죄 데이터와 결합해 범죄 발생 가능성을 예측하며, 일부 지역에서는 체포율을 30% 가까이 향상시켰다고 보고되고 있다[18,19].

그러나 이 초대규모 AI 감시 시스템은 [그림 3]에서와 같이 'Security-for-AI'의 측면에서 막대한 책임을 동반한다. 시스템 자체가 해킹되거나, 학습 데이터의 편향으로 인해 특정 집단을 잘못 추적할 경우, 그 사회적 파장은 심각할 수 있다. 따라서 이 분야는 기술의 효용과 윤리적 위험, 그리고 시스템 자체의 보안 취약점이 가장 극명하게 충돌하는 지점이기도 하다.



[그림 3] 산업별 AI-보안 융합 수준 및 데이터 민감도 매트릭스

위 사례들을 종합하면, 중국의 산업별 AI-보안 융합 사례는 하나의 공통된 진화 방향을 보여준다. 즉, 'AI-for-Security'를 통해 산업의 효율성과 통제력을 높이는 동시에, 'Security-for-AI'를 통해 그 성과물인 AI 시스템과 데이터 자산을 안전하게 가두어, 국가 주도의 디지털 생태계를 공고히 하는 것이다. 이 과정에서 데이터 주권은 선택이 아닌 필수가 되었으며, 각 산업은 자신의 특성에 맞게 이 새로운 패러다임에 적응해 가고 있다.

5. 윤리·정책·글로벌 경쟁 쟁점

중국의 AI-보안 융합 모델은 기술적 효율성과 국가 통제력 강화 측면에서 뚜렷한 성과를 거두고 있지만, 그 이면에는 해결되지 않은 심각한 딜레마와 세계적 차원의 경쟁 구도를 낳고 있다[4,5,10]. 이는 단순한 기술 정책의 차원을 넘어, 미래 디지털 세계의 질서와 가치관을 둘러싼 첨예한 대립으로 발전하고 있다[10,11]. 〈표 2〉에서와 같이 데이터 주권부터 국제 표준에 이르기까지 모든 주요 쟁점에서 중국과 미국·EU를 중심으로 한 국제사회의 접근법이 근본적으로 대립하고 있음을 확인할 수 있다[6].

〈표 2〉 AI-보안 관련 주요 윤리·규제 쟁점 비교

쟁점	중국의 접근 방식	국제사회(美·EU)의 접근 방식	잠재적 갈등 요인
데이터 주권	데이터 현지화 의무화, 국경 간 이동 엄격 통제	데이터 자유로운 흐름 강조 (개인정보 보호 전제)	글로벌 공급망 단절
알고리즘 투명성	'블랙박스' 알고리즘 허용 (국가 안보 우선)	'알 권리' 기반 설명 가능성(XAI) 요구	기술 신뢰도 저하
감시와 프라이버시	공공안전을 위한 대규모 감시 정당화	개인 프라이버시를 최우선 가치로 규정	인권 기준 차이
국제 표준	자국 기술을 국제표준(ITU 등)으로 푸시	기존 민주적 가치 기반 표준(IPEF 등) 고수	기술 패권 경쟁

데이터 주권의 강화는 글로벌 기술 생태계의 분열을 가속화하고 있다. 중국이 데이터보안법과 개인정보보호법을 통해 데이터의 국경을 넘는 흐름에 강력한 제재를 가하는 것은 국가 안보와 경제 주권을 지키려는 전략적 선택이다. 그러나 이러한 '데이터 현지화' 정책은 글로벌 기업들에게는 막대한 Compliance 부담을 지우고, 연구 공동체에게는 데이터 접근성을 제한함으로써 기술 발전의 장벽으로 작용한다. 그 결과, 미국과 EU를 중심으로 한 '데이터 자유 흐름' 진영과 중국 중심의 '통제된 데이터 생태계' 진영으로 세계 기술 공간이 양분되는 기술적 데커플링(Technology Decoupling) 현상이 심화되고 있다.

대규모 AI 감시 시스템의 확산은 사회적 신뢰와 개인 자유의 근본을 위협하고 있다. 스카이넷 프로젝트와 사회 신용 시스템의 연동은 공공안전과 사회 관리의 효율성을 정당화 논리로 삼는다. 하지만, 이 시스템이 수집하는 개인행동 데이터가 시민의 신용 점수와 직결되어 이동의 자유와 같은 기본적 권리를 제한하는 도구로 전락할 가능성을 내포한다. 특히 한족 중심으로 학습된 얼굴인식 AI가 소수민족에 대해 상대적으로 낮은 인식률을 보인다는 연구 결과는, 알고리즘의 편향이 사회적 차별을 공고히 하는 새로운 방식으로 작동할 수 있음을 경고한다. 이는 기술의 발전이 오히려 사회적 약자의 위치를 고정하는 역설적인 상황을 만들어내고 있다.

알고리즘의 불투명성은 민주적 통제와 책임 소재의 사각지대를 확대하고 있다. 중국의 규제 프레임워크는 국가 안보와 사회 안정을 최우선 가치로 둠에 따라, AI 의사결정의 내부 논리를 설명해야 하는 '알고리즘 투명성' 요구는 상대적으로 약한 수준에 머물러 있다. 이른바 '블랙박스' 상태의 AI가 공공 영역에서 중요한 결정을 내릴 때, 그 결정에 대한 시민의 '알 권리'와 이의 제기 권리는 사실상 공백 상태에 빠지게 된다. 이는 궁극적으로 "누가 AI의 실수에 책임을 지는가?"라는 근본적인 질문에 대한 답을 흐리며, 기술에 대한 민주적 감시의 사각지대를 넓혀가고 있다.

국제 표준화 경쟁은 새로운 기술 패권 다툼의 전장으로 변모하고 있다. 중국은 '중국 표준 2035' 프로젝트를 통해 ITU(국제전기통신연합)와 같은 국제기구를 주무대로 자국의 AI 안전 기준과 데이터 거버넌스 모델을 글로벌 표준으로 공식화하려는 공세를 펼치고 있다. 이에 맞서 미국은 IPEF(인도-태평양 경제프레임워크)를 통해 데이터 프라이버시와 개방성을 핵심 가치로 내세운 대항 규범을 제시하는 등, 표준을 둘러싼 패권 경쟁은 더욱 격화되는 양상이다. 이 경쟁의 승자는 단순한 기술 시장을 넘어, 미래 디지털 세계의 운영 규칙과 가치 체계를 주도할 권한을 쥐게 되기 때문이다.

결국, 중국의 AI-보안 융합 모델은 기술적 진보라는 성과와 함께 감시 확대, 불평등 심화, 글로벌 협력의 훼손이라는 세 가지 차원에서 심각한 도전에 직면해 있다. 이 모델의 지속 가능성은 기술의 발전 속도가 아니라, 이러한 윤리적·사회적 딜레마를 해결할 수 있는 정치적·제도적 역량에 달려 있다고 볼 수 있다.

6. 한국의 대응 전략 및 시사점

중국이 국가 주도의 통제 모델을 추구하는 가운데, 한국이 지속 가능한 AI-보안 생태계를 구축하기 위해서는 차별화된 접근법이 필요하다. 우리의 강점은 개방적 민주사회의 신뢰성과 기술적 역량에 기반해야 하며, 이를 통해 글로벌 가치사슬에서 필수적인 파트너 역할을 수행해야 한다.

첫째, 신뢰할 수 있는 AI-보안 기술 표준을 선제적으로 확립하는 것이 중요하다[8,9]. 중국식 모델이 가진 감시 확대와 알고리즘 불투명성 문제를 극복하기 위해, 한국은 'Security-by-Design' 원칙을 AI 개발 전 과정에 내재화해야 한다[9]. 특히 AI 시스템의 윤리적 안정성과 기술적 견고성을 검증하는 'K-AI 보안 인증제도'를 도입하고, 이를 글로벌 표준화 포럼에 적극 제안해야 한다[8,9]. 설명 가능한 AI와 차등 프라이버시 같은 기술을 조기에 상용화하여, 기술의 진보와 인권 보호가 조화를 이루는 새로운 모델을 실증해야 한다[9].

둘째, 글로벌 협력 네트워크에서 전략적 교량 역할을 강화해야 한다. 미국 중심의 IPEF와 같은 다자 협의체에서 데이터 프라이버시와 개방성 원칙을 옹호하는 동시에, 실용적 차원에서 중국과의 기술 대화 채널도 유지해야 한다. 이를 위해 한·중 AI 안전 상설 협의체를 구성해 위험 관리 방안을 논의하고, 연합학습 등 개인정보 보호 기술 분야의 공동 연구를 활성화하는 것이 필요하다. 한국만이 가진 기술적 중립성과 민주적 가치를 바탕으로 글로벌 AI 거버넌스에서 중재자적 역할을 수행해야 할 시점이다.

셋째, 국가 안보와 산업 경쟁력을 동시에 보호할 실질적인 기술 역량을 확보해야 한다. 정부는 국가 주요 인프라를 대상으로 한 합동 Red Team 운영을 정례화해야 한다. 전문 보안 요원과 AI 연구원으로 구성된 팀이 주요 AI 시스템을 대상으로 지속적인 침투 테스트와 취약점 분석을 수행하도록 해, 위협에 대한 대응 체계를 강화해야 한다. 동시에 반도체와 5G/6G 통신 등 한국의 강점 분야와 AI 보안을 융합한 'K-보안 솔루션'을 개발해 수출 전략 품목으로 육성하는 전략이 필요하다.

결론적으로 한국은 중국의 국가 통제 모델과 미국의 기술 주도 모델 사이에서 제3의 길을 개척해야 한다. 기술 패권 경쟁이 격화되는 상황에서 한국의 성공 전략은 개방적 협력과 투명한 기술 발전에 기반해야 한다. AI와 보안이 상호 발전하는 선순환 생태계를 조기에 정착시킴으로써, 우리는 글로벌 AI 공급망에서 없어서는 안 될 신뢰받는 파트너로 자리매김할 수 있을 것이다.

7. 맺음말

AX 시대의 핵심 화두는 이제 더 이상 '인공지능의 발전' 그 자체가 아니다. 진정한 도전은 '어떻게 하면 AI와 보안이 상호 선순환하는 공진화(共進化) 시스템을 구축할 것인가'에 달려 있다. 중국의 사례는 국가 주도의 강력한 통합 접근법이 기술 발전과 사회 관리에서 얼마나 놀라운 속도와 효율을 낼 수 있는지를 보여주었다. 그러나 동시에 그것이 내포한 감시 확대, 알고리즘 편향, 글로벌 기술 체제의 분열이라는 딜레마는 우리에게 신중함을 요구한다.

이제 한국이 선택해야 할 길은 명확하다. 중국의 통제 모델을 그대로 답습하는 것도, 기술 패권 경쟁의 소용돌이에 휩쓸리는 것도 아닌, '개방과 신뢰'를 새로운 경쟁력의 축으로 삼는 것이다. 우리는 AI 시스템의 설계 단계부터 보안과 윤리를 내재화하는 'Security-by-Design' 문화를 정착시키고, 설명할 수 있는 AI와 강력한 개인정보 보호 기술을 통해 세계가 신뢰할 수 있는 기술 표준을 제시해야 한다.

더 나아가 한국은 글로벌 협력의 교량이 되어야 한다. 첨예하게 대립하는 기술 블록 사이에서 우리만이 가진 기술적 중립성과 민주적 가치를 무기로, 신뢰와 협력을 기반으로 한 새로운 디지털 질서를 모색하는 데 앞장서야 할 때다. AI와 보안의 공진화는 결국 기술이 인간을 어떻게 섬길 것인지에 대한 물음이다. 기술의 속도와 효율만을 좇는 경주가 아닌, 인간의 존엄과 안전이 최우선인 '신뢰의 생태계'를 구축하는 데 한국의 미래가 놓여 있다.

8. 참고문헌

정부 및 공식 기관 문헌

[1] 중국 국무원, 「차세대 인공지능 발전계획」, 2017.

[2] 중국 국가인터넷정보판공실, 「데이터보안법」, 2021.

[3] 중국 전국인민대표대회, 「개인정보보호법」, 2021.

[4] 중국 국가데이터국, 「연차보고서」, 2024.

[5] 중국 국무원, 「디지털 중국 건설 발전 보고서」, 2023.

[6] 중국 과학기술부, 「인공지능 윤리 규범」, 2022.

단행본

[7] Xu, W., “Adversarial AI Defense Techniques,” in *Advances in AI Security*, Beijing, China: China Science Publishing, pp. 45–62, 2023.

[8] Lee, J.B.E., “AI and Cybersecurity Co-evolution Framework,” in *Journal of Intelligent ICT*, Vol. 12, No. 3, pp. 112–125, 2024.

[9] Kim, H.S., “AI Governance and Ethical Standards,” Seoul, South Korea: Korean Information Society Press, 2024.

온라인 자료

[10] Carnegie Endowment for International Peace. China’s AI Safety and Governance Landscape [Internet]. Washington, DC; 2024. Available from: https://carnegieendowment.org/research/ai/china-ai-safety

[11] Stanford University DigiChina Project. China AI Safety and Development Association Overview [Internet]. Stanford, CA; 2024. Available from: https://digichina.stanford.edu/work/china-ai-safety-and-development-association-overview/

[12] [12] Huawei Cloud. Smart City AI Security White Paper [Internet]. Shenzhen, China; 2023. Available from: https://www.huaweicloud.com/whitepaper/security/smart-city-ai-security-wp

[13] Alibaba Cloud. Cyber Threat Intelligence Platform Annual Report [Internet]. Hangzhou, China; 2024. Available from: https://www.alibabacloud.com/report/cyber-threat-intelligence-2024

[14] Tencent Med AI. Federated Learning in Healthcare: Applications and Case Studies [Internet]. Shenzhen, China; 2023. Available from: https://med.tencent.com/research/federated-learning-healthcare

[15] Baidu Apollo. Autonomous Driving Safety Architecture Technical Paper [Internet]. Beijing, China; 2024. Available from: https://apollo.auto/whitepaper/safety-architecture.html

[16] Ant Group. AI-driven Financial Risk Control Framework [Internet]. Hangzhou, China; 2023. Available from: https://www.antgroup.com/research/ai-risk-control

[17] Ping An Technology. Intelligent Risk Management Platform Technical Overview [Internet]. Shenzhen, China; 2023. Available from: https://tech.pingan.com/ai-risk-management

[18] SenseTime. Intelligent Surveillance and Security System White Paper [Internet]. Shanghai, China; 2023. Available from: https://www.sensetime.com/en/whitepaper-surveillance

[19] Megvii. AI Facial Recognition Technology and Public Security Applications [Internet]. Beijing, China; 2023. Available from: https://en.megvii.com/technology/face-recognition

[20] Zhang, L., "Federated Learning for Secure AI Model Training," *IEEE Transactions on Cybernetics* [Internet]. 2024. Available from: https://ieeexplore.ieee.org/document/10123456

주제원고

모빌리티 및 소프트웨어 정의 차량(SDV) 차원의 보안 및 인공지능 해킹 기법 현황에 대한 고찰

성신여자대학교
김준영 · 장하람

1. 서론

자동차라는 개념에서 확장된 범위로써 통칭하고 있는 모빌리티는 단순 기계 기반에서 수동적인 이동성을 제공하는 것을 넘어서 다양한 형태의 수단으로써 능동적인 이동성을 제공할 수 있는 것으로 출발하고 있다. 자동차 이외의 스쿠터, 1인 차량, 전기 자전거 등 다양한 모빌리티 수단들이 이미 전개되어서 일상생활에서 쓰이는 중이다 [1]. 이러한 수단들의 경우 기계 중심에서 소프트웨어 중심으로 전환되는 소프트웨어 정의 (Software-Defined)에 맞춰지고 있으며 현재 인공지능 (Artificial Intelligence :AI) 고도화로 인한 자율주행 기술의 지속적인 발전으로 모빌리티의 급진적인 혁신이 이루어지고 있다.

다만 이러한 소프트웨어 중심 구조와 AI 기반 기능이 확신함에 따라 소프트웨어 품질, 호환성, 복잡성 증가 등 다양한 기술적 이슈들의 발생 가능성이 커지고 있다. 보안 측면 측면에서도 치명적인 상황들이 확대되고 있으며, 특히 AI 기반 해킹의 동반 가능성이 거론되면서 소프트웨어 정의 차량 (Software-Defined Vehicle, SDV) 를 포함한 모빌리티 전반에 걸친 보안 위험이 대두되고 있다. 본 기고문에서는 이러한 모빌리티 및 SDV 관점에서의 보안 현황 및 이와 연관된 방향성에 대해서 제시하고자 한다. 특히 AI 기반 공격 기법에 대한 예시도 같이 제공하면서 앞으로 보안 측면에서 고려해야 할 사항들에 대해서도 고찰해 보고자 한다.

2. 모빌리티와 SDV: 기본 개념

2.1. 모빌리티 개념 및 주요 범위

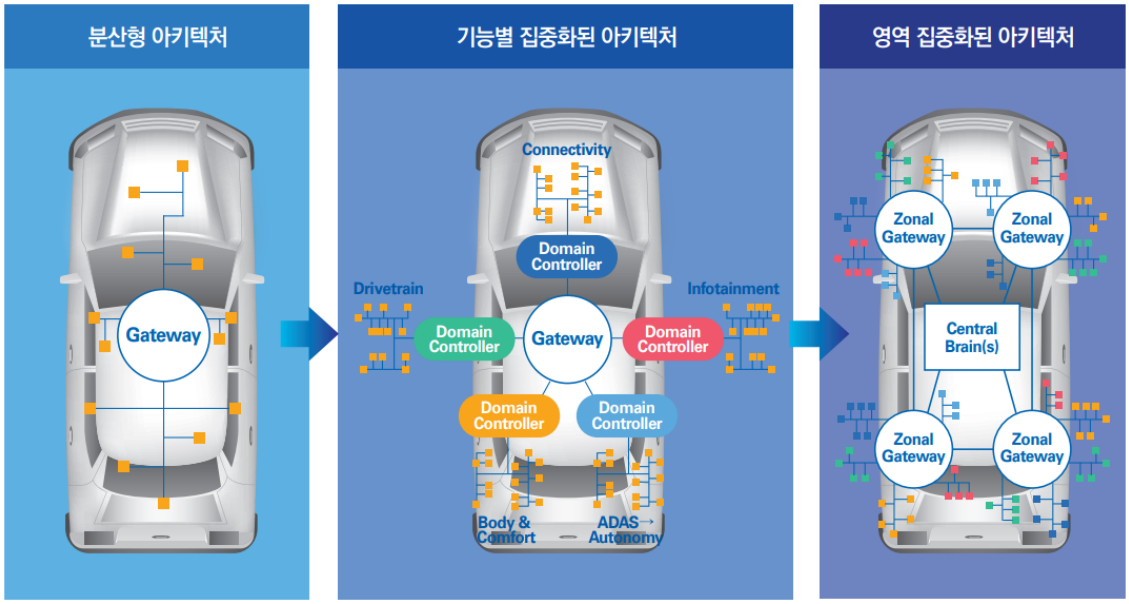
모빌리티는 이동을 제공하는 장치 및 수단으로써 단방향 및 양방향으로 사용자와 상호 작용을 한다. 기존 대중교통과 같은 단방향성을 띄는 이동 수단도 큰 모빌리티 범주에 포함되지만, 실질적으로 차량·사용자·인프라 간 양방향 데이터 교환이 가능한 자가용 자동차, 스쿠터, 자전거 등과 같은 이동 수단으로써 모빌리티가 정의된 바 있다. 다만 이러한 흐름은 2020년대 들어 모빌리티 산업이 다시 자동차 중심으로 재편되는 양상으로 나타나고 있다 [1]. 이러한 변화는 다양한 서비스 및 기능의 빠른 전개를 위한 SDV의 확산으로 이어지고 있으며 연장선상에서 자동차 산업은 선택과 집중 전략을 기반으로 커넥티비티 (Connectivity), 자율주행 (Autonomous Driving), 공유차 (Sharing) 및 전동화 (Electrification)을 포괄하는 CASE 영역에서 새로운 비즈니스 모

델들을 창출하려는 경향을 보인다 [2].

상기 4가지 기술 중 자율주행은 주변 환경 정보를 기반하여 자체적인 주행을 수행하지만, 운전 결정을 내리는 데 있어 안정적인 고속 통신 연결이 필수적이다. 이를 위해 5G 이상의 통신 기술을 지원하는 인프라가 구축돼야 하며, 자율주행 과정에서 발생하는 대용량 데이터를 처리하기 위한 데이터 센터 확충도 병행될 필요가 있다 [2]. 이를 위해 정부는 '교통 분야 3대 혁신 전략'과 더불어 “혁신 교통 서비스의 일상화” 추구를 위해 미래 모빌리티의 일상 구현 조기화 추진 및 자율주행 서비스 본격화 방안 제시도 진행하였다 [3]. 이러한 모빌리티 방향성을 실제화할 수 있는 협력 지능형 교통 체계(Cooperative Intelligent Transportation System, C-ITS)는 Vehicle-to-Everything (V2X) 기반하에 고도화를 추구하고 있으며 다양한 통신 프로토콜을 포괄하고 있다. 특히 C-ITS 서비스를 통해서 연간 약 4,300억 원의 혼잡 비용 절감 효과가 가능하다고 기대하고 있다 [4].

2.2. 소프트웨어 정의 차량 개념 및 주요 특징

과거 하드웨어 중심이었던 자동차는 소프트웨어 정의화를 통해서 차량 자체가 소프트웨어 인프라 그 자체로 전환하고 있으며 단순 이동 수단에서 실시간 정보 교환이 가능한 네트워크 기반 차량으로 진화하고 있다 [5]. 특히 소프트웨어 정의 차량 경우 기본적인 네트워크 구성과 더불어 차량 자체의 아키텍처의 변화도 가져오는 중이다. 그림 1과 같이 기존 차량 구조로 대변되는 분산형 구조에서 영역별 집중화로 진행되는 중이며 이를 통해서 소프트웨어 업데이트 및 적용이 유연하게 진행되는 형태로 SDV가 변화되는 중이다. [6] 이러한 변화가 지속해서 진행 중인 상황인 SDV의 보급률은 2021년 시점에서 2.4% 수준에 머물고 있으나 2029년에는 90% 이상으로 급증할 것으로 예상되며, 글로벌 완성차 제조기업 (Original Equipment Manufacturer: OEM)들 경우 2025년을 SDV 본격 도입 시점으로 보고 있다 [7]. 주요 기업 동향들도 비슷한 경향을 보이고 있으며 특히 HL만도 경우 SDV 시대에 대응하기 위해 AI/Cloud 플랫폼 및 생성형 AI 개발을 추진하고 있다 [7]. 차량에 특화된 자체 SLM(Small Language Model) 확보하여 공학 지식 기반 생성형 AI 서비스를 제공하고, 데이터 파이프라인 전반의 보안 강화에 HL만도는 집중하고 있다 [7].



[그림 35] SDV 아키텍처 형태 및 기본 구성 예시 [6]

3. 모빌리티 차원의 보안 현황

3.1. 주요 모빌리티별 인터페이스 및 보안 시스템

ICT와 연동된 미래 모빌리티는 상시 연결이 가능한 수단으로 변모되는바 외부와 단절되어 있거나 단방향 연결이 일반적이었던 기존 모빌리티 보다 해킹, 개인정보 유출 등 보안 위협에 훨씬 취약하며, 이는 안전과 직결되기 때문에 장치 보안 확보가 가장 최우선이 될 수밖에 없다 [8]. 특히 차량 외부 연결성이 증가하면서 공격 경로가 확대되고 있으며, 이는 공격자가 악의적으로 침투할 수 있는 다양한 공격 경로를 만드는 요인이 되고 있다 [5]. 외부 텔레매틱스부터 내부 Wi-Fi 핫스팟, 블루투스, 스마트키 및 On-Board Diagnostics (OBD) 단자까지 다양한 인터페이스가 공격 지점이 될 수 있다. 공유 모빌리티 경우 이러한 위협이 더욱 크게 나타나며, 특히 공유 전동킥보드와 같은 개인형 이동장치(Personal Mobility, PM) 형태의 수단 경우 직접적인 수단과 더불어 해당 수단 전용 앱 해킹 등을 포함하여 원격 제어, 개인정보 접근 등 다방면의 해킹 및 공격 가능성이 존재한다 [8].

3.2. 소형 모빌리티 보안 취약점 및 실제 사례

소형 모빌리티의 실제 사례는 다수 존재한다. 대표적인 경우가 원격 가속 및 급정거 제어가 있다. 2019년 보안 연구원들은 샤오미(Xiaomi) M365 전기 스쿠터의 블루투스 취약점을 이용해 원격 제어가 가능함을 확인했다. 실제 약 100m 거리 내에서 속도를 임의로 높이거나 주행 중 급정거를 유발할 수 있었으며, 이는 탑승자에게 직접적인 신체적 위험을 초래할 수 있는 수준이었다 [9][10]. 공유 킥보드 무단 이용 및 잠금 해제도 확인된 보안 문제이다. 공유 모빌리티 서비스의 인증 정보(예: 보안 토큰) 미 암호화와 미갱신 취약점을 이용해 다른 이용자로 위장해 로그인하여 무단 이용과 결제 정보 탈취가 가능함을 실제 연구로 확인했다 [11]. 또한 공유 스쿠터 앱은 사용자 이름, 연락처, 결제 정보, GPS 위치 등 민감한 정보를 수집하는데, 데이터 암호화가 미흡한 경우 공격자가 사용자 이동 경로와 방문 장소를 파악할 수 있는 위험이 존재한다 [12]. 이러한 보안 취약성은 소형 모빌리티에 국한되지 않으며, 차량 기반 모빌리티에서도 유사한 구조적 위험이 확인된다.

3.3. 자동차 주요 해킹 사례 & SDV 차원의 리스크 및 대응 방안

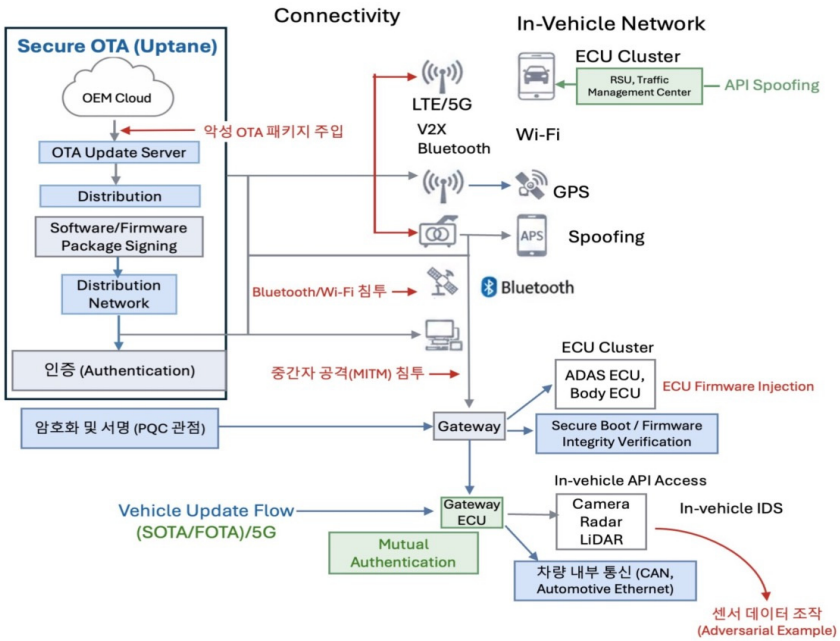
소형 모빌리티 뿐만이 아니라 자동차 역시 주요 모빌리티 보안 이슈의 핵심 대상이다. 자동차

보안 위험으로는 데이터 위/변조(21%), 자동차 시스템 조작(4%), 위치 추적(4%) 등이 확인되고 있다 [8]. 또한 자율협력 주행 및 전기차 환경에서는 악의적인 제어 권한 탈취와 시스템 오작동을 유발하는 비정상 데이터 주입 공격(Abnormal data injection attack)도 위험 요소로 지적된다 [13]. 자동차에서 발생한 주요 보안 이슈 및 해킹 사례들은 표 1을 참고한다. 표 1과 같이 대부분의 공격은 블루투스, Wi-Fi 등의 무선 통신 경로와 더불어 웹브라우저, 모바일 앱 등을 통한 Application Programming Interface (API) 연동 인터페이스를 통해서 해킹이 이루어지며, 사용자 정보 탈취와 원격 제어가 주요 피해 유형으로 확인된다. 또한 테슬라 사례에서 보이듯 자율주행 인식 시스템의 입력 데이터를 조작해 자율주행 기능을 오작동시킬 가능성도 제기되고 있다.

〈표 1〉 자동차 해킹 사건 주요 사례

#	사건명	발생 연도	주요 내용	영향/의의	문헌 번호
1	지프 체로키 원격 제어 사건	2015	• 보안 연구원 밀러·발라섹 주행 중 차량 원격 해킹 • Uconnect 셀룰러 취약점 이용 → 에어컨/ 라디오/엔진/브레이크 제어 • 원격 물리 제어 가능성 입증	• FCA 140만 대 리콜 • 차량 원격 제어 보안 중요성 부각	[14]
2	닛산 리프 원격 제어	2016	• 모바일 앱·웹 API 인증 취약점 이용 • 계기판·배터리·위치 정보 등 원격 조회·제어 가능	• 인증/모바일 연동형 차량 서비스 보안 강화 필요	[15]
3	미쓰비시 아웃랜더 해킹	2016	• 온보드 Wi-Fi 보안 취약점 • 차량 보안 알람 원격 해제	• 차량 내 Wi-Fi 진입점의 위험성 확인	[16]
4	재규어 랜드로버(JLR) 공급망 공격	2025	• 사이버 공격으로 IT 시스템 마비 • 영국 생산·물류 운영 중단	• 제조·운영 분야 OT 보안 중요성 부각	[17]
5	BMW 유료 기능 우회 해킹	2023 2025	• 전압 공급 조작으로 열선 시트 등 유료 기능 우회 • 차량 시스템 Jailbreaking 가능	• SDV 구독 모델 보안성 문제 대두 • 내부 데이터 접근 가능성 확인	[18]
6	폭스바겐 Cariad 데이터 유출	2024	• 약 80만 대 EV 관련 데이터 유출 • 공급망 전반의 취약성 노출	• 대규모 차량·부품 데이터 보호 중요성 증가	[19]
7	SiriusXM 커넥티드 서비스 취약점	2024	• 공통 API 취약점으로 차량 잠금 해제·시동·위치 추적 가능 • VIN만으로 공격 가능	• 다수 OEM 동시 위협 발생 가능성 확인 • API·플랫폼 보안 중요	[20]
8	토요타·닛산 데이터 유출	2024 2025	• 닛산 디자인 데이터 4TB 유출 • 토요타 미국 법인 240GB 데이터 다크웹 공유	• 글로벌 OEM의 랜섬웨어 취약성 증가	[21] [22]
9	테슬라 자율주행 시스템 조작	2020	• 속도 제한 표지판에 스티커 부착 → 오인식 유도 • 35→85mph 오인식으로 50mph 증가	• 카메라 기반 인지 시스템 AI 공격 취약성 증명	[23]

SDV는 차량 기능의 대부분이 소프트웨어로 구성되기 때문에 기존 자동차보다 공격 표면이 크게 확대된다. 따라서 SDV 역시 유사한 보안 위험에 노출되어 있으며 그림 2와 같이 다양한 인터페이스 및 소프트웨어 패키지를 통한 해킹 가능성이 크다. 이에 대응하기 위한 관련 기술들의 개발들도 이루어지고 있다. 예를 들어 Consumer Electronic Show (CES) 2023에서는 LG유플러스와 LG전자가 커넥티드 카 전용 양자 내성 암호 기반의 보안 기술을 보여주며 차량 내부 Audio, Video, Navigation (AVN) 및 카페이(Car Pay) 서비스의 보안 강화 가능성을 보여준 바 있다 [24]. 커넥티비티는 모빌리티 사이버보안의 핵심 키워드 중 하나로써 자율주행/정보보호까지 포괄하는 기술적 연결성을 강화하는 방향으로 진화하고 있다 [25]. SDV 측면에서도 특히 중요하게 다뤄지는 기술은 소프트웨어 무선 업데이트 혹은 원격 업데이트인 OTA (Over-The-Air)이다. OTA는 SDV의 전 생애주기에 걸쳐 보안 패치를 제공하는 핵심 수단이기 때문이다.



[그림 1] SDV 계층적 보안 아키텍처 및 위협 모델 [26]

3.4. OTA 차원의 해킹 가능성 및 실사례

OTA는 단순한 소프트웨어 업데이트가 아니라 다양한 프로토콜과 차량 구성 요소를 고려해 수행되는 복합적 업데이트 방식이기 때문에 여러 해킹 사례가 보고되고 있다. 주요 사례는 표 2에

정리돼 있다. 이러한 OTA 기반 공격 위험을 보완하기 위한 기술 연구도 진행되고 있으며, 대표적인 방안이 Secure OTA 기술이다. 제조사가 원격으로 차량 소프트웨어 업데이트와 패치를 제공하는 OTA 환경에서 보안을 확보하기 위해 Uptane과 같은 기술이 활용되고 있으며, 이를 통해 소프트웨어 공급업체에서 차량까지 이어지는 분산 소프트웨어 배포 과정의 보안성과 맞춤형 적용이 가능하다 [27].

3.5. 글로벌 기구 차원의 동향

글로벌 기구는 모빌리티 보안을 위해 규정과 표준 중심의 활동을 전개하고 있다. UN Regulation No. 155 (UN R155)는 차량의 사이버보안 및 Cybersecurity Management System (CSMS) 승인에 관한 국제 규정으로, 완성차 제조사들이 차량 라이프사이클 전반에 걸쳐 구축 및 프로세스 준수를 요구한다 [5]. ISO/SAE 21434는 도로상에서의 차량 중심의 사이버보안을 다루는 국제 표준으로, 개발·생산·운용·폐기에 이르는 전 단계에 적용된다. 또한 UN R155에서 요구하는 CSMS 인증을 위한 기술적 기반을 제공한다 [5]. 또한 ITU-T SG17은 통신 보안 표준화 경우를 진행하면서 ITS 보안 연구반(Q13)을 통해 차내 망 침입탐지시스템 방법론(X.ipscv), 차량 통신을 위한 소프트웨어 업데이트 역량(X.1373rev) 등의 차량 통신 보안 표준화를 지속하고 있다 [9][28].

4. AI 기반의 해킹 기법 및 위험성

AI(인공지능)를 이용한 자동차 해킹 사례가 존재하며, 주로 연구 단계에서 그 위험성이 입증되고 있다. 실제 악의적인 공격에 AI가 사용된 사례는 아직 드물지만, 잠재적 위협으로 간주하며 다양한 공격 기법이 연구되고 있다 [29].

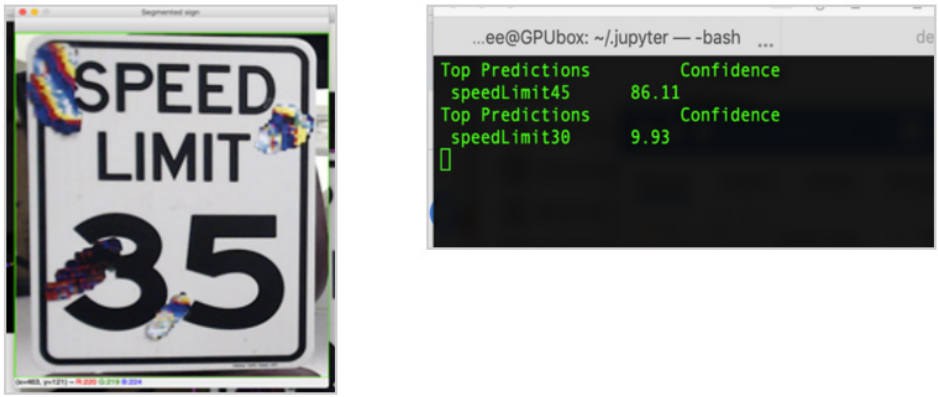
〈표 2〉 OTA 위협 항목 주요 사례

#	위협 항목	문제 요약	연구 동향 / 사례	대응 또는 특징	문헌 번호
1	업데이트 아티팩트 진정성·무결성 검증 부실	• 코드 서명 미사용 • 약한 해시·서명 검증 우회 • 키 관리 부실로 악성 펌웨어 주입 가능	• ICV OTA 보안 서베이 • OTA 보안 보증 연구내 핵심 위협 규정	• 강한 코드 서명 요구 • 서명 검증 로직 강화 및 HSM 기반 키 관리 필요	[30]
2	롤백·Freeze 공격 및 버전 관리 취약	• 공격자가 장애 ECU SW 구버전(취약 버전) 전환 • 특정 ECU의 구버전 고착 (freeze)	• Uptane 모델내 자동차 OTA 특화 치명적 공격 분류 • 메타데이터·타임 서버 기반 완화책 제시	• 메타데이터 신뢰 사슬 강화 • 타임 기반 fresh-ness 검증 필수	[31]
3	OTA 파이프라인 (Cloud/Backend등) 공급망 공격	• 빌드 서버·리포지토리·서명 인프라 침해 시 • 정식 서명된 악성 업데이트 전체 차량 배포 가능	• Scudo 등 연구 내 Uptane + in-toto 제시 • 소프트웨어 공급망 공격 대응 프레임워크 발전	• 서명 키 격리 • 빌드 파이프라인 무결성 검증(in-toto)	[32]
4	차량 내부 통신 (CAN/Ethernet) 보호 미비	• OTA 전달 경로상 내부 CAN 버스 경우 무보호 시 인젝션 ·재전송 공격 가능	• Jeep Cherokee 해킹 사례(2015) • SecOC 기반 PDU 인증 제시 (AUTOSAR)	• CAN 메시지 인증· freshness 필요 • 성능·키 관리 문제로 실 적용 난제 존재	[33] [34]
15	동반 앱·웹 API 취약점 원격 제어· 프라이버시 침해	• 앱·웹 포털 인증 취약 • 원격 문 잠금, 시동, 위치 조회 등 악용 • 사용자·차량 데이터 유출 위험	• Nissan LEAF, Hyun-dai Blue Link, Kia Connect 등 사례 존재	• 인증 강제, 세션·토큰 보안 강화 • 원격 제어 API 최소 권한화	[35]
6	OTA 실패·부분 업데이트로 인한 안전(Functional Safety) 문제	• 통신 장애·전력 문제 시 부분 업데이트 발생 • ECU 버전 불일치, 기능 상실 → 안전 문제	• ISO 26262 / SAE J3061 관점에서 OTA 아슈어런스 연구 • Fail-safe·rollback 부족 = 핵심 위험 지적	• 단계적 업데이트 검증 • Fail-safe·Rollback 전략 필수	[36]

4.1. AI 기반 모빌리티 해킹 연구 사례

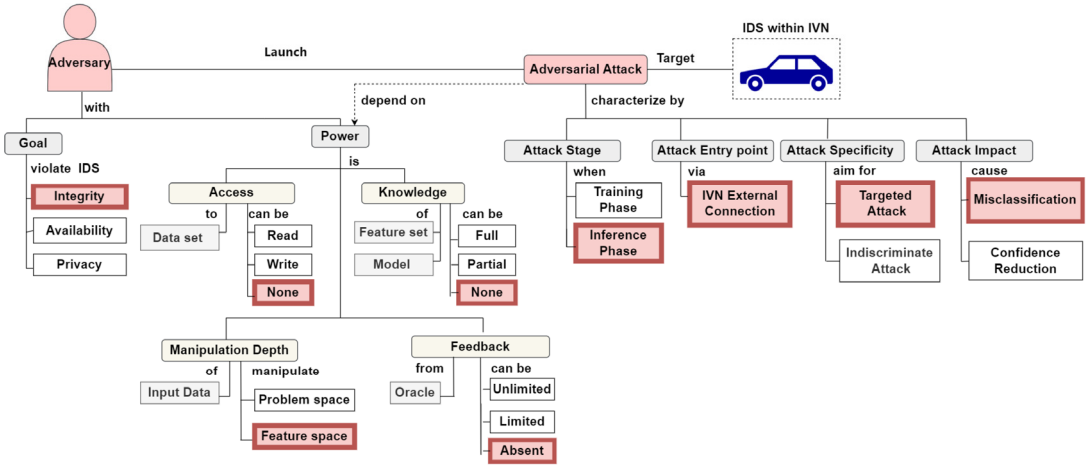
AI를 이용한 가장 대표적인 방법은 자율주행 차량의 '눈'과 '귀'에 속하는 센서 오인식을 의도적으로 발생시켜 속이는 것이다 [37]. 물리적 조작을 통한 센서 교란이 이와 비슷하게 주요 이슈 중 하나로써 앞서 표1 내에서 제시된 사례 중 하나인 2020년 McAfee 연구원은 그림 3과 같이 도로변 속도 제한 표지판에 작은 스티커 조각을 붙여 테슬라의 카메라 시스템(Mobileye EyeQ3 칩 기반)이 속도 제한을 오인식하는 데 성공했다. [38] 차량은 시속 35마일(약 56km/h) 표지판을 시속 85마일(약 137km/h)로 인식하여 실제 주행 속도를 높였으며 이는 AI 기반 이미지 인식

모델이 특정 방식으로 훈련된 구조적 취약점을 악용한 것이다. [38][39]. 이러한 오인식을 활용한 적대적 예시(Adversarial Examples) 생성도 또 다른 사례 중 하나다. 이는 AI 모델이 특정 이미지를 오분류하도록 유도하기 위해 미묘하게 변형된 이미지나 패턴을 생성하는 방식이다. 해커는 이러한 패턴을 스티커로 만들어 차량의 카메라 인식되도록 함으로써 '정지' 신호를 '진행' 신호로 오인하게 할 수 있다 [40].



[그림 2] 도로변 속도 제한 표지판의 작은 스티커 통한 속도 인식 조작 예시 [38]

최근 자동차 보안을 위해 그림 4와 같이 AI를 활용한 침입 탐지 시스템(Intrusion Detection System, IDS)이 도입되고 있으나, 공격자는 같이 AI를 활용해 IDS가 탐지하지 못하는 악성 트래픽 패턴을 생성해 방어 체계를 우회한다. IDS는 주로 알려진 공격 패턴을 학습하는 반면, AI는 새로운 유형의 공격을 생성할 수 있어 탐지 회피가 가능하다 [41].



[그림 3] 차량 내 네트워크의 IDS 상에서의 적대적 공격 대상 방식의 예시

생성형 AI의 발전은 해킹 프로그램 개발의 진입 장벽을 낮추고 있는바 자동화된 악성 코드 생성이 초보 레벨에서도 가능해지고 있다. 공격자는 AI 챗봇을 이용해 특정 시스템을 공격할 악성 코드 샘플을 요청하거나, 안티바이러스 소프트웨어에 탐지되지 않는 변종 악성 코드를 생성하도록 지시할 수 있다. 이는 차량 내부 네트워크(Control Area Network, CAN)의 취약점을 공략하는 코드를 개발하는 데 사용할 수 있다. [29][37] 또한 AI 기반 취약점 분석 및 예측 기법을 적용하면 방대한 양의 차량 데이터 분석 통한 특정 차량 모델이나 소프트웨어 버전의 잠재적 취약점을 자동으로 감지할 수 있다. 이를 통해 공격 대상의 효율적 선별이 가능하다. [38][39]

국내 경우도 모빌리티 관련된 보안 실험을 진행한 사례들이 존재한다. 대표적인 기업 중 하나로써 아우토크립트 경우 전기차와 전기차 충전 인프라간 연동 때 진행되는 보안 통신상에서 인증을 위해서 사용되는 공개키 기반 구조 (Public Key Infrastructure, PKI) 상에서 충전 이후의 인증서 폐기 메커니즘에 대한 분석을 통해서 주요 메커니즘들 차원에서의 한계점과 전기차 및 전기차 충전 인프라 상에서 발생할 수 있는 다양한 보안 시나리오들을 분석하고 이를 위한 PKI 설계도 개발하면서 앞으로 개선 방향성에 대한 제시를 한 바 있다. [42][43] 정부 차원에서도 모빌리티 보안 관련된 정책 등을 추진하고 있으며 특히 국토교통부 차원에서 2020년 12월 자율주행차 윤리 가이드라인, 자동차 사이버보안 가이드라인 및 레벨4 자율주행차 제작·안전 가이드라인들을 발표하였다. [44] 특히 자동차 사이버보안 가이드라인 경우 자동차 자체뿐만이 아닌 제작사 조직 및 전담 기관과의 정보 공유 등에 대한 지침들도 포함하여서 제시하였으며 이러한 기준을 기반으로 하여 자동차관리법 내 사이버보안 항목이 개정되어 2025년 8월 14일부터 시행 중이다. [44][45]

5. 결론 및 향후 방향성

본 기고문에서는 모빌리티와 SDV 차원에서의 보안 침해 사례들을 제시하였으며 특히 AI 해킹에 관한 실제 연구 사례 등을 토대로 하여 다각적인 측면에서의 모빌리티 및 SDV 보안을 조사하였다. 모빌리티 진화로 인해 다양한 서비스들이 도출되었고 분산형 아키텍처에서 기능별 중앙화, 영역별 중앙화로 SDV가 전환되어 가면서 모빌리티 및 SDV내 기능들도 고도화가 되어가고 있으나 이에 비례해 보안 및 해킹 이슈도 지속해서 진행되고 있다. 본 기고문에서는 모빌리티 보안 관점에서 자동차와 SDV의 핵심 기능 중 하나인 OTA를 중심으로 보안 이슈를 설명했으며, 특히 AI 기반 공격과 침투 실 혹은 연구 사례들을 함께 제시하였는데 결국 현재 AI의 고도화된 진화로

해킹에 대한 다양한 방법론들이 연구 진행하고 있으며 앞서 설명했었던 단순 API 콜 혹은 진입을 통한 해킹 대신 다양한 인터페이스 차원의 공격 등을 통한 배터리 소모 혹은 정상적 주행 방해 등의 해킹 등이 빈번하게 발생할 것으로 예측된다. 또한 정책적 차원에서도 법 제정과 더불어 이러한 보안 체계 및 고려에 기반한 기술 개발 및 표준 개발 등이 필수적일 수밖에 없다.

앞으로 모빌리티 보안 차원에서도 이러한 다양한 보안적 측면의 문제 및 다양한 인터페이스 대상 공격에 적극적이고 실질적인 방어 체계를 마련하는 동시에 구조적 차원에서 공격 자체를 차단할 수 있는 근본적인 대응 방안들을 모빌리티 및 SDV 측면에서 고려하고 연구가 추진되어야 할 것이다. 특히 이러한 기술들 경우 단순 학계 레벨이 아닌 기술 동향에 주도적인 산업계 참여가 필수적인바 실질적인 협의체 형태를 통한 선제 대응이 필요할 것이다. 또한 민간 차원에서의 대응과 더불어 실질적인 정부와 기관 차원에서의 단순 법령 제정에서 마무리되는 것이 아니라 정책 및 사업 지원 등을 통해 모빌리티 보안의 고도화를 필수적으로 동반하게끔 진행해야 할 것이다.

6. 참고문헌

[1] 김영은, 김준영, 이준세, 노송. (2024). 모빌리티 서비스 체계 내 5G 기반 응용 시스템 구현 타당성 초기 분석. 한국통신학회지(정보와통신), 41(12), 52-58.

[2] 강신남. (2024). 전문가 칼럼 미래 모빌리티 혁신으로 나아가는 여정 : 소프트웨어 기반 차량(SDV)과 자율주행기술에서의 안전성 확보가 핵심. 오토저널, 46(6), 41-44.

[3] 곽중희. (2023). CES 2023, 사람을 위한 보안·안전 기술 빛났다. 월간 Secu N,, 98-100.

[4] 박세환. (2024). C-ITS 상용화를 통한 스마트 모빌리티 도시 구축 이슈. 월간 Secu N,, 100-103.

[5] 조현재, 고현승, 오광석, 손명호, 정지현. (2025-05-21). 글로벌 차량 사이버보안 법규 및 표준 동향 분석: 규정 준수 방안. 한국자동차공학회 춘계학술대회, 제주.

[6] 김나래, 정미주, 엄이슬, 소프트웨어로 달리는 자동차, 완성차 업계가 꿈꾸는 미래, Samjong Insight Vol 88, 삼정KPMG 경제연구원, 2024

[7] 최진홍. (2025). 인터뷰 모빌리티 패러다임 시프트 SDV HL만도 비밀 무기는 “AI” : 남장우 HL만도 책임연구원. 이코노믹리뷰,, 42-44.

[8] 곽중희. (2023). 미래 모빌리티, 보안 없이는 달릴 수 없다. 월간 Secu N,, 54-57.

[9] Swati Khandelwal (2019, 2). Xiaomi Electric Scooters Vulnerable to Life-Threatening Remote Hacks, The Hacker News, Available: https://thehackernews.com/2019/02/xiaomi-electric-scooter-hack.html

[10] 이정현, (2019). "100m 밖서 가속페달"…샤오미 전기스쿠터, 해킹에 속수무책, 2019, ZDNet, Available: https://zdnet.co.kr/view/?no=20190213153931

[11] Casagrande, Marco (2024). Protocol-level Attacks and Defenses to Advance IoT Security. PhD diss., Sorbonne Université

[12] Amer Owaida (2020). e-scooters vulnerable to remote hacks, Security Middle East & Africa, Available: https://securitymea.com/2020/02/19/e-scooters-vulnerable-to-remote-hacks

[13] 송태진, 유용식, 조민제, 홍준호. (2021). 자율협력주행 전기차 환경에서 통합 사이버 보안 체계 정립 연구. 대한교통학회지, 39(4), 493-515. 10.7470/jkst.2021.39.4.493

[14] Wired (2015). Hackers Remotely Kill a Jeep on the Highway, Available: https://www.wired.com/2015/07/jeep-hack-chrysler-recalls-1-4m-vehicles-bug-fix/

[15] BBC (2016). Nissan Leaf Electric Car 'Hack' Fixed, Available: https://www.bbc.com/news/technology-35642749

[16] The Guardian (2016). Yet another car can be hacked – Mitsubishi Outlander, 2016. Available: https://www.theguardian.com/technology/2016/jun/06/mitsubishi-outlander-car-hacked-security

[17] Newsis (2025). 재규어 랜드로버 사이버 공격, Available: https://www.newsis.com/view/NISX20250918_0003334757

[18] Wired (2022). BMW's Heated-Seats-as-a-Service Model Has Drivers Seeking Hacks, Available: https://www.wired.com/story/bmw-heated-seats-as-a-service-model-has-drivers-seeking-hacks/

[19] Daum News, (2024). 폭스바겐 Cariad 80만대 데이터 유출, Available: https://v.daum.net/v/qvxTZwbYok

[20] Karn Dhingra (2022). SiriusXM hack unlocks, starts cars, Available: https://www.autonews.com/mobility-report/siriusxm-hack-unlocks-starts-cars/

[21] Dailysecu, (2024) 토요타·닛산 대규모 데이터 유출, 2024. Available: https://www.dailysecu.com/news/articleView.html?idxno=158641

[22] TCOBiz, 토요타 미국 법인 데이터 유출 추가 보도, 2024. Available: https://m.tcobiz.net/article/관련뉴스/2/501/

[23] MIT Technology Review (2020), Hackers can trick a Tesla into accelerating by 50 miles per hour, Available: https://www.technologyreview.com/2020/02/19/868188/hackers-can-trick-a-tesla-into-accelerating-by-50-miles-per-hour/

[24] 곽중희. (2023). CES 2023, 사람을 위한 보안·안전 기술 빛났다. 월간 Secu N,, 98-100.

[25] 박세환. (2024). C-ITS 상용화를 통한 스마트 모빌리티 도시 구축 이슈. 월간 Secu N,, 100-103.

[26] Li, B., Hu, W., Da, L., Wu, Y., Wang, X., Li, Y., & Yuan, C. (2024). Over-the-air upgrading for enhancing security of intelligent connected vehicles: a survey. Artificial Intelligence Review, 57(11), 314.

[27] 전상훈. (2023). 스마트 모빌리티를 위한 Secure OTA 연구 동향. 정보과학회지, 41(12), 38-46.

[28] 이상우, 전용성. (2025). 자율주행차 통신 보안 표준화 현황. 한국통신학회지(정보와통신), 42(3), 24-30.

[29] Hamza, H., et al. (2024) Cybersecurity in Autonomous Vehicles: A Comprehensive Review

Study of Cyber-Attacks and AI-Based Solutions. International Journal of Engineering Trends and Technology

[30] Uptane (2019) IEEE ISTO 6100.1.0.0 Uptane Standard, Available: https://uptane.org/papers/ieee-isto-6100.1.0.0.uptane-standard.pdf

[31] Trishank, K., Akan, B., Sebastien, A., Damon, M., Russ, B., Cameron, M., Sam L., André W. & Justin, C. (2016). Uptane: Securing software updates for automobiles. The 14th Escar Europe.

[32] NDSS Symposium (2024), Scudo: Securing Automotive OTA and Supply Chain, Available: https://www.ndss-symposium.org/wp-content/uploads/vehiclesec2024-15-paper.pdf

[33] Charlie Miller, Chris Valasek (2015, 8), Remote Car Hacking Research (Jeep Cherokee), Available: https://illmatics.com/Remote%20Car%20Hacking.pdf

[34] AUTOSAR (2019). Secure Onboard Communication (SecOC) Specification, Available: https://www.autosar.org/fileadmin/standards/R19-11/CP/AUTOSAR_SWS_SecureOnboardCommunication.pdf

[35] Andrey Fadin, (2024) Kaspersky OS, Vulnerability in Kia / Hyundai Remote Control Systems, Available: https://os.kaspersky.com/blog/vulnerability-in-kia-car-remote-control-systems

[36] Chowdhury, T., Lesiuta, E., Rikley, K., Lin, C. W., Kang, E., Kim, B. Shiraishi, S., Lawford, M. & Wassyng, A. (2018) Safe and secure automotive over-the-air updates. In International Conference on Computer Safety, Reliability, and Security (pp. 172-187). Cham: Springer International Publishing.

[37] Almutairi, S., & Barnawi, A. (2023). Securing DNN for smart vehicles: An overview of adversarial attacks, defenses, and frameworks. Journal of Engineering and Applied Science, 70(1), 16.

[38] Steve Povolny, (2020) Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles, Available: https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/

[39] Wang, N., Luo, Y., Sato, T., Xu, K., & Chen, Q. A. (2023). Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4412-4423).

[40] Ibrahum, A. D. M., Hussain, M., & Hong, J. E. (2024). Deep learning adversarial attacks and defenses in autonomous vehicles: A systematic literature review from a safety perspective. Artificial Intelligence Review, 58(1), 28.

[41] Aloraini, F., Javed, A., & Rana, O. (2024). Adversarial attacks on intrusion detection systems in in-vehicle networks of connected and autonomous vehicles. Sens (Basel Switz) 24 (12): 3848.

[42] 주승환. (2024). 전기차와 PnC 충전인프라의 사이버보안 상호운용성을 위한 PKI 설계. 한국자동차공학회 춘계학술대회, 823-824.

[43] 이종국. (2025). EVSE 및 EV 인증서 폐기 (Revocation) 메커니즘의 부재가 초래하는 보안 위험 분석. 대

한전기자동차학회 학술대회자료집, 42-43.

[44] 이용우, 강신욱, 이광범, (2020, 12). 자율주행차 윤리 · 보안 · 안전 가이드라인 3종 발표, 뉴스레터, 법무법인 세종

[45] 자동차관리법 시행규칙 (2025, 12). 국토교통부령 제1539호, 법제처 국가법령정보센터

주제원고

AX 시대, 인공지능 보안 윤리의 현주소와 과제

계명대학교
이명숙

1. 서론

오늘날 인류는 인공지능이 인간의 삶과 사회 구조를 근본적으로 재편하는 AX(Artificial Intelligence Transformation) 시대의 한가운데 서 있다. 이 시대는 단순한 디지털 전환(DX)을 넘어, 인공지능이 인간의 인식, 판단, 의사결정, 심지어 일상 언어와 감성 영역까지 깊숙이 융합되는 시대로 진입하고 있다. AI는 더 이상 도구가 아니라, 인간과 상호작용하는 '지능적 존재'로서 기능할 가능성을 가지고 있다. 이러한 변화의 중심에 보안 윤리가 있으며, 그것은 기술적 안전성 확보를 넘어 사회적 신뢰, 인간 책임, 윤리적 거버넌스 구축 등 복합적인 과제의 핵심이다.

이 논문은 AX 시대에 새로운 위협이 어떻게 출현하는지, 보안 윤리가 어떻게 재정의되어야 하는지, 책임 윤리와 거버넌스를 어떻게 설계해야 하는지, 그리고 시민과 교육 차원에서 어떤 준비가 필요한지를 알아보고자 한다.

2. 신뢰 중심의 보안 윤리, 기술을 넘어 윤리로

기존의 보안 개념은 해킹, 악성코드, 데이터 유출 등 기술적 위협을 중심으로 정의되었다. 그러나 AX 시대의 보안 윤리는 기술을 넘어 신뢰 구축이라는 윤리적 명제를 중심에 놓아야 한다. 인공지능 시스템이 대량의 데이터를 처리하여 판단을 내리는 과정은 인간의 사생활과 사회적 관계망에 깊이 연결된다. 따라서 이러한 과정 전반에서 투명성, 공정성, 책임성이 확립되어야 한다. 이것을 위해서는 다음 원칙들이 뒷받침되어야 한다.

먼저, 동의와 목적 제한의 원칙이다. 데이터는 언제나 정보 주체의 자유로운 동의 아래 수집되어야 하며, 수집 목적은 명확하게 정의되어야 한다. 또한 그 목적은 엄격히 제한되고 재사용 시 명시적 승인이 있어야 한다. 이는 GDPR(General Data Protection Regulation) 등의 개인정보보호 원칙과도 연계된다. 그러나 현재 문제는 정보 제공자가 동의하지 않으면 웹사이트 읽기 권한까지 없기때문에 동의할 수밖에 없는 구조로 되어 있다.

설명 가능성과 해석 가능성의 원칙이다. AI의 판단은 단순한 블랙박스가 되어선 안 된다. 인간이 이해할 수 있도록 설명가능한 인공지능(Explainable AI, XAI)의 원칙이 필수적이다. XAI는 모델 수준 해석, 사후 설명, 입력-출력 관계 분석, 인과관계 분석 기법 등을 포함한다[1][2]. 예컨대 NIST는 XAI 시스템이 제공해야 할 네 가지 원칙으로 '설명, 의미성, 설명 정확성, 지식 한계'를 제시했다[3]. 이러한 설명 가능성과 해석 가능성은 단순히 기술적 요구사항을 넘어, AI 활용

의 신뢰성과 책임성을 확보하는 핵심 요소이다. 따라서 XAI는 단순히 모델 성능 향상을 위한 부가 기능이 아니라, AI 시스템을 안전하고 윤리적으로 운영하기 위한 필수 조건으로 자리 잡고 있다. 다만 모든 설명이 인간의 의사결정을 반드시 향상시키는 것은 아니며, 일부 실험에서는 설명이 실제로 인간의 판단 능력을 개선시켰다는 확실한 증거가 부족하다는 연구도 있다[4]. 그러므로 설명가능한 설계는 단순한 기술이 아니라 인간 이해의 관점에서 설계된 경험이어야 한다.

편향 제거와 공정성 확보의 원칙이다. AI가 학습하는 데이터는 사회의 구조적 불평등을 반영하는 경우가 많다. 따라서 AI는 차별을 재생산하거나 확대시킬 위험을 가지고있다. 보안 윤리는 데이터수집, 전처리, 알고리즘 설계단계에서부터 편향을 감지하고 제거하는 절차를 내재화해야 한다. 특히 민감 속성(성별, 인종, 나이, 정치적 성향 등)에 대한 불공정성 경계가 엄격히 적용되어야 한다. 사례로 미국에서 가장 상태가 나쁜 환자들의 치료는 부분적으로 알고리즘에 의해 결정된다고 한다. 특히 환자의 인종이나 사회경제적 배경에 따라 치료 추천이나 진단 우선순위가 달라지는 문제가 보고되고 있다[5].

프라이버시 강화 기술 활용의 원칙이다. 민감 정보를 보호하기 위해서는 동형암호, 연합 학습, 차등 프라이버시, 보안 다자간 계산 등 프라이버시 강화 기술이 적극 활용되어야 한다[6]. 이러한 기술은 AI가 민감한 데이터를 직접 보지 않고도 학습하거나 예측할 수 있게 한다. 이처럼 신뢰 중심 보안 윤리는 기술적 통제뿐 아니라 사회적 신뢰를 내재화한 윤리 구조이다. 기술은 도구일 뿐이며, 그것을 신뢰할 수 있게 만드는 관점이 바로 보안 윤리의 핵심이다.

3. AX 시대의 새로운 위협과 보안 윤리의 확장

AX 시대에는 우리가 익숙한 위협 유형 외에도 새로운 유형의 공격과 왜곡이 출현할 수 있다. 이들 위협은 단순히 보안 사고에 그치지 않고, 사회 구조와 인간 신뢰를 무너뜨릴 수 있는 힘을 가지고 있다. 그 위험들을 몇 가지 정리해 보면 다음과 같다.

판단 조작과 오작동의 위험을 가져올 수 있다. AI 시스템의 판단이 오류를 내거나 오작동할 경우, 그 피해는 단순한 시스템 오류를 넘어서 사회적 혼란을 초래할 수 있다. 예를 들어 의료 진단 AI가 잘못된 판단을 내린다면 환자의 생명이 위태로울 수 있다. 또한 AI가 제공한 정보가 표적 공격자에 의해 조작될 가능성도 있다. 이것은 영국 NHS(국민건강서비스)가 사용하던 AI 도구가 환자의 의료기록을 잘못 생성해 당뇨병 환자가 아닌 사람을 당뇨병 스크리닝 대상으로 초대한 일을 예로 들 수 있다[7].

악의적 사용과 여론 조작이 가능하다. AI는 스팸 및 피싱 공격뿐 아니라, 정치 선전, 여론 왜곡, 가짜 뉴스 생성 같은 고도의 사회 공작에도 활용될 수 있다. AI 기반 페이크 콘텐츠, 봇 네트워크, 자동 스피어 피싱은 대규모 심리전 도구가 될 수 있다. 이 경우 보안 윤리는 사회적 안전 개념을 포괄해야 한다. 특히 보안 전문가들도 AI를 이용하여 공격을 자동화하거나 적응형 위협을 만들어 낼 수 있다. 따라서 공격과 방어를 대비하는 쪽 역시 윤리적 기준 아래 AI를 활용해야 한다는 내부 감시와 책임이 필요하다[8].

자율 시스템 간 갈등과 충돌할 수 있다. 자율주행차, 자율 드론, 자율 무기 체계 등은 여러 AI 시스템이 상호작용하는 영역이다. 이 경우 각 시스템의 판단 기준이 충돌하거나 상충할 수 있다. 예컨대, 충돌 회피 전략과 목표 달성 전략이 충돌할 수 있으며, 이때의 판단 우선순위는 윤리적으로 설계되어야 한다.

데이터 조작 및 탈취가 가능하다. AI가 학습하는 데이터가 악의적으로 변조되거나 유출되면, 악성 조작된 AI가 만들어질 수 있다. 예컨대 학습 데이터에 뒤섞인 미묘한 왜곡된 예제를 통해 AI의 판단이 몰래 조작되는 데이터 중독 공격이 가능하다. 또 통신 채널에서의 도청 및 위조 공격도 위협 요소이다. 이런 위협들에 대응하기 위해서는 보안 윤리는 단편적 기술 대응을 넘어, 법률적·제도적 윤리 체계를 포괄하는 확장된 관점을 가져야 한다. 즉, AI 시스템 설계, 운영, 평가 전 영역에 윤리기준을 내재화하는 윤리 설계 접근이 필요하다. 또한 기존의 제로 트러스트(Zero Trust) 개념은 AI 시스템에도 확대해서 적용될 수 있다. 즉, 내부 구성요소나 사용자도 기본적으로 신뢰되지 않으며, 매번의 상호작용마다 인증과 검증 절차가 필요하게 된다. 이는 기술적 보안뿐 아니라, 인간-기계 간 경계를 윤리적으로 설정하는 장치가 될 수 있다.

4. 인간 중심 책임 윤리와 거버넌스 설계

AI의 자율성이 증가할수록 판단 결과에 대한 책임 주체가 모호해지는 윤리적 공백이 발생하여 사회적 갈등과 시스템에 대한 불신을 발생시킬 수 있다. 이러한 모호성은 시스템의 오작동뿐만 아니라 악용 시 대규모 피해로 이어질 수 있는 심각한 문제이다. 이에 대처하기 위해서는 AX 시대의 보안 윤리는 '어쨌든 책임은 인간이 질 수밖에 없다'는 근본적인 이해를 바탕으로 인간 중심 책임 원칙을 핵심 축으로 삼아야 한다. 이러한 원칙을 효과적으로 구현하기 위해서는 거버넌스 설계에서 '책임의 다층적 구조', '법제도와 규제 프레임워크', '윤리기준의 내재화', '국제 거버넌스와 협력' 부분에 보다 집중하여 구체적인 설계가 이루어져야 한다.

AI 시스템 설계는 책임의 다층적 구조 형태로 설계되어야 한다. 이를 위해 AI 시스템의 기획, 설계, 개발, 운영, 폐기의 모든 단계에 걸쳐 책임 주체들을 명확히 정의해야 한다. 예컨대, 개발자는 알고리즘 설계와 모델링 과정에서 윤리기준을 내재화할 책임, 운영자와 사용자 조직은 AI 시스템을 실제 현장에 적용할 때의 모니터링과 제어에 대한 책임, 감독 기관과 정책 입안자는 윤리기준과 법제를 설계하고 감시하는 책임, 사용자와 소비자는 AI 사용에 따른 윤리적 선택과 책임, 이와 같은 다층적 책임 체계는 단순한 '누가 잘못했느냐'의 문제를 넘어서 공동 책임 체계로 작동해야 한다.

법 제도와 규제 프레임워크를 구축해야 한다. 이미 국제적으로 AI 거버넌스 논의는 활발하다. 유럽연합의 인공지능법(AI Act)은 위험 기반 접근을 도입하여 고위험 AI 시스템에 대해 엄격한 투명성, 설명성, 책임 규제를 부과한다. 이는 책임 윤리와 거버넌스 모델의 실질적 구현 시사점을 제공한다. 또한 전 세계적으로는 유네스코 회원국 194국에 AI 윤리 가이드라인 권고안이 발표되었으며, 이들은 투명성, 공정성, 비해악, 책임성, 프라이버시 보호 등을 핵심 가치로 제시한다[9][10]. 이들 가이드라인은 공통된 윤리 원칙을 제공하지만, 실질적 거버넌스 구조로 전환되기 위해선 제도적 강제력과 감시 장치가 보완되어야 한다.

윤리기준의 내재화는 AI 설계단계에서 이루어져야 한다. 보안 윤리는 사후 규제 중심이 아니라 설계단계(Stage 0)부터 윤리기준을 넣는 것으로 접근이어야 한다. 이른바 '윤리 설계'이다. 이것은 기술 설계자가 윤리적 고려를 기본 요건으로 삼는 것을 의미하며, 윤리적 영향 평가를 초기 기획 단계에서 수행, 윤리기준 체크리스트를 기술 설계, 모델 개발, 시스템 통합 등 전 단계에 배치, 설계 반복과 감시 메커니즘: 중간 점검, 외부 감리, 윤리적 감사, 탈윤리 실패 복구 구조: 오류나 악용 가능성 발견 시 신속히 대응할 책임 및 절차 마련과 같은 요소들이 포함된다. 이와 같은 구조는 단순히 기술적 규범이 아니라, 윤리적 거버넌스 체계 그 자체로 기능해야 한다.

국제 거버넌스와 협력체계로 설계되어야 한다. AI 윤리 문제는 특정 국가나 지역의 국경을 넘어선 글로벌 아젠다이다. AI 기술의 영향력은 전 세계적이므로, 국가 간의 단편적 규제로는 효과적인 통제가 불가능하다. 따라서 국제 수준의 협력체제와 거버넌스 네트워크 구축이 필수적이다. 예를 들어 IASEAI(International Association for Safe & Ethical AI)와 같은 기구는 글로벌 윤리기준 마련과 정책 협력을 촉진하는 역할을 수행하며, 이러한 민관협력 기구의 역할이 더욱 중요해지고 있다. 또한, 규제의 국제적 파편화를 방지하고 일관성을 확보하기 위해 국가 간 공조를 토대로 국제 표준 및 감시 체계를 마련해야 한다. 더 나아가 AI 안전 및 윤리 평가 기관을 상호 인정하는 체제를 구축하여, 한 국가에서 검증된 AI 시스템이 다른 국가에서도 신뢰받을 수 있도록 함으로써 글로벌 윤리 생태계를 구축하는 방향으로 나아가야 한다.

5. 교육을 통한 AI 윤리 리터러시의 확산

보안 윤리는 단지 기술자나 규제자의 몫이 아니다. 시민 개개인이 AI의 작동 원리와 윤리적 함의를 이해하고 비판적으로 사고할 수 있어야 AI가 건강한 사회적 제도로 정착할 수 있다. 그러기 위해서는 AI 윤리 리터러시가 필요하다. AI 윤리 리터러시란 단순히 'AI를 사용하는 방법'을 넘어, AI가 인간 삶과 사회 구조에 미치는 영향을 통찰하고 판단하는 능력이다. 이것은 기술적 이해, 윤리적 성찰, 비판적 사고가 결합된 역량이다. 리터러시 수준이 높을수록 이용자는 AI의 편향 가능성, 왜곡 가능성, 책임의 문제 등을 스스로 인지할 수 있다. 따라서 AI 활용이 확대될수록 윤리 리터러시는 사회적 차원의 필수 역량이 될 수밖에 없다.

또한 학교와 대학에서는 다음과 같은 방식으로 AI 윤리 교육을 통합해야 한다. 융합 교육 과정을 통해 윤리 교육이 이루어져야 한다. 지금처럼 하나의 전공에서 윤리 교육이 이루어지는 것이 아니라 융합 교육 과정을 통해 인공지능 공학, 컴퓨터 과학, 사회학, 철학, 법학을 통합한 학문 간 융합 교육이 이루어져야 한다. 프로젝트 기반 학습에서는 기존의 학습 프로세스에서 윤리 부분을 추가하여 학생들이 AI 시스템을 설계하거나 분석하는 단계에서 윤리적 쟁점을 직접 다루게 하는 방법이 있다. 케이스 스터디 중심 교육에서는 실제 AI 사고 사례, 윤리적 갈등 사례, 규제 대응 사례 등을 중심으로 한 교육으로 이루어질 수 있다.

그리고 교사(교수자) 역량 강화를 통해 교사 스스로 AI 기술과 윤리적 문제를 깊이 이해하고, 이를 학습자에게 효과적으로 지도할 수 있는 구체적 방안을 마련해야 한다. 특히 교사는 AI 윤리 교육의 방향을 설계하고, 학습자가 직면할 수 있는 다양한 윤리적 쟁점을 분석하도록 안내하며, 학생들이 실제 상황에서 윤리적 판단과 실천을 할 수 있도록 지도하는 핵심 주체이다. 따라서 교사 역량 교육은 단순한 보조적 활동이 아니라 AI 윤리 교육의 성패를 좌우하는 필수적 단계이며, 다른 교육 요소보다 우선적으로 추진되어야 한다. 이 과정에서 교사는 최신 AI 기술과 관련 법제, 윤리적 이슈에 대한 이해를 심화하고, 이를 바탕으로 교육과정 설계와 사례 기반 학습 지도 능력을 향상시켜야 한다. 이러한 체계적 역량 강화는 단기적으로는 교사의 전문성을 높이는 효과를 가지며, 장기적으로는 학습자들에게 기술 사용을 넘어 사회적·윤리적 책임을 수행할 수 있는 디지털 시민성의 기반을 제공하게 된다. 결과적으로 교사 역량 교육의 강화 없이는 AI 윤리 리터러시의 확산과 실질적 사회적 효과를 기대하기 어렵다.

그러나 윤리 리터러시는 교육만으로 실현되기는 어렵다. 사회적 문화, 미디어, 정책 담론 전반이 AI 윤리 문제를 적극적으로 다루어야 한다. 예를 들어 언론은 AI의 오류, 편향, 사고 사례를 지속적으로 보도하고, 시민 토론을 유도해야 한다. 정부와 기업은 AI 윤리기준을 공개하고, 투명

성을 확보한 소통을 해야 한다. 이처럼 윤리 리터러시는 사회적 공감대와 담론 구조까지 포함하는 개념이며, 보안 윤리가 단순 기술 논쟁을 넘어 사회적 의제로 자리 잡게 한다.

6. 미래 전망과 도전 과제

AX 시대의 인공지능 보안 윤리는 단지 현재의 문제에 대한 대응을 넘어, 미래의 기술 변화와 인간 가치의 조화를 설계하는 방향타가 되어야 한다. 다음은 주요 전망과 도전 과제이다.

첫째, 인간-기계 공생의 윤리적 패러다임 재정립이 필요하다. AI가 인간의 인지·판단 영역에 점점 깊이 개입할수록, 우리는 '인간 중심' 패러다임을 어떻게 재정립할 것인가라는 질문을 하게 된다. AI는 인간의 확장된 인지 파트너가 되어야 하며, 인간의 존엄과 창의성을 저해하지 않는 구조로 설계되어야 한다.

둘째, 설명성과 신뢰성의 긴장 관계에 놓여있다. 설명가능한 AI 기술이 진화하더라도, 복잡한 모델일수록 설명 가능성과 성능 간의 긴장이 팽팽하다. 사용자가 이해할 만한 수준의 설명을 제공하면서도 모델의 정확성을 유지하는 균형은 기술적으로 매우 어려운 과제이다. 더욱이 설명이 제공된다고 해서 반드시 신뢰로 연결되지는 않는다는 실증 연구도 있다[5].

셋째, 책임 귀속이 윤리적으로 복잡해진다. 인간-기계 공존 체계에서는 책임이 누구에게 있는지를 구별하는 것이 복잡해진다. AI의 판단 오류에 대해 누가 책임을 질 것인가? 단 하나의 책임 주체로 귀속되는 것이 적절할까? 아니면 책임을 분할하고 공동 책임 체계를 구성할 것인가? 이러한 문제는 윤리적·법률적 숙의가 필요한 쟁점이다.

넷째, 국제 표준과 감시 메커니즘을 구축해야 한다. 국제적 AI 윤리기준과 감시 메커니즘은 아직 초기 단계이다. 각국의 규제 수준은 매우 다르며, 기술 경쟁과 국가 이익 충돌이 윤리 협력을 방해할 수 있다. 따라서 국제 기준 마련, 상호 인정 체계, 독립적 평가 기관 구축은 시급한 과제다.

다섯째, 기술 인력 부족과 전문성 확보가 절실하다. AI 윤리와 보안 전문 역량을 갖춘 인력은 전 세계적으로 매우 부족하다. 기업들은 AI 윤리 전문가, 거버넌스 전문가 수요에 비해 공급이 부족하다는 우려를 표명하고 있다. 이는 윤리 기반 기술 확산의 현실적 제약이 된다.

7. 결론

AX 시대, 인공지능은 더 이상 단순한 도구가 아니다. 그것은 인간과 상호작용하며 판단하고, 사회 질서에 개입하는 지능적 주체로 부상하는 가능성을 가지고 있다. 이러한 변화는 우리에게 기술적 진보 못지않게 신뢰와 책임의 윤리적 기반을 요구한다. 보안 윤리는 더이상 기술의 부수적 고려 사항이 아니라, 인공지능 전환시대의 핵심 원리이다. 이를 위해 기술적 안전장치, 법제적 규제, 윤리적 설계, 교육과 문화적 담론이 유기적으로 결합해야 한다. 그래야만 AI가 인간의 삶을 위협하는 존재가 아니라 인간의 가능성을 확장하는 동반자로 자리매김할 수 있을 것이다.

또한, AI의 발전 속도와 범위가 급격히 확장됨에 따라, 단기적 안전장치와 규제만으로는 충분하지 않다. 사회 전체가 AI 기술의 윤리적 활용과 위험 관리에 참여하는 참여적 거버넌스 모델이 필요하다. 각국 정부, 국제기구, 산업계, 학계, 시민사회가 협력하여 공통의 윤리기준과 책임 체계를 마련하고, 이를 지속적으로 검토하고 갱신하는 체계적 노력이 요구된다. 아울러 교육과 문화적 담론은 단순한 기술 이해를 넘어, AI와 인간이 함께 살아가는 사회적·윤리적 상상력을 확장하는 역할을 해야 한다.

우리는 지금 신뢰할 수 있는 기술을 만들어야 한다. 보안 윤리는 그 신뢰의 토대이며, 우리가 AI와 공생하는 사회를 설계하는 나침반이다. 이 길이 험할지라도, 인간 존엄과 사회적 연대가 담보된 AI 미래를 향해 우리는 멈출 수 없다. 나아가, 이러한 노력이 집약될 때 AI는 단순한 도구를 넘어, 인간의 삶을 풍요롭게 하고 사회적 문제 해결에 기여하는 혁신적 동반자로 자리매김할 수 있을 것이다. 따라서 기술 개발과 윤리적 책임은 결코 분리될 수 없으며, 두 요소의 조화와 통합적 접근만이 지속 가능한 AI 사회를 가능하게 한다.

8. 참고문헌

[1] S. Ali et al., “Explainable Artificial Intelligence (XAI): What we know and what is missing.” *Information Fusion*, Vol. 99, pp. 1-52, 2023. https://doi.org/10.1016/j.inffus.2023.101805

[2] Y. Wenli, et al., “Survey on Explainable AI: From Approaches, Limitations and Applications Aspects”, *Human-Centric Intelligent Systems*, Vol. 3, No. 4, pp. 161-188, Aug. 2023. DOI: 10.1007/s44230-023-00038-y

[3] J. K. Bae, “A Study on the Legislative Bill of Artificial Intelligence Act and the Basic Principles of Explainable Artificial Intelligence (XAI),” Journal of the Korea Knowledge Information Technology Society, Vol. 18, No. 2, pp. 439-448, 2023. DOI: 10.34163/jkits.2023.18.2.017

[4] Y. Alufaisan, et al., “Does explainable artificial intelligence improve human decision-making?”, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 8, pp. 6618-6626, Feb. 2021. DOI: 10.1609/aaai.v35i8.16819

[5] Tom Simonitem, A health care algorithm offered less care to black patients, TECHNICA, Oct. 2019. https://zrr.kr/WCxm5q

[6] Wikipedia. “Trustworthy AI”. https://en.wikipedia.org/wiki/Trustworthy_AI

[7] Beatrice Nolan, UK health service AI tool generated a set of false diagnoses for one patient that led to him being wrongly invited to a diabetes screening appointment, FORTUNE, Jul. 2025. https://zrr.kr/dostUi

[8] C. Owen-Jackson, “Navigating the ethics of AI in cybersecurity”, IBM Think Insights, https://www.ibm.com/think/insights/navigating-ethics-ai-cybersecurity

[9] UNESCO, https://www.unesco.org/en/

[10] N. K. Corrêa, et al. “Worldwide AI ethics: A review of 200 guidelines and frameworks”. Patterns, Vol. 4, No. 10, pp. 1-14, Oct. 2023. https://doi.org/10.1016/j.patter.2023.100857

[11] International Association for the Study of Ethics in AI (IASEAI). https://www.iaseai.org/

우수연구실 소개

동의대학교
스마트IT연구소

연구소장	장종욱		전문분야	차량 네트워크, 데이터 통신, IoT, AI
소속	직 장 명	동의대학교 컴퓨터공학과	직위	교수
	주 소	부산진구 엄광로 176, 동의대학교 정보공학관 804호		
	전화번호	051-890-1709	E-mail	jwjang@deu.ac.kr
주요경력	- 2017 ~ 2019 : 동의대학교 대외부총장 - 1995 ~ 現 : 동의대학교 컴퓨터공학과 교수 - 1987 ~ 1995 : 한국전자통신연구원(ETRI) 연구원 - 1995 : 부산대학교 컴퓨터공학 박사			
연구실적	- 연구개발과제 : 64건 수행 - 특허(국내/국외) : 57건 (국외 2건포함)		- 기술이전 : 21건 - 논문실적(국내/국외저널) : 193편	
현재 수행 활동	- 동의대 스마트IT연구소장 (2011~現) - 부산과학기술자문단장(現) - 대한무역투자진흥공사 자문위원(現)		- (사)한국정보통신학회 상임이사(現) - (사)한국인터넷방송통신학회 자문부회장(現)	

연구실 소개

동의대학교 스마트IT연구소는 컴퓨터공학과를 기반으로 임베디드 시스템, 인공지능 기반 영상처리, 스마트 모빌리티, IoT 융합 기술 등 첨단 ICT 융합 분야를 폭넓게 연구하고 있습니다. 본 연구소는 15년 이상의 연구 경험과 다양한 산학협력, 정부 과제 수행을 통해 하드웨어와 소프트웨어, 그리고 인공지능 기술의 융합을 선도하고 있습니다.

본 연구소는 학생들에게 실무 중심의 연구 경험과 최신 기술 습득 기회를 제공하며, 졸업 후 IT, 반도체, 모빌리티, 인공지능, 로봇 등 다양한 첨단 산업 분야로 진출할 수 있는 역량을 키워줍니다. 앞으로도 첨단 ICT 융합 기술의 발전과 실용화에 기여하는 연구실로 성장해 나갈 것입니다.

〈 스마트IT연구소 연구 네트워크 〉

연구 분야

임베디드 시스템

임베디드 시스템 분야에서는 저전력, 고성능 하드웨어 설계와 소프트웨어 최적화가 중요한 연구 주제입니다. 본 연구소는 임베디드 보드에서의 딥러닝 모델 효율화, 하드웨어 자원 절감형 알고리즘 개발, 실시간 데이터 처리 시스템 구현 등

다양한 프로젝트를 수행해왔습니다. 이를 통해 스마트카, 자율주행, 스마트팜, 헬스케어 등 다양한 산업 분야에 적용 가능한 임베디드 솔루션을 개발하고 있습니다.
이러한 연구는 하드웨어와 소프트웨어의 융합을 통해 차세대 지능형 시스템의 기반을 마련하는 데 기여하고 있습니다. 실제 산업체와의 협력, 정부 과제 수행, 특허 출원 등 실용적 성과도 풍부하며, 학생들에게는 실무 중심의 연구 경험을 제공하고 있습니다.

인공지능 기반 영상처리 및 지능형 시스템

본 연구소는 인공지능(AI)과 딥러닝 기술을 활용한 영상처리 및 지능형 시스템 개발에 주력하고 있습니다. 최근에는 객체 인식, 영상 분할, 이미지 전처리, GAN 기반 데이터 생성 등 다양한 AI 영상처리 기술을 연구하고 있으며, 실제 교통, 의료, 산업 현장에 적용 가능한 솔루션을 개발하고 있습니다. 예를 들어, 차량 탑승 인원 감지, 해양 침적 쓰레기 감지, 의료영상 분석 등 다양한 응용 분야에서 성능을 입증하고 있습니다.
특히, 하드웨어 자원 제약이 있는 환경에서의 AI 모델 경량화, 실시간 처리, 엣지 컴퓨팅 기반의 영상처리 시스템 구현에 많은 노력을 기울이고 있습니다. 이를 위해 하드웨어-소프트웨어 협력 설계, 최적화된 데이터 전처리, 효율적인 신경망 구조 설계 등 다양한 기술을 통합적으로 연구하고 있습니다. 또한, 블록체인, IoT, 로봇 플랫폼 등과의 융합을 통해 지능형 시스템의 확장성과 신뢰성을 높이고 있습니다.
이러한 연구는 특허 출원, 논문 발표, 산학협력 프로젝트 등 다양한 성과로 이어지고 있으며, 실제 산업 현장에서 요구하는 실질적 문제 해결에 기여하고 있습니다. 학생들은 최신 AI 기술을 실습하고, 실제 데이터셋을 활용한 프로젝트 경험을 쌓을 수 있어, 졸업 후 다양한 IT 및 융합 산업 분야로 진출하고 있습니다.

스마트 모빌리티 및 IoT 융합 기술

스마트 모빌리티와 IoT 융합 기술은 본 연구소의 또 다른 핵심 연구 분야입니다. 차량 네트워크(OBD-II, CAN, MOST 등), 스마트 블랙박스, 어라운드뷰 시스템, 자율주행 보조 시스템 등 다양한 스마트 모빌리티 솔루션을 개발해왔으며, 실제 교통 환경에서의 데이터 수집, 분석, 제어 기술을 고도화하고 있습니다. 또한, LoRaWAN, 블록체인, 클라우드 기반의 차량 관제 및 진단 시스템 등 차세대 모빌리티 인프라 구축에도 앞장서고 있습니다.
IoT 융합 기술 측면에서는 센서 네트워크, 실시간 데이터 수집 및 처리, 엣지 컴

퓨팅, 스마트 팜, 헬스케어 등 다양한 응용 분야에 맞춘 시스템을 연구합니다. 예를 들어, 타이어 마모도 체크, 환경 모니터링, 스마트팜 로봇 플랫폼, 재난 예방 시스템 등 실제 산업 및 사회 문제 해결에 기여하는 다양한 프로젝트를 수행하고 있습니다. 이러한 연구는 정부, 지자체, 산업체와의 협력을 통해 실용적 가치를 창출하고 있습니다.

스마트 모빌리티와 IoT 융합 기술은 미래 도시와 산업의 핵심 인프라로 자리매김하고 있으며, 본 연구실은 이 분야에서의 선도적 역할을 지속적으로 확대하고 있습니다. 학생들은 다양한 실증 프로젝트에 참여하며, 현장 중심의 문제 해결 능력과 융합적 사고를 키울 수 있습니다.

수행 중인 연구 프로젝트

- CT·MVS 데이터를 활용한 AI 기반 FIP 구축 (2025.05 ~ 2025.12)
- 자체 표준 인터페이스(MXEIC)가 적용된 100kWh급 Battery Swapping ESS 및 통합관제시스템 기술개발 (2025.05 ~ 2025.12)
- 제조 현장 이동형 중장비의 중대재해 예방을 위한 온디바이스 AI 기반 실시간 충돌 방지 시스템 개발 (2025.10 ~ 2027.09)

최근 개발 기술 목록

① AI 기반 차량 탑승 인원 검지 시스템 기술
② 대비향상 기반 최적의 이미지 데이터 생성 기술
③ 다중 환경적 요인(조명, 해상도, 잡음)을 고려한 계층적 이미지 품질 개선 기술
④ 손 끼임 사고 방지를 위한 듀얼 카메라 기반 3D 위험영역 침범 판별 기술
⑤ AI 기반 불법주정차 차량 단속 시스템 기술
⑥ AI 기반 화재 감지 시스템 기술
⑦ LLM 기반 이미지 데이터셋 오토라벨링 기술
⑧ AI 기반 산업용 개인안전장비 착용 유무 판별 시스템 기술

학회 동정

하반기 주요활동

1 2025년 제3차 국제학술위원회의

- 일시 : 2025년 8월 25일(월) 오후 3시
- 장소 : 온라인(zoom)
- 안건
  - ICFICE 2025 결과 보고
  - ICFICE 2026 준비
  - 기타 안건

2 2025년 제3차 국내학술위원회의

- 일시 : 2025년 8월 26일(화) 오후 5시
- 장소 : 온라인(zoom)
- 안건
  - 2025년 추계학술대회 사전 답사 보고
  - 2026년 춘계학술대회 사전 답사 보고
  - 기타 안건

3 2025년 제3차 국문지 편집위원회의

- 일시 : 2025년 9월 17일(수) 오후 4시
- 장소 : 온라인(zoom)
- 안건
  - 과총 선정 결과 보고
  - 국문지 게재 관련 보고
  - 기타 안건

4 2025년 제3차 학회지 편집위원회의

- 일시 : 2025년 9월 23일(화) 오후 6시
- 장소 : 온라인(zoom)
- 안건
  - 학회지 현황 보고
  - 학회지 주제 선정
  - 기타 안건

5 2025년 제3차 영문지 편집위원회의

- 일시 : 2025년 9월 25일(목) 오후 4시
- 장소 : 온라인(zoom)
- 안건
  - 2025년 영문지 접수현황 보고
  - 2025년 9월 발간호 (23권 3호) 처리 현황 보고
  - 기타 안건

학회 동정

하반기 주요활동

6 2025년 추계종합학술대회 개최

- 일시 : 2025년 10월 24일(목) ~ 10월 26일(토)
- 장소 : 국립한국해양대학교
- 학술위원장 : 김세민(전주교육대학교)
- 논문 편수 : 구두 94편, 포스터 176편, 총 270편
- 우수논문 : 86편 (최우수논문 1편, 우수논문 49편, 학생우수논문 36편)
- 후원업체 : ㈜우리아이티, SK브로드밴드, 아이티센엔텍, LIG 시스템, ㈜피플랜, 대신정보통신㈜, 롯데이노베이트㈜, 메타넷디지털, ㈜아이티공간, 주식회사 케이티, LG유플러스, SK텔레콤, ㈜핸즈온테크놀러지, ㈜세오, ㈜신한항업, ㈜엠큐닉, ㈜하이제이컨설팅, 그린텍아이엔씨, 네오브릭스, 대보정보통신㈜, 세림TSG, 송암시스콤㈜, 아이씨티웨이, 엠티데이타, 올포랜드, 한국정보기술㈜, ㈜케이비아이

7 2025년 제4차 국제학술위원회의

- 일시 : 2025년 11월 25일(화) 오후 4시
- 장소 : 온라인(zoom)
- 안건
  - ICFICE 2026 준비
  - 기타 안건

8 2025년 제4차 국내학술위원회의

- 일시 : 2025년 11월 25일(화) 오후 5시
- 장소 : 온라인(zoom)
- 안건
  - 2025년 추계학술대회 결과보고
  - 2026년 춘계학술대회 준비
  - 기타 안건

9 2025년 제4차 국문지 편집위원회의

- 일시 : 2025년 12월 19일(금) 오후 2시
- 장소 : 온라인(zoom)
- 안건
  - 국문지 논문편집양식 변경 관련 보고
  - 신년회 수상 관련 보고
  - 기타 안건

10 2025년 제4차 영문지 편집위원회의

- 일시 : 2025년 12월 22일(월) 오후 4시
- 장소 : 온라인(zoom)
- 안건
  - 2025년 영문지 접수현황 보고
  - 2025년 12월 발간호 (23권 4호) 처리 현황 보고
  - 기타 안건

학회 동정

학회 갤러리

2025 창립기념일

2025 창립기념일

추계 이사회

추계 산학매칭 그린데이

초청강연

포스터 발표

JKIICE

Journal of the Korea Institute of Information and Communication Engineering

http://jkiice.org

한국정보통신학회 국문논문지는 한국연구재단 등재지로서 매달 발간되고 있습니다.

투고 분야

Information Science

- 데이터베이스, 클라우드컴퓨팅 및 빅데이터
- 인공지능 및 지능적 시스템
- 컴퓨터 비전 및 바이오 메디칼 영상
- 디지털 콘텐츠, 게임 및 멀티미디어

Communication Engineering

- 통신용 반도체
- 무선통신 및 데이터 통신
- 컴퓨터 네트워크
- 회로 및 시스템
- 정보 보호 및 보안

Information Science & Communication Engineering

- BT (바이오기술 융합) : 디지털 헬스케어, 유전체 데이터 분석, 바이오센서 데이터 전송
- CT (문화기술 융합) : VR/AR/XR, 디지털 미디어 스트리밍, 스마트 콘텐츠 제작
- IT (정보기술 융합) : AI 기반 데이터 통신, 클라우드 컴퓨팅, 블록체인 보안
- NT (나노기술 융합) : 나노센서 데이터 전송, 나노 스케일 데이터 분석
- ET (에너지기술 융합) : 스마트 그리드, 재생에너지 데이터 분석, 에너지 효율 최적화
- ST (우주기술 융합) : 위성 통신, 원격탐사 데이터 전송
- RT (로봇기술 융합) : 자율 로봇 통신, 협동 로봇 데이터 전송
- FT (금융기술 융합) : 블록체인 금융 보안, 실시간 거래 데이터 통신
- HT (헬스케어기술 융합) : 의료 영상 데이터 전송, 원격 의료
- TT (교통기술 융합) : 자율주행 차량 통신, 스마트 교통 시스템

Short Paper

- 뉴 아이디어 및 트렌드
- 정보통신일반

논문 양식

홈페이지(커뮤니티-자료실)에서 국문논문지 양식을 다운받아 작성하신 후 제출 해주시기 바랍니다.
※ 특별히 참고문헌 기술 형식을 투고 규정에 맞도록 제출하여 주시기 바랍니다.

투고 절차

1. 한국정보통신학회 신규회원가입 (http://www.kiice.org/ 참조)
2. 국문지 논문투고시스템(https://www.dbpiaone.com/jkiice/index.do)에 논문접수
   ※ 학회 홈페이지와 국문지 투고 시스템의 회원정보는 연동이 안되오니 최초 투고 시, 반드시 회원가입 후 투고해주시기 바랍니다.
3. 심사비 입금 (일반: 4만원 / 긴급: 8만원)
4. 심사 진행 (일반: 15일 / 긴급: 10일)
5. 심사통과 후 게재확정메일 발송 (이후 학회에서 게재료 메일발송)
6. 매월 말일 출판 (일반: 투고 후 3~4개월 / 긴급: 투고 시점부터 2~3개월)

문의 사항

사무국 051-463-3683/ journalokiice.org

회원여러분들의 관심과 투고 부탁드립니다.

JICCE
Journal of Information and Communication Convergence Engineering
http://jicce.org

(Indexedin Scopus)
PrintISSN : 2234-8255, OnlineISSN : 2234-8883

Call for Papers

Journal of Information and Communication Convergence Engineering (J. Inf. Commun. Converg. Eng., JICCE) is an official English journal of the Korea Institute of Information and Communication Engineering (KIICE).

It is an international, peer reviewed, and open access journal that is published quarterly in March, June, September, and December. Its objective is to provide rapid publications of original and significant contributions and it covers all areas related to information and communication convergence engineering including the following areas: communication system and applications, networking and services, intelligent information system, multimedia and digital convergence, semiconductors and communication devices, imaging and biomedical engineering, and computer vision and autonomous vehicles.

JICCE was indexed in Scopusand Scopus coverage of JICCE began in 2018.
- Scopus Cite Score 2023 of JICCE : 1.1

Authors areinvitedto submitthearticles that illustrate significant advances in theory, engineering, and application in thefield of information and communication convergence engineering.

Topics of interest include, but are not limited to, the following:

- Communication system and applications
- Networking and services
- Semiconductors and communication devices
- Intelligent information system
- Multimedia and digital convergence
- Imaging and biomedical engineering
- Computer vision and autonomous vehicles

Paper submission

All submitted manuscripts must (i) conform to JICCE formatting requirements (see “Instruction for Author”guidelines at http://jicce.org); (ii) must be original and should not have been published previously or be under consideration forpublication while being evaluated for this Journal; (iii) be submitted online at http://jicce.org.

For details of our publication please visit our website: http://jicce.org

Prof. KwangBaek Kim / Prof. Dongsik Jo
Editors-in-Chief, JICCE
Contact: journal@kiice.org

우리아이티를 통해 이루는 여러분의 Work Load

More Early I More Safely I More Fully

Partner

vmware SUSE AXGATE Security and Beyond CISCO

DELL Technologies NUTANIX Cloud

WiT (주)우리아이티
Leading company in IT convergence
부산시 동구 조방로 14 범일동, 동일타워 415호
TEL 051-637-2386 | FAX 050-5964-5555
www.wooriit.kr

SK broadband

14년 연속 국가고객만족도 1위

1등 서비스로
고객님의 사랑에 보답하겠습니다

NCSI

초고속인터넷/IPTV 부문

국가고객만족도(NCSI) 14년 연속 1위

Inspire with Technology

세상을 변화시키는 기술

우리는 고객에 대한 이해를 넘어서,
변화의 흐름을 읽고, 멈추지 않는 도전을 통해
지식과 기술로 세상을 혁신합니다.

깊은 신뢰감으로 늘 고객과 함께하고,
기술, 사람, 기업간 새로운 연결 속에서
새로운 경험과 가치를 발굴하며
아이티센은 지속가능한 변화를 실현합니다.

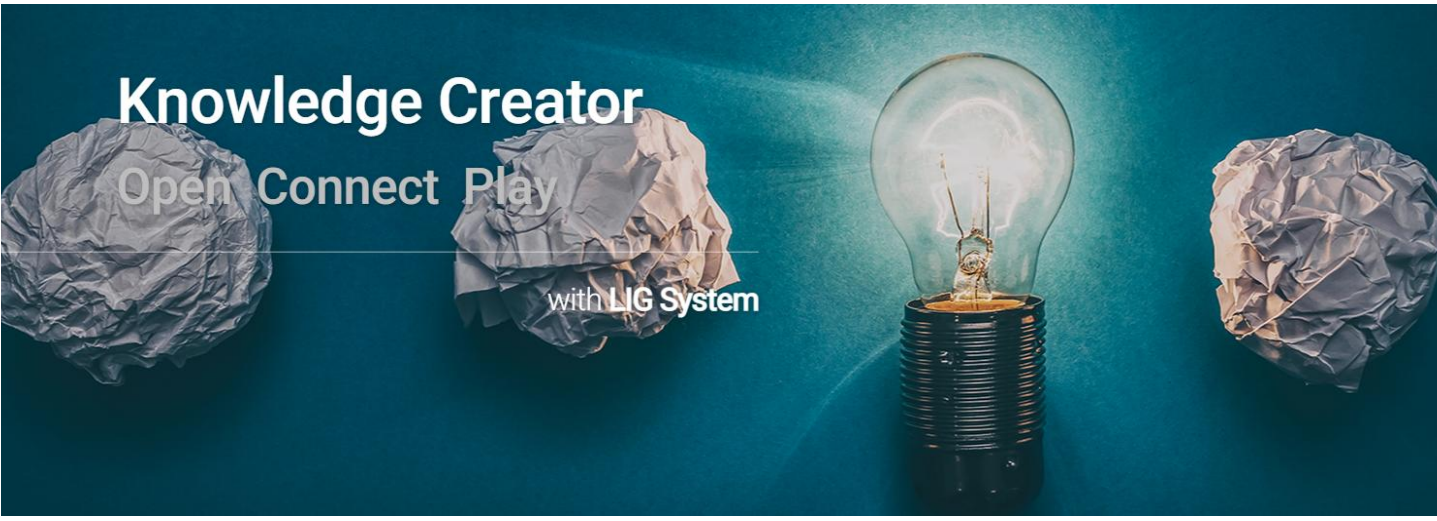
더 나은 내일을 향한 기술로
고객의 삶에 필요한 가치를 창조하는 것.
아이티센이 생각하는
우리의 역할이자 우리의 존재 이유입니다.

LIG System

프로세스 "**Innovation**"을 통해

자원의 효율성 및 생산성을 높여 고객의 성장을 보장하는 혁신 기업

주식회사 **엘아이지시스템** 입니다.



LIG

㈜엘아이지시스템
서울시 용산구 한남대로 98, 5층 (한남동, 일신빌딩) Tel. 02.6900.1600
www.ligs.co.kr

Global IT Leader!

모든 비즈니스 영역을 통합하는 통찰력으로
고객의 니즈를 완벽히 분석한 최적의 서비스로
미래를 선도하는 최첨단 기술력으로

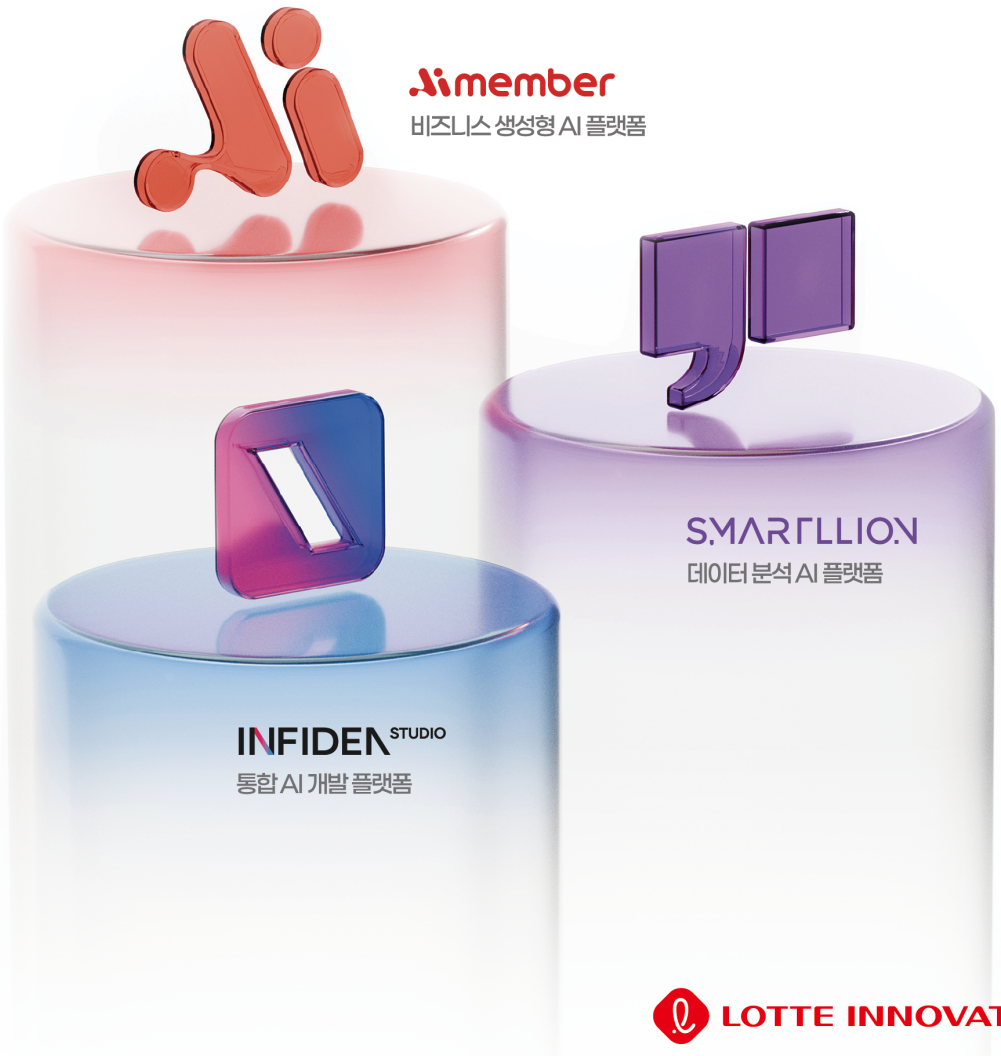
미래의 가치를 먼저 생각하는 기업

큰大믿을信
대신정보통신주식회사 Daishin Information & Communications Co., Ltd.

서울특별시 금천구 가산디지털2로 169-16, 6층 (가산동, 하우스디가산퍼스타) Tel_02-2107-5000 Fax_02-2107-0515 www.dsic.co.kr

INNOVATE AI
with LOTTE

AI **비즈니스**를 위한
Value Chain **전 영역 적용 가능**한 플랫폼



LOTTE INNOVATE

Unlocking the Value of
AI & Digital Transformation

메타넷은 AI와 클라우드 기반의 디지털 혁신을 선도하는
국내 최대 Full-Stack IT 서비스 파트너입니다.

디지털 비즈니스 플랫폼, 메타넷

Consulting | Digital | Technology | AI | Operations

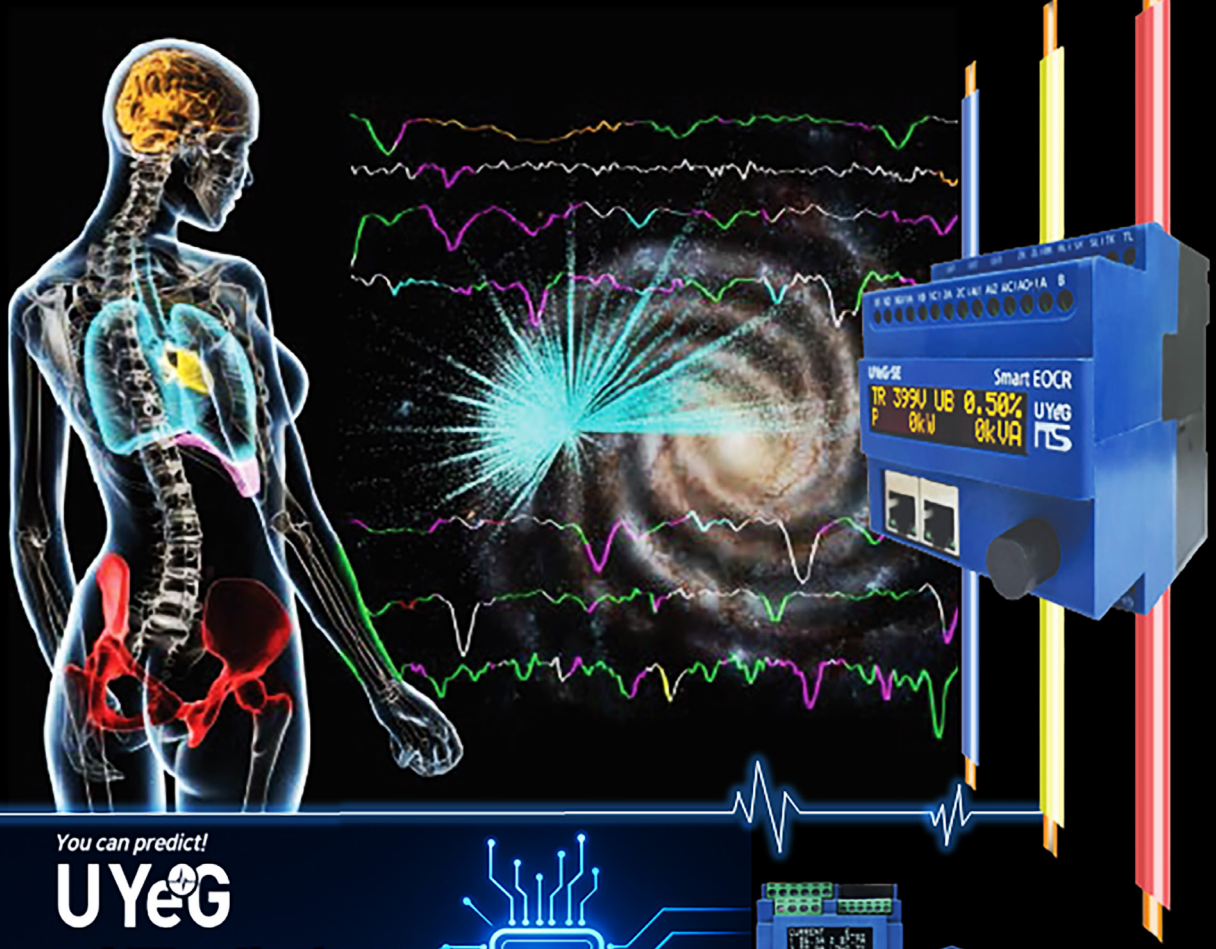
Metanet

메타넷글로벌 • 메타넷티플랫폼 • 메타넷디지털 • 메타넷사스 • 메타넷디엘 • 메타넷핀테크
락플레이스 • 유티모스트INS • 지티플러스 • 에이티앤에스그룹 • 스켈터랩스
노스스타컨설팅 • 블루칩씨앤에스 • 에미넷 • IGM세계경영연구원 • 엘릭스

인간은 혈류 血流

기계는 전류 電流

THE BEST
ESG



You can predict!

UYeG
전류 예지보전

CURRENT-BASED
PREDICTIVE MAINTENANCE

AI

사고·폭발·고장 사전 완전 차단
안전서비스 & 에너지절감

아이티공간
Intellectual Technology Space

Digital
Transformation
Partner

대한민국 기업을 위한 디지털 혁신의 시작

언택트, 디지털 뉴딜
디지털로 빠른 변화와 혁신이 요구되는 지금
당신의 기업은 어떻게 준비하고 계시나요?

KT Enterprise가
AI, Big data, Cloud의 앞선 기술과
기업 유무선 통신의 전문성으로
대한민국 기업의 디지털 혁신을 이끌어 갑니다

Digital Transformation Partner
kt Enterprise

자세한 사항은 기업전용 전화상담 (1588-0114) 또는 홈페이지 (enterprise.kt.com) 참조

AI는 언제나
사람을 향해야 하기에
기술보다 안심을 먼저 생각합니다

기술이 앞서가도
불안은 뒤에 남지 않도록
사람의 편에서 믿고 함께하는 AI

유플러스가 익시 가디언으로
모두가 안심할 수 있는
AI 시대를 시작하겠습니다

사람 중심. 안심 지능.

Assured Intelligence

LG U+

익시 가디언 ixi-Guardian
LG유플러스만의 차별적 AI 보안 기술 브랜드

Deep Fake 탐지 기술로 위·변조 음성을 판별하는 **Anti-Deep Voice**
LG AI 연구원의 초거대 AI '**EXAONE**' 기반으로 서버 통신 없이 안전하게 스마트폰 안에서 구동 되는 **On-Device AI**
국방용 보안 인증 수준의 세계적인 PUF(물리적 복제 방지) 기술을 독점 제휴한 **양자 보안(PQC)** '23년 7월~'27년 12월

SK Telecom 40th Anniversary

40th

1984 – 2001

1984. 아날로그(AMPS) 차량전화 서비스 개시
1996. CDMA 디지털 이동전화 상용화
1997. '스피드 011' 출시
1999. '스무 살의 011' TTL 출시

SPEED 011

2002 – 2010

2002. 신세기통신 합병
2002. 3G 상용화(CDMA2000 1x EV-DO)
2002. 'Be The Reds' 공동 캠페인 전개
2006. T 브랜드 출시

2011 – 2018

2011. 4G LTE 상용화
2012. 'SK하이닉스' 인수

4G LTE™ ORIGINAL

2019 – 2024

2019. 5G 상용화
2023. 나만의 AI 개인비서 'A.(에이닷)' 출시
2024. Global Telco AI Alliance 창립총회 개최

SKT 5G

초[超]시대, 생활이 되다

TBE

AI

CDMA 상용화라는 커다란 성과보다
통신 강국이라는 자부심이 더 기뻤습니다

세계 최고의 반도체 기업으로 불리는 것보다
반도체의 나라로 불리는 것이 자랑스러웠습니다

지난 40년
우리는 늘 SK텔레콤이라는 이름보다
대한민국이라는 이름으로 빛나길 바랐습니다

앞으로 40년도
글로벌 AI 컴퍼니 SK텔레콤으로서
AI 강국 대한민국의 든든한 힘이 되겠습니다

AI로 대한민국을 새롭게 하는 힘
SK telecom

안전한 미래, 융합에서 답하다

영상기반 센서융합보안기술 전문기업 (주)세오

R&D네트워크를 통한 딥 러닝 기반 영상분석기술, 레이더 센서융합기술, 통신구간 암호화 기술 등
과감한 투자 및 개발을 통해 무인교통감시장치, 실시간 암호화 영상감시시스템,
딥 러닝 기반의 종합 감시시스템 등 시민의 안전 및 재산 보호를 위한 다양한 솔루션으로
한층 진보된 사회안전망 강화에 공헌하고 있습니다.

방범 보안 시스템

도시방범, 학교, 병원, 공장, 소매점 등과 같이 CCTV를 이용한
대부분의 관제소 등에 포괄적으로 적용하여 정보유출로 인한 피해 방지

국가 보안시설

보안이 중요하게 강조되는 항만, 발전소와 같은
국가 보안이 중요하게 강조되는 곳에 적용하여 정보유출로 인한 손해 예방

가정용 보안 시스템

자녀, 반려동물 돌보기와 같이 가정에서의 CCTV 활용도가 높아짐에 따라
가정에 적용하여 사생활 노출로 인한 불안 해소와 시민의 재산보호 가능

CUBE-HIDE

해커와 제 3자로 인한 데이터 유출 시에도
영상정보를 확인 할 수 없도록 암호화 하는 장치

통신구간 영상노출, 이제그만!

- 향상된 암호화 효율성 제공
- 3차원 블록 형태의 강력한 암호화
- 기존 시스템의 변동없는 활용가능
- 네트워크 허브 및 스위치와 보안을 한번에

무인교통 감시장치

인공지능 센서융합 일체형 무인교통감시장치로 다차선,
보행자검출 등 효율적인 교통단속이 가능한 제품

다목적 다기능 통합운영 무인교통 감시장치

- 다목적/다기능 통합
- 인공지능 센서융합
- 100여개 물체 동시탐지
- 야간영상 획득 가능

AI유닛 임베디드 디바이스

영상분석 기술을 기반으로 다양한 안전취약지역 등에 적용하여
이상징후 발생 시 효율적 대처가 가능하도록 한 제품

안전관리 목적 및 AI 딥러닝 유닛 임베디드 디바이스

- AI 기반 위험상황인지
- AI 기반 산업현장 안전관리
- AI기반 교통사고 안전예방

물 관리 계장제어시스템

국가 하천 및 지방하천 등의 센서 데이터를 전송하고
신경망 알고리즘을 적용한 수위예측을 통해
자동으로 수문을 제어함으로써 홍수 피해방지 및
농업용수관리 목적의 시스템

자연 재해로부터 효율적인 스마트 물 관리 시스템

- AI 기반 수위 예측 / 3D 시뮬레이션
- 현장 센서 연동 및 원격 제어
- 실시간 정보 분석 및 모니터링

주식회사 세오
seo
www.seoitv.com

본사 경기도 안양시 동안구 흥안대로 427번길 16, 8층 (관양동, 평촌디지털엠파이어)
TEL 031-424-2345 / **FAX** 031-424-2346
광주공장/부연구소 경기도 안양시 동안구 흥안대로 427번길 16, 8층 (관양동, 평촌디지털엠파이어)
포천공장 경기도 포천시 소홀읍 소홀로 24번길 10-40

SPIKE™ Prime

LEGO education

코딩과 메이커 교육에 대한 자신감

스파이크 프라임은 다채로운 레고조립 요소들,
사용하기 쉬운 하드웨어 및 스크래치를 기반으로 하는
직관적인 드래그 앤 드롭 코딩 언어를 결합하여
쉽게 도전할 수 있는 프로젝트에서부터 파이썬의
텍스트 기반 코딩을 탐색 할 수 있는 옵션을 포함하여
무한한 창의적인 디자인 가능성에 이르기까지
학생들이 재미있게 내일의 혁신적인 마인드를
키울 수 있도록 필요한 필수 스팀과 21세기 핵심역량을
배울 수 있도록 도와줍니다!

python™

AI 카메라 키트

AI 카메라 키트에 포함된 허스키 렌즈는 머신러닝이 들어가있는
인공지능 카메라로, 물체 인식, 얼굴 인식, 라인 추적, 색상 인식,
태그 인식 등의 기능이 내장되어 있어 스파이크 프라임과
함께 사용하시면 다양한 인공지능 기능들을 체험할 수 있습니다.

제품 구성품

허스키 렌즈 PRO
(케이스 포함)

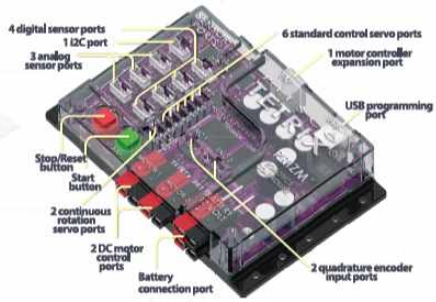
연결 케이블

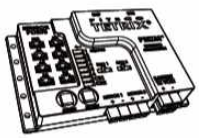
추가 브릭

핸즈온AI 조립도 및
교육자료 (PDF)

TETRIX™ PITSCO

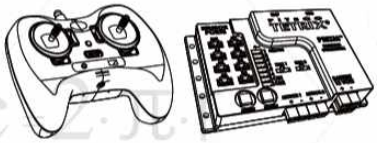
아두이노 코딩을 위한
TETRIX® MAX R/C Robotics Set





TETRIX® MAX Programmable Robotics Set

- ☐ Remote controlled
- ☑ Autonomous controlled



TETRIX® MAX Dual-Control Robotics Set

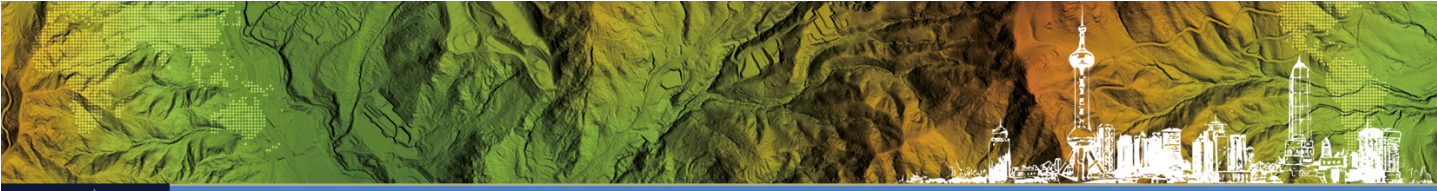
- ☑ Remote controlled
- ☑ Autonomous controlled

- PRIME 사용자는 새로운 PULSE ™ Robotics Controller 및 그래픽 TETRIX ArduBlockly 소프트웨어로 로봇을 프로그래밍 할 수 있습니다.
- PULSE는 Arduino Software (IDE) 및 PULSE 컨트롤러 Arduino 라이브러리를 사용하는 구문 기반 코딩 방법으로 프로그래밍도 가능합니다.

www.handsontech.co.kr
T : 02-2608-2633 F : 02-2608-2634

(주)핸즈온테크놀러지는 덴마크 LEGO® Education과 공식 파트너계약을 체결하고
한국 내 대학 및 초·중·고등학교에 제품을 공급하고 있습니다.
(주)핸즈온테크놀러지를 통해 구입한 제품에 한해서 공식 A/S가 가능합니다.

HandsON Technology



공간정보 사회의 글로벌 리더 ! 신한항업

www.shas.co.kr

Aerial Photography / Geodetic Survey / Digital Mapping / GIS Application / Si / Overseas Mapping

항공사진 촬영

- 당사 보유 항공기를 이용한 항공촬영
- 다목적 조사용 항공사진 촬영
- LiDAR를 이용한 3차원 공간정보 구축

측지 측량

- GPS 위성 측량
- 정밀 기준점 측량 (국가기준점)

영상도화 및 수치지도 제작

- 항공사진 수치도화
- 국가기본도 축척별 수치도화

GIS관련 프로그램 개발

- 국가기본지형도 제작 프로그램
- GIS관련 프로그램 (UIS, LIS 등)

엔지니어링 사업

- GIS를 이용한 첨단 도시계획 수립
- 관광지, 공원, 택지 등 기본계획 수립

Si (System Integration)

- 영상정보 분석, 판독 및 활용시스템 구축
- 공간정보 DB구축 및 관리·운영시스템 구축

해외사업

- 공간정보 및 관련분야 해외신규사업 발굴
- 대외 유·무상 원조사업 참여

지적

- 지적 확정 측량
- 지적 재조사 사업

주식회사 신한항업
SHINHAN Aerial Survey Co., Ltd

경기도 과천시 과천대로7길 33, 1002호(갈현동, 디테크타워A동)
Tel. 02) 6259-1000 Fax. 02) 6259-1001
Http://www.shas.co.kr Mail. shas@shas.co.kr

MQNIC

MAXIMUM QUALITY-BASED ELECTRONIC

Connected Car Service

차량 기술과 함께 연결&공유를 기반으로 미래 이동성 서비스 구축

Mobility As A Service

목적지까지 최적의 방법으로 이동할 수 있는 모든 교통수단 시스템 통합

자율주행 플랫폼

AI기반 자율주행모빌리티 통합 운영을 위한 플랫폼 구축

Digital Twin

디지털 데이터 모델로 현실 세계의 ITS를 시뮬레이션 구현

AI

수없이 생산되는 빅 데이터를 가치 있는 정보로 생성하기 위한 인공지능 분석

LBS Platform

위치기반정보와 서비스를 제공하기 위한 기본 플랫폼 구축

서울 서초구 양재천로 19 길 42 (세련빌딩) I Tel 02-521-7723
https://www.mqnic.com

MQNIC

NEOBRIX

Slogan

AUTONOMY, CONFIDENCE AND INTERACTION

Management Philosophy

행복한 구성원이 행복한 기업을 만든다. 행복한 기업이 행복한 세상을 만든다.

회사명
네오브릭스 주식회사

대표이사
현종일

임직원수 (2025년 기준)
22명(연구개발/기술인력 19명)

사업영역
소프트웨어 개발, 공급 및 유지보수
시스템 통합 자문 및 구축 관련 서비스 용역

제58회 한국정보통신학회 춘계학술대회의 성공적인 개최를 기원합니다.

㈜하이제이컨설팅

(Strategic Consensus for Customer Values)

- *Sharing successful Things*
- *Value Creating IT Consulting*
- *Base of actual Public Reference*
- *Total IT Service Cooperation*

Creative IT consulting for your Biz success

(주)하이제이컨설팅
Be creative, be sustainable

모두가 행복한 ICT 세상!

풍요롭고 행복한 세상을 만드는 ICT기업

대보정보통신

IT컨설팅에서 시스템 통합 (SI) 및 유지관리 (SM)까지 고객이 필요로
하는 최적화된 솔루션과 서비스로 고객의 성공비즈니스를 만들어 갑니다.

System Integration

최적화된 시스템을 구축합니다.

Industry Solution (전자정부, 공공, 공항, 국방, 교육 등) /
Ubiquitous Solution (U-City, ITS 등) / Network Intergration

System Management

IT시스템의 효율을 극대화합니다.

IT Outsourcing (시스템, 데이터센터, 보안서비스) / ITS 운영관리 /
장대터널 운영관리

Solutions

한발 앞선 솔루션을 제공합니다.

ITS기반 솔루션 / Smart Highway / 데이터 분석 / 검색엔진 / 보안 /
Mobile App / U-BIZ / Hi-pass

대보정보통신(주) | **본사** Add. 서울시 강남구 광평로 280 로즈데일빌딩 6층 I Tel. 02.3470.7700 I Fax. 02.3470.7719 I www.dbcs.co.kr
대보그룹 대보건설(주) 대보실업(주) 대보유통(주) 서원밸리컨트리클럽

고객의 미래가치를 실현하는 **AX전문기업**

세림TSG

Selim Technology Service Group

Realize AX for the future

AX

심층적 업무이해와 AI Agent, RAG 등
혁신기술을 접목해 AI 대전환을 선도합니다.

#범정부 초거대 AI　#국민비서 챗봇　#지능형 업무관리

Cloud

클라우드 도입을 위한 컨설팅부터 마이그레이션, 효율적 운영까지
End-to-End 서비스를 제공합니다.

#G-Cloud 운영　#Cloud Native　#MSP

Data Center

국내 최대 공공데이터센터의 경험과 노하우로 안전하고 효율적인
인프라를 설계, 구축, 운영, 유지관리 합니다.

#통합운영　#정보자원 통합구축　#SDDC 아키텍처

본사 | 대전광역시 유성구 테크노1로 62-16 (관평동 755)
대표전화 | 042.488.7900　**팩스** | 042.488.7799　**이메일** | selimtsg@selim.kr

Selim 세림티에스지(주)

전력통신 솔루션
송변전용광단말장치, WDM광모뎀

전력ICT 통합관제
지능형감시시스템, 통합관제 플랫폼

PS-LTE 시스템
전력구 비상 통신망

지능형 교통 시스템
자율주행시스템, UAM

송암시스콤주식회사

본사·공장 강원특별자치도 원주시 문막읍 동화공단로 32
연 구 소 경기도 성남시 분당구 판교로228번길 17 판교세븐벤처밸리 2단지 1동 8층

ICTWAY

아이씨티웨이

정보통신기술벤처기업

We design optimal way for ICT

IT와 관련된 특별한 서비스를 원하십니까?

저희는 ICT Infrastructure, Specialized Consulting, System Integration, Total Outsourcing에 대한 최적의 서비스를 제공합니다.

20년 이상의 다양한 프로젝트 경험과 분야별 전문기술 인력을 보유하고 있는 ICTWAY와 상의하십시오.

만족을 넘어 감동을 안겨 드리겠습니다.

Information

Communication

Technology

본 사 : 경기도 안양시 동안구 학의로 282 금강펜테리움 IT타워 A동 1405호 TEL : 070-7016-3378 FAX : 031-463-1044
대전지사 : 대전시 유성구 가정로 218 ETRI 융합기술연구생산센터 406호 TEL : 042-471-1044

차별화된 경쟁력으로
고객의 성공적인 미래를
함께 만드는 MTDATA

AI·BIGDATA
AI와 빅데이터로 비즈니스 인사이트를 실현하는 지능형 플랫폼 구축

CLOUD
클라우드 네이티브 전환/구축/운영 전 사이클의 전문 서비스 제공

SI
최신 기술 기반의 맞춤형 SI로 디지털 혁신 서비스 제공

INFRA·ITO
초대형 데이터 센터 구축 및 안정성과 효율 동시에 제공하는 운영 서비스

본사 경기도 성남시 분당구 판교로255번길9-22,514호(삼평동,우림W-CITY)
대전지사 대전광역시 서구 둔산로137번길 21, 3층 3034호(대승빌딩)
대구지사 대구광역시 동구 화랑로 9, 6층 (신천동,천우빌딩)
울산지사 울산광역시 북구 명촌7길 15, 2층(명촌동)
부산지사 부산광역시 금정구 식물원로9번길 29(장전동)

www.mtdata.co.kr

공간정보서비스 Alliance

URL http://gisalliance.co.kr

CONNECTING, SHARING & SHOWING

공간정보로 연결하고, 공유하고, 볼 수 있는 세상

네오스펙트라

올포랜드

엘티메트릭

명화지리정보

공간정보서비스 Alliance

메타버스 플랫폼 서비스

솔루션 개발, SI

데이터 가치 창출

DB구축

하늘 - 땅 - 바다 - 지하 - 메타버스 공간

주소 서울시 금천구 가산디지털 1로 145 에이스하이엔드타워 3차 1401호

T 02-855-5724 F 02-855-5746

한국정보기술 주식회사

KOREA INFORMATION & TECHNOLOGY

CHANGE THE WORLD

저희 한국정보기술 주식회사는 **미래를 만들어가는 IT전문기업이 되겠습니다.**

정보시스템

통합플랫폼을 기반으로 고객 맞춤형 정보시스템을 개발·제공합니다.

교통시스템

도로, 항공, 해상 등에 이르는 다양한 분야의 교통시스템을 제공합니다.

스마트시티

IT 기술을 기반으로 도시 전반의 영역 융합 및 안전하고 지능적인 스마트도시를 만들어 갑니다.

영상시스템

CCTV영상 모니터링 및 관련기관과 연계하여 시민의 안전확보를 최우선으로 합니다.

재난시스템

신속하고 정확한 정보수집과 상황판단으로 재난현장 컨트롤타워 기능을 수행합니다.

환경에너지

급격히 팽창할 미래산업으로 신재생에너지 사용을 활성화하여 에너지 공급 및 탄소배출저감 정책에 앞장 섭니다.

IT아웃소싱

각종 교통시스템, IT인프라 등 다양한 산업 분야의 정보시스템 설비를 운영·유지관리를 합니다.

KI&T

경기도 안양시 동안구 벌말로 126 (평촌오비즈타워) 25층 한국정보기술 주식회사

T. 031-696-0945 **F.** 031-8018-8977 **W.** www.koreaint.com

문의사항 : 기획관리팀 상무 곽정희 (연락처 : 010-8554-5994)


The Magazine of KIICE
지능정보통신
www.KIICE.org
KIICE
한국정보통신학회