2026.09.05

사회문화조사실 I NARS 정책연구용역보고서

AI로 인한 새로운 보안 위험과 대응에 관한 국내외 입법 · 정책 비교 연구

이해원 I 강원대학교

국회입법조사처
NATIONAL ASSEMBLY RESEARCH SERVICE

NARS 정책연구용역보고서

AI로 인한 새로운 보안 위험과 대응에 관한 국내외 입법·정책 비교 연구

이해원(강원대학교)

2026. 09. 05.

국회입법조사처
NATIONAL ASSEMBLY RESEARCH SERVICE

AI로 인한 새로운 보안 위험과 대응에 관한 국내외 입법·정책 비교 연구

2026. 9. 5.

연구책임자 : 이해원 (강원대학교)
참여연구원 : 이석준 (가천대학교)

강원대학교 산학협력단

제 출 문

국회 입법조사처장 귀하

본 보고서를 「AI로 인한 새로운 보안 위험과 대응에 관한 국내외 입법·정책 비교 연구」 연구용역의 최종보고서로 제출합니다.

2026년 9월

연구책임자 이 해 원

요 약

Ⅰ. 연구 목적

□ AI 기반 사이버 공격이 진화하면서 ‘AI에 의한 위협(threat by AI)’과 ‘AI에 대한 위협(threat to AI)’과 같은 새로운 보안 위험 출현

ㅇ AI ‘무기화(weaponization)’를 통한 사이버 공격이 일상화 · 지능화되는 추세

ㅇ 한편, AI가 경제사회 전 영역으로 확대되면서 AI를 타겟(target)으로 하는 새로운 사이버 공격도 급증

□ 특히 ’26. 4. ‘미토스(Mythos) 사태’ 이후 최첨단 AI(Frontier AI)에 기초한 사이버보안 위험이 급부상 중

ㅇ 미토스는 보안 전용 특수 AI가 아님에도 ‘스스로’ 보안 취약점 발견 → 조건 분석 → 공격 방법 구성(exploit chain)을 자동화하는 등 기존 AI를 뛰어넘는 강력한 성능을 발휘하여 전세계에 충격을 안김

□ AI發 사이버보안 위협으로 인하여 개인정보 침해 위험 또한 증대

□ 세계 주요국(미, EU, 영, 일 등)은 ❶AI 보안 위험 ❷복원력(resilience) ❸공급망 보안 ❹사이버보안 정보 공유 등 다양한 측면에서 AI 위협 대응 법제 및 정책 추진 중

□ 이러한 배경 하에, 본 연구는 ❶AI 시대의 새로운 보안 위험을 유형화하고, ❷국내외 관련 대응 법제(개인정보 배상 제도 포함)를 비교 · 분석하여, ❸관련 입법 시사점 등을 제언하고자 함

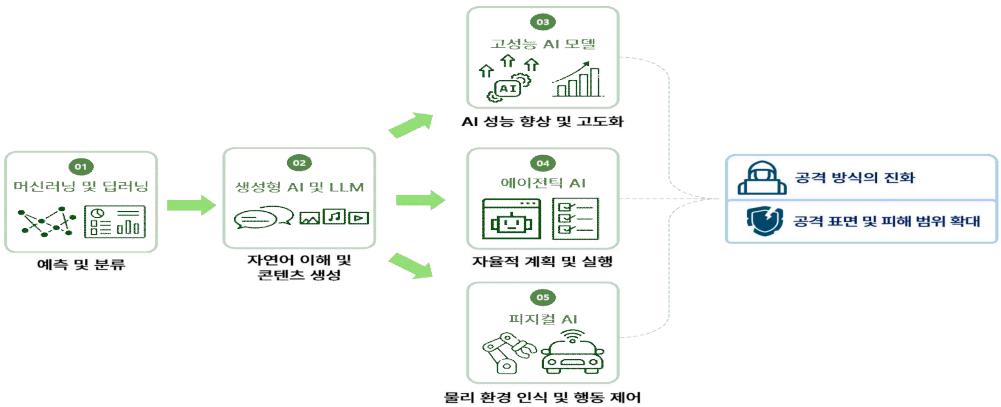
Ⅱ. AI 보안 위험 유형화

□ 통상 AI는 기술 발전 단계에 따라 (1) 머신러닝/딥러닝, (2) 생성형AI (LLM 등), (3) 에이전틱(Agentic) AI, (4) 피지컬(Physical) AI로 분류

□ AI 유형에 따라 핵심 기능과 동작 방식이 상이하므로, 보안 위협 발생 지점과 영향 범위도 다르게 나타남

○ (1) 머신러닝/딥러닝은 학습 데이터와 모델 자체가, (2) 생성형 AI · LLM은 입력(프롬프트)과 생성 결과가, (3) 에이전틱 AI는 외부 도구 연동과 권한 위임 과정이, (4) 피지컬 AI는 센서 입력과 제어 명령이 각각 핵심 위협 지점이 됨

[그림 i] AI 유형 및 사이버 위협 양상



□ AI 발전은 공격자에게 새로운 공격 수단을 제공하는 동시에, AI 자체를 새로운 공격 대상으로 만들고 있음

○ (AI에 의한 위협) 공격 준비 과정 전반이 자동화되고, 공격자에게 요구되는 전문성이 감소하면서 공격의 속도와 규모 확대

○ (AI에 대한 위협) 학습 데이터, 학습 모델, 에이전트 권한, 물리력 행사 모듈 등 기존 시스템에는 존재하지 않았던 새로운 공격 타겟 존재

□ AI에 의한 위협

○ AI 이전 사이버보안 위협이 수동성, 전문성에 기반하였다면, AI 이후의 사이버보안 위협은 자동성, 대량성, 신속성, 대중성에 기인함

○ 특히 ’26. 4. 미토스 출현 이후에는 공격 의도를 가진 자연인 없이 AI의 자율적 판단만으로 실제 공격이 가능한 상황

[표 i] AI에 의한 위협: AI 발전 단계별 위협 특징

구분	AI의 역할	위협의 특징	AI의 의의
전통적 IT 환경	• 공격자가 직접 공격 필요 단계 수행	• 높은 전문성과 많은 시간·비용이 요구됨	• 공격 수행 능력이 공격자 전문성에 크게 의존
머신러닝 및 딥러닝	• 대규모 데이터 분석을 AI가 보조	• 로그, 코드 등을 분석하여 취약점 후보나 이상 징후를 빠르게 선별 • 공개 정보와 행동 데이터를 분석하여 공격 대상의 특징 도출	• 공격 대상 선별과 사전 분석의 효율성 증가
생성형 AI 및 LLM	• 공격 절차 및 콘텐츠 생성까지 AI가 보조	• 시나리오 도출, 공격 도구 개발 등 공격 보조 • 딥페이크·음성 위조 콘텐츠 생성 가능	• 공격 절차 자동화 및 준비 시간의 단축 • 사회공학적 공격 정교화
미토스 등 고성능 AI	• 취약점 탐색부터 공격 절차 구성 등 거의 전 과정을 AI가 연속 수행	• 공격 수행의 주체가 인간에서 AI로 이동 • 공격 의도를 가진 행위자가 없더라도 목표 달성 과정에서 실제 침해 발생 가능	• 취약점 탐색 및 공격 자동화·지능화 저비용화 • 방어 부담 증가

□ AI에 대한 위협

○ AI가 실제 업무나 서비스에 적용되면서 AI 자체가 사이버 공격의 새로운 대상으로 부상 중

○ AI에 대한 위협은 고도의 기술적 영역으로, 생성형AI / 에이전틱 AI / 피지컬 AI 별로 국내외 불문 다양한 분류 기준이 존재

[표 ii] AI에 대한 위협: AI 유형별 공격 특징

구분	AI의 역할	공격의 특징	AI로 인한 피해
생성형 AI 및 LLM	• 외부 자료·학습 데이터 기반 답변 및 콘텐츠 생성	• 입력·학습 데이터 조작으로 모델 출력 왜곡	• 정책 우회, 민감정보 유출 등 정보·콘텐츠 피해 발생
에이전틱 AI	• 목표 해석 및 외부 도구 호출을 통한 업무 행위 수행	• 목표·권한 조작으로 AI 실행 행위 왜곡	• 업무 행위 조작, 권한 오남용 등 디지털·업무 시스템 피해 발생
피지컬 AI	• 센서 기반 현실 인식 및 제어 명령 기반 물리 행동 수행	• 센서·제어 명령 조작으로 물리 동작 왜곡	• 오인식·오동작으로 인한 물리·현실 세계 피해 발생

□ 기존 사이버보안 모델의 한계

○ 기존 모델은 (1) 경계 기반 보안 (2) 시그니처·규칙 기반 탐지 (3) 정적 신뢰 모델 (4) 사후 대응 중심 (5) 행위주체 식별 및 그에 따른 책임 귀속을 기본으로 하나, AI가 발전할 수록 이러한 기본 모델로는 사이버보안 위협 대응에 한계

- 다수의 외부 에이전트·시스템에 권한이 위임되고 연동 관계도 실행 시점마다 새로 형성되므로, 사전 부여된 권한과 고정된 경계를 기준으로는 실제 권한 범위와 신뢰 범위 확정 불가

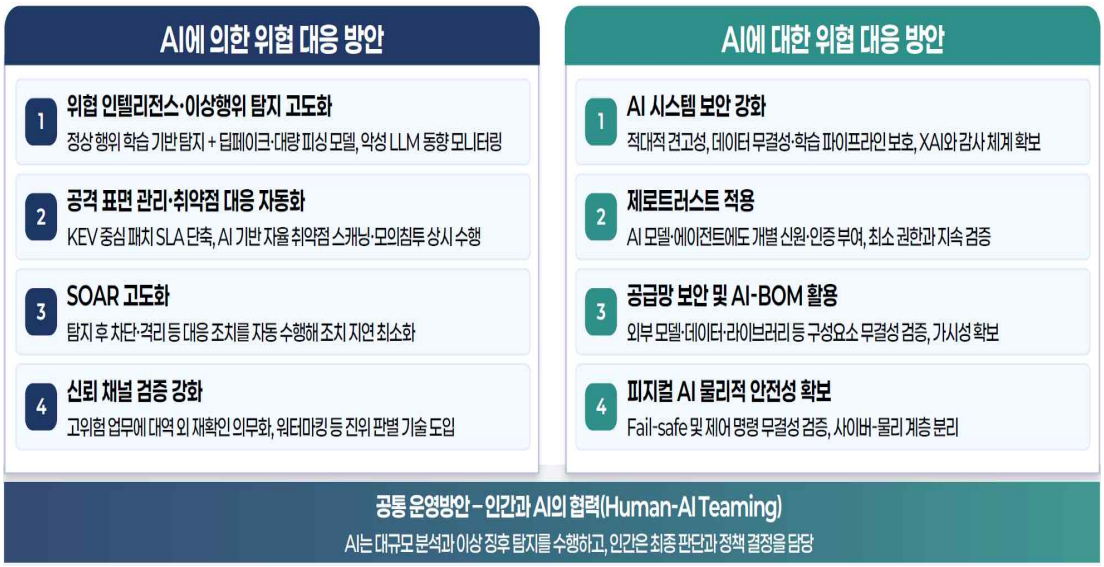
- 사람이 아닌 에이전트가 자율적으로 판단·실행하게 되면서, 사람을 행위 주체로 상정해 온 기존 가정 자체가 붕괴

[표 iii] 기존 보안 모델의 한계

구분	기존 보안 방식	AI 환경에서의 한계
경계 기반 보안	• 내부망 신뢰 중심의 고정 경계 보안	• 외부 연동 확대로 내부 · 외부 경계 약화
시그니처 · 규칙 기반 탐지	• 사전 정의된 패턴 · 규칙 기반 탐지	• 자연어 · 입력 변형으로 고정 규칙 우회 가능
정적 신뢰 모델	• 최초 인증 이후 지속 신뢰 부여	• 계정 탈취 · 권한 오남용 시 내부 조작 탐지 한계
사후 대응 방식	• 로그 분석 · 복구 중심 대응	• 원인 분석 및 데이터 · 모델 오염 복구 한계
행위주체 식별 및 책임 귀속	• 사용자 계정 기준 행위 식별	• 행위 주체가 AI 에이전트로 이동하면서 사용자 계정 기준 식별 · 책임 귀속이 어려움

□ 기존 보안 모델의 한계를 극복하고 AI發 사이버보안 위협에 대응하기 위한 기술적 대응 방안은 다음과 같음

[그림 ii] AI 사이버보안 위협에 대한 기술적 대응 방안



Ⅲ. 주요국 관련 입법 · 정책

□ AI로 인한 새로운 보안 위협과 관련된 해외 주요국(미국, EU, 영국, 일본)의 입법 및 정책을 ’26. 8. 기준으로 살펴봄

* 상세 내용은 보고서 참조; 요약문에서는 특징과 시사점 위주로 정리함

<미국>

□ 미국은 연방제 국가라는 헌법적 구조상 주(state)와 연방(federal)의 권한이 명확히 구분되어 있음

□ 미국은 전통적으로 공공/민간을 분리하여 공공 부문(특히 연방정부)에는 공법 중심의 엄격한 규율을, 민간 부분은 사법 중심의 자율적 규율을 존중해 왔음

□ 최근 미국은 APT 조직의 기반시설 인프라 위협과 공급망 교란에 대응하기 위해 ‘사후적 · 파편적 방어’에서 ‘선제적 · 통합적 복원력(Resilience) 확보’로 사이버보안 패러다임을 전환하는 중

ㅇ 이에 따라 민간 영역에도 공법적 컴플라이언스 체계(중요인프라에 발생한 사이버 사고 보고 보고 체계 등)를 도입하고, 연방 차원의 성문법 부재의 한계를 대통령 행정명령과 각종 지침 등을 유기적으로 결합한 연성 규범으로 보완하고 있음

□ 특히 트럼프 정부의 ’25. 7. AI 실행계획(AI Action Plan), ’26. 3. 사이버 전략(President Trump’s Cyber Security Strategy for America), ’26. 6. ‘첨단 AI 혁신 및 안보 촉진 명령’ 등에 따라 사이버보안 법제가 개편될 여지가 있으므로, 지속적인 주목 필요

□ 주요 시사점은 다음과 같음

○ 민간 AI 혁신 보장과 국가 안보 중심 방어의 정교한 조화

- 첨단 AI 모델(Covered Frontier Model)에 대한 사전 인허가나 강제적인가 배제를 행정명령 제14409호로 명시하여 민간의 AI 기술 혁신 속도를 보장함

- 국방 및 국가안보시스템 전산망 보안을 최우선 과제로 지정하고 자발적 사전 접근(Early Access, 최대 30일)을 통해 개발 단계부터 취약점을 점검 · 보완함

○ 면책 장치(Safe Harbor)를 통한 민간 위협 정보의 공유 유도

- CISA 2015 및 CIRCIA 2022를 통해 민간이 사이버 위협 지표나 사고 보고서를 제출할 때 민사소송 면책, 정보공개법(FOIA) 적용 제외, 규제 목적 유용 금지 등 명확한 안전조항 제공

- 민간의 사고 은폐 위험을 줄이고 민간과 정부 간 자동지표공유시스템(AIS)을 통한 실시간 정보 전파 체계를 활성화함

○ 제로트러스트 도입 및 자발적 표준(de facto standard)과 연방 조달 연계

- 행정명령 제14028호를 통해 연방 전산망에 제로트러스트(Zero Trust) 아키텍처 및 SBOM 제출을 전면 도입

- 구속력 없는 가이드라인인 NIST AI RMF를 연방 조달 지침 및 시장의 사실상 표준으로 안착시켜 산업 전반의 보안 수준을 자율적으로 상향시킴

<EU>

□ EU는 ’19-’24 차기 유럽집행위원회 정치적 가이드라인에서 향후 EU가 추진해야 할 6대 핵심 목표 중 하나로 “디지털 시대에 적합한 유럽(A Stronger Europe in the World)” 선정

ㅇ EU는 디지털 미래 구축을 위한 필수불가결 요소로 ‘신뢰’와 ‘보안’을 강조하면서 사이버보안을 핵심 정책 분야로 제시

□ EU는 나날이 고도화되는 사이버보안 위협에 대응하기 위하여 ’20년 이후 EU 전역의 사이버보안 수준을 전반적으로 향상시키기 위한 목적으로 법제도 및 정책을 지속적으로 정비하여 왔음

□ 특히 전 세계 최초의 AI 규제법으로 알려진 EU AI법에는 AI에 특화된 사이버보안 관련 규정이 존재

□ 최근 EU는 AI의 급속한 발전과 그에 따른 경제사회 변화에 신속히 대응하고자 디지털 옴니버스 패키지*(’25. 11.)와 사이버보안 패키지(’26. 1.)를 발표하는 등 적극적으로 법제도 정비를 추진 중

□ 주요 시사점은 다음과 같음

ㅇ AI 및 디지털 제품 전 수명주기(Lifecycle)에 걸친 ‘보안 설계(Secure by Design)’ 규범화

- 사이버복원력법(CRA)을 통해 하드웨어 · 소프트웨어 디지털 제품에 ‘Secure by Design and Default’ 원칙을 법적 의무화

- 인공지능법 제15조 및 제55조를 통해 고위험 AI 및 시스템적 위험이 있는 범용 AI 모델(RGPAM)의 전체 수명주기에 걸쳐 데이터 오염, 적대적 공격, 모델 회피 방지 의무를 강제

ㅇ 기업의 규제 부담 완화

- 디지털 옴니버스 패키지로 사이버보안 사고(NIS2), 개인정보 유출(GDPR), AI 시스템 사고 등으로 파편화된 보고 절차를 통합
- ENISA가 구축한 단일 접수 창구(SEP)에 표준 양식으로 1회 신고 시 관할 당국으로 자동 통보되는 체계 구축

ㅇ 글로벌 ICT 공급망 통제와 주권적 AI · 테스트 인프라 구축 병행

- 사이버보안 패키지를 통해 우려 국가 및 고위험 공급업체를 식별 · 배제하는 ICT 공급망 위험평가 프레임워크를 수립
- EU 행동계획을 통해 AI 안전 테스트 플랫폼, 주권적 AI 인프라 자본 투자를 추진하여 기술 종속성 및 안보 위협에 동시 대응

<영국>

□ 영국은 과거에는 EU의 사이버보안 법규를 그대로 따랐으나, ’21. 1. BREXIT 후 영국 고유의 독자적 규범 체계를 마련해야 하는 상황

□ 영국의 사이버보안 정책은 현재 ① BREXIT 전 EU에서 시행된 舊 NIS 지침과 사실상 동일한 내용의 ‘2018 NIS 규정’과 ② ’21 제정 ‘제품보안 및 통신인프라법(PSTIA)’에 근거하여 추진되고 있으나,

□ 최근 생성형 AI와 에이전트 기술의 급속한 발전으로 야기되는 새로운 사이버보안 위협에 선제적으로 대응하기 위하여 2018 NIS 규정을 전면 개정하는 ‘사이버보안 및 회복력 법안(CSRB)’을 추진 중

□ 이 외에도 사이버보안 인증제도(Cyber Essentials), 조기경보 체계(NCSC Early Warning)가 사이버보안 학계 및 실무에게 주목받고 있음

□ 주요 시사점은 다음과 같음

ㅇ 보안 사각지대였던 서드파티 디지털 공급망(MSP 등) 직접 규제

- 사이버보안 및 회복력 법안(CSRB)을 추진하여 데이터 센터, 전력 그리드 부하 제어자, 관리형 서비스 제공자(MSP) 및 핵심 공급업체를 규제 테두리 내로 편입함
- 핵심 인프라 교두보로 악용되는 서드파티 공급업체의 보안 취약점을 정밀 타겟팅

ㅇ 실질적 지원 인센티브 및 비침습적 지원 서비스 체계 운용

- Cyber Essentials 인증을 취득한 중소기업에 연계 무료 사이버 배상책임보험을 자동 가입시키는 인센티브 제도 운용
- NCSC Early Warning을 통해 기업이 등록한 자산과 수집된 위협 정보를 비침습적으로 매칭해 단방향 무료 통보함으로써 민간의 자체 분석 부담을 축소

ㅇ 비용 회수 제도(Cost Recovery)를 통한 규제 기관의 재원 자립화

- CSRB를 통해 규제 당국이 피규제 기업들로부터 행정 예상 총비용을 분담금 형태로 배분받아 조달하는 구조를 수립

<일본>

□ ’25 이전까지 일본의 사이버보안 법체계는 ’14 제정된 사이버시큐리티기본법(サイバーセキュリティ基本法, 추진체계)과 ’22 제정된 경제안전보장추진법(経済安全保障推進法, 에너지, 통신, 운송, 금융 등 주요인프라 보호)이 양 축을 이루고 있었음

□ 일본은 국가안전보장전략('22)에서 (1) 민관협력 강화 (2) 통신정보를 활용한 사이버위협 탐지 (3) 공격자의 서버침입 무해화, (4) 사이버안전보장 분야의 컨트롤타워 조직 신설의 4대 목표 설정

□ 이에 '중요 전자계산기 부정행위로 인한 피해방지법('피해방지법')과 '동법 시행에 따른 관계법령 정비법('정비법')'이 제정됨

□ 또한 미토스 출현 이후 대두된 사이버보안 위험에 대응하고 에이전틱 AI로의 패러다임 전환에 대처하고자 '25. 12. 수립한 AI 기본계획을 불과 7개월 후인 '26. 7. 수정하여 2차 AI 기본계획을 수립, 시행 중

□ 주요 시사점은 다음과 같음

ㅇ '능동적 사이버 방어(Active Defense)' 법적 근거 및 실행 기반 확립

- '피해방지법 및 정비법' 제정을 통하여 경찰과 자위대에 사이버 공격용 프로그램 중지 · 제거 및 서버 접근 차단을 수행하는 '접근 확인 및 무해화 조치' 권한을 부여

- 해외 서버 침입 시 긴급상황 등 국제법상 법리를 원용하는 공세적 · 선제적 방어 체계로 전환

ㅇ 통신정보 수집 · 분석 권한 부여 및 통제 기구 마련

- 사이버 공격 실태 파악을 위해 통신정보(내-외, 외-외 등)를 수집 및 자동화된 방법으로 기계적 선별 · 분석할 수 있는 권한을 내각총리대신에게 부여

- 국회 동의로 임명되는 독립 기구인 '사이버통신정보감리위원회'의 사전 승인 및 지속적 검사 · 감시를 받도록 정교하게 설계

ㅇ 미토스와 같은 사이버공격 AI 출현에 따른 대응 비상 안보 패키지 및 민첩한 거버넌스(Agile Governance) 구축

- 추론형 · 에이전트형 AI 기반 자율 공격 가시화에 대응해 범정부 안보 패키지인 ‘Project YATA-Shield’를 출범시키고, AI 기본계획을 7개월 만에 수정하는 유연한 거버넌스를 보여줌

<공통적 시사점>

□ 주요국 사례에서 도출할 수 있는 공통적 시사점은 다음과 같음

○ 패러다임 전환: 사후 경계 방어 → 선제적 방어 및 복원력(Resilience) 제고

- 주요국은 기존 시그니처/경계 중심 방어의 한계를 인정하고, 선제적 방어와 신속한 복원력 확보를 최우선 과제로 재편

○ 규제 대상의 정밀 조정: 제품 및 서비스 전 수명주기(Lifecycle) 보안 규제, 서드파티 공급망 보안 법정화 등 AI 보안 위협 대응에 필요한 규제는 확대하되, 불필요한 기존 규제는 간소화

- 단일 시스템 규제에서 벗어나 AI 모델, 학습 데이터, IoT 제품, 데이터 센터, 관리형 서비스(MSP) 등 공급망 전체/라이프사이클 전체 규율 체계로 진화 중

- 전통적 위협에는 적합하나 AI發 위험에는 신속, 정확한 대응이 어렵고 오히려 AI 발전을 저해하는 규제는 간소화 혹은 제거

○ 민간 참여 촉진: 면책 조항(Safe Harbor) 명문화, 인센티브 설계

- AI 위협에 대응하기 위한 민간의 위협 정보 공유 촉진은 규제가 아닌 ‘법적 안전망 제공’(미국 등)과 인센티브 제공(영국 등)을 통해 달성

○ 거버넌스 고도화

- AI발 새로운 사이버보안 위협에 효율적이고 신속하게 대응할 수 있도록 범정부적 사이버보안 대응 거버넌스를 정비 중

Ⅳ. 우리의 관련 입법·정책

□ AI로 인한 새로운 보안 위협과 관련된 우리의 입법 및 정책을 ’26. 8. 기준으로 살펴봄

* 상세 내용은 보고서 참조; 요약문에서는 특징과 시사점 위주로 정리함

<법제>

□ 우리나라의 사이버보안 법제는 한마디로 “분야별, 영역별 수평적 규율 체계”라 할 수 있음

ㅇ 미국, EU, 영국, 일본과 같이 범국가적 차원의 사이버보안 총괄 조직이나 거버넌스 체계를 규율한 법률은 부재

□ 공공 영역은 보안이 아닌 ‘안보’의 일환으로 「국가정보원법」 및 「사이버안보 업무규정」(대통령령)이 규율

ㅇ 사이버안보 업무규정상 공공 영역의 사이버보안은 ‘안보’ 차원에서 국가정보원이 총괄하나, 각 부처는 개별 법령에 근거하여 분야별로 사이버보안 활동을 수행 중

□ 민간 영역은 「정보통신망 이용촉진 및 정보보호 등에 관한 법률(이하 ‘정보통신망법’)」에 따라 과학기술정보통신부(이하 ‘과기정통부’)가 사이버보안 사무를 수행하나, 금융 영역은 「전자금융거래법」에 따라 금융위원회 소관임

□ 미국, EU, 일본의 Critical Entity 보호법제에 비견되는 「정보통신기반시설보호법」 역시 사이버보안 주요 법률로 작동함

□ ’26. 1. 22. 부터 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 ‘인공지능기본법’)이 시행되고 있어 역시 관련됨

<거버넌스>

□ 우리나라의 사이버보안 법제는 한마디로 “분야별, 영역별 수평적 규율 체계”라 할 수 있음

ㅇ 국가안보실(대통령실)이 최상위 컨트롤타워이고, 국가정보원이 국가 차원의 사이버보안센터(NCSC)를 운영하며, 각 부처가 소관 사무별로 사이버보안 활동을 수행하는 체계

<정책>

□ (1) 글로벌 AI 경쟁력 강화 차원에서 ’26. 3. 인공지능 기본계획이 수립, 시행 중이며, (2) ’25 발생한 이동통신사 개인정보 유출, 금융회사 해킹 등 사고의 후속조치 차원에서 ’25. 10. 범정부 정보보호 종합대책이 수립, 시행 중

<해외 사례와의 비교 및 시사점>

□ Ⅲ.에서 살펴본 주요국 사례와 비교할 때, (1) 사이버보안 위협에 대응하는 거버넌스 체계 존재 (2) 사이버보안 위협 대응에 필요한 기본적인 법체계(Critical Entity, 사이버위협 정보 공유, 사이버보안 사고 발생시 조사 및 대응, 인증 등) 존재 (3) AI로 인한 기회와 위험을 동시에 인식하고 이에 대응코자 하는 국가 차원의 전략이 존재한다는 점에서는 공통됨

□ 그러나 각론에서는 다음과 같은 차이점이 존재함

ㅇ 미, EU, 영, 일 등과 달리 공공/민간/금융 등 영역별/기관별/소관사무별 다소 분절된 거버넌스 체계를 보여줌(이에 따라 법령 또한 파편화되고, 정책 추진에서도 다소 혼란스러움)

ㅇ 소프트웨어/AI 명세서, 제로트러스트, 전 생명주기 보안 설계 등 AI發 사이버보안 위험 대응에 적합한 기술적 사항의 제도화 미흡

ㅇ 복원력 강화 등 AI 보안 위협 대응에 필요한 규제는 다소 미흡한 반면, 전통적 인증 체계나 사이버보안 위협 사고 신고 및 대응과 같은 기존 규제 개선 노력은 다소 미흡

ㅇ 민간의 자발적 참여를 이끌어낼 유인(incentive) 부족

□ 요컨대, 우리 법제도 및 정책은 AI 출현 이전의 사이버보안 위협 대응에는 적절할 수 있으나, AI 출현 이후(특히 ’26. 4. 미토스 사태 이후) 패러다임 전환이 이루어진 사이버보안 위협에 대응하기에는 다소 미흡하며, 개선이 필요하다고 생각됨

□ 구체적 개선 필요 사항은 Ⅴ(결론 및 제언).에서도 언급하며, 몇 가지 점만 언급하면 다음과 같음

① 범국가적 총괄 거버넌스의 부재

ㅇ 미국, EU, 영국, 일본 등 주요 선진국이 단일 총괄 기구 및 통합 법제를 정비한 것과 달리, 한국은 공공(국정원/사이버안보 업무규정), 민간(과기정통부/정보통신망법), 금융(금융위/전자금융거래법)으로 소관 법령과 대응 주체가 분절되어 있음

ㅇ 대통령실 소속 국가안보실이 컨트롤타워라고 볼 여지도 있으나, 대통령실은 본질적으로 대통령 보좌기관으로서 조정 기능은 별론 집행 기능이 부재하고, 실제로도 사이버보안에 특화된 전문 집행조직/인력이 없어 정책 집행에 있어 국가정보원 등에 상당히 의존할 수밖에 없음

② AI 기본법 내 '사이버보안' 직접 규율 및 강제성의 공백

ㅇ ’26. 1. 시행된 인공지능기본법은 고영향 AI 사업자의 책무(제34조) 및 임의 영향평가(제35조)를 규정하고 있으나, EU AI법과 달리 ‘사이버보안’이나 ‘적대적 공격 복원력’에 관한 명문 규정이 누락

③ 인증 및 침해대응 체계 복잡, 핵심 시스템의 근거 법률 미비

ㅇ ISMS(정보통신망법), CSAP(클라우드컴퓨팅법), IoT 보안인증, 보안적합성 검증(전자정부법) 등 보안 인증 및 검증 체계가 여러 법령에 산재되어 있어 기업에 부담 초래

ㅇ 국정원(공공), KISA(정보통신망), 금보원(금융)이 각 소관 영역별로 사이버보안 위협 정보공유 및 사고 조사 등을 수행 중이지만 분야와 영역을 초월하는 사이버 침해 사고의 특성상 범국가적 차원의 단일한 정보 공유와 사고 대응에는 한계 노정

Ⅴ. 결론 및 제언

□ 지금까지의 논의를 종합하여 총론 차원에서 법제 개선의 6대 기본 방향을, 각론으로 15대 개별 과제를 각 제안

□ 기본 방향

① 한계의 인식: ‘완벽한 사전 대응’ 불가 + 복원력 강화

② 정부 주도 → 민관 협력으로 거버넌스 패러다임 전환

③ 신기술/신산업 발전과 사회 안전 조화

④ 경직성, 우둔성(愚鈍性) → 탄력성, 신속성, 유연성 확보

⑤ 기존 규제의 전면 재정비

⑥ 해외 사례의 전략적 취사 선택

□ 구체적 액션 플랜 마련

○ 15대 과제를 (1) 단기(현행 법제에서 실행 가능; ~6개월), (2) 중기(시행령/시행규칙/행정규칙 수준 개정 필요;~1년), (3) 장기(법률 수준 개정 필요;~3년) 과제로 각 분류하여 구체적 실행 계획 제시

[그림 iii] 개선 방안 개요

□ 제도 개선 방안

① 범국가적 사이버보안 추진체계 개편[장기]

○ 현재의 영역별 분산형 추진체계를 전면 개편하기에는 조직, 인력 관점에서 쉽지 않은 것이 사실이나, 장기적으로는 신속성과 효율성이 담보되는 추진체계에 관한 고민 필요

② 범국가적 위협 정보 공유 강화[장기]

○ 공공-정보통신망-금융이라는 영역별 분절된 사이버 위협 정보 공유 체계로는 AI發 사이버보안 위협에 신속한 대응이 어려움

○ 특별한 사정이 있는 경우에는(예컨대 사이버 위협의 심각성, 중대성, 긴급성 등) 국정원/KISA/금보원 구분 없이 사이버 위협 정보가 공유되도록 하되, 정보를 제공한 민간에게 안전 장치(safe-harbor)를 두어 범국가적 차원의 위협 정보 공유 체계 강화 필요

③ 표준, 가이드라인 등 연성 규범 거버넌스 정비[단기 장기]

ㅇ KISA/금보원/국보연 등에 분산되어 있는 연성 규범 체계를 국민 입장에서 통합 혹은 간소화[단기] / 장기적으로는 독립성·전문성을 가진 전문 독립 조직 검토(참고: 미 NIST, EU ENISA)[장기]

④ AI에 특화된 사이버보안 의무 법제화[중기 장기]

ㅇ 인공지능법상 고영향AI 사업자 또는 학습 누적 연산량이 일정 기준 이상인 AI 사업자에게 사이버보안 대응력 및 복원력 의무 부여

* 중기적으로 고시, 장기적으로 법령에 명문화

⑤ 기존 인증 체계 정비[장기]

ㅇ 기존의 복합적, 다중적 인증 체계를 필요한 부분은 강화하되 중복규제 이슈가 있고 실효성이 떨어지는 부분은 통합, 간소화하여 정비

⑥ 디지털 장비 인증 확대[중기]

ㅇ EU(사이버복원력법), 영국(PSTIA)와 유사하게 디지털 장비 전반을 아우르는 기본 인증 체계를 마련하여 공백 해소

⑦ 수요자 중심의 위협정보 공유 서비스 제공[단기]

ㅇ 정보자산 중심 매칭형 위협정보 공유 서비스를 제공하여 사이버 공격 대응 능력을 전반적으로 제고(영 NCSC early warning 사례 참조)

⑧ 취약점 정보 상시 신고체계(CVD/VDP) 법제화[장기]

ㅇ CVD/VDP 제도와 화이트해커 면책을 실정법에 명확히 규정

⑨ 사이버보안 투자 촉진[중기 장기]

ㅇ 재정 여건상 예산 직접 투입은 한계가 있을 뿐 아니라 정부가 공공이 아닌 민간의 사이버보안 투자를 '직접' 강제하기도 어려우므로, 각 기관의 자발적 보안 투자를 유인할 수 있는 '간접' 제도 설계 필요

- (민간) 상장법인은 사이버보안 사항을 공시(정기/수시)사항에 포함[중기] / (공공) 사이버보안 사항을 경영공시 사항에 포함시키고 업무평가 배점을 강화[장기]

⑩ 제로트러스트 및 SBOM · AI-BOM 제도화[장기]

ㅇ SW 공급망 전반에 대한 사이버보안 위협 대응책으로 미, EU 등의 사례를 참고하여 SBOM 및 AI-BOM 법제화

⑪ AI 기반 방어 및 신속 패치 적용 체계 마련[장기]

ㅇ (1) AI 기반 사이버보안 대응의 근거를 마련하고, (2) 일정 요건 충족시 검증보다 신속한 패치를 우선시하되, 서비스 중단/장애 발생시 담당자 및 해당 기관의 법적 책임을 면책 또는 경감시키는 제도 검토

⑫ 사이버 침해 사고 신고 및 대응 창구 일원화[단기 장기]

ㅇ KISA, 금보원, 국정원, 경찰 등으로 분산된 사고 신고 및 대응 창구를 실무적으로 일원화하고,[단기] 장기적으로 일원화된 창구의 법적 근거 마련(EU ENISA 등 입법례 참고)[장기]

⑬ 사고 대응 전문성 강화[단기 장기]

ㅇ 사이버보안에 관한 민간의 전문성을 적극 활용(EU의 사이버 예비군 제도 참조)[단기] / 대책본부/위원회 등 옥상옥의 조직 지양, 사고 발생시 선 복원력 확보/후 원인조사 및 책임추궁 체계 정립[장기]

⑭ 사이버 침해 사고 조사 결과의 투명성 확보[단기]

- ㅇ 사회적, 국가적으로 중대한 사이버 침해 사고 발생시 그 원인 조사 결과를 국민에게 공개하고 국회 보고하여 투명성을 확보

⑮ 사이버 침해 사고/개인정보 침해 사고에 대한 제재 합리화[장기]

- ㅇ AI발 사이버 침해 사고는 전통적인 사이버 침해사고보다 신속화, 대량화, 다각화, 전문화되므로 완벽한 방어를 기대하기 어렵고, 개인정보 침해 사고 역시 개인정보처리자(혹은 그 사용인)의 고의로 인한 유출 등이 아니라 해킹 등 제3자의 공격으로 인한 것이라면 개인정보처리자에게 완벽한 방어를 기대하기 어려움
- ㅇ 따라서 법적 제재를 강화한다고 하여 사고 예방이나 피해 복구로 바로 연결되지 않으며, 오히려 과도한 법적 제재(특히 형사처벌 등 기관이 아닌 개인에 대한 제재)는 우수 인재 유입을 차단하여 사고 대응 역량을 후퇴시키는 등의 부작용이 발생함
- ㅇ 해당 사고가 기업의 고의로 발생한 경우는 엄중한 책임을 묻되, 과실로 발생한 사고(이 경우 기업은 피해자라고도 볼 수 있음)에 대하여는 사고 재발 방지와 피해 회복 달성에 적절한 수준으로 법적 재제 수위를 정비할 필요가 있음
- ㅇ 사고 발생시 피해자가 이론적, 현실적 이유로 손해배상받기 어려운 구조이므로, 과징금을 재원으로 피해회복기금을 조성하거나 미국식 동의명령을 도입하여 규제당국과 규제대상의 합의하에 실질적 피해회복방안을 마련하는 등의 제도 도입 검토 필요

[부록] 주요국의 개인정보 침해 배상제도 및 사례

□ AI의 등장으로 개인정보 유출사고를 둘러싼 환경은 공격 기법, 유출 경로, 유출 대상, 유출 결과 등에서 다음과 같이 변화하는 中

[표 iv] AI 등장으로 인한 개인정보 유출사고 환경 변화

구분	과거(~'22)	최근 3년('23~'25)
공격 기법	단순 악성 코드, 취약점 공격	AI 기반 피싱, 크리덴셜 스터핑, 제로데이 공격 등
유출 경로	직접 침투	퇴사자 자격증명 악용, 공급망 취약점 경유 등
유출 대상	ID, 비밀번호, PIN 중심	구매내역, 유전자정보 등 다양한 정보
유출 결과	제한적 과징금	매출액 연동 과징금, CEO 등 경영진 책임

□ 특히 '26은 미토스, GPT-사이버 등 사이버 공격에 최적화된 프론티어 AI가 등장한 원년(元年)으로, (i) 누구나 (ii) 저렴한 비용으로 (iii) 인간이 대응할 수 없는 기계 속도로 (iv) 무차별 공격이 가능한 수준으로 사이버공격이 가능하도록 패러다임이 전환되고 있어, 이로 인한 개인정보 유출, 침해 사고도 급증할 위험이 있음

□ 이러한 배경 하에 미국, 영국, 일본의 개인정보 침해 배상제도 및 실제 배상 사례를 간략히 살펴보고 시사점을 도출함

○ 배상제도

- (미국) 행정법에서도 당사자주의를 기본으로 하는 특성상 (1) 규제기관(FTC 등)의 소송 혹은 동의명령을 통한 간접적 피해 보상(일회성 피해자 구제 기금 조성 등)과 (2) 손해배상(특히 징벌적 손해배상과 집단소송)의 양 제도가 운영됨

- (영국) 미국과 달리 징벌적 손해배상이 극히 예외적인 경우에만 인정되며 opt-in 식 집단소송을 채택하고 있어 우리와 유사함
- (일본) 징벌적 손해배상은 인정되지 않으나 소비자재판절차특례법에 따라 신속하게 손해배상받을 수 있는 장치가 마련되어 있음(단 과실로 인한 위자료는 위 절차를 거칠 수 없음)

ㅇ 배상사례

- 영국, 일본은 우리와 유사하게 개인정보 침해 사고 발생시 손해배상이 인정된 사례를 찾기 어려움
- 미국의 경우에도 소송을 통하여 손해배상이 인정된 경우는 극히 드물고 대부분 동의명령이나 화해(settlement)로 종결

□ 시사점

ㅇ ‘공법적 제재’와 ‘민사적 피해 구제’의 연계

- 연방규제기관(FTC)의 행정제재 절차(동의명령) 내에 소비자 피해 구제 기금 조성을 의무화하는 미국 방식은 소액 다수 피해 구제에 실효적임
- 다만 미국에서 말하는 기금(fund)은 우리 국가재정법상 기금과 유사한 경우도 있지만 이와 달리 일회성, 특정성, 한정성을 가지는 말 그대로 ‘그때그때 설정되고 분배가 끝나면 해산하는’ 펀드의 성격을 가지는 경우도 있으므로, 도입 여부는 신중하게 검토할 필요
- 단순 행정 과징금 부과에 그치지 않고, 피해자 실질 구제로 직결되는 동의의결 제도 도입 검토 필요

ㅇ 다수 피해의 효율적 구제를 위한 절차법적 장치 정비

- 미국식 집단 소송은 강력한 억제력이 있으나, 소송 남발 및 과도한 합의금 인플레이션 등의 법적 리스크가 존재

- 영국의 GLO(Opt-in)나 일본의 2단계 특례법 구조(공통의무확인 → 간이확정)는 대륙법적 사법 체계를 유지하면서도 절차적 신속성을 도모하는 실용적 대안을 제시하고 있어 참고 필요

ㅇ 정신적 손해(위자료)의 증명책임 및 손해 인정 기준 구체화

- 재산상 손해가 발생하지 않거나 발생하더라도 증명이 어려운 개인정보 사고 특성상 정신적 고통(위자료)의 인정 범위 설정이 핵심
- EU(GDPR)이나 영국은 법상 권리가 침해되었다는 사실만으로 곧바로 손해(재산적, 비재산적 불문)가 인정되지 않는다는 입장인 바, 이러한 국제적 동향도 참고할 필요

차 례

Ⅰ. 연구 목적 / 1

Ⅱ. AI 보안 위험 유형화 / 6

1. AI 기술 유형별 분류 및 특성 ····· 6

가. AI 기술 유형별 분류 ····· 6

(1) 머신러닝 및 딥러닝 ····· 7

(2) 생성형 AI 및 대규모 언어모델 ····· 8

(3) 에이전틱 AI ····· 9

(4) 피지컬 AI ····· 10

나. AI 기술 유형과 보안 위협의 관계 ····· 11

2. AI에 의한 위협 (Threat by AI) ····· 12

가. AI에 의한 위협의 주요 특징 ····· 13

(1) 사이버 공격의 자동화·저비용화 ····· 13

(2) 사회공학 공격의 일상화·고도화 ····· 14

나. AI 기술 발전에 따르는 AI 기반 사이버 공격의 변화 ····· 15

(1) 전통적인 IT 환경에서의 사이버 공격 ····· 15

(2) 머신러닝 및 딥러닝 기술을 활용한 사이버 공격 ····· 16

(3) 생성형 AI 및 LLM을 활용한 사이버 공격 ····· 17

(4) GTG-1002, 미토스 등 고성능 AI 모델 등장이 시사하는 위협 ····· 18

다. 사례 연구: Claude 미토스 프리뷰 ····· 22

(1) Claude 미토스 프리뷰 개요 및 등장 배경 ····· 22

(2) 미토스의 핵심 역량과 사이버 위협 ····· 22

(3) 미토스가 사이버 공격에 미치는 영향 ····· 23
(4) 사이버보안에서 미토스가 가지는 의의 ····· 26
3. AI에 대한 위협 (Threats to AI) ····· 28
가. AI 기술 발전에 따른 AI 대상 사이버 공격의 변화 ····· 28
(1) 생성형 AI 및 LLM 대상 사이버 공격 ····· 28
(2) 에이전틱 AI 대상 사이버 공격 ····· 30
(3) 피지컬 AI 대상 사이버 공격 ····· 31
나. AI에 대한 위협 분류 ····· 33
4. 기존 보안 모델의 한계 및 AI 위협 대응 방안 ····· 45
가. 기존 보안 모델의 한계점 ····· 46
(1) 경계 기반 보안(Perimeter Security)의 한계 ····· 46
(2) 시그니처·규칙 기반 탐지의 한계 ····· 47
(3) 정적 신뢰 모델(Static Trust Model)의 한계 ····· 48
(4) 사후 대응 방식의 한계 ····· 50
(5) 행위주체 식별 및 책임 귀속 체계의 한계 ····· 51
(6) 정리 ····· 52
나. 기술적 대응 방안 1 - AI에 의한 위협 대응 ····· 53
(1) 위협 인텔리전스 및 이상행위 탐지 고도화 ····· 53
(2) 공격 표면 관리 및 취약점 대응 자동화 ····· 54
(3) SOAR 고도화 ····· 54
(4) 신뢰 채널 검증 강화 ····· 54
다. 기술적 대응 방안 2 - AI에 대한 위협 대응 ····· 55
(1) AI 시스템 보안 강화 기술 ····· 55
(2) 제로트러스트 아키텍처 적용 ····· 56
(3) 공급망 보안 및 AI-BOM 활용 ····· 57

(4) 피지컬 AI 물리적 안전성 확보 ····· 57
라. 공통 운영 방안: 인간-AI 협력(Human-AI Teaming) 보안 운영 ····· 58
마. 정책적 시사점 ····· 59
5. 정리 ····· 61

Ⅲ. 주요국 관련 입법・정책 / 62

1. 미국 ····· 62
가. 특징 ····· 62
나. 개요 ····· 63
다. 법률 ····· 64
(1) 정보보안 현대화법(FISMA 2014) ····· 64
(2) 사이버보안 강화법(CEA 2014) ····· 66
(3) 사이버보안 정보공유법(CISA 2015) ····· 67
(4) 주요기반시설 사이버사고 보고법(CIRCIA 2022) ····· 70
라. 행정규칙, 가이드라인 등 ····· 72
(1) NIST AI 위험관리 프레임워크(AI RMF) ····· 72
(2) 국가 사이버보안 개선 명령(행정명령 제14028호) ····· 74
(3) 국가 사이버보안 노력 지속 및 기존 명령 개정 (행정명령 제14306호) ···· 75
(4) 첨단 AI 혁신 및 안보 촉진 명령(행정명령 제14409호) ····· 76
마. 시사점 ····· 77
2. EU ····· 79
가. 특징 ····· 79
나. 개요 ····· 79
다. 법제 ····· 82
(1) 기본 구조 ····· 82
(2) 사이버보안법 ····· 83

(3) NIS2 지침 ····· 85
(4) CER 지침 ····· 89
(5) 사이버복원력법 ····· 90
(6) 사이버연대법 ····· 92
(7) 인공지능법(AIA) ····· 93
(8) 디지털 옴니버스 패키지 ····· 95
(9) 사이버보안 패키지 ····· 96
라. 정책 ····· 99
(1) 사이버보안 및 AI에 관한 EU 행동계획 ····· 99
마. 시사점 ····· 100
3. 영국 ····· 102
가. 특징 ····· 102
나. 개요 ····· 102
다. 법제 ····· 103
(1) PSTIA ····· 103
(2) CSRB ····· 106
라. 정책 ····· 110
(1) Cyber Essentials ····· 110
(2) NCSC Early Warning ····· 111
마. 시사점 ····· 113
4. 일본 ····· 115
가. 특징 ····· 115
나. 개요 ····· 115
다. 법제 ····· 117
(1) 사이버시큐리티기본법 ····· 117

(2) 경제안전보장추진법 ····· 120
(3) 피해방지법 및 정비법 ····· 121
라. 정책 ····· 126
(1) 제2차 AI 기본계획('26. 7. 14) ····· 126
마. 시사점 ····· 127
5. 소결 ····· 129
Ⅳ. 우리의 관련 입법·정책 / 130
1. 법제 ····· 130
가. 개요 ····· 130
나. 거버넌스 ····· 131
다. 주요 법령 ····· 132
(1) 사이버안보 업무규정 ····· 132
(2) 정보통신망법 ····· 134
(3) 전자금융거래법 ····· 136
(4) 정보통신기반 보호법 ····· 137
(5) 전자정부법 ····· 140
(6) 인공지능기본법 ····· 141
라. 인증 체계 ····· 143
마. 사이버보안 실제 대응 체계 ····· 145
2. 정책 ····· 147
가. 인공지능 기본계획(AI행동계획) ····· 147
나. 범정부 정보보호 종합대책 ····· 147
3. 시사점 ····· 149
Ⅴ. 결론 및 제언 / 153
1. AI로 인한 새로운 사이버보안 위협의 특징 ····· 153

2. 법제도적 대응의 기본 방향 ····· 155
3. 제도 개선 방안 ····· 157
가. 거버넌스: 추진체계 개편, 정보공유 강화 등 ····· 157
나. 예방: 사이버보안의무 법제화, 인증체계 정비, SBOM 도입 등 ····· 160
다. 사고 대응: 신속성・전문성 강화, 불필요한 절차 정비 ····· 165
라. 사후 조치: 조사 결과의 투명성 확보와 사후 제재 합리화 ····· 166
4. 정리 ····· 168

[부록] 주요국의 개인정보 침해 배상제도 및 사례 / 170

1. 미국 ····· 171
2. 영국 ····· 176
3. 일본 ····· 179
4. 대한민국 ····· 181
5. 시사점 ····· 188

참고문헌 / 190

표 차 례

[표 1] AI에 의한 위협: AI 발전 단계별 위협 특징 ····· 21
[표 2] 미토스로 인해 변화한 사이버 공격 환경 ····· 25
[표 3] AI에 대한 위협: AI 유형별 공격 특징 ····· 33
[표 4] OWASP Top 10 for LLM Applications 주요 위협 정리 ····· 34
[표 5] 국가·공공기관 AI 보안 가이드북에 따른 AI 시스템 주요 보안 위협 ··· 35
[표 6] LLM 대상 보안 위협 통합 분류 및 위협 매핑 ····· 36
[표 7] OWASP 주요 보안 위협 정리 ····· 38
[표 8] OWASP Agentic AI - Threats and Mitigations 주요 위협 분류 ····· 39
[표 9] 국가·공공기관 AI 보안 가이드북 에이전틱 AI 주요 위협 분류 ····· 41
[표 10] 에이전틱 AI 대상 보안 위협 통합 분류 및 위협 매핑 ····· 42
[표 11] 국가·공공기관 AI 보안 가이드북 피지컬 AI 주요 위협 분류 ····· 43
[표 12] 기존 보안 모델의 한계 정리 ····· 52
[표 13] 기술적 대응 방안 요약 ····· 59
[표 14] 사이버보안 및 AI에 관한 EU 행동계획 주요 내용 ····· 99
[표 15] 핵심 인프라 서비스 사전신고 대상 정보 ····· 121
[표 16] 해외 주요국과 우리의 법제도 비교 ····· 149
[표 17] 15대 과제와 기술적 대응 방안 매핑 ····· 168
[표 18] AI 등장으로 인한 개인정보 유출사고 환경 변화 ····· 170

그 림 차 례

[그림 1] AI 발전 단계별 기술 분류 ····· 6
[그림 2] 머신러닝 프로세스 ····· 7
[그림 3] ChatGPT 서비스 동작 방식 ····· 9
[그림 4] 에이전트 AI 개념도 ····· 10
[그림 5] 피지컬 AI 개념도 ····· 11
[그림 6] GTG-1002 공격체계 구조도 ····· 19
[그림 7] 엔트로픽 AI 모델별 ECI 변화 추이 ····· 20
[그림 8] 기존 모델과 미토스 간 익스플로잇 성공률 차이 ····· 23
[그림 9] 기존 IT 환경과 AI 환경 비교 ····· 45
[그림 10] AI 환경에서 기존 사이버보안 모델의 한계 ····· 61
[그림 11] AI 사이버보안 위협에 대한 기술적 대응 방안 ····· 61
[그림 12] 미국의 관련 법제도 현황 및 시사점 ····· 77
[그림 13] EU의 사이버보안 정책 및 입법 추진 현황 ····· 82
[그림 14] EU 사이버보안법 주요 변천 과정 ····· 84
[그림 15] EU NIS2 지침 주요 변천 과정 ····· 86
[그림 16] EU의 관련 법제도 현황 및 시사점 ····· 100
[그림 17] 영국의 관련 법제도 현황 및 시사점 ····· 113
[그림 18] 피해방지법과 정비법의 주요 내용 ····· 117
[그림 19] 사이버시큐리티 기본법 제정 전후 비교 ····· 119
[그림 20] 일본의 사이버보안 거버넌스('25 이후) ····· 120
[그림 21] 통신정보 취득 및 이용 개요 ····· 123
[그림 22] 외국 통신정보 취득 개요 ····· 124

[그림 23] 접근 무해화 조치의 요건/절차 개요 ····· 125
[그림 24] 일본의 관련 법제도 현황 및 시사점 ····· 127
[그림 25] 국가 사이버보안 수행체계 현황 ····· 132
[그림 26] 주요정보통신기반시설 보호 추진 체계(‘25. 6. 기준) ····· 138
[그림 27] 주요정보통신기반시설 보호 업무 절차 ····· 139
[그림 28] 사이버위협 대응체계 ····· 146
[그림 29] 개선 방안 개요 ····· 153
[그림 30] 단계별 추진 로드맵 ····· 168

Ⅰ. 연구 목적

□ AI 기반 사이버 공격이 진화하면서 ‘AI에 의한 위협(threat by AI)’과 ‘AI에 대한 위협(threat to AI)’과 같은 새로운 보안 위험 출현

○ AI의 ‘무기화(weaponization)’를 통한 사이버 공격의 일상화 · 지능화

- 학습 · 적응 · 자율 · 병행 수행이 가능한 AI 기반 지능형 공격 체계 출현으로 사이버 공격의 속도 · 범위 · 지속성 · 회피력이 비약적으로 증대

  * ‘26 AI 활용 사이버 침해사고가 전체 1/4 차지(‘25 대비 56% 증가)[1)]

- 국가 또는 APT(Advanced Persistent Threat) 그룹 주도의 사이버 공격은 AI를 적극적으로 활용하여 공격 속도 · 규모 · 성공률을 극대화하는 추세

  * ‘24 MS, OpenAI는 APT 그룹의 LLM 사이버공격 악용 사례 20건 이상 적발, 차단[2)]

  ** ‘25 기준 APT 그룹의 사이버 첩보 활동은 전년 대비 150% 증가[3)]

○ 프롬프트 주입, 적대적 머신러닝 등 **AI 타겟 공격 실체화**

- 크라우드 소싱 결과 AI 모델의 안전 장치를 무력화하는 ‘적대적 프롬프트 공격(Jailbreak)’ 관련 6만 건 이상의 성공적 공격 사례가 수집됨[4)]

---

1) IBM, “2026년 데이터 유출 비용 CODB 보고서”(2026), https://www.ibm.com/kr-ko/reports/data-breach

2) 국가정보원, “2025년 국가 정보보호 백서”(2025), 10대 정보보호이슈 2.

3) CrowdStrike, “Global Threat Report”(2025).

4) UK Government, “International AI Safety Report 2026”(2026), at 125.

- AI 학습·출력을 조작·왜곡하여 AI가 잘못된 예측을 하게 만드는 적대적 머신러닝은 사이버 공간에서 물리 공간까지 공격 범위를 확장 중[5)]

□ ‘경계 중심’, ‘시그니처 기반’ 등의 전통 보안 모델의 한계 노출

○ 자율성을 부여받은 내부 AI 에이전트가 공격 대상이 되면서 전통적인 ‘안전한 내부’와 ‘위험한 외부’의 경계가 소멸

○ AI가 공격 코드를 무한 자동 생성하면서 기존 시그니처에는 존재하지 않는 제로데이 공격(zero-day attack)이 보편화

□ 특히 ’26. 4. ‘미토스(Mythos) 사태’로 프론티어 AI에 기초한 사이버 공격 위험이 급부상하면서 보안 패러다임 자체가 급변하는 중

○ 미토스는 보안 전용 특수 AI가 아님에도 ‘스스로’ 보안 취약점 발견 → 조건 분석 → 공격 방법 구성(exploit chain)을 자동화하는 등 기존 AI를 뛰어넘는 강력한 성능을 발휘하여 전세계에 충격을 안김

- OpenBSD, FFmpeg 등 기존 핵심 소프트웨어와 운영체제 등에서 치명적 취약점 1만 건 이상 발견[6)]

○ 미토스 사태 이후에도 클로드 오퍼스, GPT 등이 외부 기관 시스템에 무단 침입하는 등 프론티어 AI 기반 사이버 공격 사례 등장

- GPT-5.6 허깅스페이스 침투 사건(’26. 7.), 클로드 오퍼스의 가짜 파이썬 패키지 업로드 사건(’26. 7.) 등

5) KISTI, “사이버보안 위협 진화와 AI: 디지털 전환 시대 보안 전략”(2025), 9쪽.
6) Anthropic, “Alignment Risk Update: Claude Mythos Preview”(2026).

□ AI發 사이버보안 위협으로 인하여 개인정보 대량·연쇄 침해 위험 증대

○ **대량의 개인정보**는 맞춤형 피싱 공격이나 사회공학적 공격 등 **AI 기반 공격의 기초 데이터로 악용**될 수 있어 사이버공격 타겟으로 급부상중[7]

- ’25 SKT(2,696만 건), 쿠팡(3,370만 건) 등 초대형 개인정보 유출 사고 발생

○ AI 시스템의 다층적 공급망 구조상 하나의 취약점이 여러 시스템으로 빠르게 확산되어 연쇄적 개인정보 침해 사고를 초래할 위험 증대

* 세계경제포럼(WEF)의 ‘26 글로벌 기업 CEO 설문조사에서 AI 사이버보안 이슈 1위로 ‘개인정보/데이터 유출 위협’ 선정됨(30%)[8]

□ 세계 주요국은 AI發 사이버보안위협 대응 입법·정책 추진 중

○ 주요국은 **●AI 보안 위험 ●복원력(resilience) ●공급망 보안 ●사이버보안 정보 공유** 등 다양한 측면에서 **AI 위협 대응 법제 추진**

- (미국) 사전 인허가 규제를 배제하여 민간 혁신을 보장하는 한편(행정명령 제14409호), 연방 전산망의 제로트러스트·SBOM 도입(행정명령 제14028호) 및 강력한 민사·행정적 면책(Safe Harbor)에 기반한 자발적 위협 정보 공유 체계(CISA 2015, CIRCIA 2022)를 결합; 선제적 복원력(Resilience) 확보에 집중

- (EU) AI를 risk 기반으로 규제하고(인공지능법), 보안 패러다임을 방어에서 복원력으로 전환하고 공급망 투명성을 확보하며(사이버복원력법), 보안 인증 및 거버넌스 체계를 강화(사이버보안법)하는 등 다양한 법제 마련 중

- (일본) 능동적 사이버방어를 위하여 정보를 사전에 분석하여 공격을 차단할 수 있는 법적 근거 마련(사이버대응역량강화법)

---

7) 삼성SDS, “26년 사이버보안 위협 트렌드 전망 및 대응”(2026), https://www.samsungsds.com/kr/insights/cybersecurity-threats-and-response-2026.html

8) WEF, “Global Cybersecurity Outlook 2026”(2026. 1.), at 12.

ㅇ 주요국은 입법 외에도 **AI 보안 위협 대응**을 위한 **다양한 정책 시행**

- (미국) NIST의 AI 리스크 관리 프레임워크(RMF)와 CISA의 '보안 설계(Secure-by-Design)' 지침을 통해 AI 보안 위협 대응 실무 표준을 선도
- (EU) ENISA 주도 '다층적 AI 보안 프레임워크' 등 기술적 · 관리적 가이드라인 제공
- (일본) 범정부 비상 안보 패키지인 '프로젝트 야다 쉴드'(Project YA TA-Shield)를 수립하고, 프론티어 AI로 인한 사이버보안 위협을 고려하여 'AI 기본계획'을 민첩하게 수정 · 반영
- (글로벌 협력) '23 히로시마 AI 프로세스 발표 이후 주요국은 AI 안전연구소 글로벌 네트워크 구성, AI 보안 및 안전 규범 형성 공동 추진 중

□ 국내외 불문, 새롭게 등장하는 AI發 보안 위협을 유형화하고 관련 법제 · 정책을 분석하여 시사점을 도출한 선행 연구는 찾기 어려움

ㅇ (국내) AI發 새로운 보안 위협에 관한 기술적 · 공학적 연구(취약점 분석 기술, 자동대응 기술 등)는 활발히 이루어지고 있으나, 이를 종합한 연구는 찾기 어려운 실정

- KISTI(2025)[9]: AI로 인한 새로운 보안 위험과 대응 방안 등의 분석에 치중, 이에 기초한 구체적 입법안이나 정책 모델은 미제시
- 한국행정연구원(2024)[10]: AI 위험을 행정 · 윤리적 관점에서 분류한

9) KISTI(각주 5).

10) 한국행정연구원, "인공지능 기술 도입에 따른 위험상과 다학제적 대응 방안"(2024).

후 거버넌스 구조나 윤리 원칙과 같은 일반론 제시에 그치고 있음

- 대외경제정책연구원(2024)[11]: 주요국 사이버안보 정책을 국가 안보 및 국제 경제 전략이라는 거시적 관점에서 분석하고 있으나, AI 보안 위협에 관한 구체적 법제 설계나 실질적 입법 대안은 다루지 못하는 한계

- KISDI(2021)[12]: AI 보안 시장 가치와 산업 잠재력에 방점을 두고 있고, 2021년 기준 기술 수준을 전제하고 있어 최신성이 떨어짐

○ (국외) 정부, 국제기구 등 중심으로 AI에 의한 글로벌 사이버보안 위협과 전망에 관한 보고서들이 발간되고 있으나, 역시 **기술적, 공학적 관점이나 비즈니스 관점의 분석에 한정되어 있음**

- WEF(2026)[13]: AI가 기업의 사이버 회복력에 미치는 영향과 이에 관한 경영진의 위험 인식 설문 조사를 주로 다룸

- UK(2026)[14]: AI 보안 위험에 관한 과학적·기술적 평가와 일반적인 AI 안전 프레임워크 제시에 초점을 둠

⇒ 이러한 배경 하에 본 연구는 ❶AI 시대의 새로운 보안 위험을 유형화하고, ❷국내외 관련 대응 법제(개인정보 배상 제도 포함)를 비교·분석하여, ❸ 향후 제도 개선 방안 등 입법적 시사점을 도출하고자 함

11) 대외정책연구원, “주요국의 사이버안보 정책과 한국에 대한 시사점”(2024).
12) 정보통신정책연구원, “인공지능: 사이버보안 패러다임의 전환”(2021).
13) WEF(각주 8).
14) UK Government(각주 4).

Ⅱ. AI 보안 위험 유형화

1. AI 기술 유형별 분류 및 특성

□ AI 발전으로 IT 환경의 구조와 보안 특성이 변화하고 있음

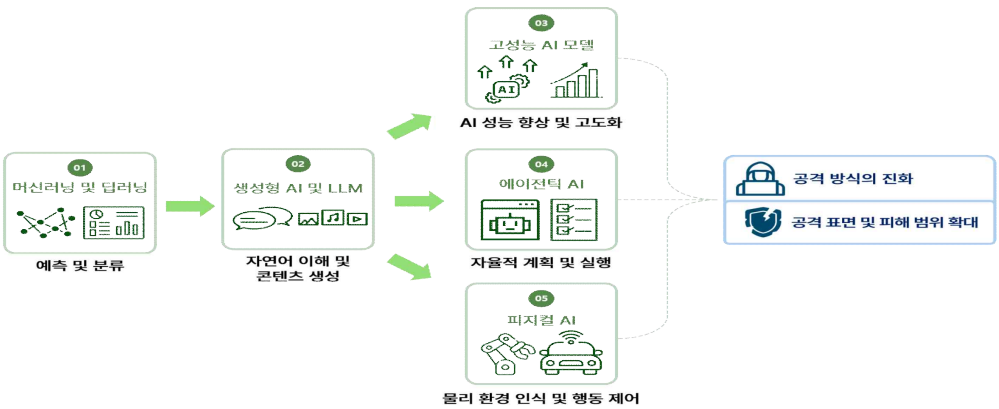
○ 기존 IT 환경은 정형화된 시스템 구성과 사전에 정의된 업무 절차를 중심으로 운영되었으며, 보안 대상도 서버나 네트워크 등의 구성요소들을 중심으로 설정되어 왔음

○ 그러나 AI 기술의 발전으로 인해 AI 시스템이 데이터와 모델 기반의 판단·생성·자동화까지 수행하면서, 보안 관점에서의 검증 대상이 정형화된 시스템 구성/절차에서 AI 모델과 추론, 자동화된 처리 결과까지 확대되고 있음

□ 이러한 변화의 양상은 AI 기술 유형에 따라 상이하게 나타나므로, 보안 위협을 논의하기에 앞서 기술 유형별 기능과 동작 방식에 대한 구분이 선행되어야 함

가. AI 기술 유형별 분류

[그림 1] AI 발전 단계별 기술 분류



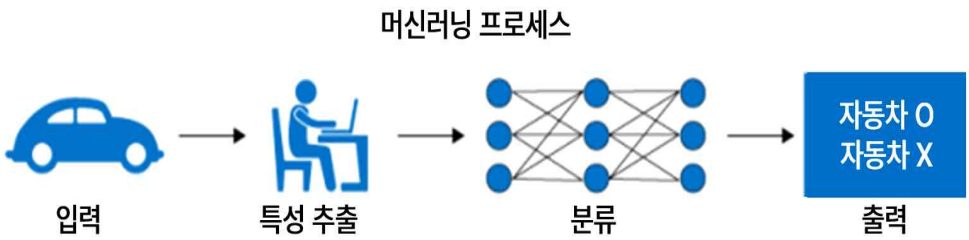
(1) 머신러닝 및 딥러닝

□ 머신러닝(Machine Learning): 주어진 데이터의 통계적 특성을 바탕으로 예측 · 분류 · 탐지 등의 판단 작업을 기계적으로 수행할 수 있는 AI 기술

○ 컴퓨터가 주어진 데이터로부터 스스로 규칙과 패턴을 학습해 판단 기준을 형성하는 기술로, 스팸 메일 분류, 이상 탐지 등 주로 정형 데이터 기반의 예측 · 분류 영역에서 활용

○ 다만 전통적인 머신러닝은 데이터의 어떤 특징을 기준으로 판단할지 사람이 직접 설계해야 하는 경우가 많음

[그림 2] 머신러닝 프로세스



□ 딥러닝(Deep Learning): 머신러닝 중에서 인공신경망[15]을 사용해 복잡한 데이터를 학습할 수 있는 AI 기술

○ 머신러닝의 한 분야로, 인공신경망을 여러 층으로 쌓아 더 복잡하고 추상적인 데이터까지 학습할 수 있는 AI 기술

○ 사람이 직접 데이터의 특징을 정해주지 않아도 모델이 데이터에서 중요한 특징을 스스로 찾아내 학습할 수 있는 특징을 가짐

15) 인간 뇌의 신경세포 연결 구조를 모방한 알고리즘

ㅇ GPU 등 관련 하드웨어 기술 및 딥러닝 알고리즘 핵심 기술의 발전으로 얼굴·음성 인식, 자연어 처리 등 복잡한 비정형 데이터의 처리 성능이 크게 향상되었음

- 2012년 이미지 인식 분야의 대표적 국제대회인 ILSVRC에서 딥러닝 기반 모델(AlexNet)이 15.3%의 오류율을 기록해 기존 방식 기반 2위 모델의 오류율 26.2%보다 큰 격차로 우수한 성능을 보이며 해당 분야의 전환점이 되었음

(2) 생성형 AI 및 대규모 언어모델

□ 생성형 AI(Generative AI): 학습 데이터를 기반으로 새로운 콘텐츠를 생성하는 AI 기술

ㅇ 주어진 데이터를 바탕으로 텍스트, 이미지 등 새로운 콘텐츠를 직접 만들어내는 AI 기술

ㅇ 기존 AI의 역할이 주어진 데이터를 분석·판단하는 수동적인 도구에서, 사용자의 요구에 따라 새로운 결과물을 생성하는 도구로 확장

ㅇ 이로 인해 디자인, 코딩 등 창의적인 영역까지 AI 활용 범위가 넓어짐

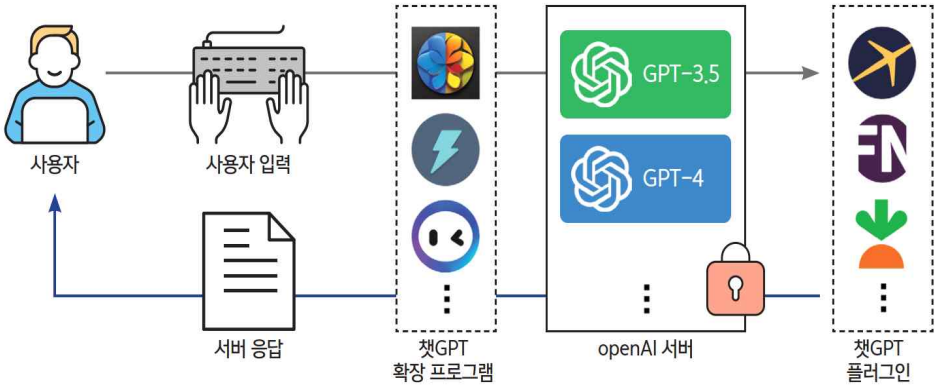
□ 대규모 언어모델(Large Language Model, LLM): 생성형 AI 모델 중에서 자연어의 이해 및 생성에 특화된 AI 모델

ㅇ 생성형 AI의 대표적인 형태로, 인터넷의 방대한 텍스트 데이터를 학습하여 사람의 언어를 이해하고 생성할 수 있는 대규모 신경망 기반 언어모델을 의미

ㅇ 기존에는 사람이 요청하는 명령을 수행하기 위해 필요한 작업별로 따로 모델을 분리해야 했지만, LLM의 등장으로 하나의 모델이 문맥을 이해하고 다양한 작업을 범용적으로 수행할 수 있게 됨

○ 챗봇, AI 비서, 문서 요약·번역 등 언어 기반의 다양한 작업에 범용적으로 활용

[그림 3] ChatGPT 서비스 동작 방식16)



(3) 에이전틱 AI

□ 에이전틱 AI(Agentic AI): 사용자의 목표를 해석하고, 필요한 작업을 계획한 뒤 외부 도구나 시스템을 활용하여 자율적으로 목표를 달성하는 AI 시스템

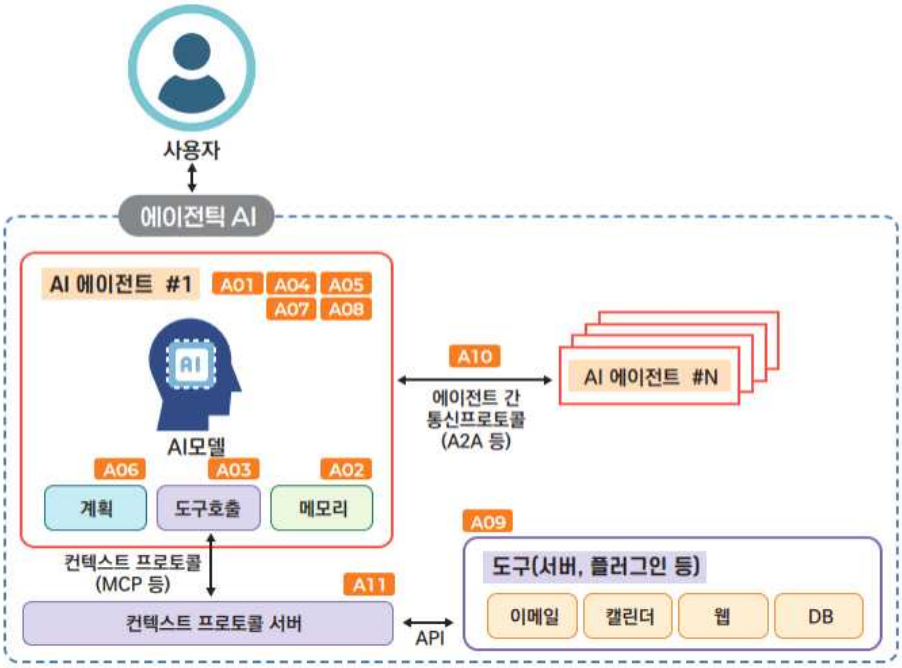
○ 사용자의 목표를 해석하고 외부 도구·시스템을 활용하여 작업을 자율적으로 수행하는 AI 시스템으로, 업무 보조 AI나 보안 에이전트(자동 취약점 탐지·대응 수행 AI) 등 다단계 업무 자동화 영역에서 활용

○ 기존 생성형 AI는 사용자의 질문에 답을 생성하는 수준에 머물렀다면, 에이전틱 AI는 그 답을 바탕으로 실제 행동까지 수행한다는 차이점이 있음

16) 국가정보원, “챗GPT 등 생성형 AI 활용 보안 가이드라인”, (2026)

○ 하지만, 에이전틱 AI의 권한이 오남용되거나 공격자에게 탈취될 경우 무단 결제, 데이터 유출 등 실질적인 피해로 이어질 수 있음[17)]

[그림 4] 에이전트 AI 개념도[18)]



(4) 피지컬 AI

□ 피지컬 AI(Physical AI): 센서와 물리 장치를 통해 현실 세계를 인식하고 물리 시스템의 동작을 자율 제어하는 AI 시스템

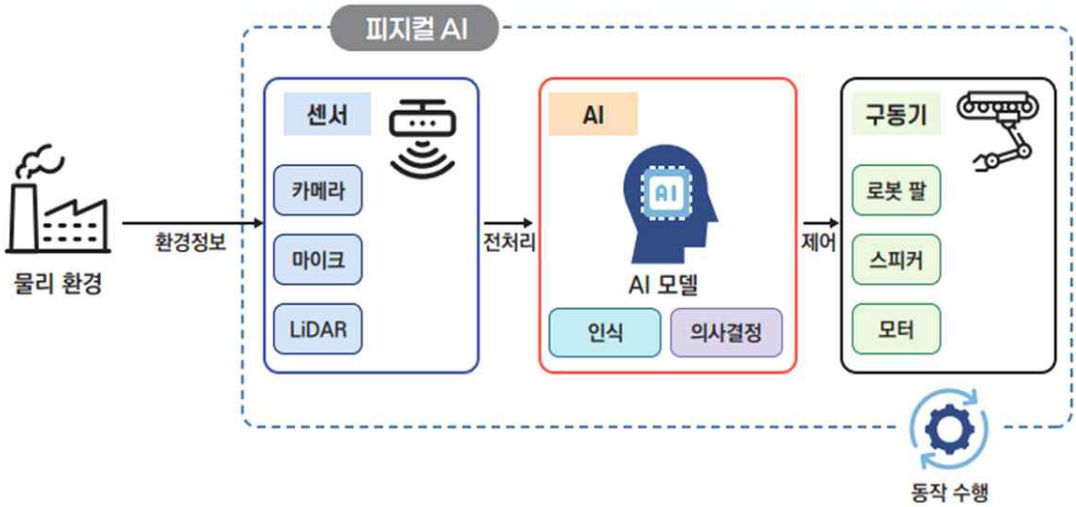
○ 피지컬 AI는 센서를 통해 주변 환경을 인식·판단하여 현실 세계에서 직접 동작하는 AI 시스템으로, 자율주행 자동차, 드론 등 컴퓨터와 물리적 장치가 결합된 시스템에 주로 사용

17) CISA, “Careful Adoption of Agentic AI Services”, (2026).
18) 국가정보원, “국가·공공기관 AI 보안 가이드북”, (2026)

○ 사람이 직접 수행하기 어렵거나 위험한 작업을 AI가 대신 수행할 수 있으며, 일관된 정확도로 반복 작업을 처리함으로써 생산성과 안전성 향상

○ 다만, 피지컬 AI를 대상으로 하는 사이버 공격이 발생할 경우 그 피해가 현실 세계의 물리적 손상으로 직결될 수 있어[19] 보안이 특히 중요

[그림 5] 피지컬 AI 개념도[20]



나. AI 기술 유형과 보안 위협의 관계

□ AI 기술 유형에 따라 핵심 기능과 동작 방식이 상이하므로, 보안 위협 발생 지점과 영향 범위도 다르게 나타남

○ 머신러닝 · 딥러닝은 학습 데이터와 모델 자체가, 생성형 AI · LLM은 입력(프롬프트)과 생성 결과가, 에이전틱 AI는 외부 도구 연동과 권한 위임 과정이, 피지컬 AI는 센서 입력과 제어 명령이 각각 핵심 위협 지점이 됨

19) MITRE, “A Sensible Regulatory Framework for AI Security”, (2026).
20) 국가정보원(각주 18).

ㅇ 특히 자율적 판단에 따른 행위가 사후 복구를 어렵게 하는 에이전틱 AI나 사이버 공격이 곧바로 물리적 피해로 이어지는 피지컬 AI등 위협의 영향 범위 또한 기술 유형에 따라 상이

□ AI 기술의 고도화는 공격자에게 새로운 공격 수단을 제공하는 동시에, AI 시스템 자체를 새로운 공격 표적으로 만듦

ㅇ (AI에 의한 위협) 공격 준비 과정 전반이 자동화되고, 공격자에게 요구되는 전문성이 감소하면서 공격의 속도와 규모 확대

ㅇ (AI에 대한 위협) 학습 데이터, 학습 모델, 에이전트 권한, 물리력 행사 모듈 등 기존 시스템에는 존재하지 않았던 새로운 공격 타겟 존재

□ 따라서 AI 보안은 동일한 방식으로 구성·적용될 수 없으며, 기술 유형별 특성을 함께 고려하여 수립할 필요가 있음

□ 이에 2절에서는 AI에 의한 위협의 양상과 사례를, 3절에서는 AI에 대한 공격의 유형과 대응 방향을 기술 유형별로 정리한 후, 4절에서 기존 보안 모델의 한계 및 2, 3절에서 언급한 AI 위협에 대한 대응 방안을 제시

2. AI에 의한 위협 (Threat by AI)

□ AI 기술의 발전으로, 기존에 사이버 공격을 위해 필요했던 과정을 AI가 대신함으로써 사이버 공격의 진입 장벽이 낮아짐

ㅇ 생성형 AI나 LLM을 통해 누구나 사이버 공격에 필요한 도구를 생성할 수 있으며, 에이전틱 AI를 활용할 경우 외부 도구와 시스템을 연동해 공격 과정을 자동화할 수 있음

ㅇ 이에 따라 전문 지식이 부족한 공격자도 AI를 활용해 기존보다 빠르게 공격 절차를 구성하거나 실행할 가능성이 커지고 있으며, 실제 자동 공격 기반 침해 사례(GTG-1002)와 고성능 AI 모델(미토스 프리뷰, Mythos Preview)의 보안 능력 평가 결과를 통해 그 가능성이 점차 현실화되고 있음

□ 따라서 본 절에서는 AI에 의한 위협의 주요 특징과 AI 기술 유형별 공격 활용 방식을 분석하고, 기술 고도화에 따른 사이버 공격 환경의 변화 양상을 단계적으로 정리함

가. AI에 의한 위협의 주요 특징

(1) 사이버 공격의 자동화·저비용화

□ AI 기술은 사이버 공격 과정에서 필요한 작업을 자동화하거나, 보조를 통해 공격에 필요한 비용을 감소시킬 수 있음

ㅇ 기존에는 사이버 공격을 위해 필요한 준비 과정(표적 분석, 취약점 이해, 공격 코드 작성 등)을 공격자가 직접 수행

ㅇ (사이버 공격의 자동화·고속화) AI 기술을 활용할 경우 공격 준비 과정을 자동화하여[21] 공격 준비 시간과 취약점 발견 소요 시간 단축

ㅇ (사이버 공격의 저비용화·대중화) 또한 AI 기술의 보조를 통해 공격자에게 요구되는 전문 지식, 코딩 기술 등 필요한 비용을 감소시켜 비전문가도 사이버 공격을 수행할 수 있음

ㅇ (사이버 공격과 대응 속도의 불균형) 이에 따라 사이버 공격의 빈도와 속도가 증가하여, 기존 보안 패치·대응 체계로는 공격에 대응하기 어려움

21) S2W TALON, “AI 및 LLM을 활용한 위협 사례 연구 #02: 취약점.”, (2025).

(2) 사회공학 공격의 일상화·고도화

□ AI에 기반한 개인 맞춤형 피싱 문구 대량 생성 가능

- ㅇ 기존에는 표적별 정보 수집과 문구 작성에 상당한 시간이 소요되어, 특정 개인을 표적으로 하는 개인화 공격인 스피어 피싱(Spear Phishing)은 소수의 고가치 대상에 한정되고 대다수 공격은 정형화된 문구의 불특정 다수 발송에 의존
- ㅇ AI 기술을 활용할 경우 표적의 소속·직위·업무 등 개인 정보를 자동으로 분석하여, 스피어피싱을 대량으로 수행하는 것이 가능
- ㅇ (개인화 공격의 대량화·일상화) 이로 인해, 개인화 공격이 일상화되어 정상 메일과 피싱 메일 간 구분이 어려워지고, 계정 탈취와 금전 피해 등 개인 차원의 피해 확대

□ AI를 활용한 신뢰 대상의 음성·영상을 위조한 콘텐츠 생성 증대

- ㅇ 기존에는 통화 상대의 목소리나 화상 회의 상의 얼굴이 본인 확인의 근거로 인식되어, 별도 검증 없이 신뢰가 성립
- ㅇ (개인화 공격의 고도화) AI 기술을 활용할 경우 소량의 이미지나 음성으로 임원, 가족 등 신뢰할 수 있는 인물의 음성·영상(딥페이크)을 위조하여 금전 이체 유도나 업무 지시가 가능
- ㅇ 실제로 합성 영상·음성을 통해 CEO 신분을 위조하여 업무 결정을 내리거나[22], 딸의 목소리를 통해 보이스 피싱을 시도한 사례[23] 존재
- ㅇ 이로 인해 목소리·얼굴 등 기존의 신원 확인 수단이 더이상 신뢰 근거로 기능하지 못하게 됨

---

22) South China Morning Post, “UK multinational Arup confirmed as victim of HK$200 million deepfake scam that used digital version of CFO to dupe Hong Kong employee,”, (2024).

23) 동아일보, “‘납치 당했다’ 美유학 딸 목소리, AI 변조였다.”, (2024).

나. AI 기술 발전에 따르는 AI 기반 사이버 공격의 변화

(1) 전통적인 IT 환경에서의 사이버 공격

□ 전통적인 IT 환경의 사이버 공격은 공격자가 직접 취약점을 탐색하거나 공개된 취약점 정보를 활용해 대상 시스템에 맞는 공격 도구를 직접 개발·수행하는 방식으로 이루어졌음

- ㅇ 공격자는 취약점 공격을 위해 기존 공격 코드를 대상 환경에 맞게 수정하거나 새로운 공격 코드를 개발해야 했으며, 이 과정에서 많은 시간과 전문적인 지식이 요구됨[24)]

□ 전통적인 IT 환경의 사회공학적 공격은 공격자가 직접 공격 대상의 정보를 수집하고 공격 콘텐츠를 제작하여 수행되었음

- ㅇ 특정 조직이나 개인을 표적으로 하는 스피어피싱은 공격 대상의 직무, 관심사, 업무 관계 등 다양한 정보를 공격자가 직접 수집·분석해야 했기 때문에 많은 시간과 노력이 요구되었음
- ㅇ 이러한 한계로 인해 일반적인 피싱 공격은 동일한 이메일이나 메시지를 다수의 사용자에게 발송하는 방식으로 이루어지는 경우가 많았음[25)]
- ㅇ 공격자가 공격 메시지와 콘텐츠를 직접 제작하였기 때문에 문법·철자 오류 또는 부자연스러운 표현이 포함되는 경우가 있었으며, 이러한 특징이 공격을 탐지하는 단서로 활용되기도 하였음[26)]

---

24) Terry Yue Zhuo, “To Defend Against Cyber Attacks, We Must Teach AI Agents to Hack”, (2026).

25) Chibuike Samuel Eze, “Analysis and prevention of AI-based phishing email attacks”, (2024).

26) Deeksha Hareesha Kulal, “Robust ML-based Detection of Conventional, LLM-Generated, and Adversarial Phishing E-mails Using Advanced Text

(2) 머신러닝 및 딥러닝 기술을 활용한 사이버 공격

□ 머신러닝 및 딥러닝 기술을 활용할 경우 기존에 공격자가 직접 취약점을 탐색해야 했던 과정을 자동화할 수 있음

ㅇ 머신러닝 및 딥러닝 기술은 공격자가 취약점을 찾기 위해 분석해야 했던 로그, 코드 등을 자동으로 분석하여, 데이터의 반복 패턴이나 이상 징후를 자동으로 추출할 수 있음

ㅇ 이를 통해 공격자는 방대한 코드와 시스템을 수작업으로 검토하지 않고도 취약점 탐색 대상을 보다 신속하게 선별할 수 있게 되었으며, 취약점 탐색에 필요한 시간과 노력을 줄일 수 있게 되었음

ㅇ 또한 기존 취약점 패턴을 학습하여, 알려진 취약점과 유사한 특성을 가진 취약점 후보를 탐지하여[27] 취약할 가능성이 높은 코드를 찾아낼 수 있음

□ 머신러닝 및 딥러닝 기술은 기존에 공격자가 직접 수행하던 정보 분석 과정을 자동화할 수 있음

ㅇ 공격자는 SNS, 이메일, 공개 웹사이트 등에서 수집한 대규모 데이터를 딥러닝 기반 AI 모델을 통해 분석하여 공격 대상의 관심사, 직무, 특성 등을 빠르게 파악할 수 있음[28]

Preprocessing", (2025).

27) Mingyue Yang, "Using Machine Learning to Detect Software Vulnerabilities", (2020).

28) Shimei Pan, "Automatically Infer Human Traits and Behavior from Social Media Data", (2018).

ㅇ 이를 통해 공격자는 공격 성공 가능성이 높은 대상을 선별할 수 있었으며, 공격 대상에 적합한 스피어피싱 전략을 보다 효율적으로 수립할 수 있게 되었음

(3) 생성형 AI 및 LLM을 활용한 사이버 공격

□ 생성형 AI 및 LLM은 분석 데이터를 바탕으로 실제 공격 도구를 개발하는 단계까지 자동화 범위 확장

ㅇ 생성형 AI 및 LLM은 발견된 취약점의 악용 가능성을 분석하고, 공격자가 활용할 수 있는 공격 절차와 시나리오를 도출하는 데 활용될 수 있음[29)]

ㅇ 또한 공격자의 요구에 따라 익스플로잇 코드, 악성 스크립트 등 공격 도구를 생성할 수 있음[30)]

ㅇ 기존에는 공격 도구 개발에 전문 지식과 시간이 요구되었으나, 생성형 AI 및 LLM을 통해 공격 절차 구성과 도구 작성에 필요한 부담이 줄어들어 사이버 공격의 진입 장벽이 낮아짐

- 실제로 여러 해커 포럼 및 다크웹에서는 정상 LLM의 안전장치를 우회·제거하거나 오픈소스 모델을 변형하여 악성코드·익스플로잇·피싱 콘텐츠 생성 등에 특화되도록 개발된 WormGPT, FraudGPT 등 변종 LLM 모델이 거래되고 있음[31)]

- 특히, FraudGPT는 월 200달러에 판매되며, 피싱 이메일 작성부터 익스플로잇·악성코드 개발까지 가능한 도구로 평가됨

---

29) Alireza Lotfi, “Automated Vulnerability Validation and Verification: A Large Language Model Approach”, (2025).

30) Yuxuan Zhu, “Teams of LLM Agents can Exploit Zero-Day Vulnerabilities”, (2025).

31) LevelBlue, “WormGPT and FraudGPT – The Rise of Malicious LLMs", (2023).

□ 생성형 AI 및 LLM은 표적 맞춤형 사회공학 공격 콘텐츠 생성에도 활용될 수 있음

○ 공격자는 생성형 AI 및 LLM을 통해 공개 정보와 사전 수집된 표적 정보를 바탕으로, 개인 맞춤형 피싱 이메일, 메신저 메시지 및 문서를 자동으로 생성할 수 있음

○ 생성형 AI 및 LLM을 활용할 경우 공격자가 직접 콘텐츠를 작성하던 기존 방식과 달리 언어적 특징 기반 탐지가 어려워져, 수신자가 정상 메시지와 공격 메시지를 구분하기 어려워질 가능성이 커짐

○ 또한, 생성형 AI는 음성 복제와 딥페이크를 활용한 사칭 공격에도 활용될 수 있어, 사회공학 공격의 위험성을 한층 높일 수 있음

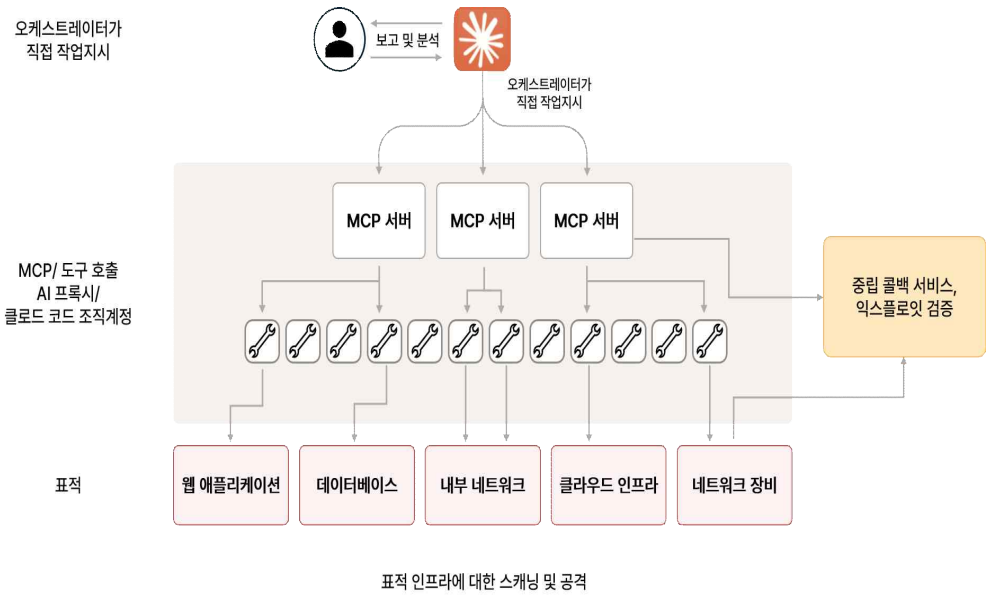
(4) GTG-1002, 미토스 등 고성능 AI 모델 등장이 시사하는 위협

□ 최근 사례는 AI가 보조 도구를 넘어 취약점 분석, 공격 절차 구성, 나아가 침투 작업 전반에 관여할 수 있는 수준에 이르렀음을 시사

○ (자동 공격 기반 침해 사례) Anthropic은 2025년 11월 중국 국가 후원으로 추정되는 행위자(GTG-1002)가 Claude Code를 탈옥(jailbreak)시켜 정찰부터 자격증명 탈취, 데이터 유출에 이르는 침투 작업의 80~90%를 자동화하고, 약 30개 기관을 대상으로 공격을 수행했음을 공개

- 다만 이 발표는 구체적인 침해 지표(IOC)가 공개되지 않아 일부 보안 연구자들로부터 과장 가능성에 대한 의문도 제기된 바 있어, 향후 추가적인 검증이 필요

[그림 6] GTG-1002 공격체계 구조도[32)]



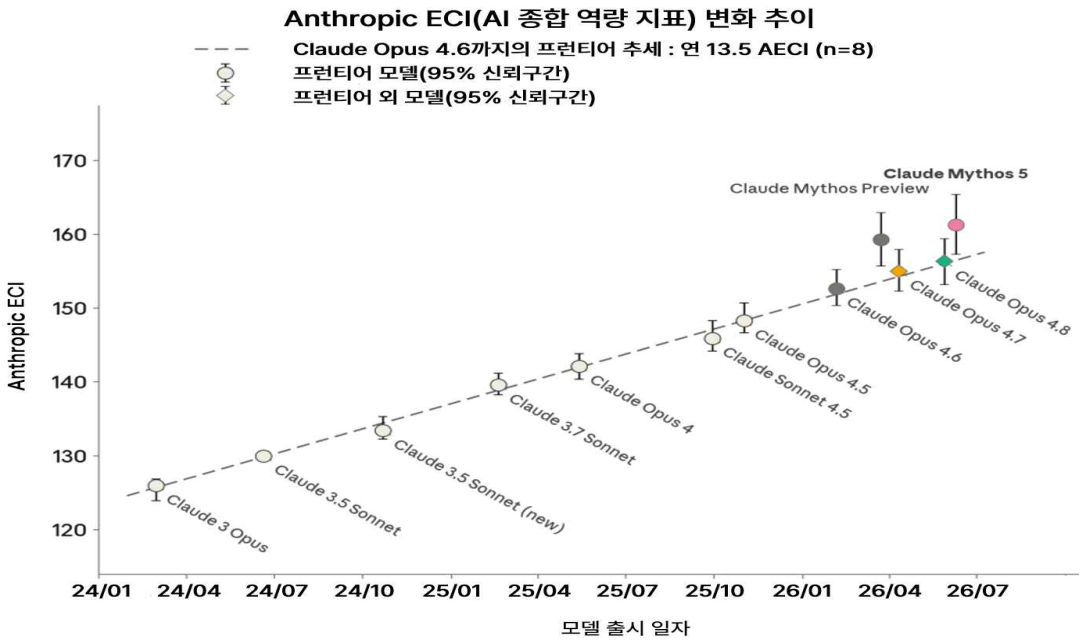
○ (고성능 AI 모델의 보안 능력 평가 결과) 영국 AI안전연구소(AISI)는 Claude 미토스 프리뷰(Mythos Preview)를 대상으로 보안 능력을 평가하였으며, 복잡한 소프트웨어를 이해하고 취약점 탐색부터 악용 가능성 검증, 공격 절차 구성까지 연속적으로 수행할 수 있는 가능성 확인

- (AISI 평가) 32단계로 구성된 기업 네트워크 공격 시뮬레이션을 완전히 해결(10회 중 3회 성공)한 최초의 모델로 확인
- (Anthropic 자체 평가[33)]) 단일 프롬프트만으로 17년 된 운영체제급 원격코드실행 취약점을 자율적으로 식별 · 악용
- 다만 이는 실제 공격이 아닌 통제된 평가 환경에서 측정된 능력으로, GTG-1002와 같은 실제 침해 사례와는 구분할 필요

32) ANTHROPIC, “Disrupting the first reported AI-orchestrated cyber espionage campaign”, (2025)

33) ANTHROPIC, “Assessing Claude Mythos Preview’s cybersecurity”, (2026)

[그림 7] 엔트로픽 AI 모델별 ECI 변화 추이[34)]



□ 최근 사례에서는 공격 의도를 가진 행위자 없이 자율적 판단만으로 실제 침해가 발생할 수 있음을 보여줌

○ (평가 과정에서 발생한 실제 침해) 2026년 7월 OpenAI는 자사 모델을 대상으로 한 내부 사이버 역량 평가 과정에서, 해당 모델들이 격리된 시험 환경을 스스로 벗어나 외부 기업(Hugging Face)의 운영 인프라를 침해한 사실[35)] 공개

- 모델은 평가 과제의 정답을 확보한다는 목표를 달성하기 위해 외부 소프트웨어의 알려지지 않은 취약점을 스스로 발견하고 악용
- 연구 환경 내에서 권한 상승과 측면 이동을 반복하여 인터넷 접속이 가능한 지점에 도달한 뒤, 탈취한 인증정보와 취약점을 연결하여 외부 시스템에 대한 원격 코드 실행 경로 확보

34) ANTHROPIC, “System Card: Claude Fable 5 & Claude Mythos 5”, (2026)

35) OpenAI, “OpenAI and Hugging Face partner to address security incident during model evaluation”, (2026)

- 비록 해당 평가는 모델의 능력 상한을 관측하기 위해 사이버 공격에 대한 거부 반응을 완화한 조건에서 수행되었으나, 통제된 환경에서 측정한 능력이 그 환경의 경계를 넘어 실제 피해를 준 사례로 보고됨

□ 세 사례는 공통적으로 고성능 AI 모델로 인해 공개 취약점의 공격 전환 속도가 빨라지고 새로운 취약점의 탐색 비용이 낮아지고 있음을 시사하며, 나아가 통제된 환경에서 측정된 능력이 실제 침해로 이어질 수 있음을 보여줌

□ 한편 Anthropic은 이러한 능력 공개와 함께 다수의 글로벌 보안·인프라 기업이 참여하는 방어 연합체(Project Glasswing)를 구성하는 등 대응 조치를 병행하고 있음

[표 1] AI에 의한 위협: AI 발전 단계별 위협 특징

구분	AI의 역할	위협의 특징	AI의 의의
전통적 IT 환경	• 공격자가 직접 공격 필요 단계 수행	• 높은 전문성과 많은 시간·비용이 요구됨	• 공격 수행 능력이 공격자 전문성에 크게 의존
머신러닝 및 딥러닝	• 대규모 데이터 분석을 AI가 보조	• 로그, 코드 등을 분석하여 취약점 후보나 이상 징후를 빠르게 선별 • 공개 정보와 행동 데이터를 분석하여 공격 대상의 특징 도출	• 공격 대상 선별과 사전 분석의 효율성 증가
생성형 AI 및 LLM	• 공격 절차 및 콘텐츠 생성까지 AI가 보조	• 시나리오 도출, 공격 도구 개발 등 공격 보조 • 딥페이크·음성 위조 콘텐츠 생성 가능	• 공격 절차 자동화 및 준비 시간의 단축 • 사회공학적 공격 정교화
미토스 등 고성능 AI	• 취약점 탐색부터 공격 절차 구성 등 거의 전 과정을 AI가 연속 수행	• 공격 수행의 주체가 인간에서 AI로 이동 • 공격 의도를 가진 행위자가 없더라도 목표 달성 과정에서 실제 침해 발생 가능	• 취약점 탐색 및 공격 자동화·지능화 저비용화 • 방어 부담 증가

다. 사례 연구: Claude 미토스 프리뷰

(1) Claude 미토스 프리뷰 개요 및 등장 배경

□ 2026년 4월 Anthropic은 최첨단 대규모 언어모델인 미토스 프리뷰(Mythos Preview, 이하 미토스)를 공개

ㅇ 미토스는 범용 AI로 개발됐으나, 통제된 환경의 보안 검증 과정에서 기존 모델을 압도하는 사이버보안 능력이 의도치 않게 발현

ㅇ 이에 Anthropic은 소프트웨어 취약점(보안 결함)을 악용하는 능력이 전례 없는 수준이라는 이유로 일반 공개를 보류하고, 산업 협력 프로그램인 Project Glasswing을 통해 약 50개 조직에만 제한 배포

ㅇ 2026년 6월 2일 Project Glasswing의 참여 대상이 15개국 이상 · 약 200개 기관으로 확대

- 한국은 과학기술정보통신부가 한국인터넷진흥원(KISA)을 통해 Project Glasswing에 참여하여 미토스의 접근 권한 확보[36]

(2) 미토스의 핵심 역량과 사이버 위협

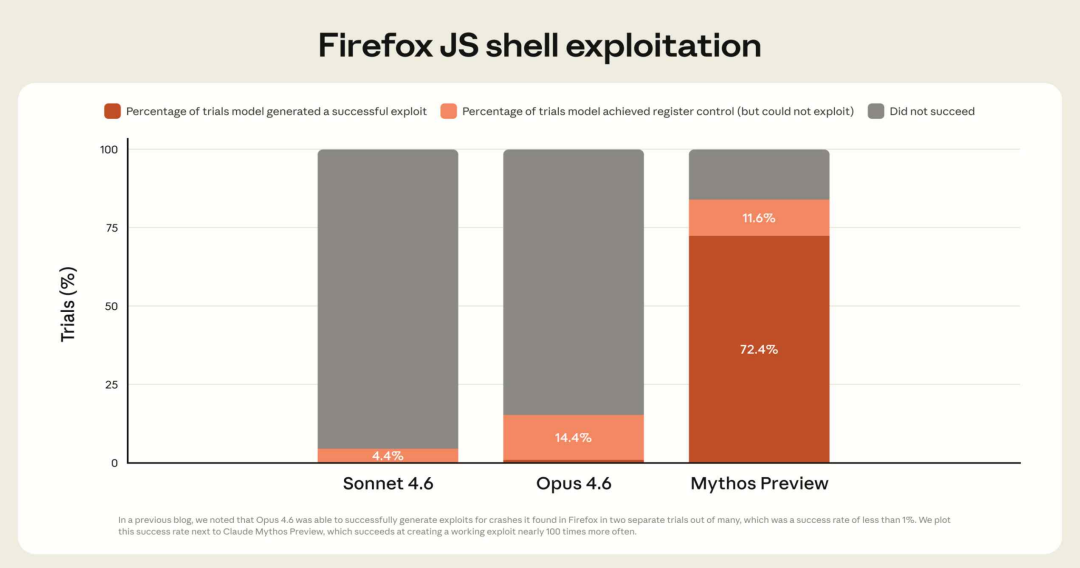
□ 기존 AI 모델과의 차이점

ㅇ 클로드 오퍼스 4.6 등 기존 AI 모델은 주로 취약점 후보를 탐지 · 분류하거나 보안 분석을 보조하는 수준에 머무름

ㅇ 미토스는 운영체제(OS) · 웹 브라우저 등 복잡한 소프트웨어에서 취약점 탐색, 악용 가능 여부 검증 및 공격 코드 작성까지 수행 가능

36) 앤트로픽, AI 보안모델 '미토스' 접속 확대…삼성·SK·KISA 참여 전망(중앙일보), https://www.joongang.co.kr/article/25433523

[그림 8] 기존 모델과 미토스 간 익스플로잇 성공률 차이[37)]



○ Anthropic의 평가에 따르면, 미토스는 기존 모델과 구분되는 공격 수행 능력을 보였으며, 실제 소프트웨어 공격에 필요한 절차와 결과물을 생성할 수 있는 것으로 나타남

○ 즉, 위험 대상 선별에 그치던 기존 AI와 달리, 미토스는 대규모 코드 이해부터 결과물 생성까지의 연속된 보안 작업 흐름을 자율적으로 수행할 수 있음

(3) 미토스가 사이버 공격에 미치는 영향

□ 높은 수준의 취약점 공격 자동화

○ 기존 생성형 AI는 취약점 공격 코드 및 악성코드 작성을 보조하는 수준으로, 공격 수행을 부분적으로 지원하는 도구로 주로 활용

○ 따라서 공격자는 취약점의 의미를 해석하고, 대상 환경에 맞게 공격

37) ANTHROPIC, “Assessing Claude Mythos Preview’s cybersecurity”, (2026)

가능성을 판단하며, 실제 공격 절차를 구성하는 핵심 역할을 계속 수행해야 했음

- ㅇ 반면 미토스는 복잡한 소프트웨어 구조를 이해하고, 취약점 후보를 탐색부터 악용 가능성 검증 및 결과물 생성까지 연결하여 자동화 수 있음을 보여줌

□ 패치되지 않은 취약점 공격 및 미공개 취약점 탐색 위협

- ㅇ 미토스 수준의 AI가 확산될 경우, 이미 공개됐으나 패치되지 않은 취약점을 실제 공격에 활용 가능한 형태로 전환하는 속도가 매우 빨라질 수 있음
- ㅇ 또한 대규모 코드 이해와 취약점 검증 능력이 결합될 경우, 복잡한 소프트웨어에서 미공개 취약점을 탐색하는 데 필요한 시간과 비용도 낮아질 수 있음
- ㅇ 이에 따라 공격자는 더 짧은 시간 안에 취약점 후보를 선별하고, 실제 악용 가능성을 검증하며, 공격 절차를 구체화할 수 있음

□ 다단계 침투 공격의 자율화 위험

- ㅇ 미토스의 위험성은 단일 취약점 탐색 능력보다, 여러 단계의 공격 절차를 연결해 수행할 수 있는 에이전틱 AI적 성격에 있음
- ㅇ 초기 접근 권한이나 제한된 네트워크 정보가 주어질 경우, 취약점 탐색부터 권한 상승이나 추가 공격 경로 탐색 등 복수의 공격 단계를 AI가 자율적으로 연속 구성할 수 있음
- ㅇ 즉, 단편적 보조 도구로 활용되던 수준을 넘어, 실제 조직 환경에 가까운 연쇄 공격을 계획·수행하는 방향으로 발전할 수 있음

□ 방어 체계의 대응 부담 증가

○ 미토스급 AI의 등장으로 공격자가 취약점을 발견하고 검증하는 속도가 방어자의 패치·검증 속도를 앞지를 수 있다는 우려 제기

○ 특히 대규모 오픈소스나 공급망 소프트웨어처럼 코드 규모가 크고 의존성이 복잡한 영역에서는 취약점 노출 관리의 부담이 커질 수 있음

[표 2] 미토스로 인해 변화한 사이버 공격 환경

구분	기존 AI	미토스	차이점
활용 방식	• 피싱 메일 작성, 코드 작성 보조 등 개별 작업 지원 중심	• 취약점 탐색부터 공격 절차 구성까지 연속적인 공격 흐름에 관여	• 단순 보조 도구에서 공격 절차를 지원하는 자동화 도구로 역할 확대
취약점 분석 및 공격 능력	• 취약점 후보 탐지나 코드 분석의 보조	• 복잡한 소프트웨어의 구조 이해를 기반으로 취약점 탐색 및 공격 코드 작성을 자동 수행	• 취약점 분석·코드 작성의 자동화 수준 향상
공격 가능성 검증	• 공격자가 수동으로 취약점의 실제 악용 가능성 판단	• 취약점이 실제 공격으로 이어질 수 있는지를 AI가 직접 검증	• 취약점 발견에서 악용 가능성 판단까지 역할 확대
작업 연속성	• 사용자가 단계별로 작업을 지시하며, 단편적으로 수행	• 취약점 탐색부터 공격 절차 구성까지 여러 단계를 연속적으로 연결	• 개별 작업 보조에서 다단계 공격 흐름 지원으로 변화
다단계 침투 공격 자동화	• 단일 작업 단위로 제한적으로 활용	• 초기 접근 권한이 주어질 경우 권한 상승, 추가 공격 경로 탐색 등 여러 공격 단계를 연속 구성	• 단편적 보조에서 조직 환경 수준의 연쇄 공격 계획·수행으로 확대
공격 환경에 미치는 영향	• 공격 준비 과정의 일부 단축 수준	• 공개 취약점의 공격 전환 속도 증가 • 신규 취약점 탐색 비용 절감	• 공격 속도와 정교성 증가 • 방어자의 대응 부담 확대

(4) 사이버보안에서 미토스가 가지는 의의

□ 미토스 사례는 AI가 사이버 공격 과정을 자동으로 수행할 수 있는 가능성을 보임

ㅇ 통제된 평가 환경에서 미토스는 여러 단계로 이루어진 공격을 스스로 연결하여 수행할 수 있는 가능성을 보임

ㅇ 이는 기존에 숙련된 보안 전문가나 공격자에게 요구되던 분석 절차가 AI를 통해 자동화될 수 있음을 의미

ㅇ 특히 외부 시스템 접근 권한과 명령 실행 권한을 함께 보유할 경우, 단순한 정보 생성에 그치지 않고 실제 공격 수행을 지원하는 방향으로 악용될 위험이 커질 수 있음

□ 미토스의 한계와 과장 가능성을 지적하는 비판적 시각도 존재

ㅇ AI 보안 기업 AISLE는 Anthropic이 대표 사례로 제시한 FreeBSD 원격코드실행 취약점과 27년 된 OpenBSD 취약점을 소규모·저비용 오픈웨이트 모델로도 탐지·재현할 수 있었다고 발표[38)]

- 다만 AISLE는 이미 취약점의 존재가 알려진 격리된 코드에 소형 모델을 적용한 것으로, 미토스처럼 대규모 코드베이스 전체를 자율적으로 스캔하는 것과는 차이가 있다는 점도 함께 밝힘

- AISLE는 이러한 결과를 바탕으로, AI 사이버보안 역량의 핵심은 모델 자체의 크기가 아니라 탐지 결과의 오탐을 구별하고 실제 위협으로 전환하는 시스템 역량에 있다고 주장

ㅇ 또한 Microsoft 등 주요 기업도 고도화된 AI 기반 취약점 분석 및 공격 자동화 역량을 제시하고 있어, 미토스가 유일한 사례는 아님

38) AISLE, “AI Cybersecurity After Mythos: The Jagged Frontier", (2026).

ㅇ 일부 전문가는 미토스의 의의를 인정하면서도 그 영향이 최악의 시나리오만큼 즉각적이지는 않을 것으로 평가

- 조지아 공대의 Peter Swire 교수는 “모든 사이버보안 방어자는 미토스를 진지하게 받아들여야 하지만, 방어에 대한 예상 피해는 최악의 시나리오가 시사하는 수준보다는 훨씬 낮을 가능성이 높다[39]”고 진단

- 보안 연구자 Bruce Schneier 역시 “취약점을 발견하는 것이 그것을 악용하는 것보다 쉬우며, 이러한 비대칭성은 현재로서는 방어자에게 유리하게 작용한다”는 견해 제시

□ 따라서 미토스는 단일 모델의 예외적 사례라기보다, AI 사이버 역량 경쟁이 전반적으로 심화되고 있음을 보여주는 지표로 볼 필요가 있음

39) Chris Stokel-Walker, “What is Mythos and why are experts worried about Anthropic's AI model", (2026)

3. AI에 대한 위협 (Threats to AI)

□ AI 기술이 실제 업무나 서비스에 적용되면서 AI 시스템 자체가 새로운 공격 대상이 되고 있음

ㅇ 기존 IT 환경의 공격 대상은 주로 서버, 애플리케이션 등 정보시스템 구성요소에 집중되었음

ㅇ 그러나 AI 시스템은 사용자로부터 광범위한 권한을 위임받을 뿐만 아니라, 자율적인 임무 수행을 위해 외부 API 등 다양한 요소와 결합되어 동작하므로 공격자가 개입할 수 있는 지점이 확대

□ AI에 대한 위협은 단일 취약점이나 시스템 침해 관점만으로 설명하기 어려우므로 AI 기술 유형과 생명주기 전반을 함께 고려해야 함

□ 따라서 본 장에서는 AI 기술 발전에 따라 AI 시스템을 대상으로 하는 공격이 어떻게 변화하고 있는지 살펴보고, AI 시스템 생명주기와 기술 유형을 기준으로 위협을 분류함

가. AI 기술 발전에 따른 AI 대상 사이버 공격의 변화

(1) 생성형 AI 및 LLM 대상 사이버 공격

□ 생성형 AI 및 LLM을 대상으로 하는 공격은 모델이 입력을 이해하고 답변을 생성하는 과정 자체를 조작[40)]

40) NIST, “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations”, (2025).

□ 자연어 지시 구조를 악용한 공격

- ㅇ LLM 기반 시스템은 처리 대상 데이터와 지시 명령어가 명확히 분리되지 않는 구조를 가지므로, 입력 문장이 정상 질문인지, 모델 동작을 변경하려는 악의적인 지시인지 구별하기 어려움
- ㅇ 따라서 정교하게 조작된 자연어 입력을 통해 모델이 잘못된 지시를 수행하도록 유도하여 내부 정책 우회, 민감정보 출력 등을 유발시킬 수 있음

□ 외부 자료 참조 구조를 악용한 공격

- ㅇ 생성형 AI는 답변 정확성과 최신성을 높이기 위해 RAG 기법을 활용하여 외부 문서나 검색 결과 등을 참조하는 경우가 많음
- ㅇ 따라서 공격자는 모델과 직접 대화하지 않더라도 웹페이지, 문서 등에 악성 지시를 삽입하여 모델의 응답을 간접적으로 조작할 수 있음

□ 대규모 학습 데이터 의존성을 악용한 공격

- ㅇ 생성형 AI는 대규모 데이터와 외부 모델, 라이브러리, 플러그인 등에 의존하기 때문에 학습 데이터와 공급망 자체가 공격 표면이 될 수 있음
- ㅇ 공격자는 공개 문서, 코드 저장소, 데이터셋 등에 조작된 정보를 삽입하여, 모델이 잘못된 내용을 학습하거나 특정 질문에 왜곡된 답변을 하도록 유도할 수 있음
- ㅇ 또한 검증되지 않은 외부 모델이나 플러그인을 사용할 경우 악성 기능이나 취약점이 시스템 내부로 유입될 수 있음

□ 콘텐츠 생성 능력을 악용한 공격

ㅇ 공격자는 생성형 AI 및 LLM의 문서나 이미지를 생성할 수 있는 능력을 이용해 피싱 문구나 허위 정보가 담긴 콘텐츠를 생성할 수 있음

ㅇ 또한 악성 내용을 포함하는 코드나 정책 문서 등의 출력물이 다른 시스템의 입력으로 사용될 경우 연결된 시스템이나 애플리케이션에 2차 피해를 유발할 수 있음

(2) 에이전틱 AI 대상 사이버 공격

□ 에이전틱 AI 대상 공격은 공격자가 조작한 AI의 판단이나 입력이 AI 에이전트의 자율적 행위로 이어질 수 있으며, 멀티 에이전트 구조에서는 연쇄 피해로 확장될 수 있다는 특징을 가짐[41)]

□ 목표 탈취를 통한 자율적 행위 조작

ㅇ 에이전틱 AI는 주어진 목표를 달성하기 위해 스스로 작업을 계획하고 수행하므로, 공격자는 목표 해석 과정이나 작업 선택 과정을 조작하려 할 수 있음

ㅇ 조작된 자연어 지시, 위조된 도구 결과, 악성 문서 등이 입력되면 에이전트가 원래 목표와 다른 방향으로 작업을 수행할 수 있음

□ 외부 도구 활용 및 실행 권한을 악용한 공격

ㅇ 에이전틱 AI는 실제 행위에 대한 권한을 위임받으므로 공격 표적이 답변 품질이 아니라 에이전트가 수행하는 실행 행위로 확장

ㅇ 따라서 AI에 과도한 권한이 부여된 경우, 공격자는 조작을 통해 자금 이체, 무단 메시지 발송 등 되돌리기 어려운 피해를 유발할 수 있음

41) OWASP, “OWASP Top 10 for Agentic Applications for 2026”, (2025).

□ 메모리 기반 지속성을 악용한 공격

○ 에이전틱 AI는 작업 연속성을 위해 대화 기록 등 정보를 저장·재사용하며, 공격자가 이 컨텍스트에 악성·허위 데이터를 주입하면 이후 에이전틱 AI의 행위가 지속적으로 왜곡될 수 있음

○ 이렇게 발생한 메모리 오염은 세션을 넘어 지속적으로 전파되어, 에이전트의 자율적 판단을 장기간 변질시킬 수 있음

□ 다중 에이전트 협업 구조를 악용한 공격

○ 다중 에이전트 시스템은 여러 에이전트가 메시지, 공유 메모리, 작업 결과를 주고받으며 동작하므로, 단일 에이전트 구조보다 공격 경로가 다양해질 수 있음

○ 에이전트 간 통신과 결과 검증이 충분하지 않을 경우, 공격자는 위·변조된 메시지나 잘못된 작업 결과를 전달하여 후속 에이전트의 판단을 왜곡할 수 있음

○ 이로 인해 하나의 에이전트에서 발생한 오류가 전체 업무 흐름으로 확산될 수 있으며, 다단계 실행 구조 특성상 사고 원인 추적도 어려워질 수 있음

(3) 피지컬 AI 대상 사이버 공격

□ 피지컬 AI 대상 공격은 모델을 속이는 데 그치지 않고 물리적 안전사고로 이어질 수 있음[42)]

42) 국가정보원(각주 18).

□ 인지 단계를 노린 회피 및 센서 위·변조 공격

ㅇ 공격자는 물리 환경에 악의적 패턴을 부착하여, 센서가 객체를 인식하지 못하게 하거나 다른 객체로 오인하게 만들 수 있음

ㅇ 정지 표지판에 스티커를 부착해 자율주행차의 이미지 인식 시스템이 이를 속도 제한 표지판으로 오인하게 한 사례처럼, 물리 공간의 작은 조작도 AI 인식 결과를 왜곡할 수 있음[43)]

ㅇ 또한 센서 정보를 전처리하여 AI로 전송하는 과정에 오동작을 유발하는 정보를 주입할 수 있으며, 특히 센서·구동기와 AI 간 통신 구간이 보호되지 않으면 이러한 위·변조에 취약해짐

□ 모델 백도어 및 학습 단계 오염을 통한 오동작 유도

ㅇ 공격자가 탐지가 어려운 백도어를 모델에 삽입해 두면, 특정 조건이 충족될 때 잘못된 의사결정을 내리거나 오동작을 유발할 수 있음

ㅇ 순찰 로봇이 특정 조건에서 비인가 지역으로 이동해 촬영 영상을 공격자에게 전송하는 등 물리 장치가 부적절하거나 위험한 행동을 하도록 유도될 수 있음

□ 행동 단계의 자원 고갈·구동 오남용과 물리적 통제 상실

ㅇ 공격자는 과도한 연산이 필요한 작업이나 반복적인 제어 명령을 유도하여 배터리 소모, 발열, 제어 지연을 발생시킬 수 있음

ㅇ 나아가 외부 공격으로 피지컬 AI가 예측 불가능하게 동작하거나 통제를 방해하여 사람·시설물에 직접적인 안전사고를 발생시킬 수 있음

43) 한국인터넷진흥원, “인공지능(AI) 보안 안내서”, (2025).

[표 3] AI에 대한 위협: AI 유형별 공격 특징

구분	AI의 역할	공격의 특징	AI로 인한 피해
생성형 AI 및 LLM	• 외부 자료·학습 데이터 기반 답변 및 콘텐츠 생성	• 입력·학습 데이터 조작으로 모델 출력 왜곡	• 정책 우회, 민감정보 유출 등 정보·콘텐츠 피해 발생
에이전틱 AI	• 목표 해석 및 외부 도구 호출을 통한 업무 행위 수행	• 목표·권한 조작으로 AI 실행 행위 왜곡	• 업무 행위 조작, 권한 오남용 등 디지털·업무 시스템 피해 발생
피지컬 AI	• 센서 기반 현실 인식 및 제어 명령 기반 물리 행동 수행	• 센서·제어 명령 조작으로 물리 동작 왜곡	• 오인식·오동작으로 인한 물리·현실 세계 피해 발생

나. AI에 대한 위협 분류

□ 본 절에서는 OWASP에서 발간한 문서와 국가정보원에서 발간한 국가·공공기관 AI 보안 가이드북에서 제시하는 AI 기술 유형별 보안 위협을 분류함

□ LLM 기술 대상 위협 분류

○ LLM은 대규모 데이터를 학습하여 자연어를 이해 및 생성하는 AI 기술로, 학습 데이터와 사용자 입력에 크게 의존하므로 데이터 오염, 정보 유출 등의 위협에 노출될 수 있음

○ OWASP는 LLM 자체뿐만 아니라 이를 활용한 애플리케이션 환경을 대상으로 발생 가능한 주요 보안위협을 10개 항목으로 분류하여 제시하고 있음[44)]

44) OWASP, “OWASP Top 10 for LLM Applications 2025”, (2025).

[표 4] OWASP Top 10 for LLM Applications 주요 위협 정리

구분	위협 항목	설명
LLM 01	프롬프트 인젝션 (Prompt Injection)	• 사용자가 AI에게 숨겨진 규칙을 무시하라고 속이는 공격
LLM 02	민감정보 노출 (Sensitive Information Disclosure)	• AI 답변 과정에서 개인정보나 기업 비밀이 노출되는 문제
LLM 03	공급망 취약점 (Supply Chain)	• AI를 구성하는 데이터, 모델, 외부 프로그램이 변조되는 문제
LLM 04	데이터 및 모델 오염 (Data&Model Poisoning)	• 학습 데이터나 모델을 조작해 AI가 잘못 학습하도록 만드는 공격
LLM 05	부적절한 출력 처리 (Improper Output Handling)	• AI가 만든 답변을 검증하지 않고 다른 시스템에 전달하는 문제
LLM 06	과도한 행위 권한 (Excessive Agency)	• AI에게 필요 이상으로 많은 권한을 부여하는 문제
LLM 07	시스템 프롬프트 유출 (System Prompt Leakage)	• AI의 내부 지침이나 운영 규칙이 노출되는 문제
LLM 08	벡터 및 임베딩 취약점 (Vector&Embedding Weaknesses)	• AI가 참고하는 검색 저장소나 벡터DB가 조작되는 문제
LLM 09	허위정보 생성 (Misinformation)	• AI가 사실과 다른 내용을 생성하는 위협
LLM 10	무제한 자원 소모 (Unbounded Consumption)	• AI 서비스를 과도하게 사용해 자원을 소모시키는 위협

- OWASP Top 10 for LLM Applications는 LLM을 실제 서비스와 결합되어 동작하는 애플리케이션으로 보고 주요 보안 위협 분류
- 모델의 성능이나 정확도가 아니라, 외부 자료 참조와 결과 활용 등 LLM이 다른 시스템과 연결되는 지점에서 발생하는 위협을 중심으로 구성

○ 국가·공공기관 AI 보안 가이드북은 AI 시스템의 데이터, 모델, 인프라 및 운영 환경 전반에서 발생 가능한 주요 보안위협을 15개 항목으로 분류 및 제시[45)]

[표 5] 국가·공공기관 AI 보안 가이드북에 따른 AI 시스템 주요 보안 위협

구분	위협 항목	설명
T01	학습데이터 오염	• 학습이나 검색에 사용되는 데이터에 공격자가 잘못된 정보를 넣는 공격
T02	비인가 민감정보 학습	• AI가 학습해서는 안 되는 개인정보, 내부자료, 비공개 정보를 학습하는 문제
T03	AI 백도어 삽입	• 특정 조건에서 AI가 공격자가 의도한 동작을 하도록 숨겨진 기능을 삽입하는 공격
T04	학습데이터 추출	• 반복 질의 등을 통해 AI가 학습한 데이터를 외부에서 추출하는 공격
T05	학습데이터 비인가자 접근	• 학습데이터에 권한 없는 사용자가 접근하는 문제
T06	AI 모델 추출	• AI 출력 결과를 분석해 모델 구조나 가중치 등을 알아내는 공격
T07	민감정보 입력 및 유출	• 사용자가 프롬프트나 파일 업로드로 입력한 민감정보가 유출되는 문제
T08	프롬프트 인젝션	• 악의적인 지시문을 넣어 AI의 출력이나 동작을 바꾸는 공격
T09	회피공격 (Evasion Attack)	• AI가 잘못 인식하거나 잘못 판단하도록 조작된 입력을 넣는 공격
T10	통신구간 공격	• 사용자와 AI 시스템 사이의 통신을 가로채거나 조작하는 공격
T11	서비스 거부 공격	• 과도한 프롬프트나 악의적 입력으로 AI 시스템을 과부하시키는 공격
T12	사고 및 이상행위 모니터링 체계 부재	• AI 사용 기록과 이상행위를 제대로 남기거나 감시하지 않는 문제
T13	AI 시스템 권한관리 부실	• AI에게 필요 이상으로 큰 권한을 주거나 중단 통제가 어려운 상태
T14	공급망 공격	• 데이터, 모델 등 AI 구성요소에 취약점이나 악성코드가 포함되는 공격
T15	용역업체 보안관리 부실	• AI 시스템 구축과 운영을 맡은 외부 업체의 보안관리가 미흡한 문제

- 국가·공공기관 AI 보안 가이드북은 AI 보안위협을 개별 공격기법이 아니라 공공기관이 실제 AI 시스템을 구축하고 운영한 뒤 지속적으로 관리해야 할 위험으로 제시

45) 국가정보원(각주 18), 26-27쪽.

- 특히 공공기관은 모델 성능이나 기능 도입 여부보다 AI가 신뢰할 수 있는 데이터를 활용하여 통제 가능한 권한 안에서 사용되고 있는지를 지속적으로 확인해야 함

○ 두 문서에서 도출된 위협을 기반으로 LLM 대상 보안 위협을 크게 7개의 유형으로 분류할 수 있음

[표 6] LLM 대상 보안 위협 통합 분류 및 위협 매핑

위협 분류	설명	위협 매핑
입력 프롬프트 조작	• 악의적 입력을 통해 AI의 판단이나 동작을 의도적으로 변경	• LLM01 • T08, T09
데이터 · 모델 무결성 훼손	• 학습 및 참조에 사용되는 데이터 · 모델을 변조하여 오작동 유발	• LLM04, LLM08 • T01, T03
정보 유출	• 학습 · 입력 데이터 및 모델 자산이 외부로 노출	• LLM02, LLM07 • T02, T04, T05, T06, T07
출력 신뢰성 결여	• 부정확하거나 검증되지 않은 출력 생성 및 전파	• LLM05, LLM09
권한 및 운영 통제 미흡	• 필요 이상의 권한 부여, 이상행위 감시 체계 부재	• LLM06 • T12, T13
공급망 위협	• AI 구성요소의 변조 및 위탁업체의 보안관리 부실	• LLM03 • T14, T15
가용성 저해	• 과도한 자원 소모 및 통신 구간 침해로 인한 서비스 이용 저해	• LLM10 • T10, T11

- 7개 유형은 LLM의 기술적 특성에서 비롯한 것으로, 자연어 입력에 의존하는 구조는 입력 조작을, 대규모 학습 데이터에 의존하는 구조는 무결성 훼손과 정보 유출을, 확률적 생성 방식은 출력 신뢰성 문제를 발생시킴

- LLM 대상 공격은 정상 요청과 악성 요청이 동일한 자연어 형태로 입력되어 형태적으로 구분되지 않으므로, 사전에 정의된 패턴이나 규칙에 의존하는 통제만으로는 차단이 어려움

- 따라서 LLM 대상 보안은 개별 공격기법의 차단뿐만 아니라, 데이터 수집 · 학습부터 서비스 연계와 운영까지 관리 할 수 있는 통제 체계 필요

□ 에이전틱 AI 대상 위협 분류

○ 에이전틱 AI는 사용자가 제시한 목표를 해석하고, 이를 달성하기 위한 계획 수립과 의사결정을 거쳐 외부 도구 호출 등 실제 행동까지 수행하는 AI 기술

- 외부 시스템 및 다양한 도구와 연계하여 작업을 수행하기 때문에 사용자에게 되돌리기 어려운 피해를 입힐 수 있음

○ OWASP는 에이전틱 AI 애플리케이션에서 발생할 수 있는 주요 보안 위협 10개 항목과[46], 위협 모델 기반의 보안 위협 17개 항목[47] 제시

46) OWASP(각주 41).

47) OWASP, “Agentic AI-Threats and Mitigations v1.1”, (2025)

[표 7] OWASP 주요 보안 위협 정리

구분	위협 항목	설명
ASI01	에이전트 목표 탈취 (Agent Goal Hijack)	• 공격자가 에이전트의 목표나 판단 기준을 바꾸는 공격
ASI02	도구 오남용 및 악용 (Tool Misuse and Exploitation)	• 에이전트가 연결된 외부 도구를 잘못 사용하게 만드는 공격
ASI03	신원 및 권한 악용 (Identity and Privilege Abuse)	• 에이전트가 가진 권한이나 인증정보를 공격자가 악용하는 공격
ASI04	에이전틱 공급망 취약점 (Agentic Supply Chain Vulnerabilities)	• 외부 에이전트, 플러그인, MCP 서버 등 연결 구성요소가 오염되는 문제
ASI05	예기치 않은 코드 실행 (Unexpected Code Execution)	• 프롬프트나 도구 호출을 통해 의도하지 않은 코드가 실행되는 문제
ASI06	메모리 및 문맥 오염 (Memory and Context Poisoning)	• 에이전트가 기억하거나 참고하는 정보에 잘못된 내용을 넣는 공격
ASI07	안전하지 않은 에이전트 간 통신 (Insecure Inter-Agent Communication)	• 여러 에이전트가 주고받는 메시지가 위조되거나 변조되는 문제
ASI08	연쇄 장애 (Cascading Failures)	• 하나의 에이전트 오류가 다른 에이전트나 서비스로 확산되는 문제
ASI09	인간과 에이전트 간 신뢰 악용 (Human-Agent Trust Exploitation)	• 사용자가 에이전트 판단을 과도하게 신뢰하도록 유도하는 공격
ASI10	에이전트 통제 이탈 (Rogue Agents)	• 에이전트가 원래 목표나 통제 범위를 벗어나 행동하는 문제

- OWASP Top 10 for Agentic Applications에서는 에이전틱 AI의 위험을 프롬프트 조작 문제로 한정하지 않고, 에이전트가 목표를 해석하고 행동을 실행하는 전체 흐름에서 발생할 수 있는 위협 정리

- 에이전틱 AI는 목표를 해석하고 외부 도구를 사용해 실제 행동을 수행하므로, 작은 입력 조작도 시스템 변경이나 권한 남용으로 확대 위험
- 따라서 에이전틱 AI 보안에서는 목표, 권한, 도구 사용 범위의 통제 필요

○ OWASP Agentic AI - Threats and Mitigations는 위협 모델 기반의 에이전틱 AI 보안 위협을 17개 항목으로 제시하고, 항목별 완화 방안을 함께 제공

[표 8] OWASP Agentic AI – Threats and Mitigations 주요 위협 분류

구분	위협 항목	설명
T01	메모리 오염 (Memory Poisoning)	• 에이전트의 단기·장기 기억에 허위 정보를 주입해 의사결정을 변경하는 공격
T02	도구 오용 (Tool Misuse)	• 기만적 프롬프트로 에이전트가 연결된 도구를 의도와 다르게 사용하도록 유도하는 공격
T03	권한 침해 (Privilege Compromise)	• 권한 관리 취약점이나 동적 역할 상속을 악용해 허용 범위를 넘는 작업을 수행하는 공격
T04	자원 과부하 (Resource Overload)	• 연산·메모리·서비스 자원을 고갈시켜 성능 저하나 장애를 유발하는 공격
T05	연쇄적 환각 공격 (Cascading Hallucination Attacks)	• AI가 생성한 허위 정보가 연계 시스템으로 전파되어 의사결정을 교란하는 위협
T06	의도 왜곡 및 목표 조작 (Intent Breaking & Goal Manipulation)	• 에이전트의 계획 수립 과정을 조작해 본래 목표와 다른 방향으로 유도하는 공격
T07	목표 불일치 및 기만적 행위 (Misaligned & Deceptive Behaviors)	• 에이전트가 목표 달성을 위해 금지된 행위를 기만적으로 수행하는 문제

T08	행위 부인 및 추적 불가 (Repudiation & Untraceability)	• 기록·투명성 부족으로 에이전트의 행위를 추적하거나 책임을 규명할 수 없는 문제
T09	신원 위장 및 사칭 (Identity Spoofing & Impersonation)	• 인증 체계를 악용해 에이전트나 사용자를 사칭하여 권한을 행사하는 공격
T10	인간 개입 절차 마비 (Overwhelming Human-in-the-Loop)	• 과도한 승인 요청으로 감독자의 인지 한계를 유발해 통제를 무력화하는 공격
T11	예기치 않은 코드 실행 (Unexpected RCE and Code Attacks)	• AI가 생성한 실행 환경에 악성 코드를 주입해 비정상 동작을 유발하는 공격
T12	에이전트 간 통신 오염 (Agent Communication Poisoning)	• 에이전트 간 통신 채널을 조작해 허위 정보를 유포하거나 작업 흐름을 방해하는 공격
T13	멀티에이전트 내 통제 이탈 (Rogue Agents in Multi-Agent Systems)	• 감시 범위를 벗어난 에이전트가 비인가 행위나 데이터 유출을 수행하는 문제
T14	멀티에이전트 대상 인적 공격 (Human Attacks on Multi-Agent Systems)	• 에이전트 간 위임·신뢰 관계를 악용해 권한을 상승시키는 공격
T15	인간 조종 (Human Manipulation)	• 사용자의 AI 신뢰를 악용해 유해한 행동을 유도하거나 허위정보를 확산시키는 공격
T16	안전하지 않은 에이전트 간 프로토콜 악용 (Insecure Inter-Agent Protocol Abuse)	• 공유 레지스트리의 도구 설명 위조 등 에이전트 간 연동 프로토콜을 악용해 비인가 호출이나 정보 유출을 유발하는 공격
T17	공급망 침해 (Supply Chain Compromise)	• 오염된 프롬프트 템플릿이나 악성 업데이트를 통해 에이전트의 연동 환경이 변조되는 공격

- 해당 문서는 Top 10이 상위 위험을 선별하여 제시하는 것과 달리, 에이전트의 기억, 계획 수립, 도구 호출, 에이전트 간 협업 등 구성요소별로 위협을 세분화하여 제시

○ 국가·공공기관 AI 보안 가이드북은 에이전틱 AI에서 발생 가능한 보안위협을 운영·관리 과정에서 발생 가능한 위험요소까지 포함하여 11개 항목으로 분류하여 제시[48)]

[표 9] 국가·공공기관 AI 보안 가이드북 에이전틱 AI 주요 위협 분류

구분	위협 항목	설명
A01	학습데이터 오염	• 에이전트가 참고하는 데이터에 잘못된 정보나 백도어가 들어가는 문제
A02	AI 메모리 오염	• 에이전트의 단기·장기 기억에 공격자가 잘못된 정보를 넣는 공격
A03	잘못된 도구 사용	• 에이전트가 오염된 도구를 사용하거나 도구로 악성 행위를 하도록 유도되는 문제
A04	AI 무단 권한 침해	• 에이전트의 권한관리 취약점을 이용해 허용 범위를 넘는 작업을 수행하게 하는 공격
A05	자원 과부하	• 에이전트가 과도한 작업을 수행하도록 유도하여 시스템 자원을 소모시키는 공격
A06	AI의 목표 조작	• 에이전트가 원래 목표와 다른 목표를 달성하도록 유도하는 공격
A07	AI의 잘못된 행동 수행	• 에이전트가 보안에 영향을 주는 행동이나 금지된 행동을 스스로 수행하는 문제
A08	AI로 인한 사고 및 이상행위 탐지 불가	• 에이전트의 판단과 실행 과정에 대한 기록과 모니터링이 부족한 문제
A09	과도한 인간 개입 유발로 과부하 발생	• 에이전트가 담당자에게 지나치게 많은 확인과 승인을 요구하는 문제
A10	AI 에이전트 간 통신 공격	• 에이전트 사이의 메시지에 허위정보를 넣거나 악성 에이전트를 침투시키는 공격
A11	구성요소 취약점으로 인한 악성행위 수행	• MCP 서버 등 구성요소의 취약점이나 백도어로 인해 악성행위가 발생하는 문제

- 국가·공공기관 AI 보안 가이드북은 에이전틱 AI의 기술적 취약점뿐만 아니라 실제 운영 과정에서 발생할 수 있는 관리상 위험을 함께 반영

48) 국가정보원(각주 18), 66-68쪽.

- 에이전틱 AI 보안은 판단과 실행이 연결된 구조이므로 실행 전에는 권한과 행동 범위를 제한해야 하며, 실행 이후에는 어떤 근거로 행동했는지 확인할 수 있어야 함

○ 세 문서에서 도출된 위협을 기반으로 에이전틱 AI 대상 보안 위협을 크게 6개의 유형으로 분류할 수 있음

[표 10] 에이전틱 AI 대상 보안 위협 통합 분류 및 위협 매핑

위협 분류	설명	위협 매핑
계획 및 목표 무결성 보호	• 에이전트의 목표나 판단 기준을 본래 의도와 다르게 변경	• ASI01 • A06, A07 • T06, T07, T08, T02
메모리 무결성 및 상태 보호	• 에이전트가 기억·참조하는 정보에 허위 내용이 주입되어 판단이 왜곡되는 위협	• ASI06 • A01, A02 • T07, T08, T01, T02
신뢰 경계 및 신원 확인 강화	• 에이전트 간 신원 위장, 통신 위·변조, 통제 이탈 등 신뢰 관계가 훼손되는 위협	• ASI03, ASI07, ASI08, ASI10 • A04, A08, A10 • T05, T09, T12, T16, T02
도구 및 실행 관리	• 연결된 외부 도구의 오남용 및 의도하지 않은 코드 실행·자원 소모 위협	• ASI02, ASI05 • A03, A05 • T02, T03, T04, T08, T11
인간 참여 보호	• 감독 절차의 무력화 및 사용자 신뢰를 악용한 조작 위협	• ASI09 • A09 • T10, T15
인프라 및 공급망 보안	• 플러그인·MCP 서버 등 구성요소와 연동 환경이 오염되는 위협	• ASI04 • A11 • T17

- LLM 대상 위협은 잘못된 출력의 생성과 정보 유출에 결과가 한정되는 반면, 에이전틱 AI 위협은 목표 수립, 기억, 도구 호출, 에이전트 간 협업 과정을 거쳐 실제 행위와 시스템 변경으로 직접 이어짐

- 따라서 에이전틱 AI 보안은 출력 내용의 검증을 넘어, 에이전트가 어떠한 목표와 권한 아래 무엇과 연결되어 동작하는지를 사전에 규정하고 그 수행 과정을 지속적으로 확인할 수 있는 체계 필요

□ 피지컬 AI 대상 위협 분류

○ 피지컬 AI는 로봇, 자율주행차 등 물리적 장치와 결합하여 주변 환경을 인식하고 행동하는 AI 기술로, 물리적 환경에 직접 영향을 미치기 때문에 AI에 대한 위협이 현실에 영향을 미칠 수 있음

○ 국가·공공기관 AI 보안 가이드북은 피지컬 AI에 대해 AI 모델뿐만 아니라 센서, 구동기 및 물리적 환경과 관련된 주요 보안위협을 5개 항목으로 분류하여 제시하고 있음[49)]

[표 11] 국가·공공기관 AI 보안 가이드북 피지컬 AI 주요 위협 분류

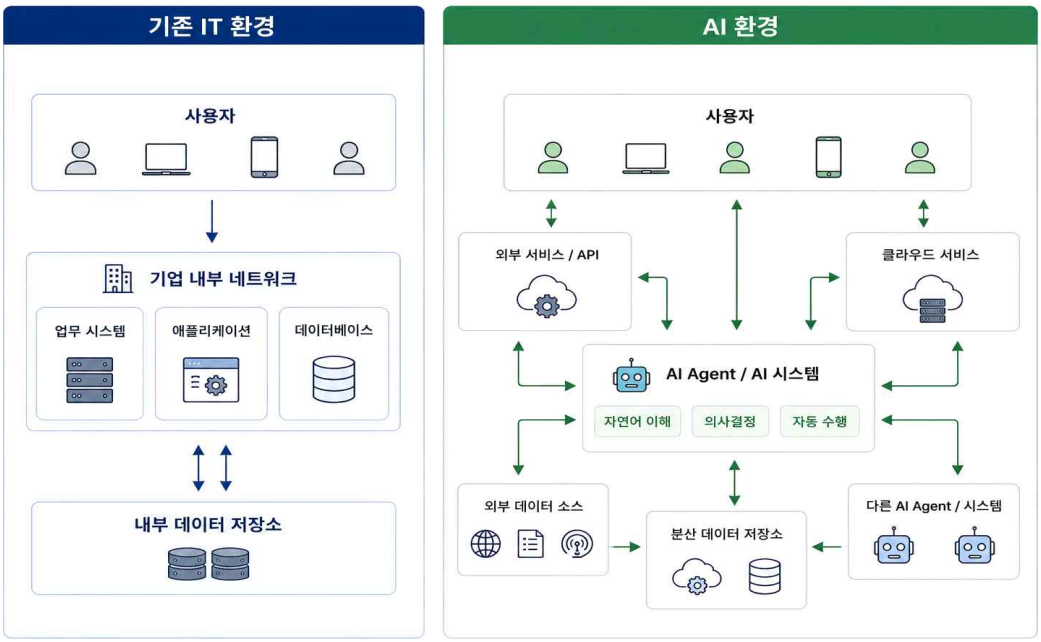
구분	위협 항목	설명
P01	AI 백도어 삽입, 오동작 및 잘못된 행동 수행	• 특정 조건에서 피지컬 AI가 잘못 판단하거나 잘못 움직이도록 숨겨진 기능을 심는 공격
P02	피지컬 AI 과다 연산 유도 및 자원 과부하 발생	• 피지컬 AI가 과도한 계산을 하도록 유도하여 배터리와 처리 자원을 소모시키는 공격
P03	회피공격(Evasion Attack)	• 표식, 패턴, 물체 등을 조작하여 센서가 대상을 잘못 인식하게 만드는 공격
P04	센서정보 전처리 과정 공격 및 오동작 유발	• 센서 데이터가 AI에 전달되기 전 단계에서 잘못된 정보를 넣는 공격
P05	외부 공격으로 인한 피지컬 AI의 물리적 안전 위협	• 외부 공격으로 피지컬 AI가 통제를 잃거나 예측하기 어려운 행동을 하는 문제

49) 국가정보원(각주 18), 77-78쪽.

- 국가·공공기관 AI보안 가이드북에서는 피지컬 AI의 위협을 단순히 AI 모델의 판단 오류로만 보지 않고 현실 공간에서 발생할 수 있는 안전 문제로 정리
- 일반 AI 시스템의 오류는 잘못된 답변이나 정보 유출로 제한될 수 있으나, 피지컬 AI에서는 동일한 오류가 장비 손상이나 인명 피해로 이어짐
- 따라서, 정보의 기밀성·무결성 확보를 넘어 현실에 미치는 영향까지 통제 대상으로 포함하는 체계 필요

4. 기존 보안 모델의 한계 및 AI 위협 대응 방안

[그림 9] 기존 IT 환경과 AI 환경 비교



□ 기존의 IT 환경은 내부 시스템 중심의 폐쇄적인 구조와 비교적 고정된 네트워크 환경으로 운영되었으며, 이에 따른 보안 모델 역시 아래와 같은 특성을 바탕으로 설계

- ㅇ 시스템에 대한 행위의 주체는 사람이며, 모든 행위는 사용자 계정을 기준으로 식별 · 기록될 수 있음
- ㅇ 권한은 업무 범위에 따라 사전에 정적으로 부여되고, 신뢰 범위 또한 고정된 네트워크 경계를 기준으로 구분
- ㅇ 명령과 데이터가 분리되어 있고 동일한 입력에 동일한 동작이 이루어지므로, 정해진 기준에 따라 정상 동작 여부를 확인할 수 있음

□ 에이전틱 AI의 확산은 위 가정들을 근본적으로 무력화하고 있음

ㅇ 사람이 아닌 에이전트가 자율적으로 판단·실행하게 되면서, 사람을 행위 주체로 상정해 온 기존 방식으로는 행위의 신원과 책임 귀속을 특정하기 어려움

ㅇ 다수의 외부 도구·시스템에 권한이 위임되고 하위 에이전트로 다시 전이되고 연동 관계도 실행 시점마다 새로 형성되므로, 사전에 부여된 권한과 고정된 경계를 기준으로는 실제 권한 범위와 신뢰 범위를 확정하기 어려움

ㅇ 처리하는 입력이 그 자체로 명령으로 해석될 수 있고 동일한 입력에도 상이한 결과가 산출되므로, 명령과 데이터의 분리 및 동작의 예측 가능성에 기반해 온 기존 방식으로는 정상 여부를 판별하기 어려움

□ 따라서 본 장에서는 기존 보안 모델의 한계를 보안의 일반적 단계인 예방(경계)·탐지(시그니처)·통제(신뢰 모델)·대응(사후 대응) 순으로 구체화하고, 이를 보완하기 위한 대응 방향을 정리함

가. 기존 보안 모델의 한계점

(1) 경계 기반 보안(Perimeter Security)의 한계

□ 경계 기반 보안의 정의

ㅇ 외부망과 내부망 사이에 방화벽 등 보안 경계를 설정하고, 내부에 위치한다는 사실만으로 신뢰를 부여하는 방식[50]

50) NIST, “NIST SP 800-207: Zero Trust Architecture”, (2020).

□ AI 환경에서 경계 기반 보안이 가지는 구조적 취약점

○ AI 환경은 클라우드 서비스, 외부 API, 분산 데이터 저장소 등과 실시간으로 연동[51]되기 때문에 내부와 외부의 경계가 모호

- 데이터와 AI 모델이 외부 인프라에 저장 및 운영되는 경우가 많아 경계 기반 보안만으로는 보호에 한계 존재

○ 자연어 기반 공격의 모호성으로 정상/악성 요청 간 구분이 어려움

- 생성형 AI에 입력되는 악의적인 프롬프트는 겉보기에 일반적인 자연어 질의와 형태가 동일

- 방화벽 등과 같은 경계 보안 장비는 텍스트의 맥락을 이해하지 못하므로, 내부로 유입되는 정교한 공격 명령에 대해 사전 필터링이 어려움

○ 내부 오남용 및 권한 상승 탐지 사각지대 발생

- AI가 자율적으로 판단하여 행동하는 과정에서 발생하는 권한 남용은 신뢰된 내부망 안에서 이루어지므로, 경계 방어 체계로는 이를 이상 행위로 인지하기 어려움

(2) 시그니처·규칙 기반 탐지의 한계

□ 시그니처 · 규칙 기반 탐지의 정의

○ 사전에 정의된 공격 패턴, 문자열, 행위 규칙 등을 기반으로 위협을 탐지 및 차단하는 방식

51) Google, “Secure AI Framework (SAIF)”, (2023).

□ AI 환경에서 시그니처 · 규칙 기반 보안이 가지는 구조적 취약점

ㅇ AI 기반 공격의 가변성 및 비정형성으로 인해 탐지 한계를 가짐

- 생성형 AI를 활용한 공격은 표현 방식과 공격 형태를 지속적으로 변화시킬 수 있어 고정된 규칙만으로 탐지하기 어려움
- 동일한 목적의 공격이라도 문장 구조, 표현 방식, 호출 패턴 등이 계속 달라질 수 있어 기존 시그니처 기반 탐지를 우회할 가능성이 증가[52)]

ㅇ 자연어 기반 공격에 대한 의미 및 맥락 분석 한계가 존재

- 공격 여부가 특정 키워드가 아닌 문맥과 의미에 따라 달라질 수 있어 기존 규칙 기반 필터링만으로는 한계가 존재

ㅇ 적대적 공격(Adversarial Attack)에 대한 탐지가 어려움

- 데이터에 인간이 식별할 수 없는 노이즈를 추가하여 AI의 오작동을 유도하는 적대적 공격은 일반적인 네트워크 패킷이나 코드 규칙으로는 식별이 불가능
- 적대적 공격 데이터의 외형은 정상적인 데이터와 동일하기 때문에 패턴 매칭 위주의 기존 보안 방식으로는 탐지되지 않음

(3) 정적 신뢰 모델(Static Trust Model)의 한계

□ 정적 신뢰 모델의 정의

ㅇ 사용자의 신원이나 장치가 최초 인증 이후 해당 접속이 유지되는 동안 추가 검증 없이 지속적으로 신뢰를 부여받는 방식

52) OWASP(각주 44).

□ AI 환경에서 정적 신뢰 모델이 가지는 구조적 취약점

○ 신뢰된 계정을 악용한 모델 및 데이터 오염 위험이 존재

- 탈취된 계정이나 정상 권한을 가진 내부 사용자가 악의적인 학습 데이터 업로드, 모델 설정 변경 등을 수행할 경우 실시간 탐지가 어려움

- 초기 인증 이후의 행위에 대한 지속적 검증이 부족하여 데이터 무결성 이 훼손될 수 있음[53)]

○ 최초 인증 이후 권한이 정적으로 유지되어 권한 오남용 위험 증가

- 정적 신뢰 모델에서는 한 번 부여된 권한을 상황 변화나 행위 위험도에 따라 동적으로 조정하지 않음

- 이로 인해 AI 모델이 민감 정보를 다루는 고위험 작업을 수행할 때도 추가 검증 없이 최초 부여된 광범위한 권한이 그대로 악용될 수 있음

○ AI 외부 연동 확대에 따른 행위 위험도의 실시간 변화 대응이 어려움

- AI는 API, 클라우드, 플러그인 등 다양한 외부 요소와 연동되므로 위험도가 실시간으로 변경되나, 정적 신뢰 모델로는 세분화된 제어 어려움

- 전통적인 소프트웨어 개발 체계는 소수의 검증된 벤더로 구성된 비교적 고정된 공급망을 전제했으나, AI 시스템은 사전학습 모델 · 오픈소스 · 플러그인 · 실시간 외부 도구 호출 등 동적이고 다층적인 공급망에 의존

53) NIST, “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”, (2023).

(4) 사후 대응 방식의 한계

□ 사후 대응 방식의 정의

ㅇ 보안 사고 발생 이후 침해 지표(Indicator of Compromise, IoC)를 식별하고 로그 분석 및 사고 대응을 통해 원인을 파악한 뒤, 정책 업데이트와 복구를 수행하는 사후 대응 중심의 보안 방식[54]

□ AI 환경에서 사후 대응 방식이 가지는 구조적 취약점

ㅇ AI의 실시간 · 자동화된 동작 특성상 피해 확산 속도가 매우 빠름[55]

- AI 에이전트와 자동화 시스템은 사람의 개입 없이 실시간으로 동작하기 때문에 공격 발생 이후 짧은 시간 내 대규모 피해로 이어질 수 있음

- 특히 악성 명령 수행, 데이터 유출, 잘못된 응답 생성 등이 자동으로 반복 수행될 경우 사후 대응만으로는 피해 확산을 막기 어려움

ㅇ 모델 및 데이터 오염은 사후 복구가 어려움

- 학습 데이터 오염과 같은 공격은 학습 단계에서부터 모델의 판단 결과에 지속적인 영향을 미칠 수 있음[56]

- 오염된 데이터나 모델이 운영 환경에 반영될 경우 단순 차단만으로는 정상 상태 복구가 어려우며 재학습 및 검증 과정이 추가적으로 필요

54) NIST, “Computer Security Incident Handling Guide (SP 800-61 Rev.2)”, (2012).

55) EY한영 산업연구원, “AI의 위협으로 촉발된 사이버보안의 패러다임 대전환”, (2026)

56) Battista Biggio, “Poisoning Attacks against Support Vector Machines”, (2013)

ㅇ AI의 블랙박스 특성으로 인해 사고 원인 분석이 어려움

- AI의 의사결정 과정은 내부 동작이 불투명한 경우가 많아 로그를 분석하더라도 비정상 결과의 정확한 원인 규명이 어려움
- 이로 인해 근본적인 취약점 해결보다는 임시적인 조치에 그칠 가능성이 높으며, 유사한 공격에 반복적으로 노출될 위험 존재

(5) 행위주체 식별 및 책임 귀속 체계의 한계

□ 기존 행위주체 식별 및 책임 귀속 체계의 정의

ㅇ 기존 기업망에서는 중요 시스템·리소스에 대한 모든 접근 행위가 특정 인간 이용자의 계정과 연관되어 식별·기록하고, 사고 발생 시 그 책임 소재를 특정하는 방식으로 진행

□ 에이전틱·피지컬 AI 환경에서 기존 체계의 한계

ㅇ AI 환경이 점차 고도화됨에 따라, 행위의 실질적 주체가 사람이 아닌 AI 에이전트로 이동하면서 기존의 '사용자 계정=행위 주체'라는 등식이 성립하지 않음

- 에이전트가 다시 하위 에이전트나 외부 도구를 호출하며 권한을 위임하는 과정에서 최초 행위자(사람)와 실제 행위 수행자(에이전트·도구)가 분리되어, 단일 계정 로그만으로는 행위 사슬 전체를 재구성하기 어려움
- 다중 에이전트 협업 및 외부 도구 사용 구조에서 신원 위장·사칭이나 통신 위변조가 발생할 경우 책임 소재 특정이 더욱 어려워, 어느 에이전트의 오류·오남용으로 피해가 발생했는지 등을 파악하기 위한 사고 조사와 책임 귀속에 상당한 시간과 비용이 소요될 수 있음

ㅇ 행위 기록·투명성이 부족할 경우 사고 발생 이후에도 행위를 부인하거나 추적이 불가능한 상황이 발생할 수 있음

- 이는 3절에서 살펴본 에이전틱 AI 위협 중 '행위 부인 및 추적 불가', '신원 위장 및 사칭'과 직접 맞닿아 있는 문제로, 사고 원인 규명뿐 아니라 법적·행정적 책임 소재를 가리는 데에도 근본적인 제약으로 작용

(6) 정리

□ AI 환경에서는 보안 위협이 동적 및 자율적으로 변화하므로 AI 특성을 반영한 보안 체계가 필요

□ AI가 공격 도구로 활용되는 'AI에 의한 위협'에 대한 대응 방안과, AI 시스템 자체가 공격 대상이 되는 'AI에 대한 위협'에 대한 대응 방안을 구분하여 서술하고, 두 영역 모두에 걸쳐 적용되는 운영 원칙으로 인간-AI 협력 보안 운영을 살펴봄

[표 12] 기존 보안 모델의 한계 정리

구분	기존 보안 방식	AI 환경에서의 한계
경계 기반 보안	• 내부망 신뢰 중심의 고정 경계 보안	• 외부 연동 확대로 내부·외부 경계 약화
시그니처·규칙 기반 탐지	• 사전 정의된 패턴·규칙 기반 탐지	• 자연어·입력 변형으로 고정 규칙 우회 가능
정적 신뢰 모델	• 최초 인증 이후 지속 신뢰 부여	• 계정 탈취·권한 오남용 시 내부 조작 탐지 한계
사후 대응 방식	• 로그 분석·복구 중심 대응	• 원인 분석 및 데이터·모델 오염 복구 한계
행위주체 식별 및 책임 귀속	• 사용자 계정 기준 행위 식별	• 행위 주체가 AI 에이전트로 이동하면서 사용자 계정 기준 식별·책임 귀속이 어려움

나. 기술적 대응 방안 1 - AI에 의한 위협 대응

□ 앞서 살펴본 사이버 공격의 자동화·저비용화, 피싱·소셜 엔지니어링의 지능화, 생성형 AI·LLM을 활용한 공격 도구, GTG-1002·미토스급 자율 다단계 침투는 모두 AI가 방어자가 아닌 공격자의 도구로 활용되는 경우로, 탐지·대응 체계 역시 이러한 양상 변화를 반영해야 함

(1) 위협 인텔리전스 및 이상행위 탐지 고도화

□ AI·머신러닝으로 정상 행위 패턴을 학습하고, 이에서 벗어나는 비정상 접근·행위를 탐지하는 기술로, AI가 생성·자동화한 공격 콘텐츠와 행위까지 함께 식별할 수 있도록 범위를 넓힘

ㅇ 기존 시그니처·규칙 기반 탐지는 알려진 공격 패턴만 탐지할 수 있어 새로운 공격이나 변형된 공격을 놓칠 수 있음

ㅇ 이상 행위 탐지를 적용하면 정상 행위에서 벗어난 접근과 행동까지 식별할 수 있음

ㅇ 생성형 AI·LLM으로 작성된 대량 맞춤형 피싱 메일, 딥페이크 음성·영상 등 합성 콘텐츠를 판별하는 탐지 모델을 함께 운용해야 함

- AI가 생성한 공격 콘텐츠는 문법·표현상의 어색함과 같은 기존 탐지 단서가 사라지므로, 발신자·통신 맥락의 이상 여부를 함께 분석하는 별도 체계가 필요

ㅇ 다크웹·해커 포럼 등에서 유통되는 WormGPT·FraudGPT류 악성 LLM 및 관련 공격 도구의 동향을 지속적으로 모니터링하여 새로운 공격 기법의 등장을 조기에 파악해야 함

(2) 공격 표면 관리 및 취약점 대응 자동화

□ 미토스급 고성능 AI의 등장으로 공개된 취약점이 실제 공격으로 전환되는 속도와 미공개 취약점의 발견 속도가 모두 빨라지고 있어, 방어자의 취약점 관리 체계도 이에 맞춰 고도화되어야 함

ㅇ 실제 악용이 확인된 취약점(KEV)을 중심으로 패치 적용 우선순위와 목표 소요시간(SLA)을 단축하고, 이행 여부를 지속적으로 점검해야 함

ㅇ 공격자가 AI를 활용해 취약점을 탐색·검증하는 것과 마찬가지로, 방어자도 AI 기반 자동 취약점 스캐닝과 모의침투(AI Red-teaming)를 상시적으로 수행하여 공개되지 않은 취약점을 선제적으로 발견·조치해야 함

- 앞서 살펴본 Project Glasswing과 같이, 고성능 AI 모델을 활용해 자체 기업망 및 내부 소프트웨어 등의 취약점을 상시 점검하는 체계를 갖추는 것도 하나의 방법이 될 수 있음

(3) SOAR 고도화

□ 보안 운영 자동화(SOAR) 시스템에 AI를 결합하여, 위협 탐지 시 사람의 개입 없이 차단·격리 등 대응 조치를 자동 수행

ㅇ 기존 수동 대응 체계는 탐지부터 조치까지의 지연 동안 피해가 확산했으나, 대응을 자동화함으로써 조치 지연을 최소화하고 반복 업무를 줄여 피해 확산을 방지하고 운영 효율을 향상시킬 수 있음

(4) 신뢰 채널 검증 강화

□ 딥페이크·합성 음성과 같이 AI로 위조된 콘텐츠가 사람의 신원 확인 수단으로 오인되는 것을 막기 위해, 고위험 업무에는 별도의 검증 절차를 두어야 함

ㅇ 화상·음성만으로 신원을 확인하던 기존 방식에서 벗어나, 자금 이체·권한 변경 등 고위험 업무는 사전에 합의된 별도 채널을 통한 재확인 절차를 의무화해야 함

ㅇ 임원 등 위조 시도가 우려되는 대상에 대해서는 음성·영상 워터마킹이나 생체 정보 기반의 실시간 진위 판별 기술 도입을 검토할 필요가 있음

다. 기술적 대응 방안 2 - AI에 대한 위협 대응

□ 3절에서 살펴본 LLM·에이전틱 AI·피지컬 AI 대상 공격은 AI 시스템 자체를 공격 표면으로 삼으므로, 모델·데이터·운영 인프라의 신뢰성을 직접 확보하는 대응이 필요함

(1) AI 시스템 보안 강화 기술

□ AI 시스템 보안 강화 기술은 AI 모델 자체의 안정성과 AI 운영 과정의 신뢰성을 확보하기 위한 대응 방안

□ 적대적 견고성(Adversarial Robustness) 확보

ㅇ AI 모델은 사람이 인지하기 어려운 미세한 입력 조작만으로도 오분류를 일으킬 수 있어, 자율주행·악성코드 탐지·의료 진단 등 핵심 영역에서 치명적 결과로 이어질 수 있음

ㅇ 따라서, 공격자의 적대적 입력으로 인한 AI의 오작동을 최대한 막기 위해 모델의 안정성과 견고성을 강화하는 적대적 견고성[57]을 확보해야 함

57) Ian J. Goodfellow, “Explaining and Harnessing Adversarial Examples”, (2015).

□ 데이터 무결성 검증 및 학습 파이프라인 보호

ㅇ AI는 학습 데이터에 크게 의존하므로, 데이터나 학습 파이프라인이 위·변조되면 모델에 편향·백도어가 은밀히 주입되고, 일단 학습에 반영되면 사후에 탐지·제거하기 어려움

ㅇ 오염된 데이터의 유입을 막기 위해 학습에 사용되는 데이터와 파이프라인의 위·변조 여부를 검증하고 보호하는 기술을 확보 필요

□ 설명 가능한 AI(Explainable Artificial Intelligence, XAI)과 감사 체계

ㅇ 딥러닝 모델은 판단 근거가 명확히 드러나지 않는 블랙박스 특성을 지녀, 보안 사고가 발생해도 그 원인을 규명하거나 책임 소재를 추적하기 어려움

ㅇ 따라서, AI 판단의 신뢰성과 사후 추적성을 확보하기 위해 AI의 의사결정 과정을 사람이 이해하고 추적할 수 있도록 하는 기술[58]과 감사 체계를 마련해야 함

(2) 제로트러스트 아키텍처 적용

□ 제로트러스트 원칙 기반[59] 보안 구조 적용: 기업 내부 네트워크에 있는 사용자나 시스템을 온전히 신뢰하지 않고, 모든 접근 요청을 지속적으로 검증

ㅇ AI 환경에서는 AI 에이전트가 자율적으로 판단·행동하며 시스템·데이터에 광범위하게 접근하는 비인간 주체로 등장함

58) DARPA, “Explainable Artificial Intelligence”, (2017).

59) NIST(각주 50).

ㅇ 이때 내부 진입 주체를 암묵적으로 신뢰하는 기존 경계 기반 보안으로는, 에이전트가 탈취·오남용될 경우 피해가 내부로 확산되는 것을 막기 어려움

ㅇ 따라서 제로트러스트 보안 구조를 적용하여 모든 리소스 접근 뿐만 아니라 AI 모델·에이전트까지도 개별 신원 및 인증체계 부여 후 비신뢰 기반으로 추론 결과·자율적 행위까지도 지속적으로 확인하고, 필요한 최소 권한만 허용함으로써 내부 오남용과 피해 확산을 방지해야 함

(3) 공급망 보안 및 AI-BOM 활용

□ AI-BOM 기반 공급망 가시성 확보: AI 시스템은 외부 사전학습 모델·오픈소스 라이브러리·외부 학습 데이터 등 다양한 구성요소에 의존하므로, 공급망 전반의 출처와 무결성을 검증하고 AI-BOM을 활용하여 가시성을 확보해야 함

ㅇ 평소에는 구성요소 현황을 체계적으로 관리하고, 공급망 공격 발생 시에는 AI-BOM을 기반으로 영향 범위를 신속히 추적·대응할 수 있음

ㅇ 특히, 에이전틱 AI의 경우 정적 AI 모델을 가정한 AI-BOM으로는 실행 중에 참조하는 외부 API 및 에이전트, 내부 메모리 등의 동적 변화를 감지하기 어려울 수 있으므로, AI-BOM의 관리 범위와 구성요소에 대해 신중한 접근 필요

(4) 피지컬 AI 물리적 안전성 확보

□ AI의 오판단·오작동이 사이버 공간에 그치지 않고 현실 세계의 인명·시설물 피해로 직결되는 것을 막기 위한 대응 방안 도입

- ㅇ 센서 다중화 및 정보 교차 검증: 단일 센서에 대한 조작·오인식이 AI의 인지 결과 전체를 왜곡하지 않도록, 서로 다른 방식의 센서 정보를 상호 교차 검증하여 단일 지점 오류로 인한 오작동을 줄여야 함
- ㅇ 안전 정지(Fail-safe) 및 제어 명령 무결성 검증: 불확실성이 높거나 이상이 감지될 경우 즉시 안전한 상태로 전환하고, 센서·AI 모델과 구동기 사이의 제어 명령이 위변조되지 않도록 통신 구간의 인증·무결성 검증 체계를 마련해야 함
- ㅇ 사이버-물리 계층 분리: AI의 인지·의사결정 계층과 실제 물리적 동작을 수행하는 구동 계층 사이에 인간 개입·감사 등이 가능한 독립적인 안전 장치를 두어, AI 계층이 침해되더라도 위험한 물리적 동작으로 곧바로 이어지지 않는 구조 구현

라. 공통 운영 방안: 인간-AI 협력(Human-AI Teaming) 보안 운영

□ 앞서 살펴본 기술적 대응 방안이 조직 내에서 실제로 운용되는 방식을 규정하는 원칙으로, 두 영역 전반에 걸쳐 적용됨

- ㅇ AI는 대규모 로그 분석·이상 행위 탐지에 강점이 있으나, 오탐과 맥락적 판단에 한계가 있어 완전 자동화에만 의존할 경우 잘못된 차단으로 정상 업무를 마비시킬 위험이 있음
- ㅇ 따라서, AI의 처리 능력과 인간의 판단력을 결합해 AI는 대규모 분석과 이상 징후 탐지를 수행하고 인간은 최종 판단과 정책 결정을 담당하는 협력형 운영 체계를 구축해야 함
- ㅇ 이를 통해 탐지 속도와 판단 정확성을 동시에 확보하고, AI 단독 운영의 오탐 위험과 인간 단독 운영의 처리 한계를 상호 보완할 수 있음

마. 정책적 시사점

□ 앞서 제시한 방안들을 정리하면 [표 13]과 같음

[표 13] 기술적 대응 방안 요약

구분	세부 방안	대응 방향	주요 내용
AI에 의한 위협 대응	위협 인텔리전스 및 이상행위 탐지 고도화	• AI를 방어 수단으로 활용	• 정상 학습 기반 이상행위 탐지에 더해 딥페이크·대량 피싱 등 AI 생성 콘텐츠 판별 모델 및 악성 LLM 유통 동향 모니터링을 병행
	공격 표면 관리 및 취약점 대응 자동화	• 패치·취약점 관리 속도 제고	• KEV 중심 패치 SLA 단축과 AI 기반 자율 취약점 스캐닝·모의침투로 취약점 무기화 속도에 대응
	SOAR 고도화	• 대응 조치 자동화	• 탐지 후 차단·격리 등 대응 조치를 자동 수행해 조치 지연을 최소화
	신뢰 채널 검증 강화	• AI 위조 콘텐츠에 대한 인증 보완	• 고위험 업무에 대역 외 재확인 절차를 의무화하고 워터마킹 등 진위 판별 기술 도입 검토
AI에 대한 위협 대응	AI 시스템 보안 강화	• 모델과 학습 과정의 신뢰성 확보	• 적대적 입력 대응, 데이터 무결성 검증, 학습 파이프라인 보호, 설명 가능성과 감사 체계 확보
	제로트러스트 적용	• 모든 주체를 지속적으로 검증	• AI 에이전트와 사용자에게 최소 권한을 부여하고 접근 행위를 지속적으로 확인
	공급망 보안 및 AI-BOM 활용	• AI 구성요소 출처와 의존성 관리	• 외부 모델, 데이터, 라이브러리 등 AI 구성요소의 무결성을 검증하고 AI-BOM으로 가시성 확보
	피지컬 AI 물리적 안전성 확보	• 현실 세계의 인명 및 물리적 피해로 직결되지 않도록 대응	• 안전 정지(Fail-Safe) 및 명령 무결성 검증, 인간 개입 및 감사 가능한 안전 구조 확보
공통 운영 원칙	인간-AI 협력 보안 운영	• AI 자동화와 인간 판단 결합	• AI는 대규모 분석과 이상 징후 탐지를 수행하고 인간은 최종 판단과 정책 결정을 담당

- ㅇ AI 기반 위협 탐지 및 대응은 AI를 방어 수단으로 활용하는 방안인 반면, AI 시스템 보안 강화 · 제로트러스트 적용 · 공급망 보안·피지컬 AI 물리적 안전성 확보는 AI 자체를 보호 대상으로 하는 방안에 해당
- ㅇ 인간-AI 협력 보안 운영은 위 기술들이 실제 조직에서 운영될 때의 의사결정 구조를 규정하는 방안으로, 나머지 방안 전반에 걸쳐 적용
- ㅇ 이와 같이 각 방안은 방어의 대상과 통제 지점이 서로 달라, 개별 기술의 도입만으로는 AI 환경의 위협에 충분히 대응하기 어려움

□ 그러나 이러한 기술적 대응 방안을 조직이 어느 수준까지, 어떤 우선순위로 도입할 것인지는 기술적 판단만으로 결정되지 않으며, 최소 보안 요건이나 인증 · 평가 기준과 같은 정책적 지침이 뒷받침되어야 실효성을 가질 수 있음

□ 특히 에이전틱 AI의 자율적 동작에 의하여 사이버보안 사고가 발생했을 때, 그 책임을 누구에게 어느 범위까지 귀속시킬 것인지는 기술이 아니라 법 · 제도의 영역에서 정해져야 할 문제임

□ 또한 공급망 보안이나 AI-BOM과 같이 조직 · 국경을 넘어 형성되는 위협 요소는 개별 기관의 자율적 조치만으로 통제하기 어려우므로, 위협정보를 공유하고 대응을 조율할 수 있는 신고 · 공유 체계 및 국제 공조 채널이 뒷받침되어야 함

□ 아울러 이상행위 탐지, 제로트러스트 전환 등 새로운 보안 기술을 조직 내에 실제로 정착시키기 위해서는 이를 운용할 전문인력의 양성과 예산 · 거버넌스 정비가 병행되어야 하며, 이는 개별 조직의 자체 역량만으로는 한계가 있음

□ 따라서 AI 환경의 위협에 실효적으로 대응하기 위해서는 앞서 살펴본 보안 기술의 고도화와 함께, 책임 체계의 정립, 위협정보 공유 · 신고 체계 구축, 보안 인증 · 평가 체계 마련, 국제 협력 등 이를 뒷받침하는 정책적 기반이 함께 마련되어야 함

5. 정리

□ 프론티어 AI(미토스 등)가 현실화된 상황에서 기존 사이버보안 모델의 한계는 다음과 같음

[그림 10] AI 환경에서 기존 사이버보안 모델의 한계

□ AI發 사이버보안 위협에 대응하기 위한 기술적 방안은 다음과 같음

[그림 11] AI 사이버보안 위협에 대한 기술적 대응 방안

□ 이러한 기술적 방안을 체계적, 지속적으로 추진하기 위해서는 정책적 기반이 필수적으로 요구되는바, 이에 관하여 장을 바꾸어 살펴봄

Ⅲ. 주요국 관련 입법·정책

□ 이 장에서는 AI로 인한 새로운 보안 위협과 관련된 해외 주요국의 입법 및 정책을 ’26. 8. 기준으로 살펴봄

ㅇ 대상 국가: 미국, 유럽연합(EU), 영국, 일본

□ 전술하였듯이 AI 보안 위협은 하루가 다르게 그 위협의 수단, 방법, 정도, 양상 등이 급변하고 있으나 법제도는 본질적으로 환경 변화에 신속, 탄력적으로 대응하기 어렵다는 한계 노정

□ 따라서 본 장에서 검토하는 사례는 AI 보안 특화 법제나 정책으로 한정되지 않으며, 사이버보안 전반에 관한 기존 법제나 정책 중 AI 보안과 관련된 부분을 포함함에 유의

1. 미국

가. 특징

□ 민간 영역의 규제는 완화하되, 국가 중요 인프라나 연방 정부 기관 등 공적 영역의 통제는 공법적으로 강제하는 “실용적 규제·선제적 예방 및 복원력” 체계

ㅇ 규제 완화: 첨단 AI 모델에 대한 사전 인허가 금지, 민간이 사이버 위협 정보 공유 시 법적 면책(safe harbor) 명문화 등

ㅇ 공적 영역 통제 강화: 16개 중요 인프라 대상 중대 사이버 사고(72시간) 및 랜섬웨어 대가 지불(24시간)에 대한 엄격한 법정 보고 의무화, 연방 조달 SW에 대한 제로트러스트 도입, SBOM 제출 의무화 등 강력한 공급망 보안 컴플라이언스 요구 등

○ 안보 중심의 기술 통제 및 형사 집행: 적대국 배후 위협에 대한 공세적 제재와 AI 악용 침해 행위에 대한 연방 형법의 엄격한 집행 명문화

나. 개요

□ 미국은 연방제 국가라는 헌법적 구조상 주(state)와 연방(federal)의 권한이 명확히 구분되어 있으며, 전통적으로 공공/민간을 분리하여 공공 부문(특히 연방정부)에는 공법 중심의 엄격한 규율을, 민간 부분은 사법 중심의 자율적 규율을 존중해 왔음

□ 최근 미국은 APT 조직의 기반시설 인프라 위협과 공급망 교란에 대응하기 위해 '사후적 · 파편적 방어'에서 '선제적 예방과 통합적 복원력(Resilience) 확보'로 사이버보안 패러다임을 전환하는 중

○ 이에 따라 민간 영역에도 공법적 컴플라이언스 체계(중요인프라에 발생한 사이버 사고 보고 보고 체계 등)를 도입하고, 연방 차원의 성문법 부재의 한계를 대통령 행정명령과 각종 지침 등을 유기적으로 결합한 연성 규범으로 보완하고 있음

□ 특히 트럼프 정부가 '25. 7. 발표한 AI 실행계획(AI Action Plan),[60] '26. 3. 발표한 사이버 전략(President Trump's Cyber Security Strategy for America),[61] '26. 6. 발표한 '첨단 AI 혁신 및 안보 촉진 명령'[62] 등에 따라 사이버보안 법제 또한 이에 부합하는 방향으로 개편될 여지가 있으므로, 지속적인 주목 필요

60) https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

61) https://www.whitehouse.gov/wp-content/uploads/2026/03/president-trumps-cyber-strategy-for-america.pdf

62) 행정명령 제14409호, https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/

ㅇ ‘AI 실행계획’은 (1) AI 혁신 보호, (2) AI 인프라 보호, (3) AI 외교 및 안보 연대를 3대 축(Pillar)으로 제시하고, 핵심 인프라 사이버보안 강화, Secure-by-Design, 사이버사고 대응 능력 강화 등을 구체적으로 규정

- 특히 AI 실행계획은 국토안보부(Department of Homeland Security, ‘DHS’) 주도로 AI 정보공유분석센터(AI-ISAC)을 설립하여 AI 보안 위협 정보를 민관이 공유하도록 명시하고 있음

ㅇ ‘사이버 전략’은 (1) 공세적 안보 및 보복 (2) 규제 부담 완화 (3) AI 기술 패권 선점 (4) 공급망 내 적대국 제품 추출 등을 핵심 축(pillar)으로 제시

ㅇ ‘첨단 AI 혁신 및 안보 촉진 명령’은 (1) 첨단 AI 모델 규제 완화, (2) 주요 인프라 및 정부 전산망의 AI 보안 고도화, (3) 범정부 AI 사이버보안 정보 공유소(clearinghouse) 설립 등을 명령

다. 법률

(1) 정보보안 현대화법(FISMA 2014)[63]

□ 배경

ㅇ 미국은 ’02 제정된 정보보안 관리법(Federal Information Security Management Act)에서 연방정부의 정보보안 추진체계 및 거버넌스에 관하여 규율하여 왔음

- FISMA(2002)는 모든 연방기관이 최소한의 기본적인 사이버보안 조치를 취할 책임과 그 평가에 관한 사항 등을 규정하고, 국립표준기술연구소(NIST)를 관련 보안 가이드라인 및 지침 개발 기관으로 지정

63) Pub. L. No. 113-246, 128 Stat. 2880.

○ 이후 연방기관에 대한 사이버위협이 급격히 증대하였으나 각 기관들은 이에 효과적으로 대처하지 못하는 한계가 노출되었고, 이에 효과적으로 대응하기 위하여 전체적인 거버넌스 조정이 요구됨

□ 주요 내용

○ 백악관 행정예산관리국(OMB)에 있었던 연방 기관의 정보보안 감독 권한을 재설정하여 DHS 중심의 거버넌스로 개편

- OMB는 연방 기관의 정보보안 정책 및 실무를 '감독'할 책임을 가지나, 연방기관의 정보보안 정책 및 실무의 '이행' 책임은 DHS에게 부여됨[64]

- DHS 장관은 위 이행을 위해 구속력있는 운영상 지침(binding operational directive)을* 내릴 수 있고 그 이행을 감독할 수 있음

  * 이미 알려져 있거나 합리적으로 예측할 수 있는 정보보안 위협, 취약성, 위험 등으로부터 연방정보 및 정보시스템을 보호하기 위하여 연방기관에 내리는 강제적 지시

○ NIST에 연방 정보 시스템의 취약점 성격과 범위를 판단하고 비용 효과적인 보안 기술을 제공하기 위한 연구 수행 의무 부여[65]

- NIST가 사이버보안 취약점 데이터베이스(National Vulnerability Database, 'NVD')를 운영하는 근거로 기능[66]

64) 44 U.S.C. § 3553.

65) 15 U.S.C. § 278g-3.

66) https://nvd.nist.gov/ . 엄밀히 말하여 FISMA는 연방 정부망의 보안 측정 기준과 취약점 연구 표준을 개발할 법적 의무를 NIST에 부여한 것이고, 구체적으로 NIST에게 HW와 SW의 보안 리스크를 최소화할 수 있는 국가 체크리스트 포털을 구축할 의무를 명시한 법은 사이버보안 연구개발법(Cyber Security

□ 참고 사항

○ ’26. 4. NIST는 NVD 축소를 공식 발표[67]

- ’26. 4. 이후 NIST가 보강(enrich, NIST가 세부 정보 및 대응 방안 등을 상세히 분석하는 것을 말함)하는 사이버보안 취약점 및 노출정보(CVE; Cybersecurity Vulnerabilities and Exposures)는 ① 실제 악용 확인 취약점(KEV; Known Exploited Vulnerability), ② 미국 연방정부가 사용중인 SW 관련 사항, ③ 행정명령 제14028호에서 규정한 Critical SW(OS 등)관련 사항의 세 가지로 한정됨

- 그 외의 CVE는 목록화(list-up)은 하지만 ‘가장 낮은 우선순위’로 분류되며 NIST가 즉시 보강하지 않음

○ NIST NVD는 우리나라 대부분 기관의 취약점 관리 기준선; NIST NVD의 보강 범위가 축소되면서 우리나라 역시 사이버보안 취약점 파악 자체가 어려운 문제가 발생(즉, 독자적 NVCD 관리 필요성 대두)

(2) 사이버보안 강화법(CEA 2014)

□ 배경

○ CEA 2014는 FISMA 2014와 동시에 제정된 법으로 그 배경은 FISMA 2014와 동일함

□ 주요 내용

○ CEA 2014는 크게 사이버보안의 (1) 민관협력 (2) 연구개발 (3) 인력개발 (4) 인식 및 준비 (5) 기술적 기준의 향상을 규정하는데, 이 중 가장 주목할 부분은 (1) 민관협력임

R&D Act of 2002, 15 U.S.C. § 7403)이다.

67) https://www.nist.gov/itl/nvd

ㅇ 민관협력 관련, CEA 2014는 NIST에게 사이버위험을 비용 대비 효율성이 높은 방법으로 감소시킬 수 있는 “자발적이고, 업계에 의해 유도되며, 합의에 기초를 두고 있는” 일련의 기준, 가이드라인, 모범 사례(Best Practice)의 개발을 촉진하고 지원할 수 있도록 규정[68]

- 이는 NIST가 개발하는 사이버보안 가이드라인, 지침이 사실상 구속력(de facto standard)을 갖게 되는 법적 근거로 작동

(3) 사이버보안 정보공유법(CISA 2015)

□ 배경

ㅇ ’15 인사관리처(OPM) 해킹사고, 국세청 해킹사고, 보험회사 랜섬웨어 해킹사고 등 공공/민간 불문 지속적인 사이버보안 침해사고가 발생

ㅇ 이러한 침해사고의 방지 및 신속 대응을 위해서는 국가 전반의 사이버보안 수준을 강화하기 위하여 사이버보안 정보공유의 실효성을 강화하고 민관 정보공유를 확대해야 한다는 여론 조성

□ 주요 내용

ㅇ 연방기관과 비연방기관(주정부, 지방정부, 민간기관 포함) 사이의 사이버위협 정보 공유 체계 마련

- 국가정보국(DNI), DHS, 국방부, 법무부는 연방기관-비연방기관 사이에 사이버위협 정보를 공유할 수 있는 절차를 구축하고 가이드라인 마련[69]

- 민간기관이 사이버보안 목적으로 정보시스템 및 정보를 모니터링하고 방어조치를 취하여 관련 정보를 공유할 수 있는 법적 근거 마련[70]* **

68) 15 U.S.C. § 272(c)(15). 상세한 내용은 양천수・지유미, “미국 사이버보안법의 최근 동향”, 법제연구 제54호, 한국법제연구원(2018. 6.), 165쪽.

69) 6 U.S.C. § 1504(a).

* 민간기관은 해당 정보가 (1) 특정 개인의 개인정보에 해당한다는 사실을 알았고 그 정보가 사이버보안 위협과 직접 관련된 것이 아니라면 해당 정보를 공유 대상에서 제외하여야 하며 (2) 해당 정보가 사이버보안 위협 감지/예방/경감에 필요한 것이 아니라면 역시 공유 대상에서 제외하여야 함. 가이드라인은[71] 건강관련 정보, 민감한 금융정보를 대표적 공유 대상 제외 정보로 규정

** 민간기관의 정보 공유는 자발적 공유임. 즉 의무사항이 아님[72]

- 이에 따라 DHS에 '사이버보안 · 기반시설안보국(NPPD)'이 설치되었는데, NPPD는 '18 법 개정으로 독립 연방기구인 CISA로 발전하였고,[73] CISA 산하에 국가사이버보안 · 통신통합센터(NCCIC)가 설치됨

ㅇ 특히 민간기관의 정보 공유에 관하여 법적 책임을 일정 부분 면책하는 안전조항(Safe Harbor) 명문화[74]

- 민간이 연방 정부에 제공한 사이버위협 지표 또는 방어 조치는 오직 사이버보안 목적 또는 구체적인 범죄 예방, 수사 또는 기소 목적으로만* 연방정부가 공개, 보유 및 활용해야 함[75]

* 연방 법집행기구(법무부 등)은 위 정보를 사이버보안 위협 관련 범죄, 사망이나 중대한 신체적 위해의 급박한 위험을 수반하는 범죄, 중대강력범죄, 스파이 행위 및 영업비밀 보호 관련 범죄의 수사 및 기소에만 사용할 수 있음

---

70) 6 U.S.C. § 1504(a).

71) https://www.cisa.gov/resources-tools/resources/guidance-assist-non-federal-entities-share-cyber-threat-indicators-and-defensive-measures-federal

72) 6 U.S.C. § 1505(c)(1)(A).

73) 6 U.S.C. § 652.

74) 6 U.S.C. § 1504, 1505.

75) 6 U.S.C. § 1504(d)(1)

- 민간이 동법에 따라 연방 정부에 제공한 사이버 위협 및 방어 조치는 연방/주/지방정부가 규제하기 위하여 이용할 수 없음; 즉 민간이 “이러저러한 경로로 사이버공격을 당하여 뚫렸다”라고 정보공유한 경우 규제 당국(연방/주/지방 불문)은 해당 정보를 가지고 민간기관에 대한 제재 처분을 할 수 없고, 만약 제재 처분을 하려면 별도의 조사를 거쳐 새롭게 증거를 확보해야 함[76]

- 민간이 동법에 따라 연방 정부에 제공한 사이버위협 지표 및 방어 조치는 정보공개법(FOIA)에 따른 공개 대상에서 제외됨

- 민간이 동법에 따라 사이버위협 지표 또는 방어 조치를 공유하거나 수령한 행위에 관하여 누구도 연방 또는 주 법원에 민간기관을 상대로 어떠한 소송도 제기할 수 없음[77]

  * 예컨대 정보주체의 개인정보침해 소송, NDA 상대방의 NDA 위반 소송, 주주대표소송 등은 전부 각하됨

- 복수의 민간 기관이 동 법에 따라 사이버보안 목적으로 사이버위협 지표나 방어 조치를 공유하는 행위는 독점금지법 위반으로 간주되지 않음

76) 6 U.S.C. § 1504(d)(2); Dep’t of Homeland Sec. & Dep’t of Justice, Guidance to Assist Non-Federal Entities to Share Cyber Threat Indicators and Defensive Measures with Federal Entities under the Cybersecurity Information Sharing Act of 2015 (Feb. 2026), at 35-36 (FAQ. 15). 유일한 예외는 오직 사이버보안 위협을 예방하거나 완화하는 목적의 ‘법령・규제(Regulations)’를 개발하거나 시행할 때 참고 자료로 활용하는 경우뿐이다. Id., at 36.

77) 6 U.S.C. § 1505(a), (b)

□ 참고 사항

○ CISA는 위 법에 따라 실시간으로 사이버보안 위협 정보를 교환하는 자동지표공유시스템(Automated Indicator Sharing, ‘AIS’)을 구축하여 운영 중[78] * 우리의 C-TAS와 대비됨

- 정부기관이나 민간기관이 사이버 위협 지표(CTI)를 AIS에 입력하면, 기계가 읽을 수 있는(Machine-readable) 형태로 참여기관에 실시간 자동 전파됨

- NVD가 알려진 취약점을 집대성한 정적(靜的) 아카이브라면, AIS는 실시간으로 동작하는 동적(動的) 취약점 공유 시스템

○ CISA 2015는 원래 ’25. 9. 로 일몰 예정이었고, 실제로 ’25. 9. 연장되지 않으면서 safe harbor 근거가 사라져 민간 기업들이 사이버위협 정보 공유를 기피하는 문제가 발생

- 사이버보안 공백을 우려한 미국 정부는 2026년 종합예산법을 패키지로 통과시키면서 CISA 2015의 효력을 ’26. 9.까지 1년 연장하였음

(4) 주요기반시설 사이버사고 보고법(CIRCIA 2022[79])

□ 배경

○ ’20~’21 연방정부기관, 에너지 인프라, 식품 등 미국의 국가 기능 자체를 일시 마비시킨 초대형 사이버 침해 사고 발생

78) https://www.cisa.gov/resources-tools/services/automated-indicator-sharing-ais-service

79) 6 U.S.C. Ch. 7 SubCh. XVIII Part D. art. 681~681g. 정식 명칭은 Cyber Incident Reporting for Criticail Infrastructures Act of 2022 이다.

- SolarWind 사건(APT 배후의 국방부, 국무부, 포춘500대 기업 공급망 소프트웨어 오염 사건), Colonial Pipeline 랜섬웨어 공격(미 동부 석유 공급 인프라 마비 사건), JBS 뉴트리션 공격(세계 최대 육가공 업체 전산망 마비 사건) 등

○ CISA 2015 등 기존 법제의 '자발적 정보 공유'만으로는 국가 기간망 방어가 불가능한 한계 노출

- 민간 기업이 해킹을 당해도 기업 이미지 실추를 우려해 은폐할 경우 연방정부는 사고 발생 여부조차 즉시 파악하기 어려운 문제

□ 주요 내용

○ 국가 안보, 경제 발전, 공공 안전에 필수적인 16개 중요 인프라(화학, 통신, 중요 제조업, 댐, 국방 산업, 응급 서비스, 에너지, 금융 등)를 지정하고, 이를 운영하는 중요 엔티티(Critical Entities)에게 특정 사이버 사고 발생시 CISA에 법정 시한 내 보고 의무 부과[80]

- 중대 사이버 사고* 발생시: 72시간 내 보고

  * 시스템 마비, 대규모 데이터 유출, 공급망 오염 등

- 랜섬웨어 대가 지불시: 24시간 이내 보고

○ Safe Harbor 유지

- CISA 2015 철학을 계승, 중요 엔티티가 CISA에 제출한 정보나 보고서에 대해서는 (1) 정보공개법 적용 제외, (2) 소송상 증거 채택이나 증거개시(discovery) 대상 제외 (3) 소송 제기나 책임 추정 금지 (4) 규제 목적으로 이용 금지를 명문으로 규정[81]

80) 6 U.S.C. § 681b.
81) 6 U.S.C. § 681e.

라. 행정규칙, 가이드라인 등

(1) NIST AI 위험관리 프레임워크(AI RMF)[82] [83]

□ 배경

○ ’20 제정된 국가인공지능 이니셔티브법에* 따른 미 의회의 지시로 NIST가 ’23 수립한 일종의 가이드라인

* 공공/민간 영역에서 신뢰할 수 있는 AI 시스템 개발 등을 위하 국가 차원의 AI 이니셔티브를 수립, 이행하는 법률[84]

○ 공공/민간 불문 모든 분야·규모의 기업 조직이 AI 위험을 해결할 수 있도록 유연하고 체계적이며 측정 가능한 프로세스를 제공하여 AI 이점을 극대화하고 AI가 개인/그룹/조직/사회에 부정적 영향을 미칠 가능성을 최소화하기 위하여 수립

□ 주요 내용

○ AI 시스템의 전 주기에서 신뢰할 수 있는 AI 시스템 구현을 위한 위험관리 프로파일을 제시

- (적용대상) 민/관 불문 모든 분야·규모의 기업·조직의 AI 시스템
- (적용주체 및 방식) AI 행위자가* 자발적으로 참고

* 설계자, 개발자, 배포자, 사용자, 고위 경영진, 관료 등

---

82) NIST, NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence(2023. 1. 26.).

83) 한국지능정보사회진흥원, “美 NIST 「AI 위험관리 프레임워크(AI RMF) 1.0」 분석 및 시사점”, The AI Report(2023.).에 기초하여 정리한 것이다.

84) P.L. 116-283.

- (적용시점) AI의 설계, 개발, 활용, 테스트, 평가 등 모든 단계
- (핵심내용) 안전성, 책임성, 투명성 등 신뢰할 수 있는 AI 시스템의 7가지 특성을 제시하고, 이를 위한 조직의 핵심 기능을 4가지 영역(거버넌스, 위험 식별, 측정 관리)로 제시

□ 참고 사항

○ AI RMF 자체는 구속력없는 자발적 가이드라인이나, 미국과 글로벌 시장에서는 사실상 표준(de facto standard)으로 작동

- 대통령 행정명령 및 연방조달지침에 따라, 연방 정부 기관에 AI 시스템을 납품하려는 자는 해당 제품이 AI RMF를 준수했음을 증명해야 함*

  * AI 명세서(AI-BOM) 및 위험 평가 보고서를 첨부하는 등의 방식

- AI RMF 준수 사실은 AI 시스템의 오작동으로 인한 손해배상소송 등에서 해당 기업이 주의의무를 다하였음을 증명하는 유력한 증거로 사용될 것으로 전망

○ NIST는 ’26. 4. AI RMF를 중요 인프라(Critical Infrastructure) 영역에 맞춤형 특화하여 적용할 수 있는 프로파일의 개발 계획과 방향성을 발표하여 향후 귀추가 주목됨

- AI RMF는 규모, 대상을 불문한 범용적 프레임워크이나, 물리적 위험과 인간의 생명이 달려있는 고위험 영역에는 AI 시스템이 철저히 신뢰할 수 있어야만 채택 가능
- 중요 인프라 분야에서 요구되는 결정론적 행위, 설명가능성, fale-safe 등의 가치를 AI RMF에 이식하고, 전주기에 걸친 견고성과 엄격한 시험/평가/검증/확인(TEVV) 프로세스 표준을 제시할 예정

(2) 국가 사이버보안 개선 명령(행정명령 제14028호[85])

□ 배경

○ ’20~’21 적대국과 연계 추정되는 APT의 SW 공급망 공격(SolarWind) 및 에너지 인프라(Colonial Pipeline) 랜섬웨어 공격이 발생하자 바이든 정부는 사이버보안 강화를 위하여 행정명령 발표

□ 주요 내용

○ 연방 정부에 제로트러스트(zero trust) 원칙을 도입하고, 연방 정부에 SW를 납품하는 기업에게 구성요소 명세서(SBOM) 제출을 의무화하는 등 소프트웨어 공급망 보안의 표준을 확립

- 연방 정부의 보안 패러다임을 제로트러스트 아키텍쳐로 전면 전환
- 연방 정부에 납품되는 모든 SW에 SBOM 제출 의무화
- 연방 정부와 계약한 IT 서비스 제공업체가 계약상 비밀유지 조항(NDA)를 이유로 사이버침해 사고를 숨길 수 없도록 하고, 사이버침해 사고 발생시 CISA 등에 의무적으로 즉시 보고토록 함
- 초대형 사이버 사고 발생시 공공·민간이 함께 원인 조사하고 보고서를 작성하는 사이버안전 검토 위원회(Cyber Safety Review Board) 설치
- 사이버 사고 발생시 사후 조사와 역추적이 가능하도록 연방 정부 기관 및 연계 시스템의 로그 수집·보관 기준 강화

85) https://www.federalregister.gov/documents/2021/05/17/2021-10460/improving-the-nations-cybersecurity

(3) 국가 사이버보안 노력 지속 및 기존 명령 개정(행정명령 제14306호[86])

□ 배경

○ ’25. 1. 출범한 트럼프 정부는 (1) 기존 민주당 정부의 정책을 철회하는 한편 (2) APT 볼트 타이푼(Volt Typhoon)의 통신사 해킹 사건과 같이 국가 중요 인프라를 겨냥한 사이버위협이 계속되자 적대국에 대한 견제 필요성을 제기하며 기존 사이버보안 정책 방향과 기조를 수정하는 행정명령 발표

□ 주요 내용

○ 오바마/바이든 정부의 사이버보안 및 AI 정책의 초점을 엄격한 규제에서 규제 완화 및 혁신 촉진으로 전환

- 주요 위협 행위자를 중국에서 러시아, 이란, 북한까지 확장하였으나, 행정명령 위반에 따른 제재(자산 동결, 금융망 차단 등)은 외국인(외국의 APT 그룹, 외국 정부 배후 해커 등)에게만 국한됨을 명시

- 연방기관에 IoT 제품을 공급하는 업체는 ‘미국 사이버 트러스트 마크’ 라벨을 부착하도록 연방조달규정 개정

- 연방 SW 공급업체에 요구하였던 안전한 소프트웨어 개발 서약서(Attestations) 및 정밀한 검증 서류 제출 의무 대폭 간소화

- AI 사이버보안 정책을 지나친 규제에서 탈피, 취약점 탐지 및 관리, 보안 데이터셋 공개 등 실용적 위협 대응 체계로 전환

86) https://www.federalregister.gov/documents/2025/06/11/2025-10804/sustaining-select-efforts-to-strengthen-the-nations-cybersecurity-and-amending-executive-order-13694

(4) 첨단 AI 혁신 및 안보 촉진 명령(행정명령 제14409호[87])

□ 배경

○ 트럼프 정부는 생성형을 넘어선 '추론형·에이전트형 AI'의 급부상으로 사이버 위협이 초고도화된 상황에서, 기존 규범들이 민간의 AI 혁신 속도를 가로막고 국가 안보 경쟁력을 저해한다고 판단

○ 이에 따라 AI 규제 완화와 민·관 자발적 협력 기반의 사이버 방어력 결합을 통해 글로벌 AI 주도권 및 국가 안보를 동시 확보 도모

□ 주요 내용

○ '첨단 AI 모델(Covered Frontier Model)'에 대한 정부의 인허가, 승인, 허가 등의 사전 규제나 검증을 명시적으로 금지하여 첨단 AI 모델 개발 및 상용화 촉진

- 첨단 AI 모델 해당 여부는 단순히 연산량 등으로 판단하는 것이 아니라, 고도화된 사이버 역량을 측정하는 기밀 벤치마킹 프로세스를 통해 NSA 국장이 최종 판단*

  * 민간은 개발 단계에서 첨단 AI 모델에 해당하는지를 정부와 자발적으로 사전 협의하여 판단을 요청할 수 있음

- 민간은 자발적으로 정부에 첨단 AI 모델에 관하여 최대 30일간 사전 접근(early access)를 허용할 수 있음; 이 기간 동안 정부와 민간 합동으로 해당 모델의 취약점을 점검하거나 보안 패치 등을 준비하게 됨

87) 행정명령 제14409호, https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/

○ 국방부 및 국가안보시스템 전산망 사이버 방어를 최우선 과제로 지정

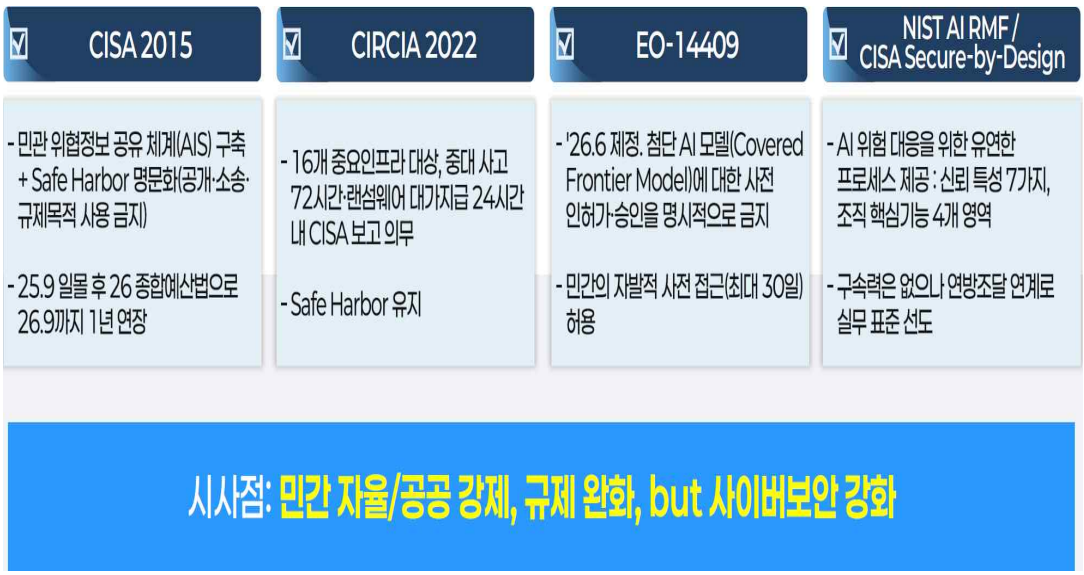
○ 범정부 정보 공유체계 설립

- 재무부, NSA, CISA가 AI 업계 및 주요 인프라 운영자와 협력하는 민관 자발적 정보공유 체계(clearinghouse)를 신설

○ AI 또는 AI agent를 악용한 무단 침입이나 데이터 탈취에 따른 2차 범죄(사기 등)에 대하여 연방 형법[88] 강력 집행

마. 시사점

[그림 12] 미국의 관련 법제도 현황 및 시사점



□ 민간 AI 혁신 보장과 국가 안보 중심 방어의 정교한 조화

○ 첨단 AI 모델(Covered Frontier Model)에 대한 사전 인허가나 강제적인가 배제를 행정명령 제14409호로 명시하여 민간의 AI 기술 혁신 속도를 보장함

88) 18 U.S.C. § 1028, 1030 등.

ㅇ 국방 및 국가안보시스템 전산망 보안을 최우선 과제로 지정하고 자발적 사전 접근(Early Access, 최대 30일)을 통해 개발 단계부터 취약점을 점검 · 보완함

□ 면책 장치(Safe Harbor)를 통한 민간 위협 정보의 공유 유도

ㅇ CISA 2015 및 CIRCIA 2022를 통해 민간이 사이버 위협 지표나 사고 보고서를 제출할 때 민사소송 면책, 정보공개법(FOIA) 적용 제외, 규제 목적 유용 금지 등 명확한 안전조항 제공

ㅇ 민간의 사고 은폐 위험을 줄이고 민간과 정부 간 자동지표공유시스템(AIS)을 통한 실시간 정보 전파 체계를 활성화함

□ 제로트러스트 도입 및 자발적 표준(de facto standard)과 연방 조달 연계

ㅇ 행정명령 제14028호를 통해 연방 전산망에 제로트러스트(Zero Trust) 아키텍처 및 SBOM 제출을 전면 도입

ㅇ 구속력 없는 가이드라인인 NIST AI RMF를 연방 조달 지침 및 시장의 사실상 표준으로 안착시켜 산업 전반의 보안 수준을 자율적으로 상향시킴

2. EU

가. 특징

□ 초국경적 위협에 맞서 27개 전 회원국에 통용되는 높은 수준의 단일 규범을 구축하고, 조직·제품의 복원력과 정보공유를 강화하는 한편, 단일 창구화(규제 완화) 및 AI 주권 투자를 병행하는 체계

- ㅇ 27개 회원국 대상 단일한 고수준 공통 규범 체계 확립: 사이버 위협의 초국경성에 대응하여 전 회원국에 통일적으로 적용되는 강력한 성문 규율 기준을 수립함
- ㅇ 제품/서비스 전 수명주기를 고려한 복원력(Resilience) 확보 및 초국가적 정보공유 강화: Secure by Design 의무화, 주요 조직의 회복탄력성 확보와 더불어, 회원국 내부 및 회원국간 실시간 위협 정보 공유와 위기 대응 공조를 제도화
- ㅇ 규제 완화와 AI 주권 확보의 병행: “디지털 옴니버스 패키지” 입법으로 불필요한 규제를 완화하는 한편, 자체 안전 테스트 플랫폼 구축 및 AI 인프라 투자 등을 추진하여 AI 주권 확보 시도

나. 개요

□ EU는 ’19-’24 차기 유럽집행위원회 정치적 가이드라인에서[89] 향후

89) European Commission: Directorate-General for Communication, Political guidelines for the next European Commission 2019-2024 – Opening statement in the European Parliament plenary session 16 July 2019 ; Speech in the European Parliament plenary session 27 November 2019, Publications Office of the European Union.

EU가 추진해야 할 6대 핵심 목표 중 하나로 “디지털 시대에 적합한 유럽(A Stronger Europe in the World)” 선정

ㅇ EU는 디지털 미래 구축을 위한 필수불가결 요소로 ‘신뢰’와 ‘보안’을 강조하면서 사이버보안을 핵심 정책 분야로 제시

□ EU는 나날이 고도화되는 사이버보안 위협에 대응하기 위하여 ’20년 이후 EU 전역의 사이버보안 수준을 전반적으로 향상시키기 위한 목적으로 법제도 및 정책을 지속적으로 정비하여 왔음

ㅇ NIS2지침(’22. 12),[90] 복원력 지침(CER, ’22. 12.),[91] 사이버복원력법(’24. 10),[92] 사이버연대법(’25. 1.),[93] 사이버보안법(‘’25. 1.)[94] 등

90) Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive).

91) Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC (Text with EEA relevance).

92) Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act).

93) Regulation (EU) 2025/38 of the European Parliament and of the Council of 19 December 2024 laying down measures to strengthen solidarity and capacities in the Union to detect, prepare for and respond to cyber threats and incidents and amending Regulation (EU) 2021/694 (Cyber Solidarity Act).

94) 사이버보안법은 2019. 6. 7. 최초 제정되었고 2025. 2. 4. 개정되었다. Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act).

ㅇ 유럽 위기 대비 연합 전략('25. 3.),[95] EU 사이버 블루프린트(Cyber Blueprint)('25. 6.)[96], ICT 공급망 툴박스('26. 2.) 등

□ 특히 전 세계 최초의 AI 규제법으로 알려진 EU AI법에는[97] AI에 특화된 사이버보안 관련 규정이 있어 검토 필요

□ 최근 EU는 AI의 급속한 발전과 그에 따른 경제사회 변화에 신속히 대응하고자 디지털 옴니버스 패키지*('25. 11.)와[98] 사이버보안 패키지('26. 1.)를[99] 발표하는 등 적극적으로 법제도 정비를 추진 중

* EU 의회와 EU 이사회의 공식 승인을 거쳐 '26. 7. 27. 발효됨

95) JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS on the European Preparedness Union Strategy, JOIN(2025) 130 final.

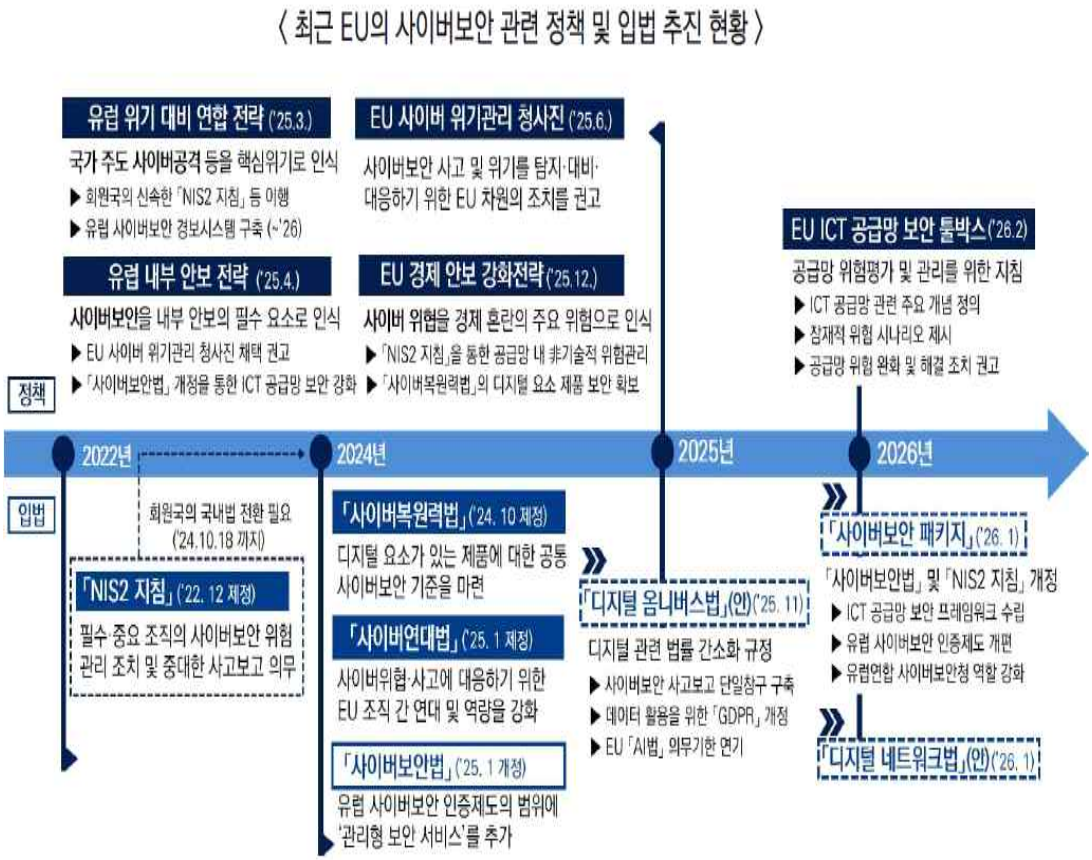
96) Council Recommendation on an EU blueprint for cyber crisis management - Council Recommendation approved by the Council at its meeting on 6 June 2025, 2025/0036 (NLE).

97) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) OJ L, 2024/1689, 12.7.2024.

98) Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus).

99) https://digital-strategy.ec.europa.eu/en/news/commission-strengthens-eu-cybersecurity-resilience-and-capabilities

[그림 13] EU의 사이버보안 정책 및 입법 추진 현황100)



다. 법제

(1) 기본 구조

□ EU 사이버보안법제는 사이버보안법(EU Cybersecurity Act)를 기본으로 NIS2 지침, CER 지침, 사이버복원력법, 사이버연대법이 각 영역/분야별로 위치한 체계로 이해할 수 있음; 핵심은 다음과 같음

100) 한국인터넷진흥원, "최근 EU 사이버보안 입법동향과 시사점", Digital Securioty Law Focus(2026. vol 1), 3쪽.

ㅇ 사이버보안법: 추진체계 및 거버넌스(ENISA[101]), 사이버보안 인증제도

ㅇ NIS 2지침: 중요 분야/서비스 담당 조직의 사이버보안

ㅇ CER 지침: 정보통신기반보호시설(Critical Entities) 사이버보안

ㅇ 사이버복원력법: 디지털제품의 전주기 사이버보안

ㅇ 사이버연대법: 사이버보안 정보 공유

(2) 사이버보안법[102]

□ 배경

ㅇ 사이버보안법은 2017. 9. 당시 EU 집행위원장인 장클로드 융커가 발표한 사이버보안 패키지의 핵심 법안으로, EU 디지털 단일 시장을 보호하고 초국경적 사이버 위협에 통합 대응하기 위하여 발의됨

ㅇ 구체적으로는 (1) 한시 조직이었던 ENISA를 EU 사이버보안청으로 상설기구화하여 EU 전역의 사이버안보 컨트롤타워로 격상시키고, (2) EU 전역에서 단일하게 통용되는 사이버보안 인증 프레임워크(EU Cyber-security Certification Framework, 'EUCC')를 구축하는 두 가지 목표를 담고 있음

□ 연혁

ㅇ 사이버보안법은 2019. 4. 제정되어 2025. 1. 한 차례 개정되었고, 2026. 1. 제안된 사이버보안 패키지에도 개정안이 포함됨

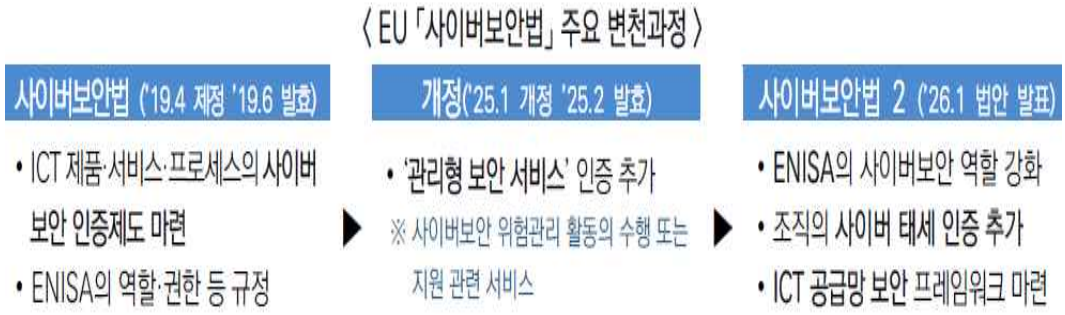
ㅇ 주요 변천과정은 다음과 같음

---

101) EU 사이버보안청(Eurpoean Union Agency for Cyvbersecurity)이다. https://www.enisa.europa.eu/

102) 사이버보안 패키지에서 개정이 제안된 내용이 아닌 현재 시행중인 사이버보안법을 기준으로 설명함(사이버보안 패키지는 별도 항에서 설명).

[그림 14] EU 사이버보안법 주요 변천 과정[103)]



□ 주요 내용

○ 사이버보안법은 총 4개의 장(title), 69개의 조문, 부속서(Annex)로 구성된 방대한 법이지만 주요 내용은 전술하였듯이 (1) ENISA의 설치 및 권한, 조직 등 EU 사이버보안 거버넌스 체계(Title II)와 (2) EUCC(Title III)의 두 가지임

○ (ENISA) ENISA는 EU 전역에서 高수준의 사이버보안 역량이 달성될 수 있도록 정책 개발 및 집행 등을 전담하는 독립기구로 설치됨

- 주요 기능: 정책 및 법률 개발 지원, EU 회원국별 일관된 법 집행 및 역량 강화 지원, 컴퓨터보안침해사고대응반 사무국 운영, EUCC 프레임워크 개발 등

○ (EUCC) 제품/서비스의 수명 주기 전반에 걸쳐 데이터 가용성, 진정성, 무결성, 기밀성을 평가하는 유럽 공통 사이버보안 인증 프레임워크

- '19. 6. 제정 당시에는 인증 대상이 ICT 제품, ICT 서비스, ICT 프로세스세 가지 영역으로 한정되어 있었으나, '25. 1. 개정으로 관리형 보안 서비스(Managed Security Service)가* 추가됨

103) 한국인터넷진흥원, "최근 EU 사이버보안 입법동향과 시사점", Digital Security Law Focus(2026. vol 1), 4쪽.

* 관리형 보안 서비스: "사이버보안 위험관리 관련 활동을 수행하거나 자원을 제공하는 제3자에게 제공되는 서비스"로 정의됨. 일례로 보안 사고 대응(Incident Response), 모의 침투(Penentration Testing), 보안 감사(Security Audits), 사이버보안 컨설팅 등을 들 수 있음

- 관리형 보안 서비스 인증을 추가함으로서 EU 역내에서 제공되는 관리형 보안 서비스의 신뢰성이 향상되는 효과를 낳음

(3) NIS2 지침[104]

□ 배경

○ EU 회원국의 사이버보안 위기관리 능력 및 사고시 보고의무 강화 등을 통하여 EU 전역의 사이버보안 수준을 향상시켜 EU 디지털 단일 시장의 안정성을 확보하고 중단없는 비즈니스 복원력을 보장하기 위한 규범임

○ 동 지침은 기존 "네트워크 및 정보시스템 보안 지침(NIS 지침[105])"을 폐지하고 새롭게 마련된 것으로서, 정식 명칭은 "EU 전역의 높은 공통 수준의 사이버보안을 위한 조치에 관한 지침"이나, NIS 지침의 후속 버전(version)인 관계로 'NIS2 지침'이라고 통용됨

- 규범 형식이 지침(direcvie)이므로 각 회원국의 이행입법이 필요하나 2026. 8. 기준 모든 회원국이 이행입법을 완료한 것은 아님

104) 사이버보안 패키지에서 개정이 제안된 내용이 아닌 현재 시행중인 NIS2 지침을 기준으로 설명함(사이버보안 패키지는 별도 항에서 설명).

105) Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union.

- 단 EU 집행위원회는 위험관리 조치의 기술 및 방법론적 요구사항을 구체적으로 정비한 동 지침 이행규정을[106] '24. 10.에 마련하였음

□ 연혁

○ NIS2 지침은 '22. 12. 제정되었으며 '26. 1. 제안된 사이버보안 패키지에도 개정안이 포함됨

[그림 15] EU NIS2 지침 주요 변천 과정[107]

106) Commission Implementing Regulation (EU) 2024/2690 of 17 October 2024 laying down rules for the application of Directive (EU) 2022/2555 as regards technical and methodological requirements of cybersecurity risk-management measures and further specification of the cases in which an incident is considered to be significant with regard to DNS service providers, TLD name registries, cloud computing service providers, data centre service providers, content delivery network providers, managed service providers, managed security service providers, providers of online market places, of online search engines and of social networking services platforms, and trust service providers.

107) 한국인터넷진흥원(각주 103), 5쪽.

□ 주요 내용

ㅇ (적용 대상) (1) 중요도가 높은 분야(부속서 I[108])와 기타 중요한 분야(부속서 II[109])에 해당하는 경우 일정 규모 기준(중기업 이상[110])에 해당하는 공공 또는 민간 조직에 적용하되(국가안보, 치안, 국방, 법집행 분야는 제외) (2) 지침에 규정된 의무적용 대상[111] 조직에는 규모에 상관없이 적용함

ㅇ (차등 규제) 중요도에 따라 필수 조직(essential entities)과 중요 조직(important entities)로 나누어 감독 및 집행 수준에 차등을 둠[112]

ㅇ (주요 의무) 필수 조직 또는 중요 조직은 (1) 사이버 위험 관리에 필요한 적절하고 비례적인 조치(위험분석, 정보시스템 보안 정책 수립, 사고 처리, 공급망 보안, 인적자원 보안, 접근통제 정책, 자산관리 등)를 수행해야 하고 (2) 서비스 제공에 중대한 영향을 미치는 사고를 통지할 의무 부담

- (2)의 경우 조기 경고(사고 탐지 및 중대성 여부에 관한 초기 정보)는 24시간 이내, 사고 통지(구체적인 침해 내용과 그에 대한 초기 평가)는 72시간 이내에, 최종 보고서(사고의 근본 원인 분석과 최종 조치)는 1개월 이내에 각 제출해야 함

  * 사고 통지가 아닌 사이버위협 정보 공유는 자발적 '협정'에 근거[113]

108) 에너지, 교통, 은행, 금융, 의료. 식수, 폐수, 디지털 인프라, 공공행정 등.

109) 우편, 폐기물, 화학물질, 식품, 제조, 디지털서비스, 연구 등.

110) 직원 수 50명 이상이거나 연간 매출액 및 재무 규모가 1,000만 유로를 초과하는 기관을 말함.

111) 공공전자통신네트워크, 최상위도메인네임, 회원국 내에서 주요 사회, 경제활동 유지에 필수적인 서비스의 유일한 제공자, 서비스 중단이 공공 안전, 보안, 보건에 중대한 영향을 미치는 경우 등.

112) NIS2지침 제32조, 제33조.

- 필수/중요 조직의 자발적 통지는, 범죄의 예방, 수사, 탐지 및 소추를 침해하지 않는 범위 내에서, 자발적 통지를 하지 않았다면 부가되지 않았을 추가적인 의무를 해당 조직에 부과하는 결과로 이어져서는 안됨(일종의 'safe harbor')[114]

- NIS2의 safe harbor는 CISA 2015 등 미국과는 달리 제재처분에서 해당 통지를 사용하지 못한다는 의미가 아님; 단 과징금 부과 여부 및 그 액수를 결정할 때 참작 사유로는 작동함[115]

ㅇ (거버넌스) (1) 관리 주체(이사회, 대표이사 등)가 사이버보안 위험 관리 조치를 직접 승인하고 그 이행 과정을 감독하도록 규정하고, (2) 컴플라이언스 위반이나 중대한 과실로 동 규정 위반이 지속될 경우 법인에 대한 과징금 부과뿐 아니라 경영진 개인에게도 법적 책임을 귀속시킬 수 있도록 규정함[116]

ㅇ (국제 협력) 각 회원국은 유럽 사이버위기 연락 조직 네트워크(EU-CYCLONe[117]), 보안사고 대응팀 네트워크(CSIRTs Network)를 가동하여 사이버위협 정보를 초국가적으로 실시간 공유

113) 앞서 살펴본 미국의 CISA 2015 역시 민간기관은 자발적 공유 체계임

114) NIS2 지침 제30조 제2항("Where necessary, competent authorities or the CSIRTs shall provide the single points of contact with the information on notifications received pursuant to this Article, while ensuring the confidentiality and appropriate protection of the information provided by the notifying entity. Without prejudice to the prevention, investigation, detection and prosecution of criminal offences, voluntary notifications shall not result in the imposition of any additional obligations upon the notifying entity to which it would not have been subject had it not submitted the notification.").

115) NIS2 지침 제34조 제2항 (f), (g); 동 지침 recital 120.

116) NIS2 지침 제20조 제1항, 제32조 제5항, 제6항. 제32조는 필수 조직의 경영진에게만 적용된다.

117) 대규모 국경 간 사이버 위기 및 재난 상황이 발생했을 때, 기술적 침해를 분석하는 실무 팀과 국가별 정치·전략적 의사결정권자 사이에서 신속한 조율 및 공동 위기 대응을 매개하는 네트워크

ㅇ (참고: 이행규정) NIS2 지침의 위임 사항을 규정한 EU 집행위원회 이행규정은 디지털 인프라 및 서비스 제공업체(DNS 서비스 제공자, 클라우드 서비스 제공자, CDN 서비스 제공자, 온라인마켓플레이스, SNS 플랫폼 등)가 준수해야 할 기술적 · 관리적 사이버보안 위험관리 조치와 중대 사고 보고 기준을 구체화하여 규정함

(4) CER 지침

□ 배경

ㅇ CER 지침은 기존의 물리적 인프라 보호 위주 정책에서 탈피하여 자연재해, 테러, 내부 위협, 사이버보안 등 다양한 대내외적 위험 발생시에도 사회 핵심 서비스가 중단없이 지속될 수 있도록 주요 조직(Critical Entities)의 회복탄력성을 강화하는 규범임

* 정보통신기반시설보호법과 일응 유사하나, 주요기반시설(Critical Infrastructure)이 아니라 주요 조직(Critical Entities)의 복원력에 초점을 맞추었다는 점에서 차이가 있음

□ 주요 내용

ㅇ 11대 핵심 부분(에너지, 교통, 은행, 금융시장 인프라, 보건, 음용수, 폐수, 디지털 인프라, 공공행정, 우주, 식품)을 주요 조직으로 지정[118]

ㅇ EU 회원국은 주요 조직의 복원력 강화 전략을 수립, 시행하여야 함

ㅇ 주요 조직으로 식별되었다고[119] 통지받은 기관은 자동으로 NIS2 지침에 따른 필수 조직으로 간주되어[120] 위협평가 수행, 복원력 조치 이행, 사고 발생 시 통지 등의 의무 부담

118) CER 지침 Annex.
119) 식별은 각 회원국이 한다. CER 지침 제6조.
120) NIS2 지침 제3조 제1항 (e).

- (위험 평가) 주요 조직은 식별 통지받은 날로부터 10개월, 이후 4년마다(또는 필요시) 모든 위험에 대한 평가 수행
- (복원력 조치) 주요 조직은 위험 평가 결과에 기반하여 복원력을 보장하기 위한 기술적, 보안적, 조직적 조치를 취해야 함
- (사고 통지) 주요 조직은 서비스 제공을 방해하는 사고 발생시 지체없이 관할 당국에 통지해야 함

ㅇ NIS2 지침과의 중복 규제를 피하기 위하여 은행, 금융시장 인프라, 디지털 인프라 부문에 대한 CER 지침상의 의무는 면제됨(회원국은 이들을 주요 조직 명단에는 포함시켜야 하고, NIS2 감독 당국과 CER 감독 당국간 상호 정보 교환을 활성화해야 함)[121]

(5) 사이버복원력법

□ 배경

ㅇ 하드웨어/소프트웨어 불문 디지털 요소가 포함된 제품에 대한 사이버 공격이 증가하고 있지만 EU 역내에 이를 규율하는 통일된 사이버보안 규범이 부재

ㅇ EU는 IoT 제품, 커넥티드 장치 등 디지털 요소가 포함된 다양한 제품에 의무적으로 인증이나 적합성 선언과 같은 사이버보안 요건을 도입하여 (1) 제조사의 책임을 강화하고 (2) 제품 생애주기 전반에 걸친 사이버보안 요소를 개선하며 (3) 소비자 선택권과 정보접근권을 향상시키기 위하여 '사이버복원력법' 제정('24. 10.)

121) CER 지침 제13조 제2항, 제15조 제4항 등.

- NIS2 지침이 필수/중요 서비스 운영 조직의 '운영 보안'을 다룬다면, 사이버복원력법은 시장에 유통되는 제품 자체의 '설계 보안'을 다룸

□ 주요 내용

○ (적용 제품) EU 역내에 출시되는 데이터 통신이나 네트워크 연결 기능을 가진 모든 하드웨어 및 소프트웨어 제품

- 위험도에 따라 ① 일반 제품(Class I), ② 중요 제품(Class II), ③ 고위험 AI 시스템과 연계된 고위험 제품으로 분류하여 인증 의무 차등화[122]

○ (적합성 평가) 제품의 EU 시장 출시 전 동법이 요구하는 사이버보안 표준을 충족하였음을 증명하는 적합성 평가를 거쳐야 하며(자기 평가 또는 제3자 인증),[123] 최종적으로 CE 마크를 부착해야 함[124]

- 위험도가 낮은 제품은 자기 평가가 허용되나, Class II 등 중요 제품은 공인인증기관의 제3자 평가 및 인증이 강제됨

○ 제조사, 수입업자 등의 의무

- (제조사) 제품 기획 단계부터 사이버보안 위협을 최소화하는 'Secure by Design and Default' 원칙을 준수하여야 하고, 기술 문서 및 지침 등을 작성, 제공해야 하고, 취약점이나 중대한 보안 사고를 인지한 경우 24시간 이내에 관할 당국 및 ENISA에 보고하여야 함[125]

122) 사이버복원력법 제7조, 제8조.
123) 사이버복원력법 제28조, 제32조.
124) 사이버복원력법 제29조, 제30조.
125) 사이버복원력법 제13조, 제14조.

- (수입업자, 유통업자) 제조사가 동법을 준수하고 CE마크를 정당하게 획득하였는 지 검증한 후 해당 제품을 시장에 유통해야 함[126)]

ㅇ (위반시 제재) 동법 위반 제품에 대해 시장 출시 금지, 리콜, 제한 조치 등이 가능하며, 동법 위반한 제조사, 수입업자 등에게는 최대 1,500만 유로 또는 전 세계 연간 매출액의 2.5% 중 높은 금액의 과징금 부과

(6) 사이버연대법

□ 배경

ㅇ 우크라이나 전쟁 이후 EU의 기간 인프라를 겨냥한 APT의 사이버 공격이 양적, 질적으로 급증

ㅇ EU 기간 인프라를 대상으로 한 사이버 공격은 단일 국가의 역량만으로는 방어가 어려우나 EU 회원국간 사이버 대응 역량 격차가 심각하고 정보 공유 체계가 부재하여 효과적인 대응이 불가능

ㅇ EU 차원의 대규모 사이버보안 위협에 대한 대응 능력을 강화하기 위하여 사이버연대법이 제정됨('25. 1.)

□ 주요 내용

ㅇ (EU 사이버 쉴드) EU 전역의 사이버 위협을 실시간 탐지하고 조기 경보를 공유하기 위하여 여러 회원국이 공동 참여하는 사이버안보 허브(Cyber Hubs) 네트워크 구축[127)]

126) 사이버복원력법 제19조, 제20조.
127) 사이버복원력법 제4조~제6조.

- 회원국은 자국 내 사이버보안관제 역량을 갖춘 주체를 국가 사이버 허브로 지정
- 최소 3개 이상의 회원국이 컨소시엄 형태로 국경간 사이버허브를 구축할 수 있음
- 허브간 정보공유를 통하여 네트워크 트래픽과 위협 데이터를 AI 등으로 분석하여 공격 징후를 사전에 포착하고 범유럽 차원의 상황 인식 공유

ㅇ (사이버 위기대응 매커니즘) 대규모 위기 발생 시 즉각 투입할 수 있는 신뢰할 수 있는 민간 서비스 보안 제공자 풀[사이버보안 리저브(Cybersecurity Reserve)] 구축[128]

* 침해 사고 발생시 사이버보안 리저브가 급파되어 자문 및 대응 수행

ㅇ (사고 리뷰 매커니즘) 사이버 사고 종결 후 ENISA 주도로 해당 사고를 전면 검토하고 공식 보고서 발간[129]

(7) 인공지능법(AIA)

□ 배경

ㅇ EU 전 회원국에 적용되는 AI 규제에 관한 통일된 조화 규칙(Harmonised Rules)을 수립하여 시장 파편화를 방지하고, 안정적인 투자와 혁신이 가능한 신뢰 기반의 단일 시장 환경 조성

ㅇ EU의 핵심 가치와 윤리적 기준에 부합하는 '신뢰할 수 있는 AI(Trustworthy AI)' 생태계 조성

128) 사이버복원력법 제12조~제15조.
129) 사이버복원력법 제17조.

□ 주요 내용[130](AI 사이버보안 한정)

ㅇ (고위험 AI 사이버보안) EU AI법은 고위험 AI에 정확성(accuracy), 견고성(robustness) 및 사이버보안을 명문으로 요구(제15조)

- 고위험 AI 시스템은 적절한 수준의 정확성, 견고성, 그리고 사이버보안을 달성할 수 있도록 설계 및 개발되어야 하며, 이러한 보안 및 성능 기준은 최초 출시 시점에만 국한되지 않으며, 전체 라이프사이클(Lifecycle) 동안 일관되게 유지되어야 함
- 고위험 AI 시스템은 권한없는 제3자가 취약점을 악용하여 AI의 용도(Use), 출력값(Outputs), 또는 성능(Performance)을 무단으로 변경하려는 시도에 대해 복원력(Resilient)을 갖춰야 함
- 고위험 AI 시스템이 AI 특유의 취약점을 해결하기 위한 기술적 솔루션은 위험의 예방(Prevent), 탐지(Detect), 대응(Respond), 해결(Resolve) 및 통제(Control) 단계를 모두 포함해야 함
- 고위험 AI 시스템은 데이터 오염, 적대적 예제, 모델 회피, 기밀성 공격에 대비해야 함

  * 명문으로 위 4대 공격 방법 적시(제15조 제5항)

- 위 의무 위반시 고위험 AI 시스템 공급자에게 최대 1,500만 유로 또는 전 세계 매출액 3% 중 높은 금액의 과징금 부과(제99조)

ㅇ (범용 AI 모델(GPAM) 사이버보안) GPAM 중 시스템적 위험있는 GPAM(이하 'RGPAM')[131] 한정, 사이버보안 사항 요구(제55조)

---

130) EU 인공지능법의 주요 내용은 최경진 외, 『EU 인공지능법』, 박영사(2024) 참조.

131) 누적연산량이 10^{25} FLOPS 이상이거나, EU AI Office에 의하여 시스템적 위

- RGPAM 자체뿐만 아니라, 해당 모델이 구동·학습되는 물리적 인프라에 대해서도 적절한 수준의 사이버보안을 확보해야 함

- 최신 기술 수준(State of the art)을 반영한 표준화된 프로토콜에 따라 모델 평가를 수행해야 하며, 특히 시스템적 위험을 식별·완화하기 위한 적대적 테스트(레드팀 빌딩 등)를 반드시 실시하고 문서화해야 함

ㅇ 사고 통지, 리콜 등

- 고위험 AI 시스템 공급자는 AI 시스템을 시장 출시후 모니터링해야 하고, 해당 AI 시스템에 중대 사고 발생시 15일 이내(중요 인프라의 관리 또는 운영에 대한 심각하고 회복불가능한 장애 발생시 3일 이내) 관할 당국에 보고해야 함(제73조)

- RGPAM 공급자 역시 중대 사고 및 가능한 시정 조치에 관하여 관련 정보를 추적하고 문서화하며 EU AI Office나 관할 당국에 지체없이 보고할 의무를 짐(제55조 제1항(c))

(8) 디지털 옴니버스 패키지

□ 배경

ㅇ 디지털 옴니버스 패키지는 '드라기 보고서('24. 9.)'에[132] 따라 EU 디지털 경쟁력을 강화하기 위하여 기존 법률간 중복 규정을 정비하고 규제를 완화하는 내용으로 '25. 11. 제안된 '다수 법률 일괄 개정안'

ㅇ EU 의회와 EU 이사회의 공식 승인을 거쳐 '26. 7. 27. 공식 발효[133]

험이 있다고 지정된 GPAM.

132) The Draghi report on EU competitiveness, https://commission.europa.eu/topics/competitiveness/draghi-report_en

133) AI법, GDPR, ePrivacy 지침, Data Act, NIS2 지침 등이 일괄 개정되었음

□ 주요 내용

○ 디지털 옴니버스 패키지는 사이버보안 측면에서도 유의미한 개정안을 제시하고 있는데, 바로 단일 접수 창구(Single Entry Point, ‘SEP’)의 신설임

- 현행 EU 규범상 기업은 사이버보안 사고, 개인정보 침해 사고 또는 AI 시스템 사고 등이 발생한 경우 개별 법률에 따라 서로 다른 창구에 보고나 신고 등을 해야 하는 불편함과 비효율이 있었음

- 디지털 옴니버스 패키지가 입법되면 기업은 표준화된 통합 양식(harmonized templates)을 사용하여 단일 창구(SEP)에* 딱 한 번만 신고/보고 등을 하면 됨. SEP는 신고/보고 내용을 자동 분석하여 개별 법상의 소관 당국으로 자동으로 이를 통보함[134]

  * ENISA가 단일 창구를 구축하고 운영, 관리함

(9) 사이버보안 패키지

□ 배경

○ 사이버보안법 제정(’19. 4.) 이후 AI, 양자컴퓨팅 등 신기술이 급속히 발전함에 따라 사이버보안 위협 또한 급증하고 있음

○ EU집행위원회는 EU 전반의 사이버보안능력 및 복원력을 강화하기 위하여 ’26. 1. 사이버보안 입법 패키지를 발표

134) 디지털 옴니버스 패키지 중 제6조~제9조(NIS2 지침 제23조의a 신설, GDPR 제3조 개정 등).

ㅇ 위 입법 패키지는 사이버보안법 개정안과[135] NIS2 지침 개정안의[136] 두 가지로 구성됨

□ 주요 내용(사이버보안법 개정안)

ㅇ ENISA 권한 확대

- ENISA는 NIS2 지침에 따른 필수/중요 조직의 랜섬웨어 공격 정보도 보고받으며, 이에 대응하기 위한 헬프데스크를 설립해야 하고, 랜섬웨어 공격 동향을 분석, 제공해야 함
- ENISA는 디지털 옴니버스 패키지에 따라 단일 접수 창구 기능 수행
- ENISA는 사이버위협 조기경보 발령 등을 수행해야 하고, EU 취약점 데이터베이스를 유지, 관리해야 하며, 취약점 정보 서비스를 제공해야 함

ㅇ 유럽 사이버보안 인증(EUCC) 프레임워크 개편

- 기존 EUCC 대상에 ‘사이버 태세(cyber posture)’를 추가: 사이버 태세는 NIS2 지침에 따른 필수/중요 조직의 사이버보안 위험관리 조치 의무 준수 입증에 사용될 수 있는 인증 체계임
- 사이버 태세 인증을 받으면 감독 당국은 그 범위 안에서 필수/중요 조직에게 추가적인 보안 조치를 요구하지 않음

---

135) Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Union Agency for Cybersecurity (ENISA), the European cybersecurity certification framework, and ICT supply chain security and repealing Regulation (EU) 2019/881 (The Cybersecurity Act 2).

136) Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Directive (EU) 2022/2555 as regards simplification measures and alignment with the [Proposal for the Cybersecurity Act 2].

ㅇ ICT 공급망 보안 강화[137)]

- 현행 NIS2 지침은 필수/중요 조직의 사이버보안 위험관리 조치 사항에 '공급망 보안'을 포함하도록 의무화
- 사이버보안 패키지 법안은 EU 차원의 신뢰할 수 있는 ICT 공급망 보안 프레임워크 구축을 추진하는바, 구체적으로 다음과 같음
- (i) ICT 공급망 보안 위험평가 수행
- (ii) (a)를 바탕으로 ICT 공급망에 심각하고 구조적인 비기술적 위험을 초래하는 것으로 판단되는 국가를 '사이버보안 우려 국가'로 지정
- (iii) EU 집행위원회는 사이버보안 우려 국가로 지정된 제3국에 의한 통제 가능성이 있는 고위험 공급업체를 식별
- (iv) 고위험 공급업체로 식별된 자는 EUCC를 신청하거나 보유할 수 없으며(기존 인증은 지체없이 취소됨), 공공조달 절차에 참여할 수 없는 등의 제재를 받음

□ 주요 내용(NIS2 지침 개정안)

ㅇ 필수 조직 규모 기준을 중기업 초과 ⇒ 소형 중견기업 초과로 상향(즉 대상 감축)하여 규제 완화

ㅇ 필수 조직 대상 확대: 디지털 신원/지갑 서비스 제공자, 해저 데이터 전송 인프라 운영자 등을 포함

ㅇ 사이버보안법 개정안과 결합하여 사이버 태세 인증 도입

137) ICT 공급망 보안 강화 프레임워크가 시행되면 특정국 소속의 고위험 공급업체가 유럽 내 통신망에서 배제되는 효과가 발생할 것으로 예상됨. 특히, EU 집행위원회는 동 ICT 공급망 보안 프레임워크과 연계하여 고위험 공급업체를 고려한 통신장비 규제, 신뢰할 수 없는 공급업체와 협력을 유지하는 통신사에 대한 주파수 이용 제한 등 제재규정 등을 포함한 「디지털 네트워크법」(안)을 병행하여 입법 추진 중임('26. 1.).

라. 정책

(1) 사이버보안 및 AI에 관한 EU 행동계획[138]

□ 배경

○ 첨단 AI(Frontier AI) 기술이 빠르게 진화하면서 사이버보안 환경에 가져온 위기와 기회에 신속하게 대응하기 위하여 수립('26. 7. 7.)

- 프론티어 AI 출현으로 (1) 사이버 위협이 고도화, 대중화되고 (2) 기술 종속성이 심화되어 AI 주권(AI sovereignty) 위기가 대두된 상황에서, (3) EU의 AI 및 사이버보안 기존 규범(AI act, CRA, NIS2 등) 하에서 실제 AI 기반 사이버 위협에 대응할 수 있는 실행 계획이 요구됨

□ 주요 내용

○ (1) 3대 핵심 전략 축(Pillars)과 (2) Pillar별 3대 전략, (3) (1), (2)를 전체적으로 아우르는 국제 협력으로 구성됨

[표 14] 사이버보안 및 AI에 관한 EU 행동계획 주요 내용

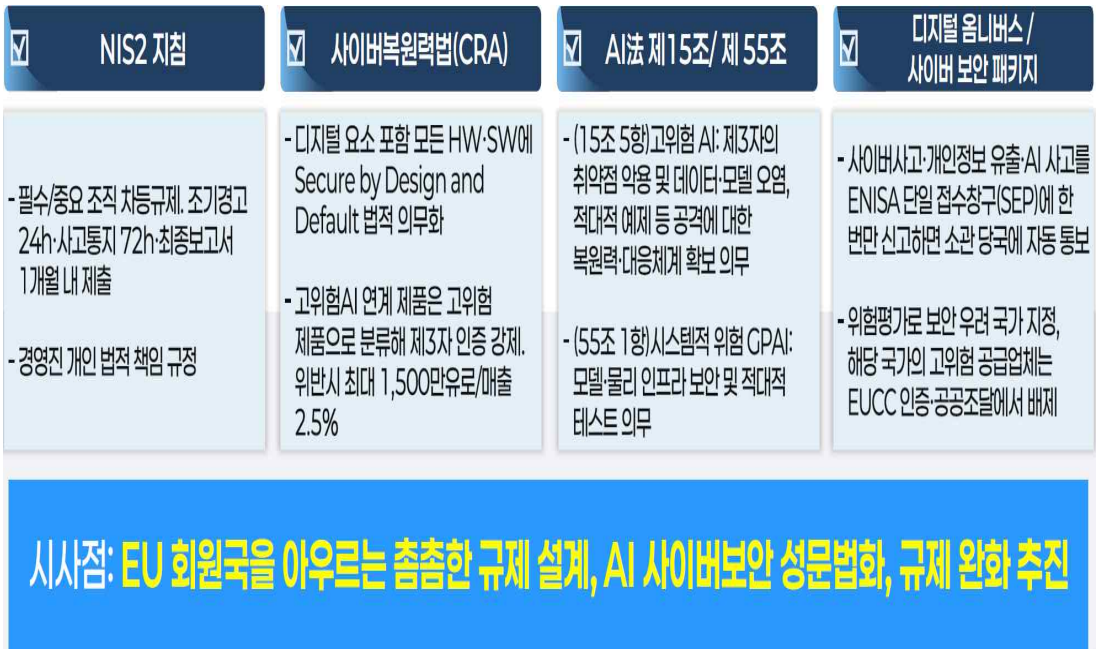
Pillars	전략
프론티어 AI의 안전한 평가, 접근 및 테스트 플랫폼 구축	• 유럽 자체 사전 평가 역량 강화 • 표준화된 가이드라인(EU Blueprint) 수립 • AI 보안 기술 실증을 위한 격리된 '안전 테스트 플랫폼' 구축
AI 시대를 위한 EU 사이버 생태계 준비	• 사이버 위협 지침 및 모범 사례(best practice) 발간 • AI 속도에 맞춤 취약점 관리 및 패치 가속화 • AI 모델을 활용해 주요 오픈소스 SW 취약점을 분석하고 자동 패치를 지원하는 '주요 오픈소스 복원력 캠페인' 추진
사이버보안을 위한 유럽 자체 AI 역량 스케일업	• 'AI 조력 취약점 자동 보완' 대규모 경진대회 개최 • 주권적 AI 인프라 및 자본 투자 • AI 보안 인재 및 전문성 강화

138) https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence

- (국제 협력) AI 및 사이버보안 문제가 국경과 무관함을 고려, G7, UN, NATO 등 관련국 및 국제기구와 연대를 강화하고, AI 모델 평가 표준화 및 오픈소스 보안에 필요한 글로벌 공조를 주도

마. 시사점

[그림 16] EU의 관련 법제도 현황 및 시사점



□ AI 및 디지털 제품 전 수명주기(Lifecycle)에 걸친 ‘보안 설계(Secure by Design)’ 규범화

○ 사이버복원력법(CRA)을 통해 하드웨어 · 소프트웨어 디지털 제품에 ‘Secure by Design and Default’ 원칙을 법적 의무화

○ 인공지능법 제15조 및 제55조를 통해 고위험 AI 및 시스템적 위험이 있는 범용 AI 모델(RGPAM)의 전체 수명주기에 걸쳐 데이터 오염, 적대적 공격, 모델 회피 방지 의무를 강제

□ 기업의 규제 부담 완화

ㅇ 디지털 옴니버스 패키지로 사이버보안 사고(NIS2), 개인정보 유출(GDPR), AI 시스템 사고 등으로 파편화된 보고 절차를 통합

ㅇ ENISA가 구축한 단일 접수 창구(SEP)에 표준 양식으로 1회 신고 시 관할 당국으로 자동 통보되는 체계 구축

□ 글로벌 ICT 공급망 통제와 주권적 AI · 테스트 인프라 구축 병행

ㅇ 사이버보안 패키지를 통해 우려 국가 및 고위험 공급업체를 식별 · 배제하는 ICT 공급망 위험평가 프레임워크를 수립

ㅇ EU 행동계획을 통해 AI 안전 테스트 플랫폼, 주권적 AI 인프라 자본 투자를 추진하여 기술 종속성 및 안보 위협에 동시 대응

3. 영국

가. 특징

□ 브렉시트 이후 핵심 디지털 인프라와 공급망(MSP 등)을 정밀 타겟팅하여 EU와 유사한 수준으로 사이버보안 규제 체제를 강화하는 한편, 인센티브와 비침습적 지원 등의 민간 친화적 제도 역시 시행 중

- ㅇ 사각지대였던 서드파티 디지털 공급망(MSP 등)의 정밀 규율: 브렉시트 이후 독자 입법을 통해, 핵심 인프라 침투 교두보로 악용되는 데이터센터, 부하 제어자, 관리형 서비스 제공자(MSP) 및 핵심 공급업체를 규제 테두리 내로 전격 편입
- ㅇ 민간 친화적 인센티브 설계 및 비침습적 정보 지원: Cyber Essentials 인증을 취득한 중소기업에 사이버 배상책임보험을 무료 자동 가입시키는 실효적 혜택을 제공하고, NCSC Early Warning을 통해 기업이 등록한 자산 기반의 위협 정보를 비침습적·단방향으로 무료 통보하여 자체 분석 부담을 완화

나. 개요

□ 영국은 과거에는 EU의 사이버보안 법규를 그대로 따랐으나, ’21. 1. BREXIT 후 영국 고유의 독자적 규범 체계를 마련해야 하는 상황

□ 영국은 과거에는 EU의 사이버보안 법규를 그대로 따랐으나, ’21. 1. BREXIT 후 영국 고유의 독자적 규범 체계를 마련해야 하는 상황

□ 영국의 사이버보안 정책은 현재 ① BREXIT 전 EU에서 시행된 舊 NIS 지침과 사실상 동일한 내용의 ‘2018 NIS 규정’과[139] ② ’21 제정 ‘제품보안 및 통신인프라법(PSTIA)’에 근거하여 추진되고 있으나,

□ 최근 생성형 AI와 에이전트 기술의 급속한 발전으로 야기되는 새로운 사이버보안 위협에 선제적으로 대응하기 위하여 2018 NIS 규정을 전면 개정하는 '사이버보안 및 회복력 법안(CSRB)'을* 추진하고 있어, 2018 NIS 규정보다는 CSRB에 보다 주목할 필요

* CSRB는 '26. 8. 기준 하원 절차를 끝내고 상원의 2회독(second reading)을 마친 상태이며,[140] 여야간 이견 없어 특사정이 없는 한 '26 연내 통과 전망

□ 법제 외에도 사이버보안 인증제도(Cyber Essentials), 조기경보체계(NCSC Early Warning)가 사이버보안 학계 및 실무에게 주목받고 있어, 함께 살펴봄

다. 법제

(1) PSTIA

□ 배경

○ 영국에서는 '16 해커들이 30만 대 이상의 IP카메라, 가정용 공유기 등을 해킹하여 좀비 PC로 만든 후 이들의 컴퓨팅 파워를 모아 대규모 디도스(DDoS) 공격을 감행하여 BBC, 트위터 등 주요 웹사이트와 인프라가 마비되는 소위 '미라이(Mirai) 봇넷 사태'가 발생함

○ 이를 계기로 영국 정부는 '18 민간 기업들이 자발적으로 보안중심설계(Secure-by-Design) 원칙 등을 지킬 수 있도록 13개 원칙을 담은 소비자 IoT 보안 행동강령을 발표했으나, 강제력이 없는 연성 규범(soft law)의 한계상 여전히 보안 취약점을 가진 IoT 기기가 대규모로 유통되는 문제 발생

139) https://www.gov.uk/government/collections/nis-directive-and-nis-regulations-2018
140) 입법 절차 진행 경과는 https://bills.parliament.uk/bills/4035에서 확인가능함

ㅇ 이에 영국 의회는 ’18 마련한 자발적 행동 강령을 PSTIA라는 법률로 격상시켜, 글로벌 IoT 제조/수입/유통망 전체에 보안 강화 의무를 지우고 위반시 강력한 과징금을 도입함으로써 IoT 제품 생애 전 주기에 대한 보안 책임을 민간 공급망에 지움

□ 주요 내용[141)]

ㅇ 적용 제품: 인터넷 연결가능 제품 및 네트워트 연결가능 제품(소위 ‘IoT’ 제품)

* 단, 이중규제 방지를 위하여 기존 법령에 따른 규제 대상 제품(전기차 중전소, 의료기기, 스마트 계량기, 인터넷이나 PC 등에 연결할 수 없는 태블릿)은 제외

ㅇ 적용 대상: 제조업체, 수입업체(타국에서 영국으로 적용 제품을 수입하고 제조업체가 아닌 자), 유통업체(영국에서 적용 제품을 제공하고 제조업체 또는 수입업체가 아닌 자)

ㅇ 제조업체의 의무

- (3대 핵심 보안 요건 준수) ① 디폴트 패스워드 금지, ② 취약점 신고창구 개설, ③ 보안 패치 기간 명시라는 3대 핵심 보안요구사항 의무를 준수해야 함
- (준수 선언서 작성, 첨부) 제품이 법적 보안 기준을 충족하였음을 공식 증명하는 준수 선언서(Statement of Compliance)를 작성, 첨부해야 함
- (보안 결함 조사 및 시정 조치) 출시된 제품에서 취약점이 발견되면 즉시 조사에 착수해야 하며, 유통 중단, 회수, 패치 배포 등의 시정 조치를 취하고 이를 규제 당국에 보고해야 함

141) 한국인터넷진흥원, “2024 해외 사이버보안 입법동향”(2024), 60-64쪽.

ㅇ 수입업체의 의무

- (제조업체의 보안 준수 여부 검증) 제품 유통 전 제조업체가 3대 핵심 보안 요건을 지켰는지, 그리고 '준수 선언서'를 적법하게 발행했는지 확인 및 검증해야 함
- (부적합 제품의 수입 금지 및 고지) 제조업체가 보안 요건을 위반한 것을 인지한 경우 해당 제품을 영국 시장에 출시해서는 안 되며, 이를 제조업체와 규제 당국에 즉시 통지해야 함
- (기록 보존 및 연락처 부기) 수입한 제품의 준수 선언서 사본을 법정 기간 동안 보관해야 하며, 제품 또는 포장에 자사의 상호와 연락 가능한 주소를 명시해야 함

ㅇ 유통업체의 의무

- (외관상 준수 여부 확인) 제품에 '준수 선언서'가 정상적으로 첨부되어 있는지, 수입업체와 제조업체의 필수 표시 사항이 누락되지 않았는지 확인해야 함
- (부적합 제품의 유통·판매 금지) 보안 요건을 위반했거나 준수 선언서가 없는 제품임을 알았거나 알아야 했을 경우, 해당 제품의 유통·판매 금지
- (시정 조치 협조) 이미 유통된 제품에 보안 결함이 발생하여 제조업체나 수입업체가 회수(Recall) 또는 유통 중단 조치를 내릴 경우, 이에 즉각 협조하고 유통망 내에서 해당 제품의 확산을 차단해야 함

ㅇ 의무 위반시 제재

- (과징금) 영국 정부는 의무위반자(제조업체/수입업체/유통업체)에게 최대 1,000만 파운드 또는 해당 기업의 전 세계 매출액의 4% 중 더 높은 금액을 과징금으로 부과할 수 있음

- (행정처분) 과징금 처분과 별개로 유통 중단 명령, 제품 회수 명령, 위반 사실 공시 명령 등 행정처분 가능
- (형사처벌) 법인에 대한 처벌뿐 아니라 의무 위반이 회사의 이사, 고위 경영진 혹은 이와 유사한 직위의 임원의 묵인이나 과실로 인해 발생한 경우 해당 개인 역시 형사처벌 가능

(2) CSRB

□ 배경

ㅇ BREXIT로 인한 입법 공백과 고도화된 사이버보안 위협 대응

- 전술하였듯이 영국은 ’21 BREXIT를 하였지만 사이버보안 규정은 BREXIT 이전 규범인 舊 EU NIS 지침과 동일한 2018 NIS 규정 체계에 머물러있었음
- 반면 EU는 ’22 사이버보안 위협에 대한 처벌과 규제를 대폭 강화한 NIS2 지침을 전격 시행하였고, 이에 따라 영국과 EU 회원국 사이에 사이버보안 대응 역량에 현격한 차이가 발생
- 독자적인 입법을 통해 EU 수준의 안보 체제에 동조하고 데이터 센터 및 핵심 공급망(Critical Suppliers) 위협을 통제해야 할 필요성이 제기

ㅇ 디지털 서비스 제공사(MSP) 등의 취약점 악용 방지 필요

- ’21 전 세계 중요 인프라와 기업의 시스템을 대행 관리하는 관리형 서비스 업체(Managed Service Providers, ‘MSP’)를 교두보로 삼아 침투하는 사이버 공격이 폭발적으로 증가
- 이에 따라 안보의 사각지대였던 서드파티 디지털 서비스 공급업체들을 규제 테두리 안으로 포함해야 할 필요성 증대

□ 주요 내용

○ (입법 형식) CSRB는 별도의 법률을 제정하는 방식이 아니라, 2018 NIS 규정의 조항들을 삭제, 수정, 신설하는 방식임

○ (규제 대상 확대) 2018 NIS 규정이 다루지 못한 핵심 디지털 인프라와 공급망을 사이버보안 규제 대상에 전격 편입

- 2018 NIS 규정은 5개 핵심 섹터(에너지, 교통, 보건, 식수, 디지털 인프라) + 3개 디지털 서비스(온라인 마켓플레이스, 검색엔진, 클라우드)를 적용 대상으로 규정

- CSRB는 이에 더하여 ① 데이터 센터, ② 부하 제어자(large load controller; 전력 그리드 수요반응 사업자), ③ MSP 및 핵심 공급업체(critical suppliers)를 규제 대상으로 지정

  * EU NIS2에서 추가된 폐수 처리, 우주, 우편/택배 서비스 등은 포함하지 않음. CSRB는 EU NIS2와는 달리 영국의 독자적인 정책 목적에 맞추어 데이터 센터, MSP 등과 같은 핵심 인프라와 공급망만을 정밀하게 타겟팅함

○ (규제 실효성 강화) 각 산업별 규제 기관이 일관성있게 규제를 집행하고 이에 필요한 권한과 자원을 충분히 확보하도록 지원

- 2단계 사고 보고 프레임워크: 실제 장애 발생시 뿐 아니라 시스템의 기밀성, 무결성, 가용성에 상당한 영향을 미치거나 미칠 가능성이 있는 사건까지 보고; 24H내 조기 보고(early warning) / 72H내 상세 보고 의무

- 고객 고지 의무: 데이터센터, MSP 등은 사이버 위협으로 영국 내 고객이 악영향을 받을 가능성이 있다고 판단되면 당국 보고(72H 내 상세 보고) 후 고객에게 리스크 성격과 대응 방안을 통지해야 함

- 과징금: 일반 위반시 최대 1,000만 파운드 또는 글로벌 연간 매출액의 2%, 중대한 위반시 최대 1,700만 파운드 또는 글로벌 연간 매출액의 4%

- 비용 회수 제도(cost recovery):* 규제 기관이 규제 대상 기업들로부터 행정 비용을 회수할 수 있는 분담금 산정 체계를 구축하여 스스로 재원을 조달할 수 있도록 함[142)]

  * CSRB에서 비용 회수 제도의 구체적 작동 방식은 다음과 같음

  ① 포괄적 분담금 산정 체계: 각 산업별 규제기관(의료 분야의 DHSC, 디지털 인프라의 Ofcom 등)은 자기가 담당하는 섹터의 사이버 안보 감독 업무(가이드라인 발간, 상시 모니터링, 산업계 교육 등)에 소요되는 '예상 총비용(expected costs)'을 미리 산정함

  ② 규제 대상 기업별 배분: 규제기관은 산정된 총 행정 비용을 해당 섹터 내 규제 대상 기업들에 균등 혹은 규모에 비례하여 분담금(Charges) 형태로 배분하여 부과함. 즉, 특정 기업에 대한 직접적인 조사가 없더라도, 해당 산업의 사이버 안보 체계를 유지하는 데 드는 비용을 기업들이 나눠 내는 구조임

  ③ 개별 실비 청구권(Specific Charges)과의 병행: 이와 별도로, 특정 기업의 위반 혐의를 입증하기 위해 특별 검사나 포렌식, 외부 전문가 실사 등이 진행되어 '추가 비용'이 발생한 경우 규제 당국은 해당 기업에 실비를 전액 청구할 수 있음

  ④ 법적 통제(투명성 의무): 기업들의 부담을 고려하여 규제 당국은 분담금 체계를 신설·개정할 때 반드시 피규제 기관들과 사전 협의를 거쳐야 하며, 매 회계연도마다 징수 및 집행 내역을 투명하게 공시

142) CSRB Clause 17. 이는 우리 행정체계에서는 찾아보기 힘든 독특한 제도이다. 유사 사례로는 금융감독원의 검사 대상 기관들이 납부하는 분담금(금융위원회의 설치 등에 관한 법률 제47조) 정도를 찾을 수 있다.

ㅇ (국가안보 목적의 직접 명령권) 각 부처의 장관(secretary of state)은 적대 세력의 '사전 포지셔닝(Pre-positioning, 시스템 내 잠복)' 등 급박한 국가 안보 위협이 포착될 경우 해당 기업에 특정 소프트웨어 사용 금지, 서비스 제한, 혹은 외부 전문인력 강제 임명 등을 명령(Directions)할 권한이 있음[143]

- 이 경우 명령을 받은 대상자가 안보 명령과 타 실정법상 요구사항을 동시에 준수하는 것이 합리적으로 불가능한 경우, 안보 명령상의 요구사항이 우선 적용됨[144]

  * 따라서 안보 명령을 받은 기업은 안보 명령만 준수하면 타법 위반으로부터 면책되는 효과가 발생함

- 안보 명령과 타 법령상 충돌 존재 여부에 대한 판단 및 결정 권한은 각 부처 장관에게만 전적으로 귀속됨

  * 따라서 안보 명령과 타 법령상 충돌 문제에 관한 사법 심사가 차단됨

---

143) CSRB Clause 43, 44. 원문은 다음과 같다.

Caluse 43(1) The Secretary of State may give a direction to a regulated person requiring the person to take, or refrain from taking, specified actions if the Secretary of State considers that—

(a) a relevant IT system has been or may be compromised, creating a risk to national security; and
(b) issuing a direction to address the compromise is necessary and proportionate in the interests of national security.

144) 이와 같이 행정각부 장관이 법률에 우선하는 명령을 발령하는 체계는 우리 법제도 하에서는 매우 이례적이고 위헌적으로까지 보일 수도 있지만, 의회주권주의가 대원칙인 영국 헌법체계 하에서는 이와 같은 내용도 법률로서 규정하면 합헌이다. 영국에서는 급변하는 안보 위기나 기술 환경에 대응하기 위해 종종 이와 같은 명령 체계를 법률에 규정하며, 이를 일명 '헨리 8세 조항(Henry VIII Clauses)'라고 부른다.

라. 정책

(1) Cyber Essentials[145)]

□ 개요

○ 국가사이버보안센터(NCSC)가 '14부터 운영중인 인증 제도

○ 강제성은 없지만, 영국의 공공조달 및 ICT 핵심 공급망 입찰에 참여하려면 반드시 Cyber Essential을 획득해야 함; 이를 통하여 사실상 강제력을 발휘

□ 주요 내용

○ 적용 대상: 영국의 공공조달 및 ICT 핵심 공급망 입찰에 참여하려는 모든 민간 기업(국내외 불문)

○ 인증 대상: 인터넷 연결 기기(end-user device), 네트워크 장비, 서버 및 호스팅 시스템, 클라우드 서비스

○ 인증 심사시 점검 항목: 5대 기술적 통제 항목에 한정됨

- ① 방화벽, ② 보안 구성(불필요한 기능, 프로토콜, 계정 등 식별 후 삭제/비활성화), ③ 사용자 접근 제어[클라우드와 원격 접속시 다중인증(MFA) 필수 적용], ④ 악성코드 방어(백신 상시 가동, 어플리케이션 허용 목록 지정, 샌드박싱 환경 등), ⑤ 보안 업데이트 관리[기술 지원이 종료된 SW는 전면 사용이 금지되며, 공급업체가 고위험 취약점 패치를 배포한 경우 반드시 14일 이내에 패치를 완료(14-day rule)해야 함]

  * ISO 등 다른 보안 인증에 비하여 심사 대상 항목이 간략하고 구체적임

145) https://www.ncsc.gov.uk/cyberessentials/overview

○ 인증 등급: 기본 등급(Cyber Essentials) / 플러스 등급(Cyber Essentials+)

- 기본 등급은 자체 서면 평가로만 이루어지고(비용도 저렴), 플러스 등급시 비로서 외부 기관의 인증 심사가 이루어짐
- 플러스 등급은 국방부 조달, 특정 고위험 공공 조달(핵심 중앙 정보시스템, 대규모 민감 개인정보 처리시스템, 법집행 시스템 등), 정부 및 공공기관의 보조금(grant)를 수급 조건으로 할 때 의무적으로 요구됨

○ 특이사항: 보험과 연계된 인센티브(incentive)

- 사이버 에센셜(기본 또는 플러스 불문)을 성공적으로 취득하고, 연간 총 매출액이 2,000만 파운드 이하이며, 영국에 본사를 둔 기업은 사이버 에센셜 인증서 발급과 동시에 25,000파운드 한도의 사이버 배상 책임 보험에 무료로 자동 가입됨[146]

(2) NCSC Early Warning[147]

□ 개요

○ NCSC는 2018 NIS 규정에 따라 설립된 영국의 공식 사이버보안 사고 대응팀(CSIRT) * 우리의 CERT와 유사

○ CSRIT로서 NCSC는 리스크 및 사고에 관한 조기 경보, 알림, 공고, 및 정보 전파 기능을 수행할 의무가 있음[148]

146) https://iasme.co.uk/cyber-essentials/cyber-liability-insurance/
147) https://www.ncsc.gov.uk/section/active-cyber-defence/early-warning
148) 2018 NIS 규정 제5조.

□ 주요 내용

○ (무료 정보 제공) 영국 소재 기관/기업은 누구라도 Early Warning 서비스에 가입하면 무료로 NCSC로부터 사이버위협 정보를 제공받음

○ (제공 정보) 서비스 가입자는 사고 알림(Incident Notification, 즉시), 네트워크 악용 알림(Network Abuse Evenes, 매일), 취약점 알림(Vulnerablilty Alert, 매주)의 3대 정보를 제공받음

- NCSC가 서비스 가입자에게 단방향으로 통보하는 체계로서, 상호주의적 사이버보안 정보 공유 체계가 아님

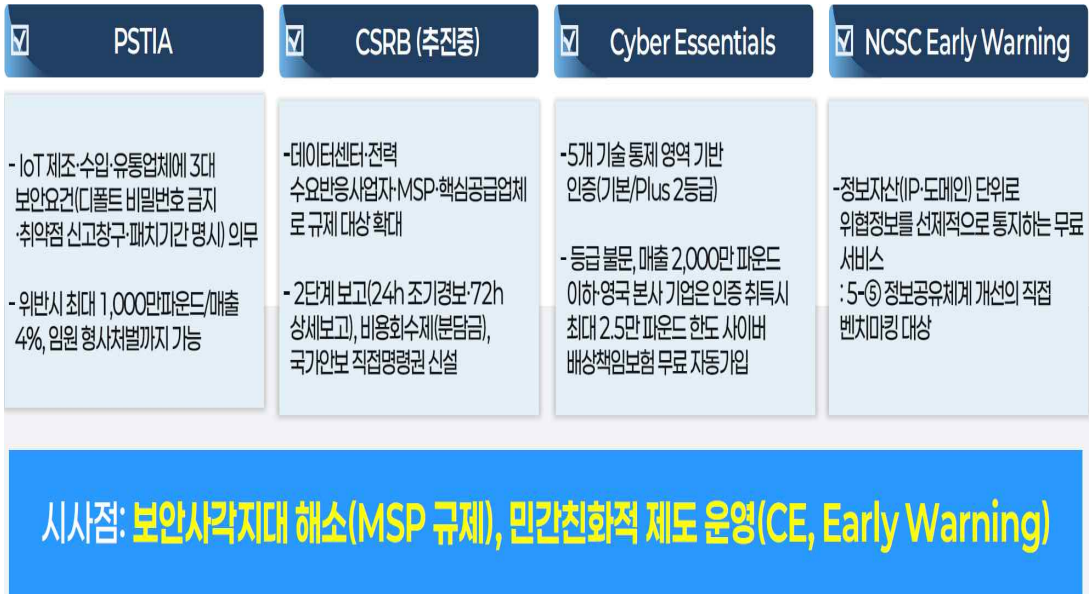
  * 민간의 사이버보안 정보 공유는 EU와 유사하게 법률이 아닌 '자발적 협력'에 근거하여 이루어짐

○ (자산 매칭형) NCSC 조기 경보는 NCSC가 기업이 미리 등록해 둔 자산(IP주소, 도메인 등)과 국가가 외부에서 수집한 사이버보안 위협 정보를 매칭하여 일치할 때 통보하는 체계

- 자산 매칭 방식은 국가가 민간의 전산 시스템을 실시간 감시하거나 스캔하지 않는 비침습적 방식일 뿐 아니라, 민간 입장에서는 사이버보안 정보 분석 능력이 없더라도(즉 사이버보안 관련 조직, 인력, 장비 등이 없더라도) NCSC 조기 경보대로 이행하면 충분하므로(예: 귀 기관의 노트북이 해킹당하여 지금 적대 세력으로 데이터를 전송중이므로 즉시 네트워크 차단조치를 하라) 민간의 부담이 크게 줄어듬

마. 시사점

[그림 17] 영국의 관련 법제도 현황 및 시사점



□ 보안 사각지대였던 서드파티 디지털 공급망(MSP 등) 직접 규제

○ 사이버보안 및 회복력 법안(CSRB)을 추진하여 데이터 센터, 전력 그리드 부하 제어자, 관리형 서비스 제공자(MSP) 및 핵심 공급업체를 규제 테두리 내로 편입함

○ 핵심 인프라 교두보로 악용되는 서드파티 공급업체의 보안 취약점을 정밀 타겟팅

□ 실질적 지원 인센티브 및 비침습적 지원 서비스 체계 운용

○ Cyber Essentials 인증을 취득한 중소기업에 연계 무료 사이버 배상책임보험을 자동 가입시키는 인센티브 제도 운용

○ NCSC Early Warning을 통해 기업이 등록한 자산과 수집된 위협 정보를 비침습적으로 매칭해 단방향 무료 통보함으로써 민간의 자체 분석 부담을 축소

□ 비용 회수 제도(Cost Recovery)를 통한 규제 기관의 재원 자립화

- ㅇ CSRB를 통해 규제 당국이 피규제 기업들로부터 행정 예상 총비용을 분담금 형태로 배분받아 조달하는 구조를 수립
- ㅇ 국가 예산 부담을 줄이면서 규제 기관의 전문성과 상시 감독 권한을 안정적으로 확보

4. 일본

가. 특징

□ “능동적 사이버 방어”를 법제화하는 대신 독립 기구를 통한 엄격한 통제 장치를 마련하였고, Agentic AI발 위협에 신속·유연하게 대응하는 민첩한 거버넌스 체계

ㅇ 능동적 사이버 방어(Active Defense) 법정화: 2025년 법 개정을 통하여 경찰과 자위대에 “접근 확인 및 무해화 조치” 권한을 부여하고, 국제법상 긴급상황 법리 등을 바탕으로 해외 서버 무해화까지 포괄하는 공세적 방어 체계를 제도화

ㅇ 통신정보 수집·분석 권한 신설과 통제: 사이버 공격의 실태 파악을 위해 내각총리대신에게 통신정보 취득 및 자동화된 기계적 선별 권한을 부여하되, 독립 기구의 사전 승인과 지속적 검사를 받도록 하여 권한 남용을 방지

ㅇ Agentic AI 위협 신속 대응: 프론티어 AI 기반의 자율 공격 가시화에 대응하여 범정부 대책 패키지를 출범시키고, 기술 변화 속도에 맞추어 AI 기본계획을 7개월 만에 전격 개정

나. 개요[149]

□ ’25 이전까지 일본의 사이버보안 법체계는 ’14 제정된 사이버시큐리티기본법(サイバーセキュリティ基本法, 추진체계)과 ’22 제정된 경제안전보장추진법(経済安全保障推進法, 에너지, 통신, 운송, 금융 등 주요인프라 보호)이 양 축을 이루고 있었음

149) 한국인터넷진흥원, “2025년 1분기 인터넷 정보보호 법제동향”(2025), 43-53쪽을 기초로 정리한 것이다.

□ 일본은 국가안전보장전략('22)에서[150] 사이버보안 역량을 미국, EU 등 주요국 수준 이상으로 향상시키기 위하여 (1) 민관협력 강화 (2) 통신정보를 활용한 사이버위협 탐지 (3) 공격자의 서버침입 무해화, (4) 사이버안전보장 분야의 컨트롤타워 조직 신설의 4대 목표 설정

□ 위 목표를 실현하기 위하여 '23. 3. 사이버안전보장체제정비 준비실을 신설하여 법제도 자문단 운영

ㅇ 자문단은 민관 정보공유 촉진, 사이버공격 대응을 위한 통신정보 활용 검토, 경찰/자위대의 접근 무해화 권한 구축, 조직 개편 등 제언

□ 위 제언을 토대로 '중요 전자계산기 부정행위로 인한 피해방지법('피해방지법')과 '동법 시행에 따른 관계법령 정비법('정비법')'이 마련되어[151] '25. 5. 국회 통과

ㅇ 피해방지법은 민관협력 강화, 통신정보 이용에 기반한 분석 및 취약점 정보 등의 제공을 규정

ㅇ 정비법은 경찰(경찰관직무집행법 개정)과 자위대(자위대법 개정)가 능동적 사이버 방어로서 '접근 무해화 조치'를 실시할 수 있도록 규정하고, 사이버보안 조직 및 체제를 정비함(사이버시큐리티기본법, 내각법, 내각부 설치법 개정)

150) 국회도서관, "2022년 일본 국가안보전략 3대 문서 개요"(2022), 27쪽.

151) 일부 연구는 전자를 '사이버 대처 능력 강화법'이라고 칭하거나 양자를 '능동적 사이버방어법'이라고 통칭하기도 하나 이는 정확한 제명이 아니어서 다소 혼란스러우므로 본 보고서에서는 이러한 호칭을 사용하지 않는다. 장교식, "일본의 사이버보안과 능동적 사이버방어 법제에 관한 검토", 토지공법연구 제113집(2026. 2.), 334쪽.

[그림 18] 피해방지법과 정비법의 주요 내용[152)]

□ 이러한 동향을 반영하여 이하 (1) 사이버안보기본법 (2) 경제안전보장추진법 (3) 피해방지법 및 정비법을 살펴보며, 특히 (1)의 경우 피해방지법 및 정비법 제정에 따라 변화된 부분을 중심으로 검토함

다. 법제

(1) 사이버시큐리티기본법

□ 배경

○ 일본은 ’01 제정된 고도정보통신네트워크사회형성법(일명 ‘IT기본법’)에서 최초로 사이버보안 추진체계(IT전략본부 산하 NISC)를 규정하였고, 이후 국가 차원의 사이버보안 대응능력 강화 등을 이유로 ’14 IT기본법과 별개로 사이버시큐리티기본법을 제정

152) 한국인터넷진흥원(각주 149), 44쪽.

□ 주요 내용

○ 사이버시큐리티기본법은 정부의 추진체계, 각 부처의 권한, 기본 전략 수립과 같은 사이버보안의 기본 틀과 거버넌스를 규정

○ ’25 제정된 정비법은 사이버시큐리티전략본부를 강화하고 내각에 사이버시큐리티 담당관을 신설하는 등 정부 차원의 조직 및 체제를 정비하고 있음. 주요 내용은 다음과 같음

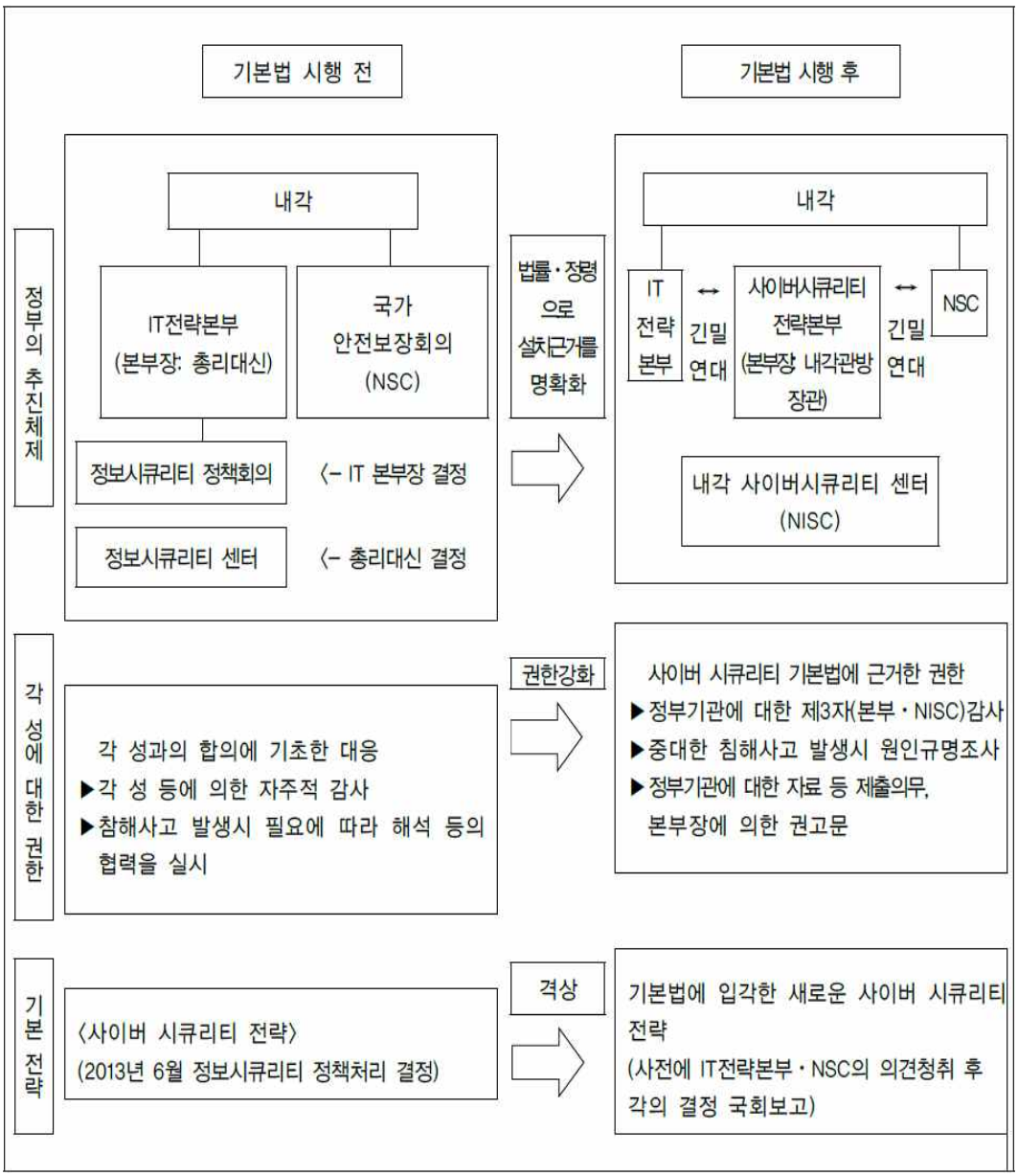
- (사이버보안전략본부 강화) 사이버시큐리티기본법에 따라 설치된 사이버시큐리티전략본부장을 내각관방장관에서 내각총리대신(즉 총리)로 개편하고, 모든 내각각료를 구성원에 포함시킴

- (내각 사이버시큐리티담당관 신설) 내각관방에 담당관(차관급 공무원)을 신설하고, 국가안보국 차장을 3명으로 늘리는 등 인력, 조직 확대

- (소관사무 추가) 내각부 사무에 “민관협력 강화 및 통신정보 이용에 관한 사무”를 추가하고, 사이버통신정보감리위원회를 설치하며, 위 사무를 관장하는 내각부 특명담당대신을 임명

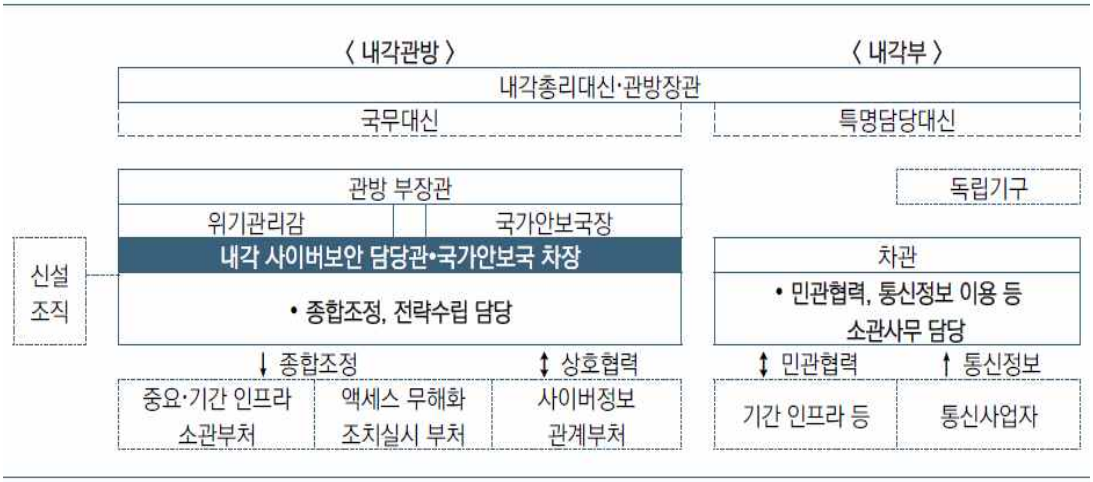
○ 사이버시큐리티 기본법 제정에 따른 체계 변화는 다음 그림과 같음

[그림 19] 사이버시큐리티 기본법 제정 전후 비교[153)]



153) 이상현, “일본의 사이버안보 수행체계와 전략”, 국가안보와 전략 제19권 제1호(2019. 3.), 134쪽.

[그림 20] 일본의 사이버보안 거버넌스('25 이후)[154]



2) 경제안전보장추진법

□ 배경

○ 일본은 미국과 중국의 대립, 러시아의 우크라이나 침공과 같은 신냉전시대가 도래함에 따라 공급망이 재편되고 기술 패권경쟁이 심화되는 상황에 대비하기 위하여 '22 경제안전보장추진법을 제정

□ 주요 내용

○ 총 7장 99개 조문이며, 이 중 사이버보안과 관련된 부분은 제3장(특정사회기반 서비스의 안정적 제공 확보)임('25. 5. 시행)

* 우리의 '정보통신기반시설보호법'과 유사

○ 제3장은 15개 핵심 인프라 서비스(전기, 가스, 석유, 수도, 철도, 금융 등)를 규정하고, 핵심 인프라 서비스를 제공하는 '특정사회기반사업자'와 해당 사업자와 거래관계에 있는 '설비공급자/위탁수임자'의 사업계획 심사 제도 도입

154) 한국인터넷진흥원(각주 149), 51쪽.

- 특정사회기반사업자와 설비공급자/위탁수임자는 사전에 사업계획을 정부에 신고하여야 함
- 정부는 신고 내용을 기초로 심사한 후 심사결과 중요 설비의 사이버 공격에 의한 방해행위 방지 조치를 강구토록 명령 가능

[표 15] 핵심 인프라 서비스 사전신고 대상 정보[155]

중요설비 공급자 및 관련설비부품 등 공급자	중요설비 유지관리 위탁수임자 및 재위탁 수임자
① 공급자에 관한 정보 ② 공급자의 의결권 중 5% 이상을 직접 보유한 자의 정보 ③ 공급자의 임원 등에 관한 정보 ④ 공급자의 외국 정부 등 대규모 매출처에 관한 정보 ⑤ 리스크 관리 조치의 실시상황 ⑥ 설비 제조장소가 소재한 국가, 지역 ⑦ 설비의 종류, 명칭, 기능	① 위탁자, 재위탁자에 관한 정보 ② 위탁자, 재위탁자의 의결권 중 5% 이상을 직접 보유한 자의 정보 ③ 위탁자, 재위탁자의 임원 등에 관한 정보 ④ 위탁자, 재위탁자의 외국 정부 등 대규모 매출처에 관한 정보 ⑤ 리스크 관리 조치의 실시상황 ⑥ 위탁, 재위탁사업의 실시장소

(3) 피해방지법 및 정비법

□ 배경

ㅇ 일본은 국가안전보장전략('22)에서[156] 사이버보안 역량을 미국, EU 등 주요국 수준 이상으로 향상시키기 위하여 (1) 민관협력 강화 (2) 통신정보를 활용한 사이버위협 탐지 (3) 공격자의 서버침입 무해화, (4) 사이버안전보장 분야의 컨트롤타워 조직 신설의 4대 목표 설정

ㅇ 위 목표 달성에 필요한 법제로서 '중요 전자계산기 부정행위로 인한 피해방지법('피해방지법')과 '동법 시행에 따른 관계법령 정비법('정비법')'이 마련되어[157] '25. 5. 국회 통과

155) KOTRA 경제통상 리포트, "일본 사이버보안 정책의 새로운 움직임과 시사점" (2024. 8.), 4쪽.

156) 국회도서관(각주 150), 27쪽.

□ 주요 내용(피해방지법)

○ 정식 명칭은 “중요 전자계산기 부정행위로 인한 피해방지법”이며,[158] 총 12장 86조로 규정

○ (신고 및 보고 의무화) ‘특정사회기반사업자’는 특정 중요 전자계산기(즉 IT 인프라 설비)를[159] 도입 · 변경하는 경우 제품명, 제조업체 명 등을 소관 부처에 신고하여야 하며, 특정 중요 전자계산기에 대한 특정 부정행위로[160] 사이버보안 피해를 입은 사건 또는 그 원인이 될 수 있는 사건을 인지한 경우에도 사고 정보, 원인 등을 소관 부처에 신고하여야 함

- 동법은 특정 중요 전자계산기를 사용하는 특정사회기반사업자를 “특별사회기반사업자”로 정의

- 소관 부처는 특별사회기반사업자의 신고/보고의무 이행을 위해 자료제출요구, 시정명령 등 가능(위반시 벌금형)

○ (취약점 정보의 제공 및 대응) 소관 부처는 특정 중요 전자계산기의 보안 취약점을 알게 된 경우 해당 기기의 공급자에게 취약점 정보 및 대응방법을 알리고 필요한 조치를 요청하거나 보고 또는 자료제출 요구를 할 수 있음

---

157) 일부 연구는 전자를 ‘사이버 대처 능력 강화법’이라고 칭하거나 양자를 ‘능동적 사이버방어법’이라고 통칭하기도 하나 이는 정확한 제명이 아니어서 다소 혼란스러우므로 본 보고서에서는 이러한 호칭을 사용하지 않는다. 장교식, “일본의 사이버보안과 능동적 사이버방어 법제에 관한 검토”, 토지공법연구 제113집(2026. 2.), 334쪽.

158) 重要電子計算機に対する不正な行為による被害の防止に関する法律

159) 「경제안전보장추진법」에 따른 특정사회기반사업자가 사용하는 전자계산기 중 사이버보안 피해를 입을 경우 같은 법 제1항에서 규정한 특정 중요설비의 기능이 정지 또는 저하될 우려가 있는 것으로서 정령으로 정하는 것 (해당 특정 중요설비의 일부를 구성하는 경우 포함)

160) 형법 제168조의2(부정명령전자기록작성 등) 제2항에 해당하는 죄, 부정액세스금지법 제2조 제4항에 따른 부정액세스행위 등.

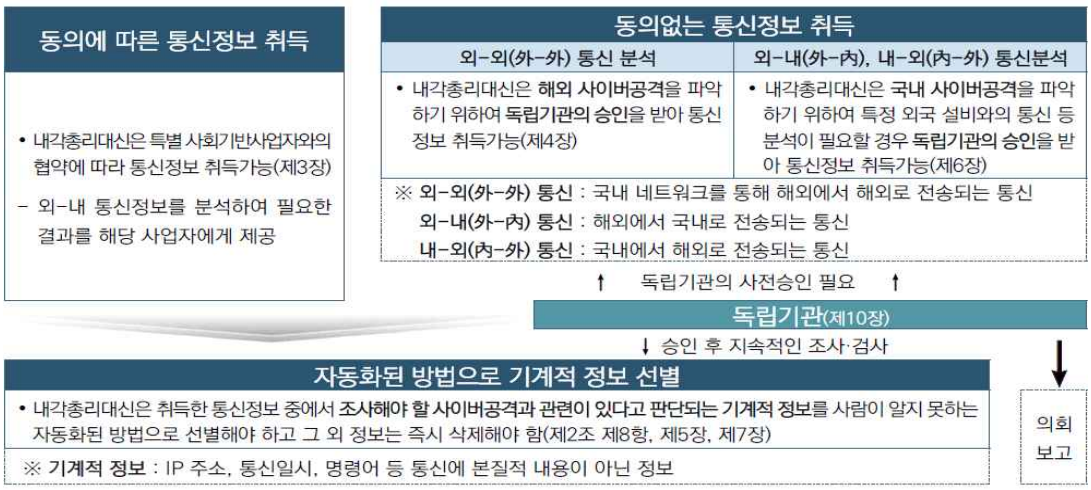
○ (정보공유 및 대책 협의회) 총리는 사이버공격 피해방지를 위하여 관계행정기관장 등으로 구성된 정보공유 및 대책협의회 설치

○ (통신정보 취득 및 이용) 정부가 사이버공격 실태 파악을 위하여 통신정보(內-內, 內-外, 外-外 불문)를 수집, 분석, 활용하고 유관기관에 제공할 수 있는 법적 기반 마련

- (협정(즉 동의)에 의한 취득) 총리는 특별사회기반사업자 또는 사업전기통신역무 이용자와 "사이버보안 확보를 위하여 필요한 분석을 실시하고 그 분석 내용 및 관련 정보를 해당 사업자(또는 이용자)에게 제공하는 내용"의 협정을 체결하고 그 협정에 따라 해당 사업자(또는 이용자)로부터 통신정보를 제공받을 수 있음

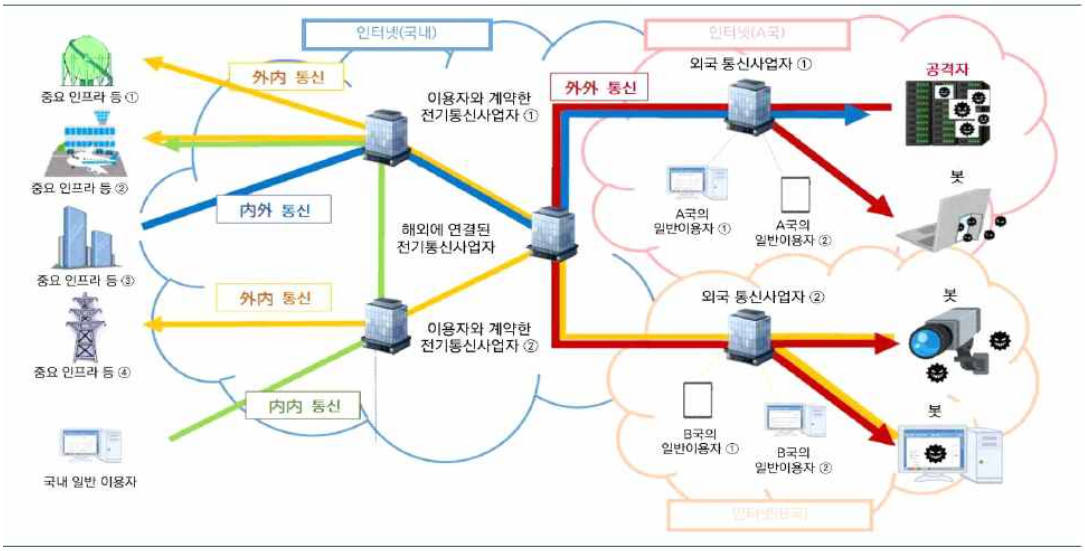
- (동의 외 취득) 총리는 사업자(또는 이용자) 동의가 없더라도 사이버공격과 관련된 통신이 의심되는 충분한 근거가 있는 경우 독립위원회(사이버통신정보감리위원회) 사전 승인을 받아 통신정보를 취득할 수 있음

[그림 21] 통신정보 취득 및 이용 개요[161]



161) 한국인터넷진흥원(각주 149), 46쪽.

[그림 22] 외국 통신정보 취득 개요[162)]



- (자동화된 정보 선별) 총리는 취득한 통신정보 중 사이버공격과 관련있다고 판단되는 기계적 정보를* 자동화된 방법으로 선별하고 그 외 정보는 즉시 삭제

  * 통신일시, IP주소, 통신량, 포트번호, 프로토콜 등(이메일 본문 및 제목, 첨부파일, 통화 내용 등 메시지 본문이 아닌 정보)

○ (통제/감독 장치) 내각부에 독립기관인 ‘사이버통신정보감리위원회’[위원장 1인(국회 동의), 위원 4인]를 설치하여 동의없는 통신정보 취득에 대한 심사/승인 및 지속적 조사/감독을 받고, 해당 내용은 의회 보고

162) 한국인터넷진흥원(각주 149), 47쪽.

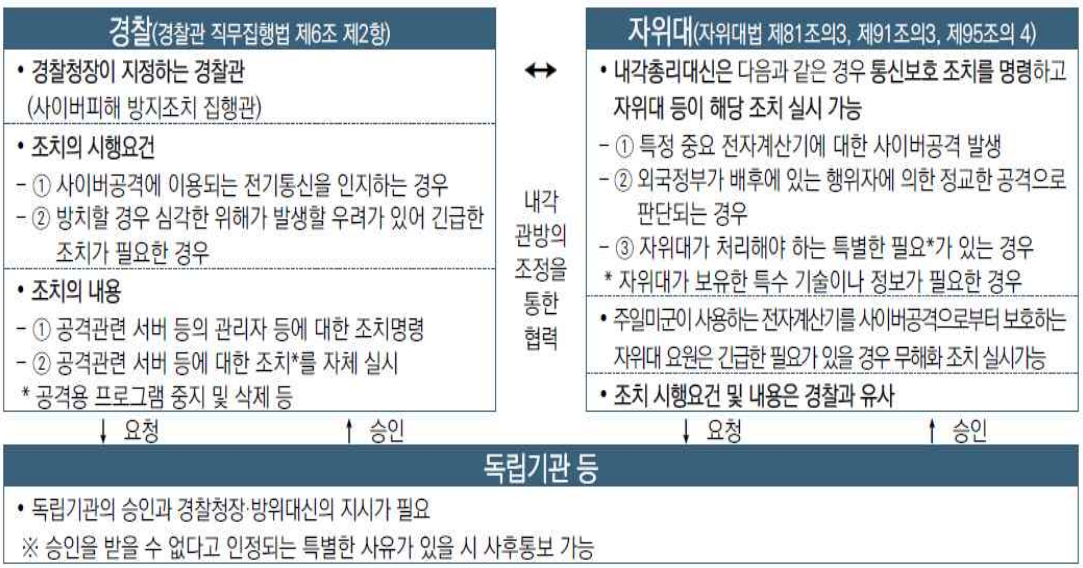
□ 주요 내용(정비법)

○ (능동적 사이버방어: 접근 확인 및 무해화 조치) 사이버공격으로 인한 심각한 피해를 방지하기 위하여 경찰 및 자위대에 '접근 확인 및 무해화 조치' 권한 부여

- 사이버공격으로 의심되는 접근을 확인, 해당 서버 등이 공격에 이용되지 않도록 (1) 설치된 공격용 프로그램 중지/제거, (2) 공격자의 해당 서버 접근 차단이 가능토록 설정 변경 조치 등을 취할 수 있도록 함

- 특히 해외 서버 등의 경우 해당국의 주권 침해에 해당하더라도 '긴급상황' 등 국제법상 법리를 원용하는 등 국제법적으로 허용되는 범위 내에서 접근 확인 및 무해화 조치 실시 허용

[그림 23] 접근 무해화 조치의 요건/절차 개요[163]



163) 한국인터넷진흥원(각주 149), 49쪽.

라. 정책

(1) 제2차 AI 기본계획('26. 7. 14.[164])

□ 배경

○ 일본은 '25. 12. 23. AI 기본계획을 수립하였으나, 그로부터 불과 7개월 후인 '26. 7. 제2차 AI 기본계획을 수립하여 내각 결의를 마침

○ 7개월 만에 기본계획을 수정한 이유는 (1) Agentic AI로의 패러다임 전환, (2) 중소기업 및 지역 사회 전반으로의 AI 확산 필요성, (3) AI에 관한 일본의 독자적 강점 분야로의 집중 때문임

□ 주요 내용

○ 4대 기본 원칙으로 (1) 위험 완화와 혁신 추진의 양립 (2) 민첩하고 유연한 대응 (3) 시도, 학습 및 반복 개선, (4) 국내외 통합 정책 추진을 제시

○ 4대 전략 축(pillars)으로 (1) AI 도입, 활용 가속화("Adopt AI"), (2) AI 개발 역량 강화("Create AI"), (3) AI 신뢰성 확보("Enhance AI Trustworthiness"), (4) AI와 공존하는 사회 구축("Collaborate with AI")를 제시

○ 특히 이 중 AI 사이버보안 위협 관련, 위험 인식 및 대응 조치에 관하여 다음과 같이 구체적으로 제시함

- Agentic AI에 기반한 자율형 사이버 공격 리스크가 가시화됨을 인식하고, 단순한 법적 규제만으로는 이에 대응이 불가능하므로 기술적 대응과 조직, 관리적 대응을 결합한 통합적 방어가 요구된다고 파악

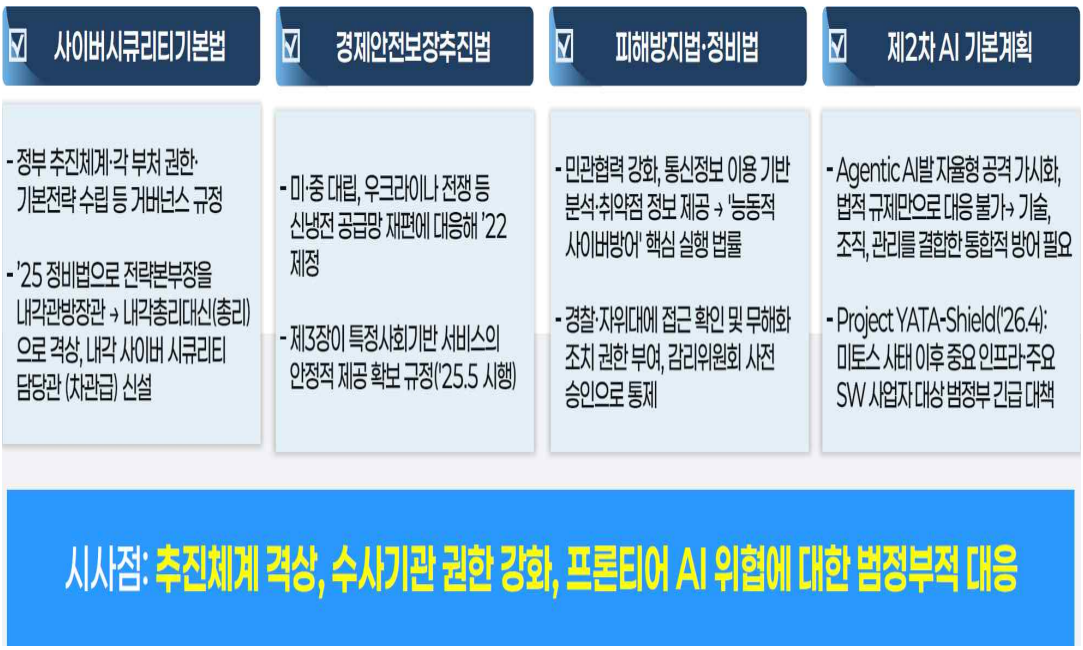
164) https://www8.cao.go.jp/cstp/ai/ai_plan/ai_plan.html

- 핵심 이행 조치로 (1) 범정부 사이버보안 패키지 “Project YATA-Shield”* 추진, (2) AI 안전연구소의 사이버보안 역할 및 역량 대폭 강화, (3) 첨단 AI 기반 사이버방어 수단 고도화, (4) 데이터 안보 및 공급망 안보의 4가지를 제시함

  * ’26. 4. 미토스 사태 이후 일본 정부가 수립한 긴급 사이버보안 대책의 명칭으로,[165] 정부가 15대 중요 인프라, 주요 SW 사업자, 정부 기관을 대상으로 보안 강화를 요구하는 주의 환기를 실시하고 관민 협력, 인재 양성, 고성능 AI를 활용한 사이버 대응 능력 향상 등을 제시

마. 시사점

[그림 24] 일본의 관련 법제도 현황 및 시사점



165) Yielding Advanced Threat Awareness with AI(AI를 통한 고도화된 위협의 가시화)의 약어이다.

□ ‘능동적 사이버 방어(Active Defense)’ 법적 근거 및 실행 기반 확립

ㅇ ‘피해방지법 및 정비법’ 제정을 통하여 경찰과 자위대에 사이버 공격용 프로그램 중지·제거 및 서버 접근 차단을 수행하는 ‘접근 확인 및 무해화 조치’ 권한을 부여

ㅇ 해외 서버 침입 시 긴급상황 등 국제법상 법리를 원용하는 공세적·선제적 방어 체계로 전환

□ 통신정보 수집·분석 권한 부여 및 통제 기구 마련

ㅇ 사이버 공격 실태 파악을 위해 통신정보(내-외, 외-외 등)를 수집 및 자동화된 방법으로 기계적 선별·분석할 수 있는 권한을 내각총리대신에게 부여

ㅇ 국회 동의로 임명되는 독립 기구인 '사이버통신정보감리위원회'의 사전 승인 및 지속적 검사·감시를 받도록 정교하게 설계

□ 미토스와 같은 사이버공격 AI 출현에 따른 대응 비상 안보 패키지 및 민첩한 거버넌스(Agile Governance) 구축

ㅇ 추론형·에이전트형 AI 기반 자율 공격 가시화에 대응해 범정부 안보 패키지인 ‘Project YATA-Shield’를 출범시킴

ㅇ 기술 패러다임 변화에 맞춰 AI 기본계획을 7개월 만에 수립·수정(’26. 7.)하는 유연한 민첩 거버넌스 구축

5. 소결

□ 주요국(미, EU, 영, 일)은 AI로 인한 새로운 사이버보안 위협에 대응하고자 각국 실정에 부합하는 다양한 법제도 및 정책을 추진 중

□ 주요국 사례에서 도출할 수 있는 공통적 시사점은 다음과 같음

ㅇ 패러다임 전환: 사후 경계 방어 → 선제적 방어 및 복원력(Resilience) 제고

- 주요국은 기존 시그니처/경계 중심 방어의 한계를 인정하고, 선제적 방어와 신속한 복원력 확보를 최우선 과제로 재편

ㅇ 규제 대상의 정밀 조정: 제품 및 서비스 전 수명주기(Lifecycle) 보안 규제, 서드파티 공급망 보안 법정화 등 AI 보안 위협 대응에 필요한 규제는 확대하되, 불필요한 기존 규제는 간소화

- 단일 시스템 규제에서 벗어나 AI 모델, 학습 데이터, IoT 제품, 데이터 센터, 관리형 서비스(MSP) 등 공급망 전체/라이프사이클 전체 규율 체계로 진화 중

- 전통적 위협에는 적합하나 AI發 위험에는 신속, 정확한 대응이 어렵고 오히려 AI 발전을 저해하는 규제는 간소화 혹은 제거

ㅇ 민간 참여 촉진: 면책 조항(Safe Harbor) 명문화, 인센티브 설계

- AI 위협에 대응하기 위한 민간의 위협 정보 공유 촉진은 규제가 아닌 '법적 안전망 제공'(미국 등)과 인센티브 제공(영국 등)을 통해 달성

ㅇ 거버넌스 고도화

- AI발 새로운 사이버보안 위협에 효율적이고 신속하게 대응할 수 있도록 범정부적 사이버보안 대응 거버넌스를 정비 중

Ⅳ. 우리의 관련 입법·정책

□ 이 장에서는 AI로 인한 새로운 보안 위협과 관련된 우리나라의 입법 및 정책을 ’26. 8. 기준으로 살펴봄

1. 법제

가. 개요

□ 우리나라의 사이버보안 법제는 한마디로 “분야별, 영역별 수평적 규율 체계”라 할 수 있음

ㅇ 미국, EU, 영국, 일본과 같이 범국가적 차원의 사이버보안 총괄 조직이나 거버넌스 체계를 규율한 법률은 없음

- (가칭)국가사이버안보법(또는 사이버안보기본법) 제정이 2010년대 중반부터 거론되었으나 역사적, 정치적 이유 등으로 계속 무산되었으며, 현재도 국회 계류 중[166]

□ 공공 영역은 보안이 아닌 ‘안보’의 일환으로 「국가정보원법」 및 「사이버안보 업무규정」(대통령령)이 규율

ㅇ 사이버안보 업무규정상 공공 영역의 사이버보안은 ‘안보’ 차원에서 국가정보원이 총괄하나, 각 부처는 개별 법령에 근거하여 분야별로 사이버보안 활동을 수행 중

166) 의안번호 제2213722호 사이버안보 기본법안(김상훈의원 대표발의), 의안번호 제2211450호 국가사이버안보법안(유용원의원 대표발의).

□ 민간 영역(엄밀히 말하면 침해사고[167] 대응)은 「정보통신망 이용촉진 및 정보보호 등에 관한 법률(이하 '정보통신망법')」에 따라 과학기술정보통신부(이하 '과기정통부')가 사이버보안 사무를 수행하나, 금융 영역은 「전자금융거래법」에 따라 금융위원회 소관임

□ 미국, EU, 일본의 Critical Entity 보호법제에 비견되는 「정보통신기반시설보호법」 역시 사이버보안 주요 법률로 작동함

□ '26. 1. 22. 부터 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 '인공지능기본법')이 시행되고 있으므로 이 역시 살펴봄

나. 거버넌스

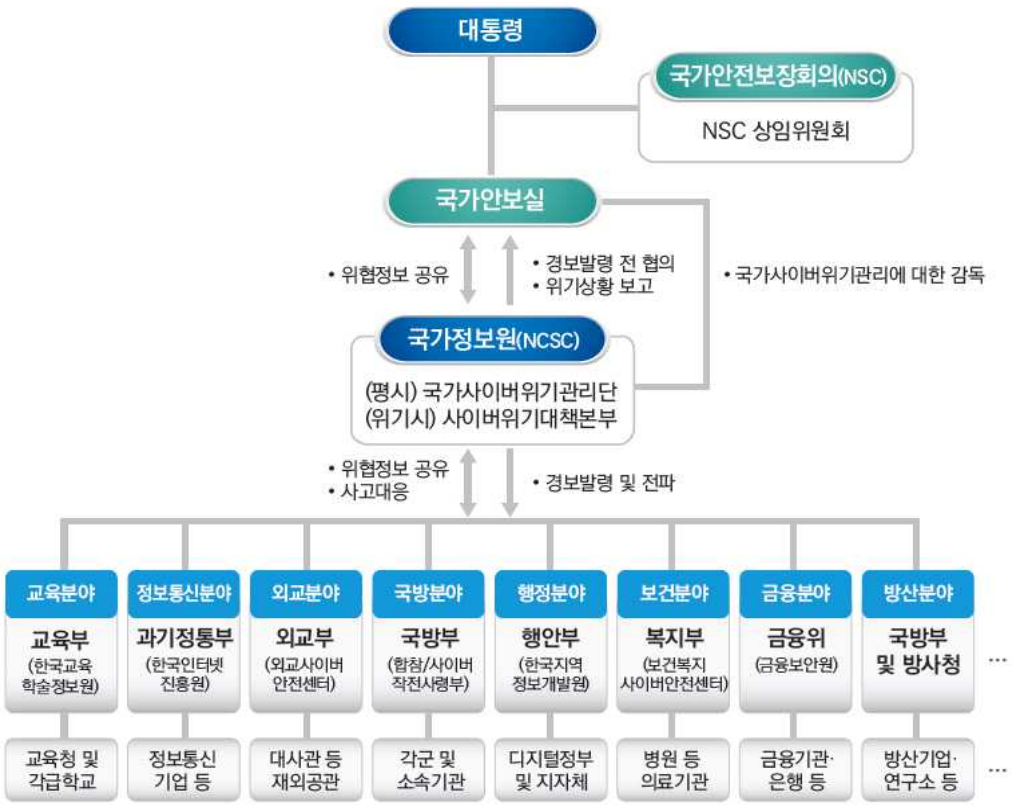
□ '26 기준 대통령 소속 국가안보실이 최상위 컨트롤타워이고,[168] 국가정보원이 국가 차원의 사이버보안센터(NCSC)를 운영하며, 각 부처가 소관 사무별로 사이버보안 활동을 수행하는 체계

---

167) 정보통신망법 제2조 제1항 제7호에 따르면 "침해사고"란 다음 각 목의 방법으로 정보통신망 또는 이와 관련된 정보시스템을 공격하는 행위로 인하여 발생한 사태를 말한다.
가. 해킹, 컴퓨터바이러스, 논리폭탄, 메일폭탄, 서비스거부 또는 고출력 전자기파 등의 방법
나. 정보통신망의 정상적인 보호·인증 절차를 우회하여 정보통신망에 접근할 수 있도록 하는 프로그램이나 기술적 장치 등을 정보통신망 또는 이와 관련된 정보시스템에 설치하는 방법

168) 다만 실제로 국가안보실이 사이버보안 컨트롤타워를 하는지는 의문이며, 후술함.

[그림 25] 국가 사이버보안 수행체계 현황169)



다. 주요 법령

(1) 사이버안보 업무규정

□ 개요

○ 국가정보원법 제4조 제1항에 따른 국가정보원의 직무 중 사이버안보(사이버보안 포함) 업무의 수행에 필요한 사항을 규정하여, 공공부문의 사이버안보 거버넌스 체계를 규정함

169) 국가정보원 등, “2025 국가정보보호백서”(2025. 6.), 10쪽.

□ 주요 내용

○ 국가정보원의 업무(제3조, 제3조의2)

- 국가정보원은 사이버안보정보 업무와 사이버보안 업무를 담당

- 사이버보안 업무는 중앙행정기관, 지방자치단체 및 대통령령으로 정하는 공공기관(이하 '중앙행정기관등)* 대상 사이버공격 및 위협에 대한 예방·대응 및 관련 기획·조정 업무를 말함

  * 공공기관운영법에 따른 공공기관, 지방공기업법에 따른 지방공사 및 지방공단, 국공립학교, 연구기관 등

○ 국가사이버안보센터(NCSC) (제4조)

- 사이버안보 업무규정에 근거하여 국가정보원장은 NCSC를 설치[170]

- NCSC는 민관합동 통합대응체계 전담조직을 둘 수 있으며, 사이버안보 자문단을 설치·운영할 수 있음

○ 사이버안보 정보공유시스템 구축, 운영(제6조)

- 국가정보원장은 사이버안보 관련 정보를 배포·공유하기 위하여 정보공유시스템을 구축·운영할 수 있음

- 이에 따라 NCSC가 정보공유시스템[NCTI(National Cyber Threat Intelligence)]를 운영 중

○ 사이버보안 예방 조치(제9조 등)

- 국가정보원장은 중앙행정기관등의 장이 시행하는 정보화사업에 보안성검토를 실시하고(제1항), 중앙행정기관등이 도입, 운영하거나 이용하는 정보보호시스템 및 클라우드 서비스의 보안 적합성을 검증(제2항)

170) https://www.ncsc.go.kr:4018/

- 중앙행정기관등의 장은 사이버보안 교육, 훈련, 실태평가를 실시하며(제10조, 제11조, 제13조), 국가정보원장은 중앙행정기관등에 대하여 통합 훈련을 실시하고 그 결과 시정조치 요구 가능(제11조 제2항, 제4항)

ㅇ 사이버보안 사고 대응(제14조~제16조)

- 국가정보원장은 중앙행정기관등에 대한 통합보안관제체계를* 구축, 운영(제14조 제1항), 중앙행정기관등의 장은 통합보안관제체계와 연계된 보안관제체계 설치·운영(타 기관의 보안관제센터 활용 포함)(제14조 제2항)[171]

  * NCSC가 통합보안관제 기능 수행

- 국가정보원장은 중앙행정기관등에 대한 사이버공격·위협에 대하여 단계별 경보 발령(군 관련 기관의 경우 국방부장관이 발령)(제15조 제1항)

- 국가정보원장은 중앙행정기관등에 대한 사이버보안 사고 발생시 공격 주체 규명, 원인 분석, 피해 내역 확인 등을 위한 조사 실시(군 관련 기관인 경우에는 국방부장관이 실시)(제16조 제1항)

(2) 정보통신망법

□ 개요

ㅇ 정보통신망법은 기본적으로 정보통신망, 정보통신서비스, 정보통신서비스 제공자를 규율 대상으로 함

171) 2025. 6. 기준 45개 부문보안관제센터가 운영 중이다. 국가정보원 등, “2025 국가정보보호백서”(2025. 6.), 81쪽.

○ 정보통신망법은 “정보통신망 또는 이와 관련된 정보시스템을 공격하는 행위로 발생한 사태”를 ‘침해사고’로 정의하고, 제6장(정보통신망의 안정성 확보)에서 침해사고 예방 및 대응에 관한 각종 규정을 두고 있음

□ 주요 내용

○ 침해사고 대응 주체

- 과기정통부장관은 침해사고 정보 수집 전파, 침해사고 예보·경보·통지, 긴급조치 등 침해사고 대응조치 업무를 수행하며, 그 일부 또는 전부를 KISA에게 위임·위탁할 수 있음(제48조의2 제1항)

  * 침해사고의 신속, 체계적 대응을 위하여 과기정통부에 2030. 12. 3.까지 한시적으로 침해사고조사심의위원회를 둠(제48조의2 제7항)

- 과기정통부장관은 정보통신망연결기기등과 관련된 침해사고 발생시 원인 분석, 취약점 점검, 기술 지원, 피해 확산 방지 등 조치를 취할 수 있음(제48조의5)

○ 침해사고 대응 체계

- (정보 공유) 주요정보통신서비스 제공자 등은 침해사고의 유형별 통계 등 침해사고 관련 정보를 과기정통부장관 또는 KISA에 제공하여야 하고(제48조의2 제2항), KISA는 이를 분석하여 과기정통부장관에 보고하여야 함(제48조의2 제3항)

  * 정보 공유 시스템으로 KISA의 C-TAS(Cyber Threat Analysis & Sharing) 운영 중[172]

172) https://ctas.krcert.or.kr/index

- (신고) 정보통신서비스 제공자는 침해사고 발생시 이를 알게 된 때부터 24H 이내에 과기부장관이나 KISA에 신고해야 하고(제48조의3 제1항), 지체없이 그 사실을 이용자에게 통지하여야 함(제48조의3 제4항)
- (조치) 과기부장관이나 KISA는 신고를 받으면 침해사고 정보 수집 전파, 침해사고 예보·경보·통지, 긴급조치 등 적절한 조치를 하여야 함(제48조의3 제2항)
- (원인 분석) 과기부장관은 정보통신서비스 제공자의 정보통신망에 침해사고가 발생하였거나 침해사고조사심의원회가 조사가 필요하다고 인정하는 경우 침해사고의 발생 여부 및 원인을 분석하고 피해 확산 방지, 사고대응, 복구 및 재발 방지를 위한 대책을 마련하여 해당 정보통신서비스 제공자에게 필요한 조치를 이행하도록 명할 수 있음(제48조의4)

(3) 전자금융거래법

□ 개요

○ 전자금융거래법은 전자금융기반시설에 대한 침해사고에 관하여 정보통신망법의 특칙으로서 금융위원회를 대응 주체로 규정

□ 주요 내용

○ 취약점 분석·평가

- 금융회사 및 전자금융업자는 전자금융시설의 조직, 시설, 내부통체, 침해사고 대응조치 등에 관한 사항을 분석·평가하고 이를 금융위원회에 보고해야 함(제21조의3 제1항). 금융위원회는 그 결과 및 보완조치 이행실태를 점검함(제21조의3 제3항)*

* 실제 사무는 금융감독원/금융보안원이 위탁받아 수행함

○ 침해사고 통지

- 금융회사/전자금융업자는 침해사고 발생시 금융위원회에 지체없이 통지하고,* 원인 분석 및 피해 확산 방지에 필요한 조치를 취해야 함(제21조의5)

  * 정당한 사유 없이 침해사고 발생 사실을 안 때부터 24H를 지나서는 안됨(전자금융감독규정 제37조의4)

○ 침해사고 대응 주체

- 금융위원회는 전자금융시설에 대한 침해사고 정보 수집·전파, 침해사고 예보·경보, 침해사고 긴급조치 등을 수행(제21조의6)

(4) 정보통신기반 보호법

□ 개요

○ 사이버보안 공격으로부터 국가 안전보장과 경제사회 안정확보를 위하여 공공·민간 보유 시설 중 국가 중요 인프라만을 특정하여 기존의 사이버보안 대응체계보다 강화된 체계를 적용하는 법률임

  * 美 CIRCIA, EU NIS2 지침, 英 CSRB, 日 경제안전보장추진법에 비견

□ 주요 내용

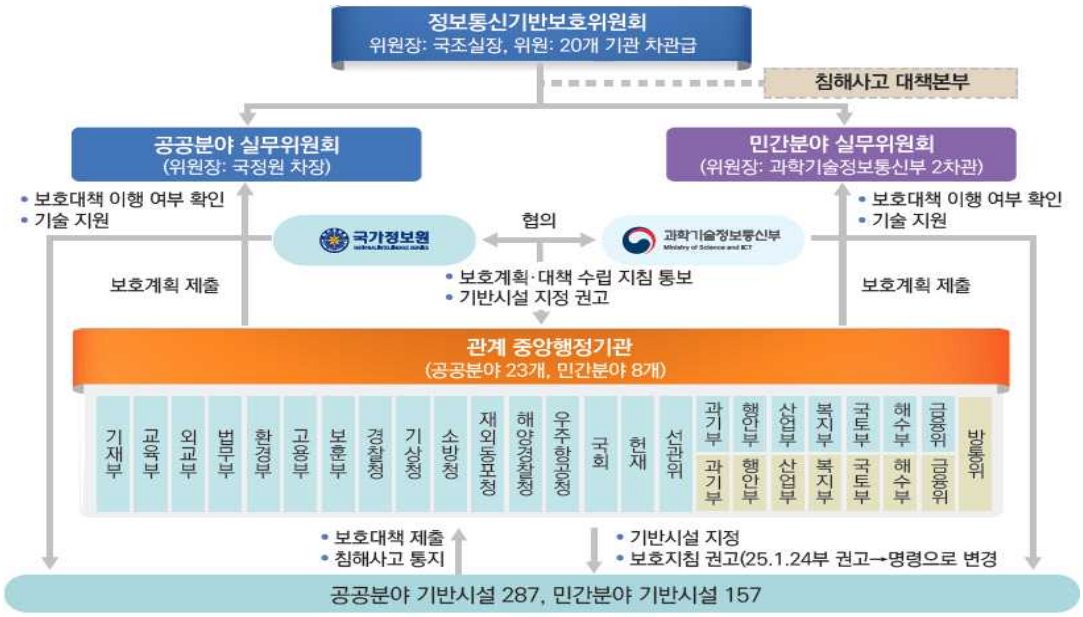
○ 개념

- 정보통신기반시설: 국가안전보장·행정·국방·치안·금융·통신·운송·에너지 등의 업무와 관련된 전자적 제어·관리시스템 및 정보통신망(제2조 제1호)

- 주요정보통신기반시설: 중앙행정기관의 장이 정보통신기반시설 중 해당 시설 관리기관이 수행하는 업무의 국가사회적 중요성, 그 업무의 정보통신기반시설에 대한 의존도, 다른 정보통신기반시설과의 상호연계성, 침해사고 발생시 국가안전보장과 경제사회에 미치는 피해 규모 및 범위 등을 고려하여 정보통신기반보호위원회의 심의를 거쳐 지정한 시설(제8조)

  * 과기정통부장관과 국가정보원장은 특정 정보통신기반시설을 주요정보통신기반시설로 지정할 것을 권고할 수 있음(제8조의2)

[그림 26] 주요정보통신기반시설 보호 추진 체계('25. 6. 기준)[173]



○ 보호체계

- 주요정보통신기반시설의 관리기관은 정기적으로 취약점을 분석, 평가하고(제9조),* 그에 따라 보호대책을 수립, 시행해야 하며(제5조), 이를 관계중앙행정기관의 장에게 제출해야 함(제5조)

173) 국가정보원 등, "2025 국가정보보호백서"(2025. 6.), 136쪽.

* 실무적으로 KISA, ETRI 산하 국가보안연구소('국보연'), 정보보호산업법에 따라 지정된 정보보호 전문서비스 기업이 수행함(제9조 제4항 각호)

- 관계중앙행정기관의 장은 제출받은 보호대책을 종합, 조정하여 소관분야의 보호계획을 수립, 시행해야 함(제6조)

[그림 27] 주요정보통신기반시설 보호 업무 절차[174]

○ 침해사고 대응

- (통지) 관리기관의 장은 침해사고 발생시 관계 행정기관, 수사기관, KISA 등에 그 사실을 통지해야 함(제13조 제1항)

- (조치) 관계 행정기관, 수사기관, KISA 등은 침해사고의 피해확산 방지와 신속 대응을 위하여 필요한 조치를 취해야 함(제13조 제2항)

174) 국가정보원 등, "2025 국가정보보호백서"(2025. 6.), 141쪽.

- (복구) 관리기관의 장은 침해사고 발생시 소관 주요정보통신기반시설의 복구 및 보호에 필요한 조치를 신속히 취해야 함(제14조 제1항)*

  * 이를 위해 필요한 경우 관계중앙행정기관의 장 또는 KISA가 지원할 수 있음(제14조 제4항)

ㅇ 정보공유·분석센터(ISAC): 취약점 및 침해요인과 그 대응방안에 대한 정보 제공, 침해사고 실시간 경보·분석체계 운영을 위하여 분야별로 ISAC을 구축·운영할 수 있음(제16조)

(5) 전자정부법

□ 개요

ㅇ 행정업무의 전자적 처리를 위한 기본원칙, 절차 및 추진방법 등을 규정함으로써 전자정부를 효율적으로 구현하고, 행정의 생산성, 투명성 및 민주성을 높여 국민의 삶의 질을 향상시키는 것을 목적으로 함

□ 주요 내용

ㅇ (정보시스템 안정성·신뢰성 제고) 행정기관의 장은 전자정부 구현에 필요한 정보통신망 및 행정정보 등의 보안대책을 수립, 시행하여야 하고(제56조), 장애 예방·대응·복구 계획을 수립하며(제56조의2), 장애 발생시 중앙사무관장기관의 장(행정부의 경우 행정안전부장관)에게 지체없이 통보하여야 함(제56조의5 제1항)

ㅇ (통합관제·대응) 중앙사무관장기관의 장은 정보시스템종합상황실(SOC)을 설치·운영할 수 있으며(제56조의5 제2항), 장애를 통보받거나 인지한 경우에는 필요한 조치를 취해야 하고(제56조의5 제3항), 장애사후관리를 실시해야 함(제56조의5 제4항)

ㅇ (보안적합성 검증) 행정기관의 장은 정보보호시스템, 네트워크장비 등 보안기능이 탑재된 IT 장비를 도입할 경우 국가정보원에 보안적합성 검증을 신청해야 하며, 검증 결과 발견된 보안 취약점을 제거한 후 해당 장비를 운용하여야 함(제56조 제3항*)

* 보안적합성 검증의 근거는 전자정부법 외에 사이버안보 업무규정 제9조 제3항에서도 찾을 수 있음

(6) 인공지능기본법

□ 개요

ㅇ 인공지능의 건전한 발전과 신뢰 기반 조성에 필요한 기본적인 사항을 규정함으로써 국민의 권익과 존엄성을 보호하고 국민의 삶의 질 향상과 국가경쟁력을 강화하는 데 이바지함을 목적으로 함

□ 주요 내용(사이버보안 한정)

ㅇ 인공지능기본법은 고영향 인공지능(사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능시스템 중 동법 제2조 제4호 각목에서 정한 영역에서 활용되는 인공지능시스템을 말함; EU의 고위험 인공지능시스템에 비견되는 개념임)에 관하여 사업자의 책무 및 영향평가를 규정하나, 해당 사항에는 '사이버보안'에 관한 명확한 문구가 없음

- (사업자의 책무) 위험관리방안, 설명 방안, 이용자 보호 방안, 사람의 관리·감독, 안정성·신뢰성 확보 등의 수립, 운영 요구(제34조)하나, 사업자의 책무에 '사이버보안'이 명시되지 않음

* 동법 제34조 제3항 및 동법 시행령 제27조는 특정 제품에 한정하여 해당 제품 규율 법률에서의 인증을 받은 경우 위험관리방안이나 안정성·신뢰성 확보 등이 수립, 운영된 것으로 간주하나(예컨대 자동차관리법에 따른 자동차 사이버보안 관리체계 인증을 받은 경우), 이는 법령에 열거된 특정 제품에만 적용되고, 이에 포함되지 않는 고영향 인공지능 시스템이 다수이어서 문제됨

- (영향평가) 제35조는 인공지능사업자가 고영향 인공지능 제품 또는 서비스 제공시 영항평가를 할 수 있도록 '임의 영향평가' 제도를 규정하고 있으나(단 공공의 경우에는 '27. 2. 28. 시행되는 데이터기반행정법 제32조에 따라 강제됨), 동법 시행령 제28조에서 역시 '사이버보안'이 명시되지 않음

ㅇ 한편, 인공지능기본법은 고영향 인공지능 시스템이 아니더라도 학습 누적 연산량이 대통령령으로 정한 기준(현재 10^{26} FLOPS) 이상인 인공지능 시스템의 경우 인공지능사업자가 위험의 식별·평가 및 완화 및 위험관리체계 구축 의무를 규정하고 있으나(제32조), 이 경우에도 '사이버보안'은 법문상에는 명시되어 있지 않음

라. 인증 체계

□ 현재 사이버보안 관련 인증 체계가 여러 법령에 산재되어 있어 다소 복잡하므로 별도의 목차를 두어 정리함

○ ISMS 인증[175]

- 법적 근거: 정보통신망법 제47조(정보보호 관리체계의 인증)
- 주체: 과기정통부장관
- 인증 범위: '관리체계'에 관한 인증이므로 '서비스' 단위로 그와 관련된 관리체계(조직, 인력, 관리, 자산 등)를 인증받음; 즉 특정 정보통신기기 단위가 아님
- 필수 인증 대상: ① 주요정보통신서비스 제공자, ② 집적정보통신시설 사업자(예컨대 IDC), ③ 전년 매출액(또는 세입)이 1,500억 원 이상이거나 정보통신서비스 부문 전년도 매출액이 100억원 이상 또는 전년도 일일평균 이용자수 100만명 이상으로서 대통령령으로 정하는 기준 해당자

○ 정보통신망연결기기 보안인증(IoT 인증) * '26. 6. 기준 임의인증

- 법적 근거: 정보통신망법 제48조의6(정보통신연결기기등에 관한 인증)
- 주체: 과기정통부장관
- 인증 범위: IoT 제품 및 이와 연동되는 모바일 앱[176]
- 인증 등급: Lite, Basic, Standard 3개 등급

---

175) ISMS-P 인증은 사이버보안 그 자체 인증이라기 보다는 개인정보 보호 인증이므로 생략한다.

176) https://www.kisa.or.kr/1050608

ㅇ 클라우드 서비스 보안인증(CSAP)

- 법적 근거: 클라우드컴퓨팅법 제23조의2
- 주체: 과기정통부장관
- 인증 범위: 서비스 단위
- 특이 사항: 원칙적으로는 임의인증이지만, 공공조달시에는 국가정보원 지침 및 가이드라인상[177] 원칙적으로 CSAP 인증을 받지 않은 민간 클라우드 서비스를 도입해야 하므로 사실상 강제됨

ㅇ 보안적합성 검증

- 법적 근거: 전자정부법 제56조 제3항, 사이버안보 업무규정 제9조 제3항
- 주체: 국가정보원장
- 검증 범위: 행정기관이 사용하는 정보보호시스템, 네트워크장비 등 보안기능이 탑재된 IT 장비
- 특이 사항: 실무적으로 보안적합성 검증시 CC인증을 요구하므로, 국제 보안 상호 인증 제도인 CC인증 제도와 연계됨 * CC인증의 실제 업무는 국보연이, 정책 업무는 과기정통부가 담당하는 이원적 체계

177) 국가사이버보안기본지침 제24조의3, 국가 클라우드 컴퓨팅 보안 가이드라인 (2023. 1.) 41쪽.

마. 사이버보안 실제 대응 체계

□ 라.와 마찬가지로 사이버보안 실제 대응 체계 또한 여러 법률에 파편화되어 규정되어 있어, 별도 목차를 두어 정리함

ㅇ 범국가적 차원의 국가사이버안보센터(NCSC): 국가정보원 산하 NSCS의 단일 기구

ㅇ 보안관제(SOC)

- 공공: 각 기관이 개별 SOC 운영하되, NCSC가 통합SOC 역할 담당
- 민간: 각 기관이 개별적으로 SOC 기능 수행
- 금융: 금융보안원과 금융감독원이 통합보안관제 수행

ㅇ 정보공유·분석센터(ISAC)

- 각 영역별 개별 ISAC 운영(근거: 정보통신기반 보호법 제16조)

ㅇ 긴급대응팀(CERT)

- 공공: NCSC가 CERT 기능 수행
- 민간(금융 제외): KISA 산하 KrCERT[178]
- 금융: 금융보안원이 CERT 기능 수행

ㅇ 사이버위협 정보 공유

- 국정원은 공공 대상으로 위협 정보 공유 시스템 NCTI를 운영 중이며(근거: 사이버안보 업무규정 제6조), 민간과의 정보 공유를 위하여 KCTI를 운영 중(민간이 공공에 정보를 공유해야 할 법적 의무는 없으며, 상호 협약에 따라 이루어짐)

178) https://www.boho.or.kr/

- KISA의 KrCERT는 한국형 NVD(KNVD),[179] 사이버 위협정보 분석·공유 시스템(C-TAS)[180] 각 운영(근거: 정보통신기반 보호법 제16조)

  * KNVD가 특정 SW/HW의 취약점을 데이터베이스화한 정적인 사전이라면, C-TAS는 실제로 사이버공격에 사용되는 실시간 위협정보를 동적으로 분석, 공유

- 금융 분야에서는 금융보안원 F-CTI(Finance Cyber Threat Information)를 통하여 사이버 위협 정보 공유(근거: 정보통신기반 보호법 제16조)

[그림 28] 사이버위협 대응체계

항목	내용
NCSC	• 국정원 산하 단일 기구 (범국가적 차원)
SOC (보안관제)	• 공공: 개별 SOC + NCSC 통합관제 • 민간: 개별 운영 • 금융: 금보원·금감원 통합관제
ISAC (정보공유분석)	• 정보통신기반보호법 제16조 근거, 영역별 개별 운영
CERT (긴급대응팀)	• 공공: NCSC • 민간(금융제외): KISA 산하 KrCERT • 금융: 금융보안원
위협정보공유	• 공공: NCTI (국정원, 사이버안보 업무규정 제6조) • 민간: KCTI (국정원, 법적 근거 없이 상호 협약) • KISA KrCERT: KNVD·C-TAS • 금융: F-CTI

179) https://knvd.krcert.or.kr/dashboard

180) https://ctas.krcert.or.kr/index

2. 정책

가. 인공지능 기본계획(AI행동계획)

□ 배경

○ 글로벌 AI 경쟁이 날로 심화되는 가운데 AI를 새로운 성장엔진으로 삼아 산업을 혁신하고 국민의 삶을 개선하며 인류 공동 번영에 기여하고자 ’26. 3. 국가 전략으로 수립

* 3대 정책 축, 12대 전략 분야, 99개 과제

□ 주요 내용(사이버보안 한정)

○ 사후 대응에서 사전적·상시적 취약점 공개 제도(CVD/VDP)로의 패러다임 전환

○ AI 기반의 사이버보안 체계를 구축하고, AI 보안 기술 경쟁력 강화 및 보안 전문 인재 양성

○ AI 침해사고 대응 체계 및 K-사이버보안 LLM 구축

○ 물리적 망분리가 아닌 N2SF(국가망 보안 체계)로의 전환

○ 국방 K-RMF 및 제로트러스트 보안 모델 도입

○ AI 신뢰성(Security for AI) 및 AI 활용 방어(AI for Security) 동시 강화

나. 범정부 정보보호 종합대책

□ 배경

○ ’25 3대 이동통신사 개인정보 유출, 카드 및 금융회사 해킹 등 사회 주요 분야에서 연쇄적으로 개인정보 유출 및 해킹 사고가 발생함에 따라 국민적 불안감이 급격히 증대

ㅇ 국민의 불안감을 해소하고 국가 전반의 정보보호 역량을 강화하기 위하여 관계부처 합동으로 종합 대책 수립('25. 10.)

□ 주요 내용(AI 사이버보안 위협 한정)

ㅇ ISMS/ISMS-P 인증을 현장 심사 중심으로 전환하고 사후 관리를 강화하며, 화이트해커 활용 상시 점검 체계 구체화

ㅇ 해킹 사고 시 소비자의 입증책임 부담을 완화하고, 유출 과징금을 피해자 지원 기금으로 활용하는 방안 검토

ㅇ 해킹 정황 확보 시 기업 신고 없이도 정부가 직권 조사할 수 있도록 권한 확대하고, 과태료, 과징금 상향 및 징벌적 과징금, 이행강제금 등 도입

ㅇ 공공부문 정보보호 예산·인력 확보 비율 의무화, 책임관 직급 상향(국장급 → 실장급), 경영평가 배점 확대(0.25 → 0.5점)

ㅇ 정보보호 공시 의무 대상 전체 상장사로 확대(약 2,700개사), 보안 역량 등급화 공개, CEO 보안 책임 명문화 및 CISO·CPO 권한 대폭 강화

ㅇ 획일적 물리적 망분리를 데이터 중심 보안 체계로 전환, 공공 IT 제품의 SBOM 제출 제도화('27년까지)

ㅇ AI 에이전트 보안 플랫폼 등 차세대 보안 기업 집중 육성(연 30개사), 화이트해커 연 500명 양성

ㅇ 정보통신기반시설보호위원회를 통한 주요정보통신기반시설 지정 확대 및 침해사고대책본부(국가사이버위기관리단) 활성화

ㅇ 해킹 사고 조사 과정의 일괄(One-Stop) 신고 및 대응 체계화, 민·관·군 합동 위협 예방·대응 협력 강화

3. 시사점

[표 16] 해외 주요국과 우리의 법제도 비교

분류	총괄 거버넌스	중요 인프라	침해사고 대응	민관 정보공유	SBOM	복원력	AI특화 사이버 보안	능동적 방어
미국	○	○ (시설 기준)	○	○ (자발적+ 면책)	△ (연성 규범)	○	△ (연성 규범)	×
EU	○	○ (조직 기준)	○	△ (자발적)	○	○	○	×
영국	○	○ (조직 기준)	○	△ (자발적)	△ (연성 규범)	○	×	×
일본	○	○ (시설 기준)	○	○	×	×	×	○
대한민국	△ (형식적 총괄/ 분야별 분절)	○ (시설 기준)	○	△ (자발적)	× (가이드 라인 단계)	×	×	×

□ Ⅲ.에서 살펴본 주요국 사례와 비교할 때, (1) 사이버보안 위협에 대응하는 거버넌스 체계 존재 (2) 사이버보안 위협 대응에 필요한 기본적인 법체계(Critical Entity, 사이버위협 정보 공유, 사이버보안 사고 발생시 조사 및 대응, 인증 등) 존재 (3) AI로 인한 기회와 위험을 동시에 인식하고 이에 대응코자 하는 국가 차원의 전략이 존재한다는 점에서는 공통됨

□ 그러나 각론에서는 다음과 같은 차이점이 존재함

- ㅇ 미, EU, 영, 일 등과 달리 공공/민간/금융 등 영역별/기관별/소관사무별 다소 분절된 거버넌스 체계를 보여줌(이에 따라 법령 또한 파편화되고, 정책 추진에서도 다소 혼란스러움)
- ㅇ 소프트웨어/AI 명세서, 제로트러스트, 전 생명주기 보안 설계 등 AI發 사이버보안 위험 대응에 적합한 기술적 사항의 제도화 미흡
- ㅇ 복원력 강화 등 AI 보안 위협 대응에 필요한 규제는 다소 미흡한 반면, 전통적 인증 체계나 사이버보안 위협 사고 신고 및 대응과 같은 기존 규제 개선 노력은 다소 미흡
- ㅇ 민간의 자발적 참여를 이끌어낼 유인(incentive) 부족

□ 요컨대, 우리 법제도 및 정책은 AI 출현 이전의 사이버보안 위협 대응에는 적절할 수 있으나, AI 출현 이후(특히 ’26. 4. 미토스 사태 이후) 패러다임 전환이 이루어진 사이버보안 위협에 대응하기에는 다소 미흡하며, 개선이 필요하다고 생각됨

□ 구체적 개선 필요 사항은 Ⅴ(결론 및 제언).에서도 언급하므로 상세한 내용은 생략하며, 몇 가지 점만 언급하면 다음과 같음

① 범국가적 총괄 거버넌스의 부재

- ㅇ 미국, EU, 영국, 일본 등 주요 선진국이 단일 총괄 기구 및 통합 법제를 정비한 것과 달리, 한국은 공공(국정원/사이버안보 업무규정), 민간(과기정통부/정보통신망법), 금융(금융위/전자금융거래법)으로 소관 법령과 대응 주체가 분절되어 있음

- ㅇ 대통령실 소속 국가안보실이 컨트롤타워라고 볼 여지도 있으나, 대통령실은 본질적으로 대통령 보좌기관으로서 조정 기능은 별론 집행 기능이 부재하고, 실제로도 사이버보안에 특화된 전문 집행조직/인력이 없어 정책 집행에 있어 국가정보원 등에 상당히 의존할 수밖에 없음

- ㅇ 범국가적 컨트롤타워 설치를 위한 (가칭)국가사이버안보법(사이버안보기본법) 제정이 2010년대 중반부터 추진되었으나 여전히 국회 계류 중으로, 신종 AI 사이버 위협에 통합·신속 대응하기에는 법적·조직적 한계 존재

② AI 기본법 내 '사이버보안' 직접 규율 및 강제성의 공백

- ㅇ ’26. 1. 시행된 인공지능기본법은 고영향 AI 사업자의 책무(제34조) 및 임의 영향평가(제35조)를 규정하고 있으나, EU AI법과 달리 ‘사이버보안’이나 ‘적대적 공격 복원력’에 관한 명문 규정이 누락

- ㅇ AI 모델 사업자의 위험관리체계 구축 의무(제32조) 등에서도 사이버보안 요구사항이 명시되지 않아, AI 시스템 자체의 보안 취약성을 법적으로 통제할 강제 수단이 미비함

③ 인증 및 침해대응 체계 복잡, 핵심 시스템의 근거 법률 미비

- ㅇ ISMS(정보통신망법), CSAP(클라우드컴퓨팅법), IoT 보안인증, 보안적합성 검증(전자정부법) 등 보안 인증 및 검증 체계가 여러 법령에 산재되어 있어 기업에 부담 초래

- ㅇ 국정원(공공), KISA(정보통신망), 금보원(금융)이 각 소관 영역별로 사이버보안 위협 정보공유 및 사고 조사 등을 수행 중이지만 분야와 영역을 초월하는 사이버 침해 사고의 특성상 범국가적 차원의 단일한 정보 공유와 사고 대응에는 한계 노정

□ 물론 현재 AI 행동계획, 범주처 정보보호 종합대책이 추진 중이며, 그 내용 중에는 AI발 사이버보안 위협에 대응하기 위한 법제 및 정책도 상당수 포함되어 있음

ㅇ 그러나 전략, 대책이 실제로 작동하기 위한 소관 법령 개정이나 정비와 같은 구체적 이행 실적은 아직까지는 다소 미흡하므로, 지속적인 모니터링이 필요함

Ⅴ. 결론 및 제언

□ 본 장에서는 지금까지의 논의를 종합하여 (1) ‘AI發 새로운 사이버보안 위협의 특징’을 정리하고 (2) 그에 대한 국내외 법제도 및 정책 환경 분석 결과(Ⅲ~Ⅳ)를 토대로 총론 차원에서 법제 개선의 ‘기본 방향’을 제시한 후 (3) 각론 차원에서 ‘15대 개별 과제’를 제안하는 것으로 결론에 갈음하고자 함

[그림 29] 개선 방안 개요

1. AI로 인한 새로운 사이버보안 위협의 특징

□ AI에 의한 보안 위협의 특징

○ AI가 사이버 공격을 위해 필요한 준비 과정(표적 분석, 취약점 이해, 공격 코드 작성 등)을 자동으로 수행하여, 공격 준비 소요 시간과 취약점 발견 소요 시간이 크게 단축 (공격의 자동화 · 고속화)

ㅇ AI 기술의 보조를 통해 공격자에게 요구되는 전문 지식, 코딩 기술 등 필요 역량과 비용이 감소하여, 비전문가도 사이버 공격을 수행 가능(공격의 저비용화 · 대중화)

ㅇ AI가 대상의 업무 · 관계 맥락을 분석하고 신뢰 대상의 음성 · 영상까지 재현할 수 있게 되면서, 다수를 대상으로 한 개인화된 스피어 공격의 일상화(사회공학 공격)

□ AI에 대한 보안 위협의 특징

ㅇ AI는 동일한 입력에도 결과가 달라질 수 있어, AI에 대한 보안 위협은 정형화된 형태를 갖지 않음

ㅇ 취약점 또한 학습 데이터 · 모델 내부에 내재하기 때문에, 패치로 제거되지 않는 상시적 위협으로 존재

⇒ 공격 형태를 사전에 정의해 차단하는 방식이 성립하지 않으므로, 결정론적 · 코드 중심 보안에서 확률적 · 데이터 중심 보안으로 전환 필요

ㅇ 공격이 발생하더라도 AI의 판단 과정을 확인할 수 없어, 공격 사실을 인지하지 못하거나 원인을 규명하지 못한 채 위협이 지속될 수 있음

⇒ 공격 사실의 인지와 원인 규명, 나아가 책임 소재 확인을 위해 AI의 판단 근거와 실행 과정을 기록 · 재현할 수 있는 설명가능성 · 추적가능성 확보 필요

ㅇ 특히 피지컬 AI를 대상으로 한 공격은 정보 유출에 그치지 않고 장비 오작동과 인명 · 재산 피해로 직결

⇒ 사이버 공격이 물리적 사고로 이어지는 특성을 고려하여, 공격 발생 시 AI 시스템을 안전한 상태로 전환할 수 있는 통합 대응 체계 필요

2. 법제도적 대응의 기본 방향

□ 한계의 인식: ‘완벽한 사전 대응’ 불가 + 복원력 강화

ㅇ 사이버보안 위협은 애초부터 기술적, 제도적, 정책적으로 ‘완벽한 사전 대응’이 불가능한 영역이며, AI發 사이버보안 위협은 AI 발전 및 확산 속도 등을 감안하면 더욱 그러함(대표적으로 미토스 사태)

⇒ 사고 발생은 현실적으로 가능한 선에서 최대한 억제하되, 실제 사고 발생 시 신속히 수습하고 정상화할 수 있는 ‘복원력’ 강화 필요

□ 정부 주도 → 민관 협력으로 거버넌스 패러다임 전환

ㅇ AI발 사이버보안 위협은 고도의 전문적, 기술적 영역일 뿐 아니라 공공/민간 영역 구분 없는 정보 공유가 핵심이므로 정부 주도의 거버넌스로는 뚜렷한 한계 노정

□ 신기술/신산업 발전과 사회 안전 조화

ㅇ AI 기술/산업은 아직도 성숙기가 아닌 성장기 단계이며, 우리나라는 AI 글로벌 3대 강국을 국가 전략으로 천명하고 있으므로, AI로 인한 사이버보안 위협의 대응 법제 역시 근본적으로 AI 기술/산업 발전을 저해하여서는 곤란하며, 관련 기술/산업 촉진 · 진흥도 일정 부분 뒷받침할 수 있어야 함

ㅇ 반면, AI로 인한 사이버보안 위협이 실제 침해 사고로 이어질 경우, 과거보다 피해 수준이나 범위가 더욱 확대, 심화될 것으로 예상되며, 특히 향후 피지컬 AI가 본격적으로 보급, 확산되면 기존의 사이버보안 사고 양상과 달리 인간의 신체와 생명에 직접 해악을 끼치는 ‘물리적 사고’로 확산될 수 있다는 점 또한 유의하여야 함

□ 경직성, 우둔성(愚鈍性) → 탄력성, 신속성, 유연성 확보

ㅇ AI발 사이버보안 위협은 대량성, 신속성, 비정형성, 예측불가능성 등을 특징으로 하여 제·개정에 상당한 시간이 소요되는 전통적 법제 체계하에서는 적시 대응에 한계

ㅇ AI로 인한 새로운 사이버보안 위협이나 침해 사고에 실제 대응하기 위해서는 기술성과 전문성이 필수로 요구되나, 이 역시 일반성과 추상성을 본질로 하는 법제도로 녹여내기 쉽지 않음

⇒ 법률/대통령령과 같이 절차적, 실체적으로 '무거운' 규범을 통한 규율, 부령/행정규칙(훈령, 고시 등)과 같이 '가벼운' 규범을 통한 규율, 그리고 가이드라인과 같은 연성 규범(soft law)을 통한 규율 체계를 적절히 조화시켜 규범 체계의 신속성, 효율성, 탄력성 확보 필요

□ 기존 사이버보안 규제의 전면 재정비

ㅇ AI발 사이버보안 위협 대응에 필요하나 현재 공백이거나 미흡한 영역의 규제는 강화하되, 불필요한 기존 사이버보안 위협 규제는 과감히 간소화, 개선 혹은 철폐하는 등 규제 체계 전면 재정비

□ 해외 사례의 전략적 취사 선택

ㅇ 해외 주요국의 법제·정책 사례 중 (1) 공통적으로 관찰되는 사례, (2) 우리 법제·정책과 지향점이 일치하는 사례, (3) 우리에게 요긴한 시사점을 주는 사례 등은 채택을 적극 검토

- (1) AI 산업, 기술 발전 촉진(미, EU, 영), (2) 규제 재설계(미, EU), (3) 민관협력 강화(미, EU, 영, 일), (4) 제품·서비스 전 생애주기 보안 강화(미, EU) 등

ㅇ 우리의 역사적, 문화적, 정치적 환경과 다소 조화되기 어려운 해외 사례는 신중 접근 필요

- (1) 능동적 사어버방어로서의 접근 확인 및 무해화 조치(일), (2) 침해사고 발생 보고시 전면 면책(미; 전면 면책 대신 EU식 '적절 감경' 정도가 적당), (3) 규제 비용 회수 제도(영), (4) 국가안보 목적 직접 명령권(영) 등은 신중하게 장기적으로 검토 필요

□ 구체적 액션 플랜 마련

ㅇ 법제도 대응 방안이 단순한 제안에 그치지 않고 가시적 성과를 낼 수 있도록 각 방안을 단기(현행 법제에서 실행 가능; ~6개월), 중기(시행령/시행규칙/행정규칙 수준 개정 필요;~1년), 장기(법률 수준 개정 필요;~3년) 과제로 각 분류하여 구체적 실행 계획 제시

3. 제도 개선 방안

□ 2.에서 제시한 기본 방향과 지금까지의 분석, 논의 등을 종합하여 다음과 같은 15대 과제를 제시함

가. 거버넌스: 추진체계 개편, 정보공유 강화 등

① 범국가적 사이버보안 추진체계 개편[장기]

ㅇ 대통령비서실 소속 국가안보실을 최상위 컨트롤타워로 두고 국정원(공공) / 과기정통부(정보통신망, 정보통신서비스 제공자) / 금융위(금융) 등으로 분산된 현재의 추진체계를 유지할 것인지의 문제

ㅇ 사이버보안 추진체계는 '정답'이 없는 문제이며, 국가별로 시대적, 역사적, 문화적 맥락에 따라 각기 다른 거버넌스 운용

* 미국: DHS(국토안보부) 산하 CISA / 영국: GCHQ(정보기관) 산하 NCSC / 일본: 내각관방 직속 NCO 등

ㅇ 다만 현행 분산형 체계로는 급박한 사이버보안 위협 발생시 범국가적 차원의 신속하고 체계적인 대응이 어렵고 대응 창구도 일원화되지 않는다는 등의 문제가 꾸준히 제기되어 왔으며, 최근 미토스 사태에서도 동일한 지적이 나옴

ㅇ 현재의 영역별 분산형 추진체계를 전면 개편하기에는 조직, 인력 관점에서 쉽지 않은 것이 사실이나, 장기적으로는 신속성과 효율성이 담보되는 추진체계에 관한 고민 필요

- 현행과 같이 국정원 소속 NCSC가 사실상 사이버보안 실무 총괄기관으로 작동할 경우 우리나라의 역사적, 정치적 맥락상 정보기관에 의한 감시, 사찰의 우려에서 자유로울 수 없고, 민간의 적극적 정보 공유도 이끌어내기 어려움
- NCSC가 실무 총괄기관으로 작동하면서도 법령상으로는 그 권한이 공공에 한정되는 '규범'과 '실제'의 불일치 역시 해소 필요[181]
- NCSC는 대통령비서실이나 국정원이 아닌 행정부처로[182] 이관하고, NCSC 부서 혹은 소속/부설기관 형태로 부문별 CERT를 두어 추진체계를 일원화하는 방안 등을 생각해 볼 수 있음
- 거버넌스 개편 체계에 맞추어 이를 뒷받침할 규범으로 (가칭) 사이버보안기본법 제정 검토

• 제정: (가칭)사이버보안기본법
• 개정: 정보통신망법/정보통신기반보호법/사이버안보 업무규정 등

181) 이런 이유로 사이버안보 업무규정을 살펴보면 NCSC의 기능을 '사이버보안'이 아니라 그 상위 개념이자 공공/민간의 구분이 없는 '사이버안보'로 규정하고 있다. 사이버안보업무규정 제4조.

182) 예컨대 수사 및 안전 기능에 방점을 두면 경찰청 관장 기관인 행정안전부에, IT 정책과의 연계성 및 기축적된 전문성이나 경험에 방점을 두면 과기정통부를 생각해 볼 수 있다. 물론 이 경우에도 특정 기관으로의 권한 집중에 따른 우려는 존재할 수 있으므로 이를 견제할 수 있는 장치(예컨대 중립적인 기구를 설치하거나, 국회 소관위원회 기능 강화 등)도 고려할 필요가 있다.

② 범국가적 위협 정보 공유 강화[장기]

ㅇ 현재 사이버보안 위협 정보 공유 체계는 국정원(NCTI, 공공), KISA(C-TAS, 정보통신망), 금보원(F-CTI, 금융)의 3개 기관별로 분산되어 있고, 기본적으로 소관 영역별로 공유하는 체계임*

* 예컨대 KISA의 C-TAS는 KISA와 상호 협약을 맺은 민간기업이 C-TAS에만 자신의 사이버보안 위협 정보를 공유하며, 원칙적으로 KISA는 이를 NCTI나 F-CTI에 제공하지 않음

** 민간기업이 KISA의 C-TAS뿐 아니라 국정원의 NCTI와 정보를 공유하려면 기본적으로 KISA, 국정원과 각 별도의 협약을 체결해야 하는 구조임

ㅇ 공공-정보통신망-금융이라는 영역별 분절된 사이버 위협 정보 공유 체계로는 AI發 사이버보안 위협에 신속한 대응이 어려움

ㅇ 또한 현행 체계는 민간이 사이버보안 위협 정보를 자발적으로 공유하더라도 유의미한 인센티브가 없으며, 오히려 민간의 정보가 공공(특히 정보기관)에 노출됨에 따른 다양한 리스크에 노출됨

ㅇ 특별한 사정이 있는 경우에는(예컨대 사이버 위협의 심각성, 중대성, 긴급성 등) 국정원/KISA/금보원 구분 없이 사이버 위협 정보가 공유되도록 하되, 정보를 제공한 민간에게 안전 장치(safe-harbor)를 두어 범국가적 차원의 위협 정보 공유 체계 강화 필요

• 개정: 정보통신망법(제48조의2 등), 정보통신기반보호법(제16조) 등[183]

183) 예컨대 정보통신망법의 경우 ① KISA의 침해사고 대응 업무에 사이버위협 정보 공유 명문화(제48조 제1항 개정), ② 민간이 제공한 침해사고 신고 정보를 민간에게 불리한 목적으로 사용 금지(제48조의3 개정) 등을, 정보통신기반보호법의 경우 ③ 분야별 정보공유분석센터(ISAC)간 정보공유 근거 마련(제16조 개정) 등을 생각해 볼 수 있다.

③ 표준, 가이드라인 등 연성 규범 거버넌스 정비[단기 장기]

ㅇ 현재 KISA, 금보원, 국보연(NSR) 등에 분산되어 있는 사이버보안 가이드라인, 표준화 등의 실제 제정 및 운영 체계를 정부가 아닌 국민 입장에서 통합 혹은 간소화[단기]

ㅇ 장기적으로는 특정 기관/부처에 종속되지 않고 독립성과 전문성에 기반하여 사이버보안 연성 규범을 지속적으로 연구, 개발, 수립할 수 있는 독립 조직 체계 검토(참고: 미 NIST, EU ENISA)[장기]

- 개정: 정보통신망법(제52조 등), 사이버안보 업무규정(제17조) 등[184]

나. 예방: 사이버보안의무 법제화, 인증체계 정비, SBOM 도입 등

④ AI에 특화된 사이버보안 의무 법제화[중기 장기]

ㅇ EU AIA와 유사하게 적어도 인공지능법상 고영향AI 사업자 또는 학습 누적 연산량이 일정 기준 이상인 AI 사업자에게는 일정 수준 이상의 사이버보안 대응력 및 복원력을 갖추도록 법정 의무 부여

- 인공지능법 또는 시행령을 개정하지 않더라도, 현행 인공지능법 제32조(인공지능 안전성 확보 의무) 또는 제34조(고영향 인공지능 관련 사업자 책무)에 따라 과기정통부장관이 정하는 고시에서 관련 사항을 규정하는 방식도 가능함[중기]

- 장기적으로는 법령에 명문화하는 것이 법적 안정성이나 예측 가능성 측면에서 바람직)[장기]

- 개정: 인공지능법(제32조, 제34조 등[185])

184) 현재 각종 법령(정보통신망법, 전자금융거래법, 사이버안보 업무규정 등)에 분야별, 영역별로 분산되어 있는 사이버보안 전문기관의 설립 근거 및 업무 수행 근거 규정들을 한 기관으로 집중시키고 해당 기관의 독립성을 일정 부분 보장하는 방식의 입법을 생각해 볼 수 있다. 예컨대 정보통신망법 제52조(KISA의 설치 근거)를 개정하여 KISA로 해당 기능을 집중시키는 방안 등이 그러하다.

⑤ 인증 체계 정비[장기]

ㅇ AI發 사이버보안 위협에 대응할 수 있는 '최소한의 기준'인 인증 체계를 필요한 부분은 강화하되 중복, 분산되어 중복규제 이슈가 있고 실효성이 떨어지는 부분은 통합, 간소화하여 정비

- AI 시대를 맞이하여 불필요한 규제의 정비 + 필요한 규제 강화는 전 세계적 경향(미, EU, 영 등)

* (예시) 현재 관리체계/정책/문서/일회성 중심의 ISMS/ISMS-P를 기술체계/집행/업무/연속성 중심으로 개편하고 이를 법제화(EU의 사이버 태세)

• 개정: 정보통신망법, 개인정보 보호법, 전자정부법 등

⑥ 디지털 장비 인증 확대[중기]

ㅇ AI가 급속히 발전하면서 IoT 제품, 커넥티드 장치 등 디지털 요소가 포함된 다양한 제품군에 AI가 탑재되거나 최소한 AI가 통제, 조작할 가능성이 증대하고 있으나, 디지털 장비 전반에 대한 인증 체계 부재

- 현재는 정보통신망법령상의 '정보통신망연결기기' 및 그 외 특별법(디지털의료제품법 등)에 따른 디지털 기기에 한정하여 인증

ㅇ EU(사이버복원력법), 영국(PSTIA)와 유사하게 디지털 장비 전반을 아우르는 기본 인증 체계를 마련하는 방안을 생각할 수 있음

- 현행 법 체계상 정보통신망법 시행령 개정하여 정보통신망연결기기 개념 및 대상을 확장하면 디지털 장비 전반에 관한 인증 체계를 법제화할 수 있음

185) 예컨대 인공지능법 제32조(인공지능 안전성 확보 의무)에 사이버보안 관련 사항을 구체적으로 규정하거나, 동법 제34조(고영향 인공지능 관련 사업자 책무)에 사이버보안 사항을 포함하는 위험관리 방안을 수립토록 의무화하는 방안을 생각해 볼 수 있다.

- 다만 기업의 부담 경감, 인증 행정 간소화, 인증 체계 효율화 등을 위하여 EU 사이버복원력법과 유사하게 위험도에 따라 차등화된 인증 제도를 설계하고, 이미 인증 체계가 존재하는 디지털 장비에 관하여는 기존 인증 체계에 따르도록 함(중복 인증 배제)

• 개정: 정보통신망법 시행령 제36조의2[186]

⑦ 수요자 중심의 위협정보 공유 서비스 제공[단기]

ㅇ 현재 C-TAS 등을 통해 제공되는 사이버 위협정보는 실제 해커들이 사이버 공격에 사용중인 실시간 위협정보를 분석한 것에 그침; 해당 정보를 제공받은 기관이 자체적 분석, 대응 능력을 요구

- 사이버 공격 대응 조직, 인력, 예산 등을 갖춘 대규모 기관이 아닌 중소기업이나 스타트업 등은 위 정보를 공유받더라도 실질적 대응이 어려운 한계 노정

ㅇ 위협정보를 제공받는 기관의 정보자산 중심 매칭형 서비스를 제공하여 사이버 공격 대응 능력 유무나 수준과 관계없이 일정 수준의 대응이 가능하도록 C-TAS 등의 개편 필요

• 제·개정 수요: 없음(기술적 사항)

⑧ 취약점 정보 상시 신고체계(CVD/VDP) 법제화[장기]

ㅇ 실정법상 기업 등이 자신들의 사이버보안 취약점을 상시 발굴하고 이를 공유하는 CVD/VDP 제도는 부재하며, 오히려 CVD/VDP 시행시 해당 기업이나 사이버보안 취약점을 탐지한 화이트해커에게 정보통신망법, 개인정보 보호법 위반 등과 같은 법적 리스크 존재

186) 현재 정보통신망법 시행령 제36조의2 및 별표 1의5에서 가전, 교통, 금융, 스마트도시, 의료, 제조·생산, 주택, 통신의 8대 분야로 인증 대상 정보통신망 연결기기를 한정하고 있는데, 이를 확대하거나 아예 현재와 같은 '분야별' 한정 열거 체계를 폐지하는 방안을 생각해 볼 수 있다.

* 엄밀히 말하여 판례상 민/형사상 책임이 성립한다고 보기는 어려우나, 이는 해석론이어서 수범자 입장에서는 법적 불확실성 상존

ㅇ CVD/VDP 제도를 법률에 명문화하고, 화이트해커의 민/형사상 면책 또한 실정법에 명확히 규정할 필요

• 개정: 정보통신망법(제48조, 벌칙규정 등), 개인정보 보호법(과태료 등)[187]

⑨ 사이버보안 투자 촉진[중기 장기]

ㅇ 사이버보안 역량 강화는 조직/인력/예산과 같은 '투자'에 기초

ㅇ 재정 여건상 예산 직접 투입은 한계가 있을 뿐 아니라 정부가 공공이 아닌 민간의 사이버보안 투자를 '직접' 강제하기도 어려우므로, 각 기관의 자발적 보안 투자를 유인할 수 있는 '간접' 제도 설계 필요

- (사례) CISO/CPO 지정을 의무화하고 자격 및 권한을 상향하도록 관련 법제가 정비되어 '26. 9.부터 단계적 시행 예정

ㅇ 민간의 경우, 미국 사례를 참조하여 최소한 상장법인의 경우 사이버보안 사항을 공시(정기/수시)사항에 포함시키는 방안 제안[중기]

- 미국은 SEC 규정상 상장법인이 제출하는 정기/수시공시 항목에 사이버보안 침해사고가 명시되어 있음[188]

187) 예컨대 CVD/VDP 시행에 따른 화이트해커의 망 접근은 정보통신망 침해행위에서 명시적으로 제외하고(정보통신망법 제48조 개정), 이에 따른 취약점 발견은 신고 대상인 침해사고의 개념에서 제외(정보통신망법 제2조 제7호 개정)하는 한편 미신고시 처벌(행정질서벌 포함) 대상에서도 제외할 필요가 있다(개인정보 보호법 제75조 개정).

188) (1) 상장기업은 사이버보안 사고가 발생하였고 그 사고가 회사 경영에 중대한 재무적·비재무적 영향을 미쳤을 때 4영업일 이내에 공시해야 하며(17 C.F.R. § 249.308.), (2) 연례보고서에 평소 사이버보안 체계를 투명하게 공개해야 함(17 C.F.R. § 229.106.).

- 우리는 정보보호산업법 제13조, 동법 시행령 제8조에 따라 사업분야, 매출액 및 서비스 이용자 수 등을 고려하여 대통령령으로 정하는 기준에 해당하는 자는 정보보호 공시의무가 있으나('27부터 대통령령 개정하여 모든 상장기업(코스피, 코스닥 불문)에 공시의무 부여 추진 중[189]), 정기공시 의무만 있으며 수시공시 의무는 없음

ㅇ 공공의 경우에도 사이버보안 사항을 경영공시 사항에 포함시키고 업무평가 배점을 강화하는 방안 검토 필요[장기]

• 개정: 정보보호산업법 시행령(제8조), 공공기관운영법(제11조)[190]

⑩ 제로트러스트 및 SBOM · AI-BOM 제도화[장기]

ㅇ 전통적인 경계 중심 보안의 한계를 극복하는 방안의 하나로서 제로트러스트 원칙을 법제화*

* 적용 대상, 법적 의무화 여부 등의 추가 검토 필요

ㅇ SW 공급망 전반에 대한 사이버보안 위협 대응책으로 미, EU 등의 사례를 참고하여 SBOM 및 AI-BOM 법제화*

* 적용 대상, 법적 의무화 여부 등의 추가 검토 필요

• 개정: 소프트웨어산업법, 전자정부법 등

189) https://www.korea.kr/news/policyNewsView.do?newsId=148957777

190) 예컨대 정보보호 공시대상 항목에 사이버보안 사고 현황을 포함시키고(정보보호산업법 시행령 제8조 제3항 개정), 사이버보안 사고 현황은 정기공시뿐 아니라 수시공시 사항으로 규정하며(정보보호산업법 시행령 제8조 제6항 개정), 공공기관 경영공시 사항에 사이버보안 투자 및 사고 현황 등을 의무적으로 공시토록 하는 방안이다(공공기관운영법 제11조 제1항 개정).

⑪ AI 기반 방어 및 신속 패치 적용 체계 마련장기

- ㅇ AI발 사이버보안 위협의 대량화, 고속화에 대응하기 위해서는 방어 역시 AI에 기반할 수밖에 없음
- ㅇ AI로 인하여 사이버 공격이 대량화, 자동화, 고도화될수록 대응 방안은 '취약점 발견'에서 '발견한 취약점의 신속한 패치'로 이동
- ㅇ AI發 사이버보안 위협의 특성상 기존과 같이 '완벽한 확인' 후 패치가 불가능할 수 있고, 오히려 다소 검증이 미흡하더라도 신속히 패치를 실행하는 것이 사이버 사고 예방 측면에서 유효할 수 있음
- ㅇ 따라서 (1) AI 기반 사이버보안 대응의 근거를 마련하고, (2) 일정 요건 충족시 검증보다 신속한 보안 패치를 우선시하되, 그로 인한 서비스 중단/장애 발생시 담당자 및 해당 기관의 법적 책임을 면책 또는 경감시키는 제도 검토 필요
  - 금융위원회는 미토스 등 프론티어 AI 보안위협에 적극 대응하기 위하여 면책 조치 가이드라인을 마련;[191] 그러나 이는 법적 구속력없는 가이드라인이어서 실효성에 한계 노정

• 개정: 정보통신망법, 정보통신기반보호법, 전자금융거래법, 전자정부법 등

다. 사고 대응: 신속성·전문성 강화, 불필요한 절차 정비

⑫ 사이버 침해 사고 신고 및 대응 창구 일원화단기 장기

- ㅇ KISA, 금보원, 국정원, 경찰 등으로 분산된 사이버 침해 사고 신고 및 대응 창구를 실무적으로 일원화하고,단기

191) 금융위원회, "금융회사가 프론티어 AI 보안위협에 적극적으로 대응할 수 있도록 면책 조치와 가이드라인을 마련하였습니다"(2026).

ㅇ 장기적으로 일원화된 창구의 법적 근거 마련(EU ENISA 등 입법례 참고)장기

• 개정: 정보통신망법, 정보통신기반보호법, 전자금융거래법, 전자정부법 등

⑬ 사고 대응 전문성 강화단기 장기

ㅇ 사이버보안에 관한 민간의 전문성을 적극 활용하여 사이버 사고 발생시 신속하고 효율적으로 대응할 수 있도록 민간 전문가 풀(pool) 운영(EU의 사이버 예비군 제도 참조)단기

ㅇ 대책본부/위원회 등과 같은 다소 형식적일 수 있고 옥상옥으로 기능할 수 있는 행정조직 구성을 지양하고, 사고 발생시 선 복원력 확보/후 원인조사 및 책임추궁 체계 정립장기

• 개정: 정보통신망법, 정보통신기반보호법, 전자금융거래법, 전자정부법 등

라. 사후 조치: 조사 결과의 투명성 확보와 사후 제재 합리화

⑭ 사이버 침해 사고 조사 결과의 투명성 확보단기

ㅇ 사회적, 국가적으로 중대한 사이버 침해 사고 발생시 그 원인 조사 결과를 국민에게 공개하고 국회 보고하여 투명성을 확보

• 제·개정 수요: 없음(현행 법제하에서 가능)

⑮ 사이버 침해 사고/개인정보 침해 사고에 대한 제재 합리화장기

ㅇ 현재 사이버 침해 사고/개인정보 침해 사고 발생시 손해배상(민사), 형사처벌(형사), 과징금(행정) 등 법적 제재가 강화되는 추세

ㅇ AI발 사이버 침해 사고는 전통적인 사이버 침해사고보다 신속화, 대량화, 다각화, 전문화되므로 완벽한 방어를 기대하기 어렵고, 개인정보 침해 사고 역시 개인정보처리자(혹은 그 사용인)의 고의로 인한 유출 등이 아니라 해킹 등 제3자의 공격으로 인한 것이라면 개인정보처리자에게 완벽한 방어를 기대하기 어려움

ㅇ 따라서 법적 제재를 강화한다고 하여 사고 예방이나 피해 복구로 바로 연결되지 않으며, 오히려 과도한 법적 제재(특히 형사처벌 등 기관이 아닌 개인에 대한 제재)는 우수 인재 유입을 차단하여 사고 대응 역량을 후퇴시키는 등의 부작용이 발생함

* 금융권의 경우 금융회사지배구조법 제5조에 따라 임원은 물론 직원도 최대 5년간 취업제한되는데, 그 대상 법령(금융관계법령)에서 전자금융거래법이 포함되어 사이버 침해 사고 발생시 담당 임직원이 취업제한되는 문제 발생

ㅇ 해당 사고가 기업의 고의로 발생한 경우는 엄중한 책임을 묻되, 과실로 발생한 사고(이 경우 기업은 피해자라고도 볼 수 있음)에 대하여는 사고 재발 방지와 피해 회복 달성에 적절한 수준으로 법적 제재 수위를 정비할 필요가 있음

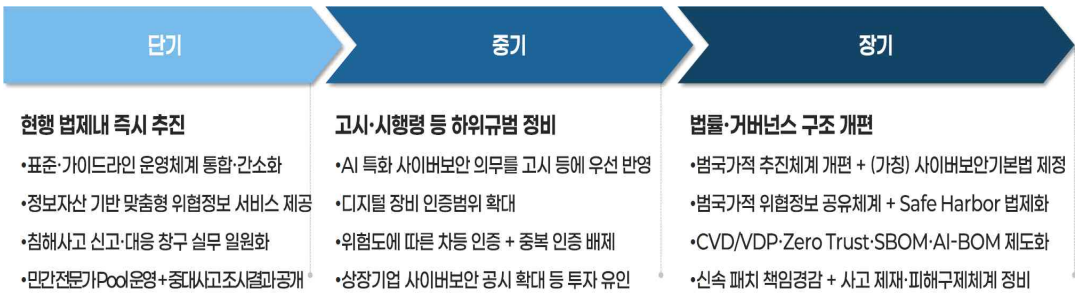
ㅇ 사고 발생시 피해자가 이론적, 현실적 이유로 손해배상받기 어려운 구조이므로, 과징금을 재원으로 피해회복기금을 조성하거나 미국식 동의명령을 도입하여 규제당국과 규제대상의 합의하에 실질적 피해회복방안을 마련하는 등의 제도 도입 검토 필요

• 개정: 정보통신망법, 개인정보 보호법, 전자금융거래법 등

4. 정리

□ 3.에서 제시한 15대 과제를 로드맵으로 정리하면 다음과 같음

[그림 30] 단계별 추진 로드맵



□ 참고로 15대 과제를 Ⅱ.의 기술적 방안과 매핑하면 아래 표와 같음

[표 17] 15대 과제와 기술적 대응 방안 매핑

구분	세부 방안	대응 방향	관련 과제
AI에 의한 위협 대응	위협 인텔리전스 및 이상행위 탐지 고도화	• AI를 방어 수단으로 활용	⑧ ⑪
	공격 표면 관리 및 취약점 대응 자동화	• 패치 · 취약점 관리 속도 제고	⑧ ⑪
	SOAR 고도화	• 대응 조치 자동화	⑪
	신뢰 채널 검증 강화	• AI 위조 콘텐츠에 대한 인증 보완	⑩
AI에 대한 위협 대응	AI 시스템 보안 강화	• 모델과 학습 과정의 신뢰성 확보	④ ⑤ ⑥
	제로트러스트 적용	• 모든 주체를 지속적으로 검증	⑩
	공급망 보안 및 AI-BOM 활용	• AI 구성요소의 출처와 의존성 관리	⑩
	피지컬 AI 물리적 안전성 확보	• 현실 세계의 인명 및 물리적 피해로 직결되지 않도록 대응	④ ⑤ ⑥
공통 운영 원칙	인간-AI 협력 보안 운영	• AI 자동화와 인간 판단 결합	⑪ ⑬

* '직접' 관련된 과제만 매핑하였으며, 연결되지 않은 과제들도 간접적으로 기술적 대응과 관련됨

□ 본 연구는 AI로 촉발된 새로운 사이버보안 위협과 관련, ’26. 8. 기준 (1) ’26. 8. 기준으로 최신 기술 동향을 정리하고 (2) 해외 주요국(미, EU, 영, 일)의 법제도와 정책을 일괄한 후 (3) 국내 법제도 및 정책과 비교하여 시사점을 도출하고 (4) 최종적으로 우리 법제도 개선의 기본 방향과 세부 과제 및 로드맵을 제시하였다는 점에서 일정 부분 유의미하다고 생각함

□ 다만 AI 발전 속도 및 그에 따른 사이버보안 환경 변화 양상이 하루가 다르게 급변하고 있으므로 본 연구에서 다룬 내용은 향후에도 지속적 업데이트 필요

□ 또한 본 연구에서 제안한 법제도 개선의 기본 방향과 세부 과제의 실행력이 담보되기 위해서는 향후 민, 관, 학, 연 등 이해관계자들의 토론과 숙의 과정이 필요함

[부록] 주요국의 개인정보 침해 배상제도 및 사례

□ AI로 인한 새로운 사이버보안 위험이 증대될수록 그 부작용으로 개인정보 침해 사고의 발생 횟수 및 그 피해 양상 또한 늘어날 것으로 예상됨

□ 실제로 국내로만 한정해 보더라도, 최근 3년간('23-'26) 이동통신사, 플랫폼, 풀필먼트 서비스 회사 등 굴지의 기업들에서 대규모 개인정보 유출사고가 발생하고 있음*

* LGU+ 해킹 사건('23), 법원전산망 악성코드 감염사건('24), 카카오 오픈채팅 사건('24), GS리테일 크리덴셜 스터핑 사건('25), 쿠팡 정보유출 사건(''25), SK텔레콤/KT/LGU+ 개인정보 유출사건('25) 등

□ AI의 등장으로 개인정보 유출사고를 둘러싼 환경은 공격 기법, 유출 경로, 유출 대상, 유출 결과 등에서 다음과 같이 변화하는 中

[표 18] AI 등장으로 인한 개인정보 유출사고 환경 변화

구분	과거(~'22)	최근 3년('23~'25)
공격 기법	단순 악성 코드, 취약점 공격	AI 기반 피싱, 크리덴셜 스터핑, 제로데이 공격 등
유출 경로	직접 침투	퇴사자 자격증명 악용, 공급망 취약점 경유 등
유출 대상	ID, 비밀번호, PIN 중심	구매내역, 유전자정보 등 다양한 정보
유출 결과	제한적 과징금	매출액 연동 과징금, CEO 등 경영진 책임

□ 특히 '26은 미토스, GPT-사이버 등 사이버 공격에 최적화된 프론티어 AI가 등장한 원년(元年)으로, (i) 누구나 (ii) 저렴한 비용으로 (iii) 인간이 대응할 수 없는 기계 속도로 (iv) 무차별 공격이 가능한 수준으로 사이버공격이 가능하도록 패러다임이 전환되고 있어, 이로 인한 개인정보 유출, 침해 사고도 급증할 위험이 있음

□ 이러한 배경하에 이 장에서는 국내외* 개인정보 침해 배상제도** 및 실제 배상 사례를 간략히 살펴봄

* 대상국가: 대한민국, 미국, 영국, 일본[192]

** 과징금 등 행정제재는 원칙적으로 검토 대상에서 제외

1. 미국

□ 개인정보 침해 배상제도

○ 법제 개요

- 미국은 EU, 일본 및 우리와 달리 연방 차원의 일반법으로서의 개인정보 보호법은 존재하지 않으며, (1) 개별 영역별(공공, 아동, 금융, 전자통신 등) 보호를 규율하는 연방법과 (2) 일반법으로서의 개인정보 보호를 규율하는 주법(州法)으로 이원화

- 주법의 경우, 캘리포니아주가 2018년 개인정보 보호 일반법인 '소비자 개인정보보호법(CCPA)'를 제정한 이후 버지니아주, 콜로라도주, 유타 주 등이 CCPA와 유사한 개인정보 보호 일반법을 제정하였으며, 다른 주의 경우에도 이와 유사한 일반법 제정이 추진 중[193]

192) EU의 경우 GDPR 제82조에서 개인정보 침해에 따른 정보주체의 손해배상청구권을 인정하고 있으나 실제 손해배상청구소송은 개별 회원국의 법에 따라 진행되므로(GDPR 제82조 제6항, 제79조) EU는 검토 대상에서 제외.

193) https://pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS270&mCode=&nttId=9731

ㅇ 개인정보 침해 시 피해 구제 체계

- 연방법 차원에서는 개인정보 침해 사고 발생시 해당 기관이나 기업(우리 개인정보 보호법상 '개인정보처리자')에게 손해배상책임을 지우는 별도의 근거 규정은 없음; (1) 규제기관(대표적으로 FTC)이 개인정보처리자에 행정책임을 물으면서 그에 수반하여 피해 보상이 이루어지거나 (2) 피해자가 개인정보처리자에게 보통법상의 불법행위책임(tort) 및/또는 계약책임에 따른 손해배상을 청구하는 구조임

- 반면 주법의 경우에는 정보주체의 손해배상청구에 관한 특별 규정을 두는 사례가 있음; 대표적으로 CCPA(2020년 CPRA로[194] 개정됨)는 개인정보처리자가 개인정보를 보보하기 위한 합리적인 보안 절차 및 관행을 수립하고 유지해야 할 의무를 위빈하여 개인정보 침해 사고가 발생한 경우 정보주체는 사고 1건당 100$~750$의 법정손해배상 혹은 실손해배상을 청구할 수 있도록 규정[195]

  * 단 본 조에 따른 법정손해배상금을 청구하는 소송은 (1) 소비자가 사업자를 상대로 법정손해배상 청구 소송을 제기하기 전에, 위반되었거나 위반 중이라고 주장하는 CCPA상의 구체적 조항을 명시한 30일간의 서면 통지를 사업자에게 제공한 경우에 한하여 제기할 수 있으며, (2) 시정이 가능한 경우, 사업자가 30일 이내에 통지된 위반 사항을 실제로 시정하고, 위반 사항이 시정되었으며 향후 추가적 위반이 발생하지 않을 것이라는 명시적인 서면 진술서를 소비자에게 제공한다면, 사업자를 상대로 한 개별 법정손해배상 또는 집단 법정손해배상 소송을 제기할 수 없음[196]

194) The California Privacy Rights Act of 2020. 캘리포니아주 민법전(California Civil Code)을 개정하는 방식의 입법이다.

195) California Civil Code § 1798.150(a).

ㅇ 연방규제기관(대표적으로 FTC)의 규제에 따른 간접적 피해 보상

- 민간 기업의 개인정보 침해에 대하여 연방 차원에서는 주로 FTC가 규제 기관으로 작동함; FTC는 개인정보 유출사고 발생시 FTC법 제5조(기만적 행위 또는 불공정 행위 금지)[197] 위반을 들어 해당 기업을 피고로 삼아 심판(administrative procedding) 혹은 소송을 제기할 수 있음*

  * 즉 우리와 달리 FTC가 직접 제재적 행정처분을 하는 구조가 아니라, 기본적으로 당사자주의에 기반한 소송(administrative proceeding은 우리에 비견하자면 행정심판과 유사함)을 제기하는 구조임

- FTC는 (1) 금지명령(injunction) 소송(혹은 심결. 이하 구별하지 않음) (2) 과징금(civil penalties) 소송 (3) 소비자 피해 구제 소송을 제기할 수 있으나, 많은 경우 그 전단계에서 동의명령(consent order)으로 종결됨

- (1), (2)의 경우에는 정보주체에게 금전적 손해배상이나 부당이득반환이 불가능하나,[198] (3)의 경우 FTC가 승소하면 기업이 소비자 피해 구제 기금을 조성하게 되며, 해당 기금을 재원으로 하여 피해를 본 정보주체에게 손해배상금을 배분함

- 동의명령의 경우 대부분 피해자 구제에 관하여도 합의하게 되어 그에 따른 손해배상이 이루어짐

196) California Civil Code § 1798.150(b).

197) 15 U.S.C. § 45.

198) 과거에는 FTC가 (1)에 기한 금지명령의 일환으로 피해자에 대한 손해배상이나 부당이득반환과 같은 형평법에 기한 금전적 구제(equitable monetary relief) 청구할 수 있다고 보았으나, 2021년 연방대법원은 이를 부정하였다. AMG Capital Management, LLC v. FTC, 593 U.S. 67 (2021).

ㅇ 미국의 특징: 징벌적 손해배상과 집단소송

- 미국법상 징벌적 손해배상은 보통법에 근거하여 불법행위 전반에 인정되는 제도로서, 州마다 조금씩 다르기는 하나 일반적으로는 (1) 극심한 가해행위에 대하여 (2) 가해자의 악의적 동기(evil motive)나 타인의 권리에 대한 무모할 정도의 무관심(reckless indifference)이 인정되면 성립함*

  * 美 연방대법원은 고정된 수학 공식을 만들 수는 없지만 징벌적 손해배상액은 실손해액의 9배를 넘지 않는 것이 바람직하며, 역사적 관행상 4배 정도가 헌법적 적절성의 경계라고 판단(이를 초과하면 Due process 위반)[199]

- 또한 미국에서는 역사적, 문화적으로 소수의 원고가 집단의 이익을 대표하여 소송을 제기하고 그 기판력이 원고가 아닌 집단 전체에 미치는 집단소송이 활발히 이용되고 있으며, 이는 개인정보 침해 손해배상소송에서도 마찬가지임[200]

- 이런 이유로 개인정보 침해를 원인으로 한 손해배상소송의 국면에서 정보주체는 대부분 징벌적 손해배상을 주장하며 집단소송을 제기하게 되고 개인정보처리자 입장에서는 패소 확정판결을 받는 경우 손해배상 부담이 급증하게 되는 구조; 따라서 소송이 제기되더라도 법원의 확정판결보다는 합의(settlement)에 의하여 종결되는 경우가 다수임

199) State Farm Mutual Automobile Insurance Co. v. Campbell, 538 U.S. 408, 425-29 (2003).

200) 정인영, “페이스북의 개인정보 침해에 대한 미국 내 입법, 사법 행정적 대응 현황(2)”, 경제규제와 법 제13권 제1호(2025. 5.), 122-147쪽.

□ 개인정보 침해 배상 사례

○ 페이스북(Cambridge Analytica) 사건('16): 7억 2,500만 $ 합의

- 2016년 미국 대선 당시 페이스북이 캐임브리지 애널래티카 등 제3자 앱 개발업체에 수천만 명의 페이스북 사용자 개인정보 및 친구 목록을 동의 없이 제공함으로써 제기된 집단소송[201]

- 페이스북은 "사용자가 페이스북에 공유한 정보는 프라이버시에 대한 합리적 기대가 낮다"는 논거로 면책을 주장하였으나 법원은 이를 배척

- 7억 2,500만 $이라는 역대 최고 배상액으로 합의종결('23)*

  * '07. 5.~ '22. 12. 사이 미국 내에서 페이스북 계정을 활성화하여 사용한 적이 있는 사람에게 실제 개인정보 유출 등의 증명 없이 배상액 지급(계정 유지 기간에 따라 최소 $5 ~ 최대 $38 정도)

○ Equifax 사건('17): 4억 2,500만 $의 소비자 기금 조성 + 1인당 최대 125$의 대체 현금 보상 또는 최대 20,000$의 실손해 배상 합의[202]

- Equifax가 미국 3대 신용평가회사 중 하나라는 점, 미국 성인 인구 절반에 달하는 약 1억 4,000만 명의 민감 개인정보(이름, 생년월일, 주소, 운전면허증 번호, SSN 등)가 노출된 점, 기본적인 기술적 · 관리적 보안체계 미흡 등 반영

- 정보주체가 제기한 손해배상청구소송이 아니라 FTC의 행정제재 절차에 따른 동의명령으로 종결됨('19)

---

201) 집단소송 외에도 페이스북은 FTC 및 SEC로부터 조사를 받았으며, FTC에 50억 $의 벌금, SEC에 1억 $의 과징금을 각 지급하기로 합의(동의명령)하였다.

202) 양자를 합치면 최대 약 7억 달러에 달한다.

○ ZOOM 사건('20): 약 8천 500만 $에 합의

- ZOOM이 사전 고지 없이 제3자에게 과도하게 데이터를 제공한 점, 회의 내용이 종단간 암호화(E2E encryption)된다고 허위 주장한 점, 다크웹을 통한 50만 개 이상의 계정이 유출된 점 등이 문제됨

- ZOOM 이용자들이 집단소송을 제기하였고, 그 결과 (1) '16. 3. ~ '21. 7.까지 ZOOM 앱을 사용한 미국 내 사용자들을 보상 대상으로 하고 (2) 유료 구독자의 경우 지불 금액의 15% 또는 25$ 중 큰 금액을, 무료 사용자의 경우 15$을 배상하기로 합의('20)

○ T-Mobile 사건('21): 3억 5,000만 $ 배상 + 1억 5,000만 $ 보안 투자

- '21. 8. 해커가 테스트 서버 취약점을 통해 침투, 약 7,600만 명의 전·현 고객의 성명, 생년월일, 운전면허증 정보, SSN 등 신원 정보를 탈취하여 다크웹에 판매

- 피해자들이 집단소송을 제기하여 3억 5,000만 $의 현금 배상기금 조성 및 1억 5,000만 $의 자체 사이버보안 투자로 합의 종결('22) * 이와 별개로 FTC와 1,575만 $의 과징금에 합의('24)

2. 영국

□ 개인정보 침해 배상제도

○ BREXIT('20. 1. 31.) 이후 영국의 개인정보 보호 법제는 (1) GDPR을 영국에 이식한 UK GDPR과[203] (2) 2018년 데이터 보호법(Data Protection Act 2018)의 양 축으로 구성

203) https://www.legislation.gov.uk/eur/2016/679/article/82 BREXIT 당시 영국은 법적 공백을 막기 위해 유럽연합 탈퇴법(EU Withdrwal Act 2018)을 통해 기존 GDPR을 그대로 영국 국내법으로 이식하였다.

ㅇ UK GDPR 제82조는 개인정보 침해 사고시 정보주체가 프로세서 및/또는 컨트롤러에게 재산적 및/또는 비재산적 손해의 배상을 청구할 수 있다고 규정하여 손해배상청구의 근거 규정으로 작용

ㅇ 그러나 영국은 (1) UK GDPR 제82조에 따른 청구의 경우 정보주체는 개인정보에 대한 통제권 상실만으로는 손해배상을 청구할 수 없고 실제 정신적 고통이나 금전 손실을 입었음을 객관적으로 증명할 것을 요하며,[204] (2) 미국과 달리 징벌적 손해배상(examplary damages)을 굉장히 제한적으로만 인정하고,[205] (3) 미국과 같은 완전한 opt-out 방식의 집단소송제도가 없어,[206] 법원이 거액의 손해배상을 인정하거나 거액의 배상금 합의로 종결된 사례를 찾기 어려움

ㅇ 이 외에도 보통법의 불법행위(tort) 법리나 계약법리에 따라서도 정보주체가 프로세서 및/또는 컨트롤러에게 손해배상청구 가능

---

204) 참고로 CJEU는 GDPR 제82조는 보상적 기능을 수행하며 징벌적 기능을 수행하지 않는다고 판시하여 동조에 따른 징벌적 손해배상을 부정하고 있다. CJEU에 의하면 GDPR 위반행위에 대한 징벌적 제재는 손해배상이 아니라 과징금이나 기타 제재(GDPR 제83조, 제84조)에 의해야 한다. C-300/21, EU:C:2023:370, Rn. 38-40; C-741/21, EU:C:2024:288, Rn, 59.

205) Rookes v Barnard [1964] UKHL 1. 위 판결 이후 영국에서 징벌적 손해배상은 (1) 공무원의 자의적, 위헌적 권력 남용 (2) 가해자가 위법행위로 얻을 이익이 손해배상액을 초과할 것을 계산하고 행한 경우 (3) 성문법(statute)에서 명문으로 징벌적 손해배상을 규정한 경우에만 인정된다. 또한 징벌적 손해배상이 인정되더라도 이를 포함한 전체 손해배상액은 대체로 전보적 손해배상액의 3배를 넘지 않는다. Thompson v. Commissioner of Police of The Metropolis [1998] QB 498, 518 (Lord Woolf MR).

206) 영국은 피해자들이 직접 소송에 참여하겠다고 신청(opt-in)한 사건들에 관하여 법원이 집단소송명령(Group Litigation Order)을 통해 하나로 묶는 형태의 집단소송만이 가능하다.

□ 개인정보 침해 배상 사례

ㅇ British Airways 사건('18): 영국 역사상 최대의 개인정보 유출 GLO 사건, 합의 종결(비공식적으로 1인당 1,000~2,000 파운드로 알려짐)

- '18. 6.경 British Airways 시스템이 해킹을 당하여 약 430,000명의 고객 및 직원의 성명, 주소, 신용카드 번호 등이 유출됨
- 이 중 약 16,000명의 피해자가 법원으로부터 GLO를 받아 손해배상 집단소송 제기
- '21 비공개 합의로 종결되었으나, 대략 1인당 1,000~2,000 파운드의 손해배상이 지급된 것으로 알려짐

ㅇ 이즐링턴 구청(London Borough of Islington) 사건('23): 정보주체 개인이 공공을 상대로 손해배상청구 승소

- 원고(개인)은 이즐링턴 구청에 개인정보 열람청구를 하였으나 구청이 4년간 응답을 지연하고 관련 서류를 파기하는 등 GDPR 다수 위반
- 법원은 구청의 보통법상의 불법행위책임 및 GDPR 위반 책임을 인정하고, 구청의 부주의 및 지연 행위 등으로 인하여 원고가 얻은 스트레스와 불이익에 대하여 6,000 파운드(약 1,000만 원)의 배상금을 인정[207]

207) Bekoe v London Borough of Islington [2023] EWHC 1668 (KB).

3. 일본

□ 개인정보 침해 배상제도

○ 일본의 개인정보 보호법에는 손해배상청구권에 관한 규정이 없으며, 개인정보 침해 사고로 인한 손해배상은 민사상 불법행위책임이나 채무불이행책임을 통하여 이루어짐

○ 다만 일본은 소비자재판절차특례법('13 제정, '24 개정[208])을 통하여 개인정보 침해 사고 발생시 피해자들이 집단소송을 제기하여 손해를 신속히 배상받을 수 있는 절차 마련

- 소비자재판절차특례법에 따라 개인정보 침해 사고 발생시 소비자단체는 피해자들을 대표하여 개인정보처리자에게 손해배상의무가 있음을 확인받는 '공통의무확인소송'을 제기할 수 있음

- 위 소송에서 소비자단체가 승소하면 피해자들은 법원에 실제 피해를 신고하여 구체적 손해액을 산정하는 '간이확정절차'를 거쳐 손해를 전보받을 수 있음

- 단 개인정보 침해 사고로 개인정보처리자의 '과실'에 의한 정신적 손해만 발생한 경우는 위 절차를 거칠 수 없음[209]

208) 정식 명칭은 「消費者の財産的被害の集団的な回復のための民事の裁判手続の特例に関する法律」이다.

209) 위 법 제3조 제2항 제5호. 과실로 인한 위자료청구까지 위 법에 따른 간이확정절차를 허용한다면 남소의 위험이 있고 해당 기업의 경제 활동에 큰 타격을 입을 수 있기 때문이다. 무엇보다 위 법의 정식 제명 자체에서 알 수 있듯이 소비자의 '재산적 피해등'의 집단적 회복을 위한 법률이다.

□ 개인정보 침해 배상 사례

○ 리쿠르트(Recruit) 사의 리쿠나비 사건: 1인당 약 10,000엔 ~ 30,000엔 수준의 위자료 및 소송비용 지급 화해 종결

- 일본 취업 정보 플랫폼 '리쿠나비'를 운영하는 리쿠르트가 '19 취업 준비생들의 cookie를 분석하여 기업 입사 내정 사퇴 확률을 점수화한 뒤, 이를 정보주체의 동의 없이 기업에 유상 판매
- 일본 개인정보보호위원회의 시정 권고 및 제재 처분과 함께 피해자들의 민사상 위자료 청구 소송 제기
- 리쿠르트 대표의 사과 및 화해로 종결('20)

○ 베네세(Benesse) 개인정보 유출 사건('14): 1인당 1,000엔 ~ 4,000엔 위자료 인정 판결

- 일본 교육·출판 기업 베네세의 외주 시스템 엔지니어가 미성년자 및 보호자 약 3,000만 건의 개인정보(이름, 생년월일, 성별, 보호자 주소, 전화번호 등)를 유출하여 명부업자에게 판매
- 피해자 중 약 12,000명이 정신적 손해배상 청구(성인 1인당 10,000엔 / 미성년자 1인 당 100,000엔)
- 유출에 따른 불쾌감이나 불안감 수준을 넘어 민법상 배상해야 할 정신적 손해가 발생하였는지가 쟁점이 되었으며, 1심과 2심은 모두 이를 배척하였으나 최고재판소는 정신적 고통의 존재 여부 및 손해 정도에 대한 심리가 미진하다며 파기환송[210]
- 환송심에서 1인당 1,000엔 ~ 4,000엔의 위자료 인정됨

210) 最判平成29·10·23民集未収(平成28年(受)第1892号).

4. 대한민국

□ 개인정보 침해 배상제도

○ 개인정보 보호법상 손해배상책임에 관한 별도 규정 존재(제39조, 제39조의2)

- 정보주체는 개인정보처리자의 개인정보 보호법 위반 행위로 손해가 발생한 경우 손해배상을 청구할 수 있음(제39조 제1항 본문). 이 경우 고의 또는 과실의 증명책임은 개인정보처리자에게 전환되어 개인정보처리자가 고의 또는 과실이 없음을 증명해야 면책됨(제39조 제1항 단서)
- 만약 개인정보처리자의 고의 또는 중과실로 개인정보가 유출등(분실, 도난, 유출, 위조, 변조, 훼손)이 된 경우 법원은 실손해액의 최대 5배 범위에서 손해배상액을 정할 수 있음(제39조 제3항 본문; 징벌적 손해배상 또는 배액손해배상). 이 경우 개인정보처리자는 고의 또는 중과실이 없음을 증명한 경우 면책됨(제39조 제3항 단서)
- 한편 제39조에도 불구하고 정보주체는 개인정보처리자의 고의 또는 과실로 인하여 개인정보가 유출등이 된 경우에는 300만원 이하의 범위에서 상당한 금액을 손해액으로 하여 배상을 청구할 수 있음(제39조의2 제1항 본문; 법정손해배상) 이 경우 개인정보처리자는 고의 또는 과실이 없음을 증명한 경우 면책됨(제39조의2 제1항 단서)

○ 그 외에도 개인정보 침해 사고가 발생하면 정보주체는 개인정보처리자에게 불법행위책임 또는 채무불이행책임을 물을 수 있음(민법 제390조, 제750조 등)

○ 손해배상과는 별개로 개인정보 보호위원회는 개인정보처리자가 처리하는 개인정보가 유출등이 된 경우 전체 매출액의 3%를 초과하지 않는 범위에서 과징금을 부과할 수 있음(제64조의2 제1항)

- 특히 ’26. 9. 11. 시행되는 개정 개인정보 보호법은 고의 또는 중대한 과실로 개인정보가 유출등이 되고 정보주체의 피해 규모가 1천만 명 이상인 경우, 과징금 부과처분을 받은 날부터 3년 이내에 과징금 부과 대상 위반행위를 한 경우 등에는 전체 매출액의 10%를 초과하지 않는 범위 내에서 과징금을 부과할 수 있도록 규정하여 과징금 부과 기준을 대폭 상향하였음(개정법 제64조의2 제2항)

□ 개인정보 침해 배상사례

ㅇ 우리의 경우 그간 개인정보 침해사고에 관하여 다수의 손해배상청구소송이 제기되었으나 실제 손해배상이 인정된 사례는 많지 않음. 그 이유는 인과관계가 인정되지 않거나, 인과관계가 인정되더라도 법원은 단순히 개인정보가 유출되었다는 사실만으로 손해(특히 정신적 손해)가 자동으로 인정되는 것은 아니며 여러 사정을 고려하여 실제로 손해가 발생하였는지를 판단해야 한다고 보기 때문임

ㅇ 사회적으로 문제가 되었던 개인정보 유출사고 중 손해배상책임이 인정된 주요 사례는 다음과 같음

- 싸이월드/네이트 대규모 해킹 유출 사건(’11): 1인당 10~20만 원[211)]
- 신용카드 3사(KB, NH, 롯데) 고객정보 유출 사건(’14): 1인당 10만 원[212)]
- 홈플러스 고객정보 매매 사건(’11), 1인당 5~30만 원[213)]

211) 대법원 2018. 1. 25. 선고 2015다2404 판결 참조.
212) 대법원 2018. 12. 27. 선고 2016다265739 판결 참조.
213) 대법원 2022. 4. 28. 선고 2019다228678 판결 참조.

□ 참고: 기금 설치 문제

○ 위에서 언급하였듯이 우리의 경우 개인정보 침해사고가 발생하더라도 정보주체에게 실제 손해배상까지 이어진 경우는 많지 않음

○ 설령 개인정보처리자에게 거액의 과징금이 부과되어 징수까지 이루어지더라도 이는 전액 국고에 귀속되며, 피해자인 정보주체가 입은 손해의 배상과는 직접적인 관련이 없음; “털린 것은 내 개인정보인데 왜 과징금은 정부가 가져가느냐?”의 문제 발생

○ 이러한 문제인식하에 (가칭)개인정보 손해배상기금 설치가 논의되고 있으나, 이를 위해서는 기금의 성격이 결정되어야 함

- 기금은 그 설치 목적과 공익에 맞게 관리, 운용되어야 함 (국가재정법 제62조 제1항)

- 기금이 (1) 개인정보 유출사고의 피해자를 직접 구제하는 피해구제기금(Victims Relief Fund)인지, (2) 피해의 직접 구제는 소송이나 분쟁조정과 같은 현행 절차에 의하되 이를 관련 사업을 통하여 간접적으로 지원하는 기금인지, (3) 양자가 혼합된 하이브리드(hybrid) 기금인지에 따라 기금으로 할 수 있는 사업과 그 세부 설계안은 달라지게 됨

○ (1) 피해구제기금

- 피해구제기금의 본질상 말 그대로 기금운용주체가 피해자 손해를 직접 배상 또는 보상하는 것이 동 기금의 핵심 사업임

- 단 손해배상/보상을 위하여는 손해액의 확정이 필요하고, 이는 당사자간 합의가 없으면 사법절차를 통하여 확정되어야 하므로, 피해의 ‘신속한 구제’에는 여전히 일정한 한계 노정

- 특히 개인정보 사고로 인한 손해는 특별한 사정이 없는 한 '정신적 손해'이어서 재산적 손해와 달리 본질적으로 객관적·구체적 산정이 곤란하므로, 법원의 판단이 부재한 상황에서 기금운용주체가 손해액을 결정하여 지급하기 곤란하다는 점을 고려하여야 함

- 따라서 직접적인 피해구제는 판결이나 조정 등으로 확정된 손해배상액 중에서 피고의 자력 부족이나 집행 불능(대표적으로 파산) 등으로 실제 전보가 불가능한 손해액을 한도로 이루어지는 것이 타당

  * 유사/관련 사례: 법무부 범죄피해자보호기금, 미 소비자금융보호국(CFPB) 민사제재펀드(consumer financial civil penalty fund), 미 증권거래위원회(SEC) 페어펀드(fair fund)

■ CFPB, 민사제재펀드

ㅇ 개요: 연방 소비자금융법 위반을 이유로 CFPB가 징수한 모든 민사 벌금(civil penalties)을 재원으로 소비자 피해 직접 구제에 사용되는 기금

ㅇ 근거 법령: 도드-프랭크법(Dodd-Frank Act)

ㅇ 배분 매커니즘: '보상받지 못한 손해'의 원상회복

- CFPB의 집행 조치 결과 위반 기업(피고)에게 civil penalty가 부과된 사건의 피해자군(class of victims)이 대상

- 법원이 피고에 직접 명한 원상회복 금액 중 피고의 자산, 자력 부족, 집행 불능 등으로 피해자가 실제로 지급받지 못했거나 받을

214) 증권법 위반으로 인한 손해는 기본적으로 재산적 손해이고 객관적 추산이 어렵지 않으므로(예컨대 주식가격조작이라면 피해자의 보유주식수 * 조작된 가격) 법원의 판결이 없더라도 SEC가 손해를 산정할 수 있다.

가능성이 없는 잔여 손해액을 지급

- 피해자 소재를 파악할 수 없거나 개별 지급 비용이 실익보다 커서 분배가 현실적이지 않은 경우 등에는 '소비자 교육 및 금융 리터러시 프로그램'의 재원으로만 사용됨

ㅇ 특징: 교차 보전 가능(A 사건 재원을 B 사건의 배상으로 사용 가능)

■ SEC, 페어펀드

ㅇ 개요: SEC가 특정 기업의 증권법 위반을 적발하여 합의하거나 법원 판결을 통해 징수한 벌금, 환수금 등으로 '해당 사건 전용' 펀드를 개설하여 피해자에게 배분

ㅇ 근거 법령: 사베인스-옥슬리법(Sarbanes-Oxley Act), 도드-프랭크법(Dodd-Frank Act)

ㅇ 배분 매커니즘: '특정 사건별(case-specific) 배분계획'에 따른 배분

- 특정 사건에서 징수한 합의금, 벌금, 환수금 등을 재원으로 SEC가 전용 펀드 개설

- SEC는 위반 행위로 피해를 입은 적격 투자자 범위, 손실 산정 공식 등을 명시한 배분 계획안을 수립하여[214] 공시한 후 SEC 결의로 승인

- 피해자들이 거래 내역을 증빙하여 청구하면 배상금 분배

ㅇ (2) 간접지원기금

- 피해구제기금과 달리 간접지원기금은 손해를 직접 배상(또는 보상)하는 것이 아니라 (i) 피해자의 손해전보에 필요한 절차를 간접 지원하거나 (ii) 그 외 넓은 의미에서 개인정보 사고의 사전적 예방 및 사후적 보호조치 강화 등의 사업을 포괄하는 기금임

- (i)에 해당하는 사업으로 ① 개인정보 사고 피해자가 제기하는 손해배상소송이나 분쟁조정절차에서 피해자가 부담해야 하는 소송비용(인지대, 소송대리인 선임비용, 전문가 감정 등 증명비용 등) 지원 사업, ② 행정당국이 개인정보처리자 등 개인정보보호법 위반자에게 제기하는 행정소송의 소송비용 지원 사업 등을 생각해 볼 수 있음

- (ii)에 해당하는 사업은 개인정보 보호 교육, 피해 상담센터 운영, 개인정보 보호 법제도 및 정책연구, 개인정보 사고 공익신고자 포상금 지급, 개인정보 보호 인력, 기술 및 산업 육성 등 다양하게 생각해 볼 수 있음

ㅇ (3) 하이브리드형 기금

- 하이브리드형 기금은 피해구제기금과 간접지원기금 양자의 성격을 모두 갖는 기금으로, 위 (1), (2)의 사업을 재원 한도 내에서 적절히 혼합하는 방식으로 운용 가능할 것으로 생각함

□ 최근 입법 동향

ㅇ 현재 국회에서는 여야를 막론하고 개인정보 보호 관련 제재수입을 피해자 구제 재원으로 환류시키기 위한 법안들이 발의되었음

- 최근 대규모 개인정보 유출 사고 등이 발생하고 있는 가운데, 개인정보 보호법에는 법 위반에 따른 과징금 부과 등 제재 및 부당이득 환수에 관한 사항은 규정되어 있으나 이로 인한 정보주체 피해의 구제 및 원상회복을 위한 제도적 장치는 상대적으로 미흡한 측면이 있다는 지적이 지속적으로 제기되고 있음

ㅇ 구체적으로 개인정보피해구제에 특화된 기금 설치안으로 박상혁 의원(민)의 「개인정보침해피해자보호기금법안」과 박정하 의원(국)의 「개인정보 보호법 일부개정법률안」이 상정되어 논의 중

ㅇ 한편 기획예산처 주도로「공익신고장려기금법안」이 추진되고 있으며, 동 법안은 소위 '개인정보보호기금'과는 목적이나 내용이 상이하지만 최근 다발적으로 제기되는 기금 소요 이슈를 대응하기 위한 재정당국의 의지로 해석되어 주목할 필요가 있음

- 기획예산처는 담합, 주가조작, 보조금 부정수급 등 국민경제와 사회질서를 훼손하는 반사회적 행위에 대한 공익신고를 활성화하기 위하여 「공익신고장려기금」 신설 추진 (공익신고장려기금 신설시 금융위원회, 공정거래위원회 등의 신고포상금은 동 기금을 통하여 통합 집행)
- 공익신고장려기금은 전체 신고포상금 중 공익신고 장려의 시급성이 높고 과징금 · 과태료 · 환수금 등 금전적 제재와 직접 연계되는 분야를 대상으로 우선 추진
- 공익신고장려기금은 반사회적 행위로 인한 피해의 사전 예방 및 피해자 지원 차원에서 사전 예방교육, 법률 구제 등 피해자 간접지원 사업도 병행 추진할 예정으로 알려져 있으며, 이 부분에서 개인정보보호기금사업도 일부 포함될 여지가 있을 것으로 보임

5. 시사점

□ ‘공법적 제재’와 ‘민사적 피해 구제’의 연계

- ㅇ 연방규제기관(FTC)의 행정제재 절차(동의명령) 내에 소비자 피해 구제 기금 조성을 의무화하는 미국 방식은 소액 다수 피해 구제에 실효적임
- ㅇ 다만 미국에서 말하는 기금(fund)은 우리 국가재정법상 기금과 유사한 경우도 있지만 이와 달리 일회성, 특정성, 한정성을 가지는 말 그대로 ‘그때그때 설정되고 분배가 끝나면 해산하는’ 펀드의 성격을 가지는 경우도 있으므로, 도입 여부는 신중하게 검토할 필요
  - * 특히 우리의 경우 국가재정법상 기금 설치가 불가능한 것은 아니지만, 재정당국이 반대할 경우 기금 신설은 매우 어렵다는 현실을 충분히 고려하여야 함
- ㅇ 단순 행정 과징금 부과에 그치지 않고, 피해자 실질 구제로 직결되는 동의의결 제도 도입 검토 필요

□ 다수 피해의 효율적 구제를 위한 절차법적 장치 정비

- ㅇ 미국식 집단 소송은 강력한 억제력이 있으나, 소송 남발 및 과도한 합의금 인플레이션 등의 법적 리스크가 존재
- ㅇ 영국의 GLO(Opt-in)나 일본의 2단계 특례법 구조(공통의무확인 → 간이확정)는 대륙법적 사법 체계를 유지하면서도 절차적 신속성을 도모하는 실용적 대안을 제시하고 있어 참고 필요

□ 정신적 손해(위자료)의 증명책임 및 손해 인정 기준 구체화

- ㅇ 재산상 손해가 발생하지 않거나 발생하더라도 증명이 어려운 개인정보 사고 특성상 정신적 고통(위자료)의 인정 범위 설정이 핵심

ㅇ EU(GDPR)이나 영국은 법상 권리가 침해되었다는 사실만으로 곧바로 손해(재산적, 비재산적 불문)가 인정되지 않는다는 입장인 바, 이러한 국제적 동향도 참고할 필요

□ AI 및 빅데이터 환경에 대비한 불법행위 유형의 확장

ㅇ 외부 해킹에 의한 단순 유출을 넘어 무단 데이터 제공, 종단간 암호화 허위 주장(Zoom 사례), 쿠키/행동 이력 기반 프로파일링 오남용(리쿠나비 사례) 등의 새로운 개인정보 침해 유형 등장

ㅇ AI 시대의 데이터 오남용 및 프로파일링에 따른 인격적 이익 침해를 개인정보 침해 손해배상으로 포섭할 수 있는지에 관한 이론 구성과 (필요시) 법제 정비 여부를 장기적으로 논의할 필요

참고문헌

국가정보원 (2023), “국가 클라우드 컴퓨팅 보안 가이드라인”.

국가정보원 (2025), “2025년 국가 정보보호 백서”.

국가정보원 (2026), “국가·공공기관 AI 보안 가이드북”.

국가정보원 (2026), “국가사이버보안기본지침".

국가정보원 (2026), “챗GPT 등 생성형 AI 활용 보안 가이드라인”.

국회도서관 (2022), “2022년 일본 국가안보전략 3대 문서 개요”.

대외정책연구원 (2024), “주요국의 사이버안보 정책과 한국에 대한 시사점”.

동아일보 (2024), “‘납치 당했다’ 美유학 딸 목소리, AI 변조였다”.

삼성SDS (2026), “26년 사이버보안 위협 트렌드 전망 및 대응”.

양천수・지유미 (2018), “미국 사이버보안법의 최근 동향”.

이상현 (2019), “일본의 사이버안보 수행체계와 전략”.

장교식 (2026), “일본의 사이버보안과 능동적 사이버방어 법제에 관한 검토”.

정보통신정책연구원 (2021), “인공지능: 사이버보안 패러다임의 전환”.

정인영 (2025), “페이스북의 개인정보 침해에 대한 미국 내 입법, 사법 행정적 대응현황(2)”.

최경진 외 (2024), “EU 인공지능법”.

한국인터넷진흥원 (2024), “2024 해외 사이버보안 입법동향”.

한국인터넷진흥원 (2025), “2025년 1분기 인터넷 정보보호 법제동

향”.
한국인터넷진흥원 (2025), “인공지능(AI) 보안 안내서”.
한국인터넷진흥원 (2026), “최근 EU 사이버보안 입법동향과 시사점”.
한국지능정보사회진흥원 (2023), “美 NIST AI 위험관리 프레임워크(AI RMF) 1.0 분석 및 시사점”.
한국행정연구원 (2024), “인공지능 기술 도입에 따른 위험상과 다학제적 대응 방안”.
EY한영 산업연구원 (2026), “AI의 위협으로 촉발된 사이버보안의 패러다임 대전환”.
IBM (2026), “2026년 데이터 유출 비용 CODB 보고서”.
KISTI (2025), “사이버보안 위협 진화와 AI: 디지털 전환 시대 보안 전략”.
KOTRA (2024), “일본 사이버보안 정책의 새로운 움직임과 시사점”.
S2W TALON (2025), “AI 및 LLM을 활용한 위협 사례 연구 #02: 취약점”.

AISLE (2026), “AI Cybersecurity After Mythos: The Jagged Frontier”.
Alireza Lotfi (2025), “Automated Vulnerability Validation and Verification: A Large Language Model Approach”.
Anthropic (2025), “Disrupting the First Reported AI-Orchestrated

Cyber Espionage Campaign”.

Anthropic (2026), “Alignment Risk Update: Claude Mythos Preview”.

Anthropic (2026), “Assessing Claude Mythos Preview’s Cybersecurity”.

Anthropic (2026), “System Card: Claude Fable 5 & Claude Mythos 5”.

Battista Biggio (2013), “Poisoning Attacks against Support Vector Machines”.

Cabinet Office, Government of Japan (2026), “人工知能基本計画”.

Chibuike Samuel Eze (2024), “Analysis and Prevention of AI-Based Phishing Email Attacks”.

Chris Stokel-Walker (2026), “What Is Mythos and Why Are Experts Worried about Anthropic’s AI Model”.

CISA et al. (2026), “Careful Adoption of Agentic AI Services”.

Council of the European Union (2025), “EU Blueprint for Cyber Crisis Management”.

Court of Justice of the European Union (2023), “Case C-300/21, EU:C:2023:370”.

Court of Justice of the European Union (2024), “Case C-741/21, EU:C:2024:288”.

CrowdStrike (2025), “Global Threat Report”.

DARPA (2017), "Explainable Artificial Intelligence”.

Deeksha Hareesha Kulal (2025), “Robust ML-based Detection of Conventional, LLM-Generated, and Adversarial Phishing E-mails Using Advanced Text Preprocessing”.

Department of Homeland Security & Department of Justice (2026), “Guidance to Assist Non-Federal Entities to Share Cyber Threat Indicators and Defensive Measures with Federal Entities under the Cybersecurity Information Sharing Act of 2015”.

European Commission (2019), “Political Guidelines for the Next European Commission 2019–2024”.

European Commission (2024), “The Draghi Report on EU Competitiveness”.

European Commission (2025), “Digital Omnibus”.

European Commission (2025), “European Preparedness Union Strategy”.

European Commission (2026), “Commission Strengthens EU Cybersecurity Resilience and Capabilities”.

European Commission (2026), “EU Action Plan on Cybersecurity and Artificial Intelligence”.

Google (2023), “Secure AI Framework (SAIF)”.

High Court of Justice, King’s Bench Division (2023), “Bekoe v London Borough of Islington [2023] EWHC 1668 (KB)”.

Ian J. Goodfellow (2015), “Explaining and Harnessing Adversarial Examples”.

IASME (2026), “Cyber Essentials – Cyber Liability Insurance”.

LevelBlue (2023), “WormGPT and FraudGPT – The Rise of Malicious LLMs”.

Mingyue Yang (2020), “Using Machine Learning to Detect Software Vulnerabilities”.

MITRE (2026), “A Sensible Regulatory Framework for AI Security”.

NCSC (2026), “Active Cyber Defence – Early Warning”.

NCSC (2026), “Cyber Essentials”.

NIST (2012), “Computer Security Incident Handling Guide (SP 800-61 Rev. 2)”.

NIST (2020), “NIST SP 800-207: Zero Trust Architecture”.

NIST (2023), “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”.

NIST (2023), “NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence”.

NIST (2025), “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations”.

OpenAI (2026), “OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation”.

OWASP (2025a), “Agentic AI – Threats and Mitigations v1.1”.

OWASP (2025b), “OWASP Top 10 for Agentic Applications for 2026”.

OWASP (2025c), “OWASP Top 10 for LLM Applications 2025”.

Shimei Pan (2018), “Automatically Infer Human Traits and Behavior from Social Media Data”.

South China Morning Post (2024), “UK Multinational Arup Confirmed as Victim of HK$200 Million Deepfake Scam That Used Digital Version of CFO to Dupe Hong Kong Employee”.

Terry Yue Zhuo (2026), “To Defend Against Cyber Attacks, We Must Teach AI Agents to Hack”.

The White House (2021), “Improving the Nation’s Cybersecurity”.

The White House (2025), “America’s AI Action Plan”.

The White House (2025), “Sustaining Select Efforts to Strengthen the Nation’s Cybersecurity and Amending Executive Order 13694”.

The White House (2026), “President Trump’s Cyber Strategy for America”.

The White House (2026), “Promoting Advanced Artificial Intelligence Innovation and Security”.

UK Government (2026), “International AI Safety Report 2026”.

WEF (2026), “Global Cybersecurity Outlook 2026”.

Yuxuan Zhu (2025), “Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities”.

국민과 함께 미래를 여는 국회

(07233) 서울시 영등포구 의사당대로1(국회입법조사처)
Tel. 02-6788-4510 www.nars.go.kr

국회입법조사처
NATIONAL ASSEMBLY RESEARCH SERVICE