사이버안보연구, 3집 1호 2026 여름, 049-087

국내 소버린 AI 검증 프레임워크 도입 방안 연구

김도연 | 서울여자대학교
조휘정 | 이화여자대학교

논문접수일 2026년 5월 15일 | 논문심사일 2026년 6월 8일 | 게재확정일 2026년 6월 25일

논문 요약

본 연구는 사이버 공간이 상시적 위협 환경으로 고착된 상황에서 데이터·컴퓨팅·모델·인력의 결합재로서 AI가 초래하는 주권·안전·회복탄력성의 공백을 해소하기 위해 한국형 소버린 AI 6-Gate 검증 체계(K-SAF)를 제안한다. 클라우드·플랫폼·공급망의 초연결로 단일 실패지점과 불투명성이 연쇄 위험을 증폭시키는 가운데, 대규모 학습·배포에 대한 외부 벤더 의존은 정보주권과 감사 가능성의 결손을 키운다. K-SAF는 정책 적합성, 주권 거버넌스, 안전성·신뢰성, 운영 복원력의 네 게이트를 통해 법·관할·접근 통제와 적대적 강건성·프라이버시·품질 지표를 정량화하고, 실시간 방어·자가치유·즉시 격리 절차를 표준화한다. 각 게이트는 임계치 기반 통과/중단과 재시도 메커니즘을 결합해 하나라도 미달 시 생애주기를 강제 중단하는 안전망을 구현하며, 조달·감사 체계와 연결되어 탐지-격리-분석-복구의 국가 대응 체인을 실행 가능하게 한다. 제안 체계는 개방성과 상호운용성을 유지하면서 핵심 의사결정권과 책임성을 국내에 고정시켜, 국가 인프라로서의 AI 운용을 지속 가능하고 책임 가능한 수준으로 끌어올리는 최소 요건을 제공한다.

주제어 소버린 AI, 사이버 안보, 거버넌스, 검증 프레임워크, 데이터 주권

Ⅰ. 머리말

현대 국가안보는 디지털 인프라의 가용성과 신뢰성을 포괄하는 복합적 개념으로 재편되고 있다. 사이버 공간은 낮은 진입 비용·익명성·초연결성으로 인해 공격자에게 비대칭적으로 유리한 환경을 형성하며(U.S. Department of Defense 2011), 상호연결된 인프라는 단일 실패지점(SPoF)의 파급을 사회 전반으로 확산시킨다(Li and Liu 2021). AI 시스템은 전 생애주기에 걸쳐 공격 표면이 기하급수적으로 확장되고(Eeet al. 2024), 핵심 인프라를 외부 벤더에 의존하는 구조는 회복탄력성과 정보주권의 공백으로 직결된다. 미국·EU·중국·일본 등 주요국이 소버린 AI를 핵심 전략 의제로 채택하고 검증 체계를 병행 제도화하는 반면, 국내에는 데이터 주권·연산 경계·모델 안전성·운영 복원력을 전주기·다계층으로 판정하는 국가 단위 프레임워크가 부재하다. 이러한 공백을 배경으로, 본 연구는 6개 게이트로 구성된 한국형 소버린 AI 검증 프레임워크(K-SAF)를 제안하고, 집행 가능하고 재현 가능한 검증 기준선을 제도화하는 실천적 근거를 제공하고자 한다.

1. 연구 방법

본 연구는 한국형 소버린 AI 검증 프레임워크의 설계 근거를 확보하기 위해 체계적 문헌 검토와 비교사례분석을 결합한 방법론을 적용하였다.

1) 체계적 문헌 검토

국내외 AI 거버넌스 프레임워크 및 사이버안보 표준 문헌을 체계

적으로 검토하여 K-SAF 설계의 이론적 기반을 도출하였다. 검토 범위는 ① AI 생애주기 거버넌스 관련 국제 표준(ISO/IEC 22989:2022, ISO/IEC 42001, NIST AI RMF 1.0), ② 주요국 AI 규제·정책 문서(EU AI Act, OMB Memorandum M-25-21·M-25-22, 일본 디지털청 「생성AI 조달·이용 가이드라인」), ③ AI 보안 아키텍처 관련 기술 문헌(MITRE ATLAS, NIST AI 100-2e2023, ENISA 위협지형 보고서), ④ 국내 법령 및 가이드라인(「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」, KISA 「인공지능(AI) 보안 안내서」, 개인정보보호위원회 AI 관련 안내서)으로 설정하였다. 분석 기준은 각 문헌이 다루는 통제 영역, 법적 구속력 수준, 기술적 집행 수단, 생애주기 커버리지의 네 항목으로 구조화하였으며, 이를 통해 기존 프레임워크의 공통 요소와 미비 영역을 식별하였다.

2) 비교사례분석

소버린 AI 검증 체계를 제도화하고 있는 미국·EU·중국·일본의 사례를 공통 분석 기준에 따라 비교하였다. 분석 기준은 ① 거버넌스 층위(법령·표준·조달·연산 인프라), ② 법적 구속력(강제 규정·권고·자율 준수), ③ 기술 집행 수단(정책 엔진·인증·등록·감사), ④ 생애주기 커버리지(조달 전·운영·폐기 단계 포함 여부)의 네 가지로 설정하였다. 각 사례에서 나타나는 제도적 접근의 공통점과 차이점을 분석함으로써, 한국의 제도 환경에서 도입 가능한 요소와 국내 특수성을 반영한 보완 요소를 도출하였다.

3) 6-Gate 도출 논리

위 두 방법론의 결과는 다음의 3단계 논리로 K-SAF 설계에 연결된다. 첫째, 체계적 문헌 검토를 통해 AI 거버넌스의 이론적 원칙과 국

제 프레임워크의 공통 통제 영역을 확인하였다. 둘째, 비교사례분석을 통해 해외 주요국 프레임워크에서 상대적으로 미비한 두 영역을 한국 제도 환경의 공백으로 규명하였다. 셋째, 이 두 단계의 분석 결과를 통합하여, 기존 국제 프레임워크의 공통 통제 4개 영역에 한국형 신설 2개 게이트를 추가한 6-Gate 구조를 도출하였다. 이 과정은 이론적 원칙에서 국제 비교를 거쳐 한국 특화 설계로 이어지는 연역적·귀납적 혼합 접근에 해당하며, 각 단계의 근거는 III장에서 구체적으로 제시된다.

Ⅱ. 국내/외 소버린 AI 동향 분석

1. 소버린 AI 시대의 도래

디지털 전환으로 데이터·연산·모델·플랫폼이 결속되면서 AI는 국가 핵심 인프라가 되었다. 전 생애주기가 네트워크·클라우드·API로 연결되어 공격 표면이 급격히 확대된 가운데, ENISA(2025)는 2024-2025년 수천 건의 사이버 사고가 EU 디지털 인프라 회복력을 시험하고 있음을 지적한다. 따라서 소버린 AI의 도래는 기술 확산의 귀결이 아니라 사이버 안보 차원의 필연적 대응이다.

1) 소버린 AI의 개념

(1) 소버린 AI의 개념의 등장과 선행 논의

소버린 AI(Sovereign AI)는 특정 기술 스택의 국산화나 데이터 현지화와 구별되는 개념으로, AI 시스템의 전 생애주기에 걸쳐 법적 관할권과 기술적 실효 통제권을 동시에 확보하는 것을 목표로 한다. 주요국의 입법 동향을 비교·분석한 선행 연구는 AI 주권의 확보가 기술 자립 그 자체보다 법적 규율권의 확립에 달려 있음을 밝힌 바 있으며(김태훈·이성엽 2025), 이는 소버린 AI가 인프라 보유의 문제를 넘어 제도적 통제권의 문제임을 시사한다. 이러한 맥락에서 소버린 AI는 정책·기술·산업의 삼중 결합을 통해 규범 준수, 공급망 리스크 완화, 공공·민간 경쟁력 제고를 동시에 추구하는 국가 인프라 전략으로 규정할 수 있다.

(2) 소버린 AI의 구성 축과 상호의존 구조

소버린 AI는 상호의존적인 네 개의 통제 축으로 구성되며, 하위 층위의 통제가 상위 층위의 실효성을 결정하는 계층적 의존 구조를 형성한다. 데이터 주권이 확보되지 않으면 연산 주권이 실질적 의미를 갖지 못하며, 연산 주권 없이는 모델 주권이 공허해지고, 이 세 축이 결속되지 않으면 플랫폼 주권은 형식적 선언에 머문다.

첫째, 데이터 주권은 데이터의 수집·이동·보관·학습·추론 전 과정에서 국내법 우선의 관할을 확립하고 국경 간 이전 시 통제력을 유지하는 것을 의미한다. EU GAIA-X는 연합형 데이터 인프라를 통해 이용자의 자율적 통제 모델을 제시하며(ETRI 2021), 한국은 「개인정보보호법」상 민감정보 국외 이전 제한이 법적 기반을 형성하나 AI 학습 데이터 전주기 거버넌스는 아직 제도화 단계에 머물러 있다. 둘째, 연산 주권은 대규모 모델 학습·추론에 요구되는 HPC/AI 컴퓨팅 파이프라인을 국내에서 자립적으로 운용하는 역량을 의미하며, 역외 의

존은 수출통제·공급망 충격 시 AI 서비스 연속성을 직접 위협한다. EU는 EuroHPC 엑사스케일 시스템과 19개 AI 팩토리 네트워크로 연산 자원 접근성을 보장하는 반면(EuroHPC Joint Undertaking 2024), 한국은 국가 슈퍼컴퓨팅 인프라와 민간 클라우드 GPU 자원이 분절된 구조로 운용되어 통합적 설계가 요구된다. 셋째, 모델 주권은 자국 데이터로 모델을 안전하게 학습·운영하고 위험관리·감사 체계를 설계 단계부터 내재화하는 것을 의미한다. EU AI Act는 GPAI 제공자에게 기술문서 작성·학습 데이터 요약 공개·저작권 준수 정책 수립을 의무화하며(Regulation (EU) 2024/1689), 모델 주권의 실현은 가중치 보유를 넘어 학습·평가·배포 이력 전반의 추적가능성 확보를 전제로 한다.

(3) 소버린 AI와 사이버안보의 교차점

소버린 AI의 전략적 중요성은 기술 자립 논리를 넘어 사이버안보 관점의 구조적 필요성에서 도출된다. AI 시스템은 데이터 파이프라인·모델 아티팩트·API 인터페이스·운영 인프라가 상호 연결된 복합 공격 표면을 형성하며, 이 중 어느 하나라도 역외 사업자에 의존하는 경우 공급망 공격·모델 오염·추론 탈취에 대한 국가 대응 역량이 구조적으로 제한된다. ENISA AI 위협 지형 보고서(ENISA 2025)는 소버린 AI 통제 역량의 부재가 국가 핵심 인프라의 회복탄력성을 약화시키는 직접적 요인임을 지적하며, 이는 소버린 AI의 네 축이 산업 경쟁력 지표가 아닌 국가 사이버안보 아키텍처의 필수 구성 요소임을 시사한다.

2) 소버린 AI 도입의 필요성

소버린 AI 도입의 필요성은 회복력, 전략적 자율성, 산업·공공가치

의 세 측면에서 도출된다.

첫째, 국가 회복력과 규범 준수의 내재화다. AI 생애주기는 계층 간 상호의존으로 단일 실패지점(SPoF)과 연쇄 위험이 증폭되기 쉽다(Li and Liu 2021). 소버린 AI는 데이터·연산 경계에서 위험을 분할·완충하고, 감사 가능한 로깅과 개방형 표준을 통해 사후 규명 가능성을 높인다. 또한 역내 통제 가능한 데이터·연산·플랫폼을 전제로 규제 준수 절차를 설계 단계에 내재화함으로써, 의료·공공·에너지 등 핵심 분야에서 회복탄력성과 규범 준수를 단일 아키텍처 안에서 동시에 달성한다.

둘째, 지정학·공급망 리스크 완화다. 현대 AI 인프라의 공급망은 소수 글로벌 사업자에 집중되어, 수출통제·제재 같은 역외 요인 발생 시 운영 연속성이 즉각 위협받는다. 2022년 이후 미국의 대중국 AI 반도체 수출통제 강화가 그 위험성을 실증한다. 소버린 AI는 역내 연산 역량과 연합형 데이터 인프라로 공급 충격을 흡수하고, 멀티벤더 설계로 의존도를 분산시킨다. 반도체 설계·제조 역량을 보유하면서도 AI 연산용 GPU의 외산 의존도가 높은 한국의 구조적 비대칭은 이를 해소할 정책 설계를 요구한다.

셋째, 산업 경쟁력과 공공가치의 동시 달성이다. 영어·서구 맥락에 최적화된 범용 모델은 한국어의 특수성과 의료·법률·행정 도메인 규범을 충분히 반영하지 못해 성능 저하와 법적 리스크를 유발한다. 자국 데이터로 학습된 소버린 AI는 이를 구조적으로 해소하며, 데이터 출처 관리·추론 감사·책임 귀속 구조를 역내에서 직접 통제함으로써 신뢰성과 책임성의 기술적 제도화를 가능하게 한다. 이러한 국가 단위 AI 생태계는 공통의 신뢰 기반 위에서 협력하는 혁신 환경을 조성하여 장기적 국가 경쟁력으로 이어진다.

2. 소버린 AI 해외 동향 분석

1) 미국

미국은 포괄적 AI 기본법 없이 행정명령·표준·조달 지침·연산 인프라를 유연하게 결합하여 소버린 AI 역량을 구축해 왔다. 강제적 법률 대신 조달과 사용사례 공개 같은 실질적 집행 수단을 통해 시장에서 작동하는 규범을 형성한다는 점이 핵심이다.

규범 층위에서는 바이든 행정부의 Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence(2023.10)이 AI 안전성·투명성·국가안보를 포괄하는 상위 프레임을 제시하였고, 트럼프 행정부가 이를 폐기·재정렬했으나 행정명령 중심의 거버넌스 운영 방식은 유지되고 있다. 표준·평가 층위에서는 NIST AI RMF(2023)와 GPAI 특화 가이드라인이 위험 관리 참조체계를 제공하며, AISI와 AISIC 컨소시엄이 시험·검증 역량의 국내 정착을 추진하고 있다.

공공부문 집행 층위에서는 OMB 메모랜덤을 통해 CAIO 지명·AI 거버넌스 위원회·사용사례 인벤토리 공개가 제도화되었다. 2025년 4월 OMB Memorandum M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust(2025.4)와 OMB Memorandum M-25-22: Driving Efficient Acquisition of Artificial Intelligence in the Federal Government(2025.4)가 이를 대체하였으나 핵심 거버넌스 구조는 유지되어, 정권 교체에도 "규범→평가→책임→조달"의 정책 순환 구조가 안정적으로 작동하고 있다. 연산 인프라 층위에서는 El Capitan·Frontier·Aurora 등 엑사스케일급 HPC가 공공 연산 자산으로 기능하고, CHIPS and Science Act가 약 527억

달러 규모의 반도체 R&D 지원으로 공급망 국내 경로를 강화한다.

미국 모델의 핵심 함의는 입법 부담 없이도 행정명령과 OMB 메모를 매개로 책임성과 복원력을 제도화할 수 있다는 데 있다.

2) 중국

중국의 소버린 AI 전략은 콘텐츠 규율·데이터 관할·평가 표준·연산 자립·시장 등록의 다섯 층위를 일관된 국가 통제 구조로 결합하는 데 특징이 있다.

규범 층위에서는 2023년 8월 시행된 「생성형 AI 서비스 관리 잠정 조치」(CAC 2023)가 공개형 생성형 AI 서비스 제공자에게 윤리·데이터 보안·콘텐츠 규범 준수와 보안평가·표시 의무를 부과한다. 데이터 주권 층위에서는 CSL·DSL·PIPL 3법이 중요데이터와 민감정보의 국내 보관 의무·국외 이전 제한을 통해 국가 관할 중심의 통제 체계를 구축한다. 2024년 3월 시행된 「국경 간 데이터 이전 촉진·규제 규정」은 준수 부담을 일부 완화하였으나, 중요데이터·CIIO에 대한 엄격한 요건은 유지된다.

표준·평가 층위에서는 TC260의 「생성형 AI 서비스 기본 안전 요구」가 학습 데이터 안전·모델 안전·안전 조치·안전성 평가의 네 축을 제시하며, "표준–평가–등록"의 연결 구조가 제도화되고 있다. 연산 층위에서는 화웨이 Ascend 계열 NPU를 중심으로 국산 학습·추론 스택을 구축 중이나, 수출통제에 따른 공정 수율 한계·HBM 조달 제약 등 구조적 한계가 병존한다. 시장 집행 층위에서는 CAC가 생성형 AI 서비스의 사전 등록과 표시를 의무화하여 "등록 없이는 공개 서비스 불가"라는 진입 통제 메커니즘을 운영한다.

중국 모델의 핵심은 시장 진입 자체를 정책 도구로 제도화한다는

데 있다. 법령·표준·등록을 통한 규제 정밀화와 국산 연산 생태계 확장을 병행하는 이중 전략은, 규제 준수·감독을 용이하게 하면서 공급망 리스크를 국가 시스템 내부에서 흡수하려는 시도로 평가된다.

3) EU

EU는 구속력 있는 법률을 중심축으로 데이터 주권·클라우드 규범·공공 연산·조달 표준을 단계적 집행 체계로 정렬하는 법제 중심형 접근을 취한다. 위험 기반 분류 체계로 규제 강도를 차등화하고, 기술문서화·적합성 평가·공식 등록·사후 모니터링의 절차적 폐루프와 결합하는 것이 핵심 특징이다.

규범 층위에서 2024년 8월 발효된 EU AI Act(Regulation (EU) 2024/1689)는 금지 관행(2025년 2월)·GPAI 의무(2025년 8월)·전면 적용(2026년 8월)의 단계적 시행 일정을 명시하며, EU AI Office가 가이드·코드·집행을 총괄한다(European AI Office 2025). 데이터 주권 층위에서는 Data Act(2025년 9월 적용)와 DGA가 데이터 접근·공유·이동·재이용의 규범적 연속성을 EU 관할 아래 정립하는 상호보완적 구조로 작동한다.

클라우드·연산 층위에서는 EUCS의 최고 보증 등급에 대한 주권 요건 포함 여부가 쟁점으로 부상한 가운데, EuroHPC JU의 엑사스케일급 슈퍼컴·AI Factory 프로그램이 스타트업·중소기업의 연산 자원 접근 경로를 제공하고(EuroHPC Joint Undertaking 2024), European Chips Act가 2030년까지 430억 유로 이상의 투자로 칩 공급망 내재화를 추진한다. 조달 층위에서는 EU AI 모델 계약조항(MCC-AI)(European Commission 2025)이 AI Act 요건·감사·책임 배분을 계약 단계에 내재화하여 법-조달-운영 간 실행 간극을 축소한다.

위험 기반 분류→기술문서화→적합성 평가→등록→사후 모니터링으로 이어지는 절차적 폐루프는 소버린 AI 검증 프레임워크 설계에 가장 직접적인 참조 모델을 제공한다.

4) 일본

일본은 대외적으로 G7 히로시마 AI 프로세스를 주도하여 '신뢰가능한 AI'의 행동규범을 다자 합의로 끌어올림으로써 규범 주권의 기반을 확보했다(G7 2023). 대내적으로는 2024년 2월 J-AISI를 설치하여 프런티어 모델에 대한 안전성 평가 방법론·벤치마크·레드팀 프로토콜의 국내 허브를 구축했다.

조달·운영 층에서는 2025년 5월 디지털청이 「생성AI 조달·이용 가이드라인」(デジタル庁 2025)을 공표하여 각 부처의 CAIO 지정, 생애주기 전 단계의 준수책임, 분기별 운영 보고를 제도화했다. 고위험 도입의 경우 조달 단계에서 사전 보고와 자문 절차를 거치도록 하여 사실상의 사전심사가 작동한다. 이로써 J-AISI의 평가 프레임과 디지털청 가이드라인이 결합되어 "국내 기준을 충족한 AI만 공공시장에 진입"하는 구조가 제도화됐다.

연산 인프라 측면에서는 ABCI를 2025년 3.0으로 업그레이드하여 공공 연산 경로를 확장하고, 가버먼트 클라우드의 국내 사업자 경로를 분기·월 단위로 점검하고 있다. 이는 규범·평가 보증·조달 집행·연산 복원력을 통합 설계함으로써 정책적 자율성과 기술적 회복탄력성을 동시에 제도화한 사례다.

5) 소결 : 해외 사례 비교분석과 K-SAF 설계 논리

미국·EU·중국·일본의 소버린 AI 거버넌스를 거버넌스 층위, 법적

구속력, 기술 집행 수단, 생애주기 커버리지의 네 기준으로 비교하면 공통점과 차이점이 뚜렷하게 나타난다. 공통적으로 4개국 모두 공급망 적합성, 안전성 검증, 운영 복원력 영역에서 명시적 요건을 갖추고 있으며, 공공조달을 핵심 집행 수단으로 활용하고 있다. 반면 두 가지 영역에서 공통적인 제도적 공백이 확인된다. 첫째, 데이터·연산 주권의 기술적 집행 측면에서 중국은 국내 잔류 의무를 법령으로 규정하고 있으나, 미국·EU·일본은 정책 엔진 기반의 런타임 집행까지는 나아가지 않고 있다. 둘째, 폐기·전환 단계는 EU가 사후 모니터링 의무를 부과하나 구체적 절차는 4개국 모두에서 미비하다. K-SAF는 4개국의 공통 통제 요소를 흡수하되, 이 두 공백을 각각 Gate 2(데이터·연산 주권)와 Gate 6(폐기·전환 보증)으로 신설하여 보완함으로써 6-Gate 구조를 도출하였다.

3. 소버린 AI 국내 동향 분석

국내에서도 소버린 AI 도입을 위한 시도가 본격화되고 있다. 정부는 독자 AI 파운데이션 모델 프로젝트를 공모·선정하고 데이터 공급기관을 병행 모집하는 등 연산·데이터·인재를 묶는 지원 체계를 가동했으며(과학기술정보통신부 외 2025), 공공부문 AI 서비스의 목록화·공개 등 거버넌스 정비도 예고되었다. 그럼에도 국내에는 소버린 AI를 체계적으로 검증할 국가 단위 프레임워크가 아직 없다. 현행 사업 공고·과제관리·기술평가는 개별 프로젝트 심사에 머물러, 데이터 주권·연산 경계·모델 안전성·운영 복원력을 전주기로 판정하고 사후까지 추적하는 제도적 페루프가 부재하다. 반면 EU는 위험기반 분류에 따라 기술문서화·적합성평가·등록·사후 모니터링의 절차적 순환을

법으로 고정하였고, 영국은 AI assurance 지침과 플랫폼으로 검증 생태계의 표준화를 추진하는 등(DSIT 2024) 해외에서는 구축과 검증의 제도화가 병행·의무화되고 있다. 그동안 국내 논의가 도입 자체에 집중되었다면, 이제는 "도입된 것을 어떻게 신뢰할 것인가"가 동일한 무게로 다뤄져야 한다. 데이터·연산·모델·운영이 촘촘히 결박된 AI 인프라에서는 상시적·증거기반 검증이 전제되지 않으면 조달·감독·책임 추궁이 작동하지 않기 때문이다. 따라서 정책 적합성·주권 거버넌스·안전성·운영 복원력을 연속적으로 점검하는 국가 단위 검증 프레임워크가 요구되며, 이는 구축 사업과 공공·민간 조달을 검증 가능한 운영체계로 연결해 국제 규범 정합성과 시장 규율, 국가 회복력을 확보하는 기반이 된다. 지금 필요한 것은 "더 많은 도입"이 아니라, 도입의 신뢰를 입증하는 제도화다.

III. 한국형 소버린 AI 검증 프레임워크 도입 정책 제안

해외 주요국은 소버린 AI를 국가 인프라로 규정하고, 위험 기반 분류·기술문서화·적합성 평가·사후 모니터링의 절차적 순환을 제도화하고 있다. 이는 모델 개발 성공만으로는 공공 신뢰와 규제 적합성을 확보할 수 없으며, 검증 가능한 운영 체계 없이는 조달·배치·확산이 지속될 수 없음을 시사한다. 이에 본 논문은 공급망·정책 적합성(Gate 1), 데이터·연산 주권(Gate 2), 인적·조직 통제(Gate 3), 안전성·윤리 검증(Gate 4), 운영 복원력(Gate 5), 폐기·전환 보증(Gate 6)의 6단계로 AI 생애주기를 검증하는 한국형 소버린 AI 검증 프레임워크를 제안한다.

1. 게이트 구조의 이론적 기반

K-SAF의 6-Gate 구조는 세 가지 이론적 원칙에 기반한다. 첫째, 심층 방어 원칙이다. 독립적 다층 통제를 직렬 배치하는 이 원칙은 사이버보안 아키텍처의 핵심 설계 원리로, AI의 복합 공격 표면에도 동일하게 적용된다(Ee et al. 2024). K-SAF는 이를 생애주기에 투영하여 공급망 진입부터 폐기까지 독립적 검증 관문을 설치하며(Leon et al. 2026), 한 게이트라도 미충족 시 전체 진행이 차단되는 구조는 후속 층위가 보완적으로 작동한다는 심층 방어 논리의 구현이다. 둘째, 제로 트러스트 원칙이다. "명시적 검증, 최소 권한, 침해 가정"을 핵심으로 하는 이 원칙은 데이터 수집에서 추론·운영에 이르는 AI 파이프라인의 각 신뢰 경계에서 지속적 검증을 요구하며(Microsoft Security 2026), K-SAF의 deny-by-default 설계와 게이트별 증적 제출 의무가 그 제도적 표현이다. 셋째, 전주기 거버넌스 원칙이다. ISO/IEC 22989:2022는 AI 생애주기를 7단계로 규정하고(ISO/IEC 2022), NIST AI RMF는 GOVERN·MAP·MEASURE·MANAGE의 순환 구조를 제시한다(NIST 2023). 다만 공급망 진입 이전과 서비스 종료 이후 단계는 거버넌스 공백이 발생하기 쉬워(European AI Office 2025), K-SAF는 이를 각각 Gate 1과 Gate 6으로 포괄해 책임 연속성을 제도화한다.

K-SAF의 6개 게이트는 이 세 원칙과 함께 기존 국제 프레임워크의 통제 범위를 분석하여 한국 제도 환경의 공백을 추가 반영하는 방식으로 도출되었다. 아래 표 1은 주요 국제 프레임워크와 K-SAF의 대응 관계를 정리한 것이다.

〈표 1〉 기존 AI 거버넌스 프레임워크의 통제 범위와 K-SAF 게이트 선정 근거 비교

통제 영역	EU AI Act	NIST AI RMF	ISO/IEC 42001	K-SAF
공급망·조달 적합성	o	△	△	Gate 1
데이터·연산 주권	△	△	△	Gate 2 (신설)
인적·조직 통제	o	o	o	Gate 3
안전성·윤리 검증	o	o	o	Gate 4
운영 복원력	o	o	△	Gate 5
폐기·전환 보증	△	△	△	Gate 6 (신설)

(○: 명시적 요건 존재, △: 부분적 언급 또는 미비)

Gate 2와 Gate 6은 기존 프레임워크에서 미비한 영역을 한국 제도 환경에 맞게 보완한 것이다. Gate 2는 「개인정보 보호법」상 민감정보 국외 이전 제한과 AI 학습 데이터 거버넌스 공백을 기술적으로 해소하며, Gate 6은 서비스 종료 이후의 책임 공백과 개인정보 잔존 위험을 구체적 절차로 다룬다. 이처럼 K-SAF는 심층 방어·제로 트러스트·전주기 거버넌스를 공통 기반으로 삼되, 국제 프레임워크의 공백을 한국 법제도 환경에서 식별해 추가 게이트로 구체화했다. 각 게이트는 독립적 법적·기술적 체크포인트로 작동하며, 한 게이트라도 미충족 시 다음 단계 진행을 차단한다.

2. 핵심 게이트 구성과 선정 근거

K-SAF는 정부·공공부문이 주도하는 법·정책형 보증체계를 전제로, 후보 레이어 중 한국 환경에 최적화된 6개 게이트를 선정하였다.

각 게이트는 AI 생애주기 상의 독립적 법적·기술적 체크포인트로 작동하며, 한 게이트라도 미충족 시 다음 단계 진행을 차단한다. 구성은 다음과 같다.

1) Gate 1: 공급망·정책 적합성

AI 시스템을 구성하는 모든 컴포넌트의 출처·무결성·보안 상태를 조달 진입 이전에 검증하고, 국내 법령·윤리기준과의 적합성을 확인한다. 이 게이트는 악성 코드 삽입, 학습 데이터 오염, 서드파티 취약점이 후속 단계로 전파되는 것을 차단하는 최초의 통제 지점이자, 공공조달의 영향력을 시장 규율로 전환하는 핵심 관문이다.

조달 진입 단계의 핵심 위협은 ML 공급망 침해다. 외부에서 조달되는 오픈소스 라이브러리/사전학습 모델/학습 데이터는 악성 의존성 주입, 백도어가 삽입된 모델 가중치, 데이터 오염, CVE를 후속 생애주기로 전파시키는 통로가 된다. 이 위협은 진입하면 하위 게이트의 동제많으로는 제거가 어려우므로, 진입 이전 차단이 공학적으로 가장 효율적이다. 아래 네 검증 항목은 각 위협에 직접 대응하도록 구성된다.

구체적 검증 항목은 네 가지다. ① SBOM·MBOM의 완전성을 암호학적 서명으로 확인한다. ② 오픈소스 라이브러리·사전학습 모델의 알려진 취약점을 스캐닝하고, 고위험 미패치 항목이 존재하면 진입을 차단한다. ③ 학습 데이터의 출처를 추적하여 오염·무단 수집·라이선스 위반 여부를 검토한다. ④ 모델카드·PIA 결과의 완전성을 자동 교차검증하고, 「AI 기본법」상 고영향 AI의 경우 사전 신고·평가 이행 여부를 확인한다.

국제적으로도 같은 방향의 제도화가 진행 중이다. 미국은 OMB

M-25-22(OMB 2025b)까지 연방 조달 전반에 SBOM 요건을 통합하였고, EU는 AI Act와 정합되는 고위험 AI 조달용 모델 계약조항(MCC-AI-High-Risk)을 공식화하였다(European Commission 2025). 일본 디지털청의「생성AI 조달·이용 가이드라인」(デジタル庁 2025) 역시 조달 단계에서 AI 거버넌스 요건을 통합하는 동일한 구조를 채택하고 있다.

2) Gate 2: 데이터·연산 주권

데이터의 수집·이동·보관·학습·추론 전 과정에서 국내법 우선의 관할을 기술적으로 강제하고, AI 연산 인프라의 보안 상태와 국내 관할 준수 여부를 함께 검증한다. Gate 1의 문서상 적합성을 런타임 환경에서 실제로 집행하는 단계다. 이 단계의 핵심 위협은 두 축에서 발생한다. 데이터 측면에서는 역외로의 비인가 유출과 데이터 잔류 위반이, 연산 측면에서는 학습 및 추론 환경에 대한 과다 권한, 연산 인프라 침해를 통한 모델/추론 탈취 및 부채널 노출 등이 핵심 위협이다.

데이터 주권 측면에서는 국내 잔류율 100%를 설정하고, OPA 등 정책 엔진을 통해 외부 이전을 deny-by-default로 차단한다. 개인정보·민감정보의 가명·익명화 적용과 최소권한 원칙 준수를 확인하며, 국경 간 이전 시 법적 근거의 유효성을 검증한다. 연산 자원 측면에서는 GPU 클러스터·클라우드 인스턴스의 접근 통제, 학습 환경의 네트워크 격리, TEE 적용 여부를 점검하고, 인프라의 물리적 위치와 운영 주체의 국내 관할 여부를 확인한다.

국제적으로는 EU Data Act·EuroHPC AI Factory·일본 ABCI 3.0·중국 CSL·DSL·PIPL이 같은 방향의 제도적 기반을 제공한다. 데이터 주권과 연산 주권은 검증 방법과 담당 주체가 상이하지만 AI 생애주

기상 동일한 단계에서 병렬 수행되므로 단일 게이트로 통합 관리하는 것이 행정적으로 합리적이다.

3) Gate 3: 인적·조직 통제

AI 시스템의 개발·운영에 관여하는 인력과 조직 구조의 보안 적합성을 검증한다. 기술적 통제가 아무리 정교해도 내부자 위협과 거버넌스 공백은 전체 체계를 무력화할 수 있어, 이 게이트는 인프라 구축과 안전성 검증 사이에서 인적 층위의 통제 기반을 확립한다. 인적 층위의 핵심 위협은 내부자 위협, 과다 권한으로 인한 단일 인원의 파이프라인 장악 등이 있다.

검증 항목은 세 측면이다. 조직 구조 측면에서 CAIO 또는 동등 직위 지정과 AI 윤리·위험관리위원회의 운영 실태를, 접근 통제 측면에서 역할 기반 접근 제어(RBAC)·최소권한·직무 분리(SoD) 이행을, 비상 대응 측면에서 인간 개입 절차의 문서화·훈련과 내부자 위협 탐지 체계를 점검한다.

EU AI Act Article 14는 고위험 AI에 대한 인간 감독 체계를 법적 의무로 명시하며(Regulation (EU) 2024/1689), NIST AI RMF의 GOVERN 기능은 거버넌스 구조·책임·문화를 핵심 요소로 제시한다(NIST 2023). 현행 국내 AI 검증 체계는 기술적 통제에 집중되어 인적·조직 층위가 사실상 부재하며, 이는 ISMS-P 인증의 인적 보안 통제항목과의 정합성 측면에서도 보완이 요구된다.

4) Gate 4: 안전성·윤리 검증

개발된 AI 시스템이 유해성·편향·취약성 측면에서 정량적 안전 기준을 충족하는지를 독립적 레드팀 시험과 표준화된 벤치마크로 검증

하고, 배포가 사회적 취약계층·민주주의적 절차·환경에 미치는 영향을 평가한다. 기술적 안전성 검증과 사회적 영향 평가를 단일 게이트로 통합하여 배포 전 검증의 완결성을 확보한다.

안전성 검증은 국제 벤치마크 기반 임계치를 적용한다. ASR 5% 미만(HarmBench, 공공 도메인 3% 미만)(Mazeika et al. 2024), 유해 발화율 2% 미만(ToxiGen)(Hartvigsen et al. 2022), PII 노출율 0.1% 미만(Nakka et al. 2024), 프롬프트 인젝션 탐지 F_1 90% 초과(Liu et al. 2023)를 권고 임계치로 설정하며, 이 수치들은 선행 벤치마크 연구에서 관측된 분포와 안전 등급 기준을 준거로 도출한 초기 권고값으로, 국내 공공 도메인 실증사업을 통해 주기적으로 갱신되어야 한다.

설명가능성은 LIME·SHAP 기반으로 상위 특성 기여도를 검증한다. 모든 시험은 Gate 2에서 확립된 국내 관할 인프라 환경에서 수행된 것만 유효하다. 윤리·사회적 영향 평가로는 알고리즘 편향 정량 측정, 선거·공공의료·치안·금융 등 핵심 공공기능 영역의 오남용 시나리오 평가, 허위정보 생성 가능성 시험과 합성 콘텐츠 워터마킹 확인, 에너지 소비·탄소 발자국 문서화를 수행한다. ISO/IEC 42001·NIST AI RMF·EU AI Act와의 정합성을 점검하여 해외 조달 시장 진입 시 중복 평가 비용을 최소화하며, K-SAF 준수 기업이 해외 인증 취득 시 그 산출물을 증거로 활용할 수 있는 구조를 제도화한다.

5) Gate 5: 운영 복원력

배포 이후 실시간 모니터링·이상 탐지·자동 격리·복구·감사 로그 관리를 통해 운영 단계의 안전성을 지속적으로 유지한다. Gate 4에서 확정된 안전성 임계치를 운영 모니터링의 알람 기준으로 연계하여 개발·운영 단계 간 검증의 연속성을 확보한다.

핵심 검증 항목은 네 가지다. ① 이상 탐지 시 자동 격리 완료 시간을 시스템 특성에 맞게 설정하고 실측으로 확인한다. ② 트래픽 차단→롤백→킬스위치의 단계적 대응 절차와 각 완료 시간이 해당 서비스의 RTO 이내인지 점검한다. ③ 감사 로그를 변조 방지 저장소에 표준 스키마로 축적하여 사후 추적 가능성을 확보한다. ④ 운영 중 새롭게 관측된 공격 패턴을 Gate 4의 레드팀 시나리오에 즉시 반영하여 검증 체계를 상시 갱신한다.

ISO 22301의 업무연속성 관리, NIST SP 800-61의 침해사고 대응 가이드, EU AI Act Article 72·73의 사후 모니터링·중대 사고 보고 의무(Regulation (EU) 2024/1689)가 국제적 정합성의 근거를 제공한다. 한국 금융·공공 서비스의 짧은 RTO 기준은 자동화된 복원력 체계의 필요성을 한층 부각시킨다.

6) Gate 6: 폐기, 전환 보증

AI 시스템의 서비스 종료, 버전 전환, 공급업체 변경 시 데이터·모델·로그의 안전한 처리와 서비스 연속성을 검증한다. 그동안 상대적으로 주목받지 못한 단계이지만, 통제가 충분하지 않으면 개인정보 유출, 책임 소재 공백, 공급업체 종속으로 이어질 수 있다. 검증 항목은 다음과 같다. 첫째, 학습 데이터·모델 가중치·중간 산출물의 안전한 삭제 또는 이관을 암호학적으로 검증하고, 잔존 개인정보의 완전 삭제를 확인한 뒤 삭제 증명서를 발급한다. 둘째, 후속·대체 시스템으로의 전환 절차의 적절성과 데이터·모델 이식성을 점검한다. 셋째, 감사 로그의 법정 보존기간 준수와 만료 후 안전 폐기 절차를 확인한다.

GDPR과 「개인정보 보호법」의 삭제권이 AI 학습 데이터에 어떻게

적용되는지는 아직 국제적 합의가 없는 실무 쟁점이며, EU AI Act도 폐기 단계의 구체적 요건은 충분히 다루지 않는다. K-SAF Gate 6은 이 영역의 절차적 기준을 구체화하는 것을 목표로 한다.

3. 6-Gate의 인과관계와 상호작용

K-SAF의 6-Gate는 개별 심사 단계의 단순한 합을 넘어, 정책—기술—인적 통제—검증—운영—폐기가 서로 연결되는 순환적 피드백 구조를 형성한다. 이는 한 단계의 규범 결정이 다음 단계의 기술적 적용으로 이어지고, 운영과 폐기 단계에서 관찰된 결과가 다시 정책 및 조달 요건으로 반영되는 역동적 순환 구조를 의미한다.

1) Gate 1 → Gate 2: 정책 강제력의 기술적 구현

Gate 1에서 결정된 거버넌스·조달 요건은 Gate 2에서 기술적 수단을 통해 실제로 적용된다. 예를 들어 "의료 데이터는 국내 서버에서만 처리한다" 또는 "가명처리된 국내 데이터만 학습·추론에 사용한다"와 같은 문서상의 요건은 Gate 2에서 OPA(Open Policy Agent) 등 정책 엔진의 규칙으로 변환되어 실행 환경에서 자동으로 작동한다. 이 구조에서는 담당자의 재량적 판단이나 사후 점검에 의존하지 않으며, 위반이 발생하면 해당 데이터 흐름이 즉시 차단되고 위반 로그가 자동으로 기록된다. deny-by-default 체계가 기본 원칙으로 작동하는 것이다.

한국의 정책 환경에서는 대규모 공공조달이 이 관문에 실질적 시장 규율 기능을 부여한다. Gate 1의 기준을 충족하지 못한 사업자는 입찰 단계에서 배제되므로, 민간 기업은 시스템 설계 초기부터 정책

적합성을 반영해야 할 강력한 유인을 갖게 된다. 다만 문서나 증적의 미비로 탈락한 경우에는 기한 내 보완을 전제로 한 제한적 재심사를 허용하여, 제도의 절차적 정당성과 현장의 실행 가능성을 함께 확보한다.

2) Gate 2 → Gate 3: 기술 통제 기반 위에서의 인적·조직 검증

Gate 2에서 확립된 데이터·연산 주권 통제는 Gate 3의 인적·조직 통제가 실효성을 갖기 위한 기반을 제공한다. 최소권한 원칙과 네트워크 격리 같은 기술적 접근 통제가 먼저 마련되어야, Gate 3에서 다루는 역할 기반 접근 제어(RBAC), 직무 분리(SoD), 내부자 위협 탐지 체계의 실효성을 실질적으로 검증할 수 있다. 반대 방향의 흐름도 작동한다. Gate 3에서 특정 직군에 과도한 데이터 접근 권한이 부여되어 있거나 비상 절차에 공백이 발견되면, 해당 사항은 Gate 2의 정책 엔진 규칙으로 즉시 반영되어 기술적 접근 차단 항목에 추가된다. 이러한 양방향 피드백을 통해 기술적 통제와 인적 통제가 정합성을 유지하며 유기적으로 결합된다.

3) Gate 3 → Gate 4: 조직 역량이 안전성 검증의 신뢰성을 담보

Gate 3에서 확립된 거버넌스 구조와 인력 통제 체계는 Gate 4의 안전성·윤리 검증이 신뢰성을 갖추는 토대가 된다. 제3자 레드팀 운영, 윤리위원회의 사회적 영향 평가, 이해관계자 참여 절차는 모두 Gate 3의 조직 역량과 직무 분리 체계가 작동할 때 실질적 기능을 한다.

금융, 의료처럼 규제 강도가 높은 영역에서는 “국내 관할 하에서 국내 데이터와 인프라를 사용하여 수행된 평가만 인정한다”는 원칙

이 핵심이며, 해외 처리·저장 환경에서 도출된 시험 성적은 인정되지 않는다. 국내 요건을 다시 충족한 뒤 평가를 새롭게 진행해야 한다. 한편 Gate 4의 레드팀 시험에서 데이터 경로의 취약성이나 프롬프트 인젝션을 통한 외부 API 호출 같은 구체적 위험이 발견될 경우, 해당 패턴은 즉시 Gate 2의 정책 규칙으로 반영되어 실행 환경의 자동 차단 항목에 추가된다.

4) Gate 4 → Gate 5: 검증된 안전성의 운영 보증

Gate 4에서 합격 판정을 받은 모델의 임계값은 Gate 5의 운영 단계 모니터링 기준으로 그대로 연계된다. 유해 발화 지표, 개인정보 노출 탐지율, 해석가능성과 재현성 등 개발 단계에서 확정한 기준값이 운영 모니터링 시스템의 알람 기준으로 직접 연결되는 것이다. 실서비스 중 특정 유형의 질의에서 유해도 지표가 기준을 초과하면, 수동 개입 없이 경보·격리·롤백으로 이어지는 킬스위치 절차가 자동으로 작동하고, 모든 조치와 그 맥락이 변조 방지 감사 로그에 축적된다. 동시에 운영 중 새롭게 관측된 신종 공격 프롬프트나 우회 패턴은 즉시 Gate 4의 레드팀 시나리오에 반영되어, 다음 버전 평가의 기본 항목으로 자리잡는다. 이처럼 사전 시험과 실시간 감시가 결합되어 만들어내는 연속적 보증 체계는, 짧은 RTO를 요구하는 금융 및 공공 서비스 환경에서 사회적 리스크의 발생 가능성을 효과적으로 줄인다.

5) Gate 5 → Gate 6: 운영 이력의 안전한 종료 관리

Gate 5에서 축적된 감사 로그, 이상 징후 통계, 모델 버전 이력은 Gate 6 폐기·전환 단계의 핵심 입력이 된다. 서비스 종료 또는 버전

전환 시, Gate 5에서 관리된 운영 데이터의 법정 보존 여부 확인과 안전 폐기 절차 이행이 Gate 6에서 최종 검증된다. 학습 데이터·모델 가중치의 암호학적 삭제 증명, 후속 시스템으로의 이식성 보장, 공급업체 변경 시 잔존 데이터의 완전한 회수가 이 단계에서 확인된다. 폐기 완료 증명서는 Gate 1의 다음 사업 주기 입찰 적격 판단 근거로 활용되어 생애주기 전체의 책임 연속성을 확보한다.

6) Gate 6 → Gate 1: 폐기·전환 인텔리전스의 정책 환류

Gate 5와 Gate 6에서 축적된 운영 로그, 이상 징후 통계, 폐기 과정에서 발견된 취약 유형은 Gate 1의 정책 및 조달 요건을 주기적으로 갱신하는 토대가 된다. 다수 기관에서 공통된 보안 취약 패턴이나 폐기 단계의 문제가 반복적으로 관측되면, 다음 연도 사업 공고에 해당 위험을 다루는 신규 요건이 추가된다. 이러한 요구사항은 다시 Gate 2의 정책 규칙으로 변환되고, Gate 4의 시험 항목으로 표준화되며, Gate 5의 운영 모니터링 기준에 임계값으로 반영된다. 결과적으로 K-SAF는 운영과 폐기 단계에서 축적된 정보를 정책으로 흡수하여, 분기 단위로 신속하게 갱신될 수 있는 적응형 규제 체계로 작동하게 된다.

4. Fail-Retry 메커니즘과 "강제 안전망"

K-SAF는 각 단계에서 발생한 결함이 다음 단계로 이어지지 않도록 설계된 연속적 게이트 구조를 통해, 개발·검증·운영·폐기 생애주기 전반에 걸친 실질적 안전망을 구현한다. 핵심 원리는 다음과 같다. 어느 하나의 게이트라도 기준을 충족하지 못하면 전체 생애주기의 진

행이 즉시 중단되며, 원인이 제거되고 보완이 확인될 때까지 문제가 발생한 단계로 되돌아간다. 이는 문제를 사후에 벌금이나 권고로 수습하는 정적인 규제 방식이 아니라, 위반 자체가 실행되지 못하도록 사전에 차단하는 기술 기반의 집행 구조다.

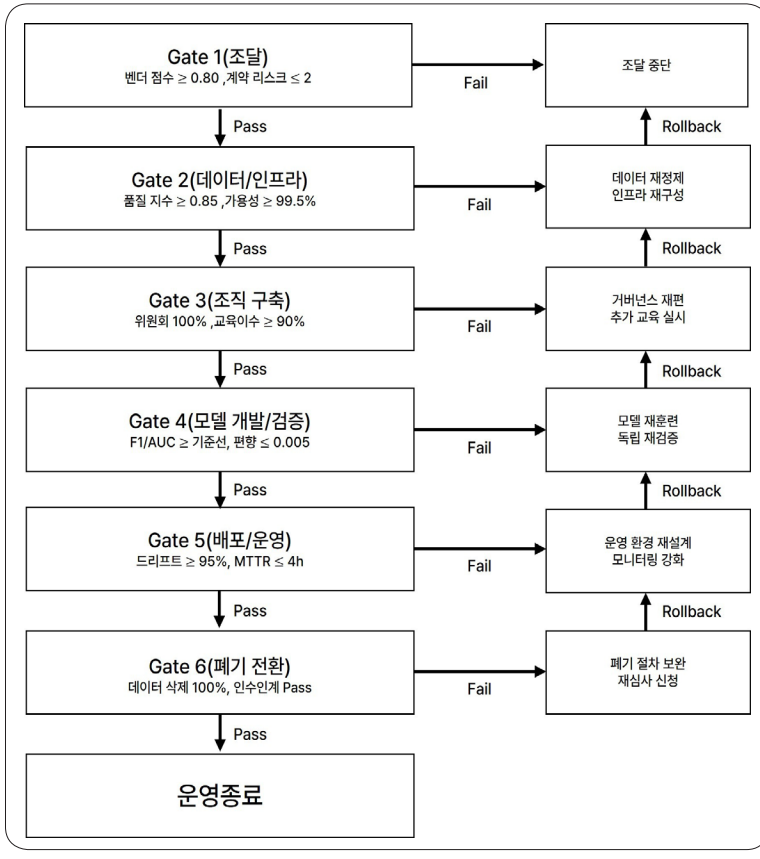
1) 연속 게이트의 작동 원리

K-SAF의 6개 게이트는 AI 시스템 생애주기에 직렬 배치된다. 그림 1을 보면, 정방향(→) 경로에서는 게이트를 순차 통과해 운영·폐기에 이르고, 역방향(←) 경로에서는 기준 미달 시 발견 지점에서 즉시 정지하여 필요한 선행 단계로 롤백한다. 이때 각 게이트는 통과 여부를 판단하는 정량 임계값과 증빙 요건을 보유하고, 실패 시 취해야 할 필수 보완 조치와 재심사 절차가 함께 정의된다. 결과적으로 저품질·고위험 AI가 최종 사용자 환경에 도달할 확률을 구조적으로 0에 수렴시킨다.

2) 게이트별 통과 기준, 실패 조치, 재시도 전략

표 2 에 제시된 바와 같이, K-SAF의 각 게이트에는 명확한 통과·실패 판정 기준이 사전에 정의되어 있으며, 실패 시 교정 조치와 재시도 허용 범위 또한 절차적으로 규정되어 있다. 이를 통해 평가의 일관성과 재현성을 확보하고, 교정·재검증의 운영 부담을 예측 가능하게 관리한다.

〈그림 1〉 K-SAF 6-GATE 순차 검증 및 Fail-Retry 흐름도



〈표 2〉 K-SAF 게이트별 통과 기준

게이트	통과 기준	실패 시 조치	재시도 정책	한국형 특수 고려 사항
Gate 1	요구 서류 100% 제출 및 적합성 확인. 긴급(Critical) 등급 CVE 미패치 항목 0건	보완 요구 → 재심사	최대 2회 재심사 후 미충족 시 사업 참여 제한	나라장터 연계 자동 검증, 전자서명·타임스탬프 부착 원본 요구

게이트	통과 기준	실패 시 조치	재시도 정책	한국형 특수 고려 사항
Gate 2	데이터 주권 위반 0건, 데이터 국내 잔류 100%, 연산 인프라 국내 관할 확인	파이프라인 및 아키텍처 재설계, 위반 원인 제거 후 정책 엔진 재검증	제한 없음	국내 보안 인증과의 정합
Gate 3	CAIO 또는 동등 직위 지정 확인, RBAC 및 SoD 설정 적정성 확인, human-in-the-loop 절차의 문서화 및 훈련 이행 확인	조직 구조 보완, 접근 통제 재설정, 절차 재수립 후 재심사	최대 2회 재심사 후 미충족 시 개발 착수 불허	ISMS-P 인적 보안 통제항목과의 연계 자동 점검, 금융 AI 내부통제 의무 매핑
Gate 4	ASR 〈 5%(공공 도메인 〈 3%), PII 노출율 〈 0.1%, 프롬프트 인젝션 탐지 F_1 〉 90%, 설명가능성 등 사전 합의 임계치 충족. 레드팀 다중 시나리오 중대한 취약점 0건	재학습 및 방어 대책 강화 후 재평가, 취약점 패치 계획 제출, 필요 시 외부 자문 활용	최대 3회 권장(초과 시 프로젝트 지속 여부 재검토)	KISA 생성형 AI 보안 가이드라인 자동 매핑, 국내 관할 인프라에서 수행된 시험만 유효로 인정
Gate 5	자동 격리 완료 30초 이내 실측 확인, 롤백 및 비상조치 시뮬레이션 통과, 변조 방지 감사 로그 축적 체계 가동 확인	킬스위치 즉시 발동, 문제 모델 격리 및 안전 버전 전환, Gate 4 레드팀 시나리오 갱신	제한 없음	전자금융감독규정에 따른 복구 목표 시간 준수, 대형 서비스의 무중단 및 Blue-Green 배포 표준화
Gate 6	학습 데이터 및 모델 가중치의 삭제, 잔존 개인정보의 완전 삭제 및 삭제 증명서 발급. 감사 로그의 법령상 보존기간 준수 확인. 후속 시스템 이식성 확보	미삭제 데이터 재처리, 이관 절차 재수립 후 재검증	해당 없음	「공공기록물관리법」 보존기간 연동 자동 알림, 클라우드 서비스 계약 종료 시 데이터 반환·삭제 의무 이행 확인

〈그림 2〉 K-SAF 게이트별 정량 임계값 및 재시도 정책

재시도 정책은 게이트의 성격에 따라 차등적으로 적용된다. Gate 2, 5, 6은 지속적 개선 원칙에 따라 재시도에 별도의 상한을 두지 않으며, Gate 1, 3, 4는 행정 부담과 시험 비용, 프로젝트 일정 관리를 고려하여 재시도 횟수에 상한을 부여한다. 이러한 차등 구조는 무분별한 반복 시도를 억제하는 동시에, 초기 단계부터 기준을 충족하려는 유인을 강화한다. 결과적으로 기업은 시스템 설계 초기부터 기준 충족을 목표로 하게 되며, 전체 검증 과정의 효율성과 품질이 함께 개선된다.

3) '하나라도 실패하면 전체 중단' 원칙의 함의

K-SAF는 사전 차단을 핵심 원칙으로 삼아, 불확실성이 해소되지 않은 배포를 제거하고 통과 판정을 규칙 기반으로 자동화함으로써 재량적 판단의 한계를 줄인다. 공공·금융 부문이 AI 시장 수요의 상당 부분을 차지하는 한국의 구조에서 K-SAF 기준은 사실상 시장 접근 요건으로 작동하며, 미충족 시 입찰 배제·계약 해지로 이어져 업계가 설계 초기부터 정책 적합성을 반영할 유인을 형성한다. 6-Gate 체계는 기존 4-Gate 구조에 인적·조직 통제(Gate 3)와 폐기·전환 보증

(Gate 6)을 더해 기술 중심 통제의 한계와 생애주기 말단의 책임 공백을 함께 해소한다. 특히 Gate 6은 폐기 완료 증명을 차기 조달의 필수 요건으로 삼아 서비스 종료 이후의 책임 연속성을 제도화하며, 이러한 유인·제재 구조는 권고 수준의 규율을 실질적 보증으로 끌어올린다.

5. K-SAF 도입 시, 기대 효과

1) 사이버 안보 측면에서의 이점

(1) 공격면 축소와 공급망 무결성 보장

K-SAF는 Gate 1에서 모델카드·SBOM/MBOM·데이터 프로비넌스 등 증적 기반 요건을 강제함으로써, 네트워크·클라우드·컨테이너·API 전 계층에서 최소권한 원칙, 마이크로 세분화, 외부 통신 통제가 설계 산출물로 구체화되도록 한다. 학습 데이터부터 가속기 드라이버에 이르는 전체 공급망은 전자서명·원격 신뢰 검증으로 연결되어 배포 이전에 무결성이 자동 확인되며, 고위험 CVE 미패치 항목이 존재할 경우 다음 단계로의 진입 자체가 차단된다. Gate 3의 인적·조직 통제는 이러한 기술적 통제를 인력과 절차의 층위로 확장하여 내부자 위협과 권한 남용의 여지를 추가로 축소하고, Gate 6은 폐기·이관 과정의 무결성을 최종 검증함으로써 공급망 위협이 생애주기 말단에서 재발하는 것을 차단한다.

(2) 사이버 위협에 대한 회복력과 상시 대응 체계 구축

Gate 2의 코드 기반 정책은 데이터 국내 잔류와 암호키 관할 일치

성을 집행하여, 공격 발원지가 해외이더라도 국내 핵심 자산의 역외 조작·유출 여지를 구조적으로 제거한다. Gate 5는 자동 격리·킬스위치·롤백을 조달 의무 조항으로 격상하여 침해 발생 시 지체 없는 복구가 표준 절차로 작동하도록 하며, 탐지·격리·분석·복구의 전 단계를 게이트별 지표와 증적 제출 의무로 연결한다. Gate 5에서 수집된 모니터링 데이터는 Gate 4의 레드팀 시나리오로, Gate 6의 폐기 단계 취약 유형은 Gate 1의 조달 요건으로 각각 환류되어, 신종 위협에 대한 탐지와 완화의 지연 시간이 단축되고 동일 유형 사고의 재발 가능성이 낮아진다.

2) 기술적 이점

(1) 규제 준수의 자동화와 재현 가능한 보증

정책과 규정의 변경 사항을 Gate 2의 정책 엔진과 파이프라인 규칙으로 자동 변환함으로써, 담당자의 수기 점검에 의존하던 준수 활동이 시스템 차원에서 상시적으로 수행된다. Gate 1의 나라장터 연계 자동 검증, Gate 3의 ISMS-P 통제항목 연계 점검, Gate 4의 KISA 생성형 AI 보안 가이드라인 자동 매핑이 함께 작동하면 요구사항 해석의 자의성이 줄어들고 사전 단계에서부터 위반 탐지와 교정이 가능해진다. 각 게이트의 통과 산출물은 암호학적 서명과 함께 변조 방지 저장소에 기록되어, Gate 1 조달부터 Gate 6 폐기까지 생애주기 전 구간의 검증 이력이 사후 감사 증거로 안정적으로 보존된다.

(2) 적응적 방어와 지속적 모델 보강

K-SAF는 운영 현장에서 축적되는 위협 정보를 검증 체계로 즉시

환류하는 폐쇄 피드백 루프를 구현한다. Gate 5에서 수집된 오탐·미탐지 패턴, 이상 징후, 장애 로그는 Gate 4의 레드팀 시나리오와 안전성 벤치마크에 반영되고, Gate 6의 폐기 단계에서 발견된 취약 유형은 다음 주기의 Gate 1 조달 요건 개정으로 이어진다. 이 양방향 피드백 구조를 통해 모델·정책·보안 필터가 점진적으로 강화되며, 신종 위협에 대한 탐지와 대응 속도가 지속적으로 개선된다.

3) 사회·정책적 이점

(1) 국민 신뢰의 제도화

K-SAF의 게이트 통과 사실을 공적 등록부와 표시 제도를 통해 제공하면, 이용자와 시민은 검증 가능한 안전·신뢰 정보를 바탕으로 서비스를 선택할 수 있게 된다. Gate 3의 이해관계자 참여 절차 이행 확인은 시민사회가 AI 거버넌스 과정에 제도적으로 개입할 수 있는 경로를 보장한다. 정기적인 재검증과 개정 이력의 공개는 사회적 감시와 자율적 규율을 결합하여, AI 기술 도입에 대한 사회적 수용성을 점진적으로 높인다.

(2) 데이터 주권과 프라이버시의 실질적 구현

K-SAF는 데이터의 국내 잔류, 접근 통제, 계보 관리, 정책 위반의 완전한 차단 등 주권 관련 지표를 Gate 2의 코드 기반 정책으로 집행하며, Gate 6의 삭제 증명 발급을 통해 생애주기 전 구간에서의 데이터 주권 집행을 완성한다. 이로써 관련 법률과 지침이 선언적 규범에 머무르지 않고, 시스템 아키텍처, 운영 절차, 감사 로그를 통해 실제로 집행 가능한 통제로 전환된다. 국경 간 데이터 이전이나 역외 저장

이 불가피한 경우에도, 데이터와 연산 경계에 대한 세분화된 통제와 감사 경로가 확보되므로, 국제적 적정성(adequacy) 평가와 형평성 논의에서 신뢰 가능한 준거를 제시할 수 있다.

(3) 사회 기반 영역의 복원력 및 책임 연속성 확보

Gate 4의 안전성 임계치와 Gate 5의 자동 복구·격리·롤백 절차가 결합되면, 오류, 오남용, 침해 사고가 발생하더라도 피해 규모와 복구 시간을 예측 가능한 범위 안에서 관리할 수 있다. 금융, 의료, 행정 등 사회 핵심 영역에서 서비스의 연속성이 강화되며, 사고가 발생한 이후에는 원인 규명과 재발 방지 조치가 지표와 증적에 기반하여 수행된다. Gate 6은 서비스 종료 이후에도 책임의 연속성을 보장함으로써, 공급업체 변경이나 시스템 전환 과정에서 발생할 수 있는 거버넌스 공백을 사전에 차단한다. 부처와 기관 간의 지표 호환성과 보고 양식의 표준화는 사고 대응의 상호운용성을 높여, 다기관 공조 체계를 절차와 기술 차원에서 뒷받침한다.

IV. 맺음말

본 연구가 제안한 6-Gate 검증 체계는 사이버 안보 관점에서 필수적이다. AI 시스템은 네트워크·클라우드·API로 상호 연결되어 공격 표면이 비약적으로 확장되며, K-SAF는 공급망 취약성 차단(Gate 1)부터 데이터·연산 경계 격리(Gate 2), 인적·거버넌스 통제(Gate 3), 안전성 지표 수치화(Gate 4), 실시간 방어·복구(Gate 5), 생애주기 말단의 책임 연속성 보장(Gate 6)까지 유기적 방어 구조를 형성한다.

한국의 핵심 과제는 구축된 소버린 AI를 어떻게 증명하고 신뢰를 지속할 것인가로 수렴된다. K-SAF는 국제 규범과의 정합성을 확보하면서도 국내 제도 환경에 적응 가능한 구조로, 전 생애주기에 걸친 책임의 연속성을 제도화한다. 다만 Gate 4 임계치의 국내 실증·갱신, 국제 규범과의 교차 인정, Gate 6의 모델 가중치 암호학적 삭제 검증 등은 심화 연구 과제로 남으며, 민간 확산 전략과 기술 인프라 개발이 병행되어야 한다.

이러한 과제의 점진적 해소를 통해 한국은 신뢰 가능한 AI 거버넌스의 선도 국가로 자리매김할 수 있다. K-SAF의 도입은 선택이 아니라, 국가 인프라로서의 AI를 지속 가능하고 책임 있게 운영하기 위한 최소 요건이다.

〈참고문헌〉

과학기술정보통신부·정보통신산업진흥원·한국지능정보사회진흥원·정보통신기획평가원. 2025. “독자 AI 파운데이션 모델 프로젝트 참여 정예팀 공모 공고”.

「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(법률 제20676호, 2026. 1. 22. 시행).

「개인정보 보호법」(법률 제19234호).

한국전자통신연구원(ETRI). 2021. 「GAIA-X 분석 및 데이터 댐 발전 방향」. 대전: 한국전자통신연구원.

한국인터넷진흥원(KISA). 2023. 「인공지능(AI) 보안 안내서」. 나주: 한국인터넷진흥원.

ENISA. 2025. ENISA Threat Landscape 2025 (ETL 2025). European Union Agency for Cybersecurity.

김태훈·이성엽. 2025. 「소버린 AI(Sovereign AI) 구축의 글로벌 법제 동향에 관한 연구」. 『경영법률』 36(1): 195-223.

European AI Office. 2025. General-Purpose AI Code of Practice. European Commission. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

European Commission. 2025. Model Contractual Clauses for the Public Procurement of High-Risk AI (MCC-AI-High-Risk), v2.0. Luxembourg: Publications Office of the European Union.

EuroHPC Joint Undertaking. 2024. “JUPITER Officially Propels Europe into the Exascale Era.” https://eurohpc-ju.europa.eu.

Ee, Shaun, Joe O’Brien, Zoe Williams, Amanda El-Dakhakhni, Matthew Aird, and Alaina Lintz. 2024. “Adapting Cybersecurity Frameworks to Manage Frontier AI Risks: A Defense-in-Depth Approach.” arXiv:2408.07933.

G7. 2023. “Hiroshima AI Process: International Guiding Principles and Code of Conduct.” Joint statement (October 30; December 1).

Hartvigsen, Thomas, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection." In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3309–3326. Dublin: Association for Computational Linguistics.

International Organization for Standardization, and International Electrotechnical Commission (ISO/IEC). 2022. ISO/IEC 22989:2022 Information Technology – Artificial Intelligence – Artificial Intelligence Concepts and Terminology. Geneva: ISO.

International Organization for Standardization, and International Electrotechnical Commission. 2023. ISO/IEC 42001:2023 Information Technology – Artificial Intelligence – Management System. Geneva: ISO.

Leon, Florin, et al. 2026. "Lifecycle-Based Governance to Build Reliable Ethical AI Systems." Systems Research and Behavioral Science 43(3): 1116–1132. https://doi.org/10.1002/sres.70014.

Li, Yuchong, and Qinghui Liu. 2021. "A Comprehensive Review Study of Cyber-Attacks and Cyber Security: Emerging Trends and Recent Developments." Energy Reports 7: 8176–8186.

Liu, Yi, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. "Prompt Injection Attack against LLM-Integrated Applications." arXiv:2306.05499.

Mazeika, Mantas, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. "HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal." arXiv:2402.04249.

Microsoft Security. 2026. “Zero Trust for AI Reference Architecture.” https://www.microsoft.com/en-us/security/blog/2026/03/19/.

Nakka, Krishna Kanth, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. “PII-Scope: A Benchmark for Training Data PII Leakage Assessment in LLMs.” arXiv:2410.06704.

National Institute of Standards and Technology (NIST). 2023. AI Risk Management Framework 1.0 (NIST AI 100-1). Gaithersburg: NIST.

U.S. Department of Defense. 2011. Department of Defense Strategy for Operating in Cyberspace. Washington, DC: Department of Defense.

U.S. Office of Management and Budget (OMB). 2025a. Accelerating Federal Use of AI through Innovation, Governance, and Public Trust (M-25-21). Washington, DC: OMB.

U.S. Office of Management and Budget (OMB). 2025b. Driving Efficient Acquisition of Artificial Intelligence in Government (M-25-22). Washington, DC: OMB.

U.S. White House. 2023. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Washington, DC: The White House.

U.S. White House. 2025. Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence. Washington, DC: The White House.

UK Department for Science, Innovation and Technology (DSIT). 2024. Introduction to AI Assurance. London: HM Government.

Regulation (EU) 2024/1689 (Artificial Intelligence Act).

デジタル庁(일본 디지털청). 2025.「生成AI の調達・利用に関するガイドライン」(생성AI 조달·이용 가이드라인). 東京: デジタル庁.

全国网络安全标准化技术委员会(SAC/TC260, 전국 사이버보안 표준화 기술위원회). 2025. GB/T 45654-2025:「网络安全技术 生成式人工智能服务安全基

本要求」(사이버보안 기술 — 생성형 인공지능 서비스 보안 기본 요구). 北京: 国家市场监督管理总局·国家标准化管理委员会.

中华人民共和国国家互联网信息办公室(중화인민공화국 국가인터넷정보판공실, CAC). 2023.「生成式人工智能服务管理暂行办法」(생성형 인공지능 서비스 관리 잠정조치).

A Study on Introducing a Sovereign AI Verification Framework for Korea

Kim Doyeon | Seoul Women's University
Cho Hwijung | Ewha Womans University

This study proposes the Korea Sovereign AI Framework (K-SAF), a six-gate verification system designed to address critical gaps in sovereignty, safety, and resilience that arise as AI becomes a composite national asset integrating data, compute, models, and human capital. As hyperconnectivity across cloud platforms and supply chains amplifies cascading risks arising from single points of failure and opacity, growing dependence on external vendors for large-scale AI training and deployment undermines information sovereignty and auditability. K-SAF addresses these challenges through sequential gates spanning supply-chain and policy compliance, data and compute sovereignty, human and organizational control, safety and ethics verification, operational resilience, and decommissioning and transition assurance. Across these gates, the framework quantifies legal-jurisdiction, access-control, adversarial-robustness, privacy, and quality metrics, while standardizing real-time defense, self-healing, and rapid-isolation procedures. A threshold-based pass/halt mechanism with retry policies ensures that any non-conforming component triggers a mandatory lifecycle suspension, enabling a national response chain of detection, isolation, analysis, and recovery integrated with procurement and audit systems. The framework preserves openness and interoperability while anchoring core decision-making authority and accountability domestically, providing the minimum requirements to raise the operation of AI as national infrastructure to a sustainable and accountable level.

Keyword Sovereign AI, cybersecurity, governance, verification framework, data sovereignty