사이버안보연구, 3집 1호 2026 여름, 261-298

에이전틱 AI 공급망 위협과 국가 대응 전략:

동적 AIBOM 기반 능동적 사이버 방어 체계 제언

최윤성 | 고려대학교

논문접수일 2026년 5월 16일 | 논문심사일 2026년 6월 1일 | 게재확정일 2026년 6월 16일

논문 요약

최근 인공지능(AI) 공급망에서 잇따른 침해 사건과 LLM 에이전트에 의한 제로데이 양산 능력의 실증은, 인간 방어자가 기계의 속도를 따라잡지 못하는 구조 자체의 비대칭이 이미 현실로 자리 잡았음을 보여준다. 이에 대응하여 미국·EU·G7을 중심으로 빠르게 결정화되고 있는 AI 자재명세서(AIBOM, AI Bill of Materials) 관련 규범은, AIBOM을 산업계의 자율 권고를 넘어 시장 진입을 결정짓는 '수출 장벽(Export Barrier)'이자 광의의 '사이버 무기 체계'로 격상시키고 있다.

기존의 정적(Static) SBOM으로는 끊임없이 변하는 에이전트 환경과 '바이브 코딩' 시대의 보이지 않는 종속성을 다루기 어려우며, 빌드·배포·런타임 전 단계를 잇는 통합된 신뢰 명세와 능동적 평가·검증 인프라가 함께 갖춰져야 한다. 이러한 변화의 함의 위에서 필자는 한국이 단순한 수범자(受範者)가 아닌 '대체 불가능한 안보 파트너(Indispensable Partner)'로 자리매김하기 위한 차등화된 AIBOM 설계, Compliance-as-Code 생태계, 능동적 사이버 방어(Active Cyber Defense) 인프라의 세 정책 방향을 제언한다.

특히 본고는 미국과 EU가 각각 FY26 NDAA와 EU AI Act를 통해 AI 시스템의 공급망 투명성, 추적성, 기술문서 의무를 제도화하는 방향에서, AIBOM을 차세대 AI 보안과 조달, 규제 준수를 아우르는 인프라로 정착시키는 정책적 동력으로 작용하는 흐름에 대응하여, 한국이 단순한 규제 도입을 넘어 신뢰 체인 전반을 실시간 차단하는 '국가 기술 인프라(National Technical Infrastructure)'로서의 AIBOM 체계를 구축해야 한다는 점을 강조하고자 한다.

주제어 에이전틱 AI, 공급망 보안, 동적 AIBOM, 능동적 사이버 방어, 대체 불가능한 안보 파트너

Ⅰ. 서론

기계의 속도로 진화하는 AI 생태계에서, 연결의 가장 약한 고리는 이제 코드가 아닌 '의존성(Dependency)' 그 자체이다. AI 기술은 파운데이션 모델, 미세조정(Fine-tuning) 데이터셋, 런타임 플러그인, 그리고 외부 API를 호출하는 에이전트(Agent) 인프라 등 전 세계에 분산된 다단계의 복잡한 공급망 생태계(Ecosystem)를 이룬다. AI 공급망의 사이버보안 위험은 단순히 소프트웨어 코드의 결함을 넘어, 모델 가중치(Weights)의 오염, 악의적인 프롬프트 인젝션, 그리고 자율형 에이전트의 권한 오남용에서 발생한다. 가트너(Gartner)는 2028년까지 기업 침해 사고의 25%가 외부 또는 내부 악의 행위자에 의한 AI 에이전트 오용(AI agent abuse)으로부터 발생할 것으로 전망하면서, AI 에이전트가 이미 가시화되지 않은 공격 표면을 빠르게 확대하고 있다고 지적했다(Gartner 2024).

2025~2026년에 걸쳐 보고된 일련의 사건들은, 단일 패키지의 침해가 수천 개의 하위 종속성으로 전파되는 '에이전틱 공급망 공격(Agentic Supply Chain Attack)'이 체계적 위협 양상으로 부상했음을 보여준다. 이러한 환경 변화의 단면은 OpenAI의 공동창립자인 안드레이 카파시(Andrej Karpathy)가 2025년 2월 자신의 X(구 트위터) 계정을 통해 '바이브 코딩(Vibe Coding)'이라는 신조어를 제시한 데에서도 확인된다. 그는 "코드의 흐름에 몸을 맡기고, 지수적 성장에 올라탄 채, 코드의 존재 자체를 잊어버리는 새로운 형태의 개발 방식"을 묘사했는데(Karpathy 2025), 해당 트윗은 약 450만 회 이상 조회되며 콜린스(Collins) 사전의 2025년 올해의 단어로 선정되었다(Collins 2025). 정작 카파시 본인은 이를 주말 프로토타입을 위한 가벼운 개념으로 제안

했지만, 실무 현장에서는 AI 코딩 도구가 추천하는 오픈소스 패키지를 검토 없이 설치하는 관행으로 빠르게 확산되었으며(Ng 2025; Willison 2025), 이는 '보이지 않는 종속성(Invisible Dependency)' 위협의 토양이 되었다.

이러한 위협에 대응하여 미국은 2026 회계연도 국방수권법(FY26 NDAA)을 통해 국방부 조달 AI 시스템에 대한 공급망 투명성과 생애주기 보안, 추적성 강화를 추진하고 있으며, SBOM 개념을 AI 시스템까지 확장하는 방향을 제시했다. 이에 따라 향후 DoD 조달 체계에서 AIBOM(AI Bill of Materials) 기반 요구사항이 확대될 가능성이 커지고 있다. 한편 유럽연합(EU)은 EU AI Act를 통해 고위험 AI 시스템에 대해 기술문서, 추적성, 기록관리 및 투명성 의무를 부과하고 있으며, 이는 AI 공급망 구성요소와 모델 계보(lineage)를 구조적으로 관리하는 AIBOM 체계의 제도적 필요성을 강화하고 있다.

본 연구와 직접 관련되는 국내 선행 연구는 SW 공급망 보안과 SBOM 정책 영역에서 축적되어 왔다. 최윤성(2022)은 미국 행정명령 EO 14028을 중심으로 SBOM 제도의 도입 배경과 구성 요소를 정리하였고, 최윤성(2024)은 이를 발전시켜 미국 SW 조달 개선 흐름이 한국의 공급망 보안 관리체계에 갖는 시사점을 도출하였다. 두 연구는 SBOM을 '정적 문서로서의 구성요소 명세'로 다루는 공통점을 가지며, 분석 대상도 전통적 소프트웨어에 한정된다.

본고는 이 토대 위에서 세 가지 방향으로 논의를 확장한다. 첫째, 분석 대상을 SW에서 AI 시스템으로 확장한다. 둘째, 명세의 성격을 빌드 시점의 정적 스냅샷에서 빌드·배포·런타임을 잇는 동적 신뢰 명세로 재정의한다. 셋째, 선행 연구가 다루지 않은 명세와 방어 활동의 결합, 즉 AIBOM과 능동적 사이버 방어의 환류 구조를 정책 설계

의 중심에 둔다.

이를 위해 최근 발생한 AI 공급망 피해 사례에 기반한 공격 특징 분석, 미국 및 글로벌 표준 기구를 중심으로 전개되는 AIBOM 도입 이슈와 정책 동향, 그리고 국내 AI 공급망 보안 정책 수립을 위한 시사점을 차례로 도출한다.

본 연구는 정책 제안형으로서 두 가지 분석 방법을 결합한다. 첫째, 사례 분석이다. 2025~2026년 공개된 침해 분석 보고서와 위협 정보 기업의 관측 자료를 토대로 여섯 건의 에이전틱 공급망 침해 사례를 피해 규모·확산 경로·자격증명 재사용의 세 기준으로 교차 검토하여, AI 공급망 위협의 구조적 특징을 도출한다(Ⅱ장). 둘째, 정책 비교 분석이다. G7·ISO의 다자 규범, 미국·EU의 입법·조달 조치, 일본·중국의 거버넌스 모델, 영국·싱가포르의 평가 인프라를 (i) 다자 규범, (ii) 시장 진입 요건, (iii) 거버넌스 모델, (iv) 평가·검증 인프라의 네 차원으로 유형화하여 비교한다(Ⅳ장). 이 두 분석의 교차점 위에서 한국의 현황을 진단하고 정책 방향을 제언한다(Ⅴ장).

Ⅱ. AI 공급망 위협의 정책 진단

1. 무엇이 변했는가: 위협의 무게중심 이동

〈표 1〉은 주요 침해 사례를 피해 규모와 확산 속도의 관점에서 요약한 것인데, AI 공급망 위협의 무게중심은 개별 소프트웨어의 코드 결함에서 종속성 그래프와 신뢰 사슬, 자격증명 캐시로 옮겨가고 있

다. 2025년부터 최근까지 공개된 다수의 침해 분석 보고서와 위협 정보 회사의 관측이 공통으로 가리키는 방향이 그렇다.

〈표 1〉 2025~2026 주요 AI 공급망 공격 사례

Case (Date)	Damage (피해 규모 및 확산)
① Hugging Face nullifAI 캠페인 (2025.01~)	세계 최대 AI 모델 허브에 의도적으로 손상된 PyTorch Pickle 모델이 업로드되어 다운로드 즉시 RCE가 발생(7z 압축으로 보안 스캐너 'Picklescan' 우회). 더 큰 위협은 저장소 자체의 신뢰성 붕괴로, 학술 연구상 악성 모델 업로드는 전년 대비 5배 증가했고 100건 이상의 악성 사례가 보고됨. 페이로드는 단순 리버스 쉘에서 자격증명 절도·시스템 핑거프린팅까지 진화
② Shai-Hulud Worm 1.0 (2025.09)	세계 최초의 자가복제 npm 웜. rxnt-authentication@0.0.3(Patient Zero)을 시발점으로 약 500여 개 npm 패키지(@ctrl/tinycolor 등 주당 220만 다운로드 포함)에 전파. TruffleHog를 악용해 npm·GitHub·AWS·GCP 토큰을 탈취하고, 피해자 계정의 다른 패키지에 postinstall 훅을 주입해 자동 재발행. 공개 GitHub 저장소 'Shai-Hulud'에 탈취 자격 증명을 데이터 형태로 게시
③ Trivy & KICS Supply Chain Attack (2026.02~03)	전 세계 컨테이너 보안 표준 스캐너인 Trivy의 GitHub Actions 저장소 77개 중 76개 버전 태그가 단 하룻밤 사이에 일제히 악성 커밋으로 교체(Tag Poisoning)되어, 기존 워크플로우 파일을 변경하지 않은 상태에서 모든 다운스트림 CI/CD가 자동 감염됨. 약 1,000개 이상의 기업 SaaS 환경이 영향권에 들었고, 약 12시간의 노출 시간 동안 GitHub Actions Runner 메모리에서 클라우드 시크릿·GPG 키·DB 자격증명이 직접 탈취됨. '보안 도구 자체가 무기화'된 정교한 사례
④ LiteLLM Attack (2026.03)	월 9,700만 회 다운로드의 AI 통합 API 라이브러리 LiteLLM (v1.82.7~1.82.8)이 침해되었을 때, 이 패키지를 의존성으로 끌어오는 패키지가 1,705개에 달했으며, 그 매개체에는 dspy (월 500만), google-cloud-aiplatform (월 1억 8천만) 등 초대형 패키지가 포함됨. '바이브 코딩' 시대 종속성 폭발의 치명성을 정량적으로 증명한 사례

Case (Date)	Damage (피해 규모 및 확산)
⑤ Telnyx PyPI Attack (2026.03.27)	LiteLLM 침해 단 3일 만에 월 100만+ 다운로드의 Telnyx Python SDK가 동일 공격자에 의해 PyPI에 직접 푸시됨(WAV 스테가노그래피로 페이로드 은닉). PyPI는 6.5시간 만에 격리했지만, 짧은 노출 시간에도 텔레포니·SMS·음성 통신 자격증명이 고밀도로 수집됨. '한 번의 침해가 다음 침해의 자격증명을 공급하는' 자격증명 연쇄 공격이 실증된 사건
⑥ Mini Shai-Hulud (2026.04~05)	2026년 4월 29일과 5월11일 두 차례 파상 공격을 통해 TanStack(라우터), Mistral AI(프랑스 LLM 기업의 SDK), UiPath, OpenSearch, guardrails-ai 등 170여 개 npm 패키지의 404개 악성 버전과 2개 PyPI 패키지가 동시 침해. pull_request_target 트리거의 단명 OIDC(OpenID Connect) 토큰을 스크래핑하는 새 침해 경로 사용

출처: 공개 보고서를 저자가 종합·재구성[1)]

먼저, 침해가 시작되는 지점이 코드 결함이 아니라 종속성과 태그, 자격증명이라는 신뢰 인프라 자체이다. 그리고 단일 침해의 영향이 보안 도구·CI/CD·통신 인프라까지 연쇄 확산된다(후술할 블래스트 레디어스). 마지막으로 침해와 침해 사이의 시간 간격이 단축되면서 한 사건의 자격증명 캐시가 다음 사건의 초기 침투 벡터로 즉시 재투입되는 양상이 관찰된다. 이 세 가지 패턴은 본고가 AIBOM을 정책 도구로 주목하는 근거이다.

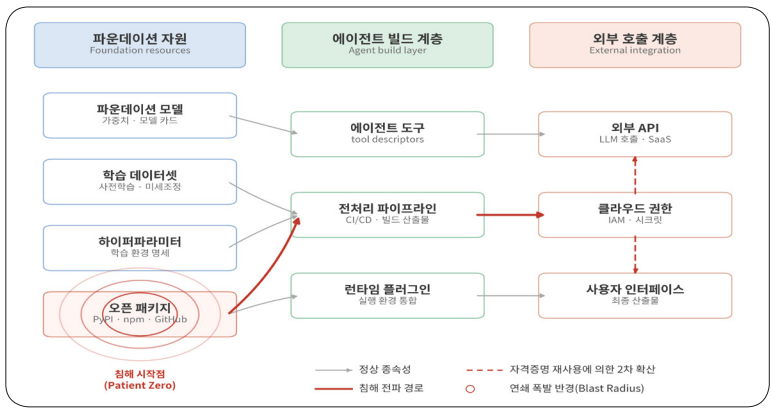
〈표 1〉에 정리된 사례들은 OpenAI·Google·Anthropic의 폐쇄형

1) ① ReversingLabs (2025a), Unit 42 (2026a). ② ReversingLabs (2025b). ③ Socket.dev (2026), Microsoft Security Blog (2026), Arctic Wolf (2026), Aqua Security (2026). ④ Trend Micro (2026a), GitGuardian (2026). ⑤ Trend Micro (2026b), Endor Labs (2026), StepSecurity (2026). ⑥ Unit 42 (2026b).

상용 API, Hugging Face 생태계의 오픈 웨이트 모델, 그리고 PyPI·npm·GitHub Actions·Docker Hub의 패키지 인프라가 거미줄처럼 얽힌 다중 생태계 환경에서 발생했다. 이 사례들이 시사하는 바는 프롬프트 단계의 가드레일만으로는 예방이 어렵다는 점, 그리고 외부 구성 요소의 출처·무결성을 가시화하지 못한 채로는 어떤 방어 전략도 작동하기 어렵다는 점이다.

특히 깃허브 액션(GitHub Actions), 도커 허브(Docker Hub), npm, PyPI를 아우르는 전방위적 공격에서 보듯, 단 하나의 취약한 패키지나 에이전트 도구가 감염되면 통합 환경 전체로 취약점이 확산되는 무한대의 '블래스트 레디어스(Blast Radius)'를 갖는다. 에이전틱 생태계에서의 블래스트 레디어스란, 단일 컴포넌트의 침해가 해당 에이전트에게 위임된 클라우드 제어 권한(IAM)과 연결된 모든 외부 API 인프라의 연쇄적인 붕괴로 직결되는 치명적인 '연쇄 폭발 반경'을 의미한다.

〈그림 1〉 단일 컴포넌트 침해의 연쇄 폭발 반경(Blast Radius)



출처: 저자 작성

〈그림 1〉은 에이전틱 AI 공급망의 다층 구조와, 단일 컴포넌트의 침해가 어떻게 좌측 파운데이션 자원에서 중앙 에이전트 빌드 계층을 거쳐 우측 외부 호출 계층까지 연쇄적으로 파급되는 양상을 보여준다. 그림의 적색 경로는 〈표 1〉의 침해 사례에서 공통으로 관찰된 전파 패턴이다. 이에 안전한 AI 시스템 구축을 위해서는 이토록 파편화된 빅테크 API와 오픈 웨이트 모델 등 외부 구성 요소를 완벽히 가시화하고 제어하는 AIBOM 기반의 에이전틱 거버넌스가 필수적이다.

2. 방어 비대칭의 확대: 대중화·자동화의 격차

Anthropic의 프런티어 레드 팀 연구원 니콜라스 카를리니(Nicholas Carlini)가 2026년 3월 공개한 사례는 이러한 변화의 단면을 실증적으로 보여준다. 그는 별도의 정교한 분석 도구 없이, 단지 코드 저장소를 내려받은 뒤 모든 소스 파일에 대해 동일한 프롬프트("이 프로젝트에서 익스플로잇 가능한 취약점을 찾아 보고서를 작성하라")를 반복 입력하는 단순한 셸 스크립트만으로 Claude Opus 4.6 모델이 검증된 고위험(severity:high) 취약점 500건을 산출하도록 만들었다(Anthropic 2026). 후속 검증 단계에서 동일 모델로 익스플로잇 가능성을 재확인했을 때 그 성공률은 거의 100%에 달했으며, 인기 콘텐츠 관리 시스템 'Ghost'에 대해서는 광범위하게 익스플로잇 가능한 SQL 인젝션 취약점을 즉석에서 도출하기도 했다.

1) '엘리트 관심 부족(Scarcity of Elite Attention)' 가설의 붕괴

Sockpuppet의 토머스 프태섹(Thomas Ptacek)은 자신의 글 "Vulnerability Research Is Cooked"(2026.03)에서, 그동안 일반 인프라가

안전했던 이유는 단지 소수의 최정상급 해커가 시간 자원을 크롬·iOS·안드로이드 등 고가치 표적에만 한정 투입했기 때문이라고 진단한다. 다시 말해, 지방 은행과 종합병원의 운영체제·라우터·심지어 인터넷에 연결된 가전제품에 이르기까지, 패치를 위해 누군가가 차를 몰고 현장에 가야 하는 수많은 약점은 그저 '엘리트의 무관심'이라는 외피에 의해 보호되어 왔을 뿐이라는 것이다. 그런데 이 외피는 지치지도 잠들지도 않는 AI 에이전트가 한계비용 거의 0에 가까운 가격으로 세상의 모든 소스 트리를 동시에 훑기 시작하면서 빠르게 벗겨지고 있다.

2) Bitter Lesson의 보안 영역 재현

리처드 서튼(Richard Sutton)이 2019년 제시한 'Bitter Lesson' 명제, "결국 인간의 도메인 전문 지식보다 더 많은 데이터와 더 큰 연산이 승리한다"는 관찰은 이제 소프트웨어 보안 분야에도 동일한 강도로 적용되고 있다. 과거 취약점 연구자들이 폰트 라이브러리의 메모리 레이아웃이나 C++ 가상 함수 테이블처럼 지엽적이지만 결정적인 지식을 축적하는 데 자기 시간의 80%를 소진해야 했다면, 이제는 그 누적 지식 전체가 프런티어 모델의 가중치에 이미 잠재되어 있다. 결과적으로 공격자는 무수한 에이전트를 병렬로 동원하여 취약점 보고서를 쏟아내는 반면, 인간 방어자(패치 담당자)는 기계의 처리 속도를 결코 따라잡지 못하는 '구조적 비대칭(Structural Asymmetry)' 단계로 진입하고 있다(Mandiant 2026).

3) 정책적 함의: 능동적 사이버 방어로의 패러다임 전환 요구

이러한 비대칭은 Ⅴ장에서 후술하는 '능동적 사이버 방어(Active

Cyber Defense, ACD)' 전환의 핵심 근거가 된다. 즉, 인간 방어자만으로 기계의 속도에 맞설 수 없는 이상, 우리도 동일한 종류의 '방어용 AI 군집(Defensive Swarm)'을 사전 배치하여 우리 인프라의 취약점을 적보다 먼저 발굴·차단해야 한다는 것이다. 단순한 패치 자동화를 넘어선 능동적 취약점 사냥(Active Bug Hunting)이 국가 사이버 안보의 새로운 기본값이 되어야 한다. 이 함의는 Ⅲ장 3절과 Ⅴ장 1절 3)에서 다룬다.

Ⅲ. 정책 도구로서의 AIBOM

1. AIBOM은 무엇이며 무엇을 가능케 하는가?

AIBOM(AI Bill of Materials)은 미국 바이든 정부의 행정명령(EO 14028)으로 도입된 SBOM 개념을 AI 시스템에 확장한 것으로, 모델 메타데이터와 학습 데이터셋, 하이퍼파라미터, 전처리 파이프라인, 에이전트 도구 명세, 평가 결과를 포괄하는 동적 명세서를 가리킨다. 일반 SW 영역에서 SBOM은 Linux Foundation이 관리하는 SPDX와 OWASP가 주관하는 CycloneDX의 두 표준 형식을 중심으로 자리잡았으며(Linux Foundation 2024; OWASP 2024), AIBOM 역시 이 두 표준의 확장 작업을 통해 구체화되고 있다. 〈그림 2〉는 AIBOM의 핵심 구성 요소를 학습 단계와 배포·운영 단계로 구분하여 정리한 것이다.

〈그림 2〉 AIBOM의 구성 요소

출처: 저자 작성

본고가 AIBOM을 정책 도구로 주목하는 이유는 세 가지이다. 첫째, Ⅱ장에서 정리한 위협의 무게중심 이동(코드 결함 → 종속성·신뢰 사슬·자격증명)에 대응하기 위해서는 외부 구성 요소의 출처와 무결성을 가시화할 표준 명세가 필요한데, 기존 SCA(Software Composition Analysis) 도구는 AI 시스템 고유의 비(非)코드 자산(모델 가중치, 학습 데이터셋, 에이전트 도구 명세)을 인식하지 못하는 가시성 공백을 갖는다. 가트너도 2029년까지 기업 AI 침해 사고의 60%가 소프트웨어 취약점이 아니라 손상된 제3자 데이터셋이나 모델 제공자로부터 발생할 것으로 전망하면서, 기존 SBOM과 SCA 도구가 모델 가중치를 '읽어내지' 못하는 본질적 한계를 가진다는 점을 강조한 바 있다(Gartner 2026). 둘째, AIBOM은 산업 자율 권고가 아니라 계약·조달 단계의 통과 요건으로 격상되면서 글로벌 시장 진입의 기준선이 되고 있다. 이 제도화 흐름이 본 장에서 정의하는 동적 AIBOM 개념을 요구하게 된 경위는 Ⅳ장에서 상술한다. 셋째, AIBOM은 단일 문서가 아니라 빌드·배포·런타임을 잇는 신뢰 명세의 결합체로 발전 중이며, 이 결합 양

식이 본고가 '동적 AIBOM (Dynamic AIBOM)'이라 부르는 개념의 핵심이다. 다만 AIBOM은 만능 도구가 아니라 보완재가 필요한 정책 도구다.

2. AIBOM의 한계와 기술적 심층 과제

산업 표준 논의에 비추어 보면, AIBOM이 실제 보안 보증으로 기능하기 위해서는 세 가지 보완재가 필요하다. 첫째, 빌드 시점의 정적 스냅샷 한계이다. 기존 SBOM은 빌드가 완료된 시점의 구성요소 목록을 한 차례 기록할 뿐, 이후 같은 버전 태그가 다른 산출물로 교체되는 형태의 공격에는 본질적으로 취약하다. 이 한계를 메우기 위해 산업 진영에서 제시되는 접근이 미국 뉴욕대학교 연구진이 주도하고 현재 CNCF 프로젝트로 운영되는 인-토토(in-toto) 프레임워크로, 소스 체크아웃부터 빌드·패키징·배포에 이르는 각 단계의 수행자와 산출물을 암호학적 증명서로 연결한다(CNCF 2024). 이를 AIBOM이 단일 문서가 아닌 '동적 증거 사슬'로 진화하기 위한 첫 번째 보완재로 본다.

둘째, 취약점 경고의 맥락 부족이다. AIBOM이 구성요소의 취약점을 보고하더라도 실제 운영 환경의 통제 상태(가드레일, 외향 트래픽 필터링, 최소 권한 원칙)에 따라 익스플로잇 가능 여부가 달라지므로, 모든 CVE를 동일 강도의 경고로 전환하면 '경고 피로(alert fatigue)'가 발생한다. 이를 해소하기 위한 표준이 미국 CISA가 2022년 제안한 '취약점 익스플로잇 가능성 교환(VEX)'이며, 이를 AI 에이전트 환경에 맞게 확장한 'AI-VEX'를 AIBOM의 두 번째 보완재로 제안한다(CISA 2022). 한편 산업 진영에서는 OpenSSF의 OSV-Scanner와 같은 오픈소스

취약점 데이터베이스 연동 도구가 SBOM-VEX 결합의 자동화 기반을 이미 제공하고 있어(OpenSSF 2024), AI-VEX의 구체적 구현은 이러한 기존 생태계를 토대로 발전시킬 수 있다.

셋째, 런타임 단계 위협의 사각이다. AIBOM과 동적 증거 사슬, AI-VEX는 빌드·배포 단계까지의 무결성을 다루지만, 에이전트가 실제로 가동되는 런타임 단계에서 발생하는 메모리 스크래핑이나 학습 데이터로부터의 정보 누출까지 자동으로 차단하지는 않는다. Ⅱ장에서 살펴본 자격증명 수집기형 공격이나 학습 데이터의 멤버십 추론 위협에 대응하기 위해서는 실행 환경의 격리와 학습 데이터의 보호를 함께 보증하는 런타임 무결성 계층이 AIBOM의 세 번째 보완재로 요구되며, 산업 진영에는 이를 뒷받침하는 다양한 기법이 이미 존재하나 성능·비용 측면에서 보편적 적용이 어렵다는 한계도 분명하다.

이상의 세 보완재를 결합한 형태가 곧 '동적 AIBOM'이다. 즉, 빌드 사슬·익스플로잇 가능성 상태·런타임 격리 증명을 함께 충족하는 통합 신뢰 명세이다.[2] 이러한 결합 모델이 어느 시점에 어느 조직에서 실현 가능한지는 자산의 보안 등급과 조직의 성숙도에 따라 달라지며, Ⅴ장 2절의 정책 제언이 이 차등화를 다룬다. 그리고 이 통합 신뢰 명세가 실제 방어 활동과 어떻게 결합하는지가 다음 절의 주제이다.

2) 본고가 다루는 동적 AIBOM 기반 사전 검증 체계와 관련하여, '에이전틱 인공지능 공급망 보호를 위한 동적 자재명세서 기반 사전 검증 시스템 및 그 방법'에 관한 특허가 출원 중이다. 구체적 데이터 모델과 구현은 해당 출원 및 별도의 기술 표준화·실증 연구에서 다룬다.

3. 동적 AIBOM과 능동적 사이버 방어의 환류 구조

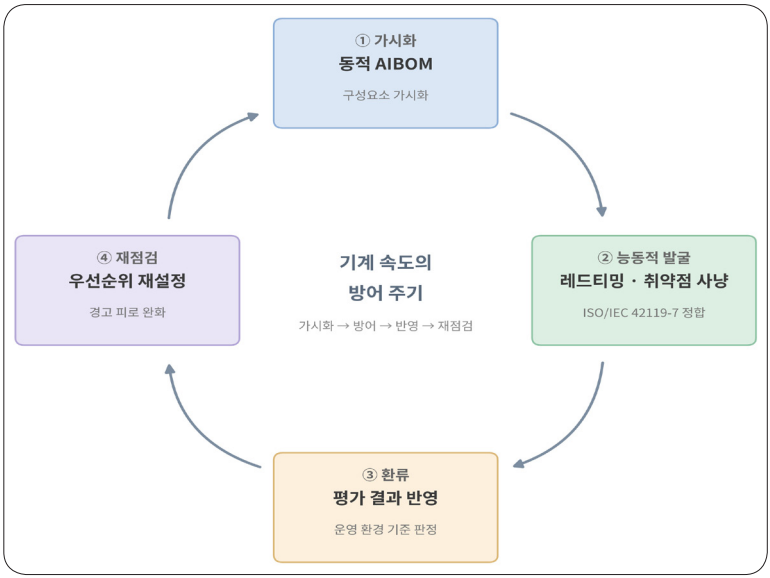
앞 절까지의 논의는 AIBOM이 "무엇을 기록하는가"에 관한 것이었다. 본 절은 한 걸음 나아가, 그 기록이 Ⅱ장 2절에서 진단한 구조적 비대칭에 어떻게 대응하는 방어 활동으로 전환되는지를 다룬다. 동적 AIBOM과 능동적 사이버 방어(ACD)는 별개의 정책 수단이 아니라 하나의 순환 구조를 이루는 두 국면이다.

첫째, AIBOM은 능동적 방어의 전제 조건이다. 능동적 사이버 방어란 침해가 발생한 뒤에 대응하는 수동적 방어와 달리, 자기 인프라의 취약점을 공격자보다 먼저 발굴·차단하는 일련의 선제 활동을 가리킨다. 그런데 이 활동들은 모두 "무엇을 점검할 것인가"라는 대상 정의에서 출발하며, 그 대상은 곧 자산과 종속성의 전체 목록이다. Ⅱ장에서 확인한 블래스트 레디어스가 보여주듯, 에이전틱 환경에서 점검 대상은 자기 코드가 아니라 모델 가중치·데이터셋·에이전트 도구·외부 API로 뻗어 있는 종속성 그래프 전체이다. AIBOM이 이 그래프를 가시화하지 못하면 레드티밍은 보이는 자산만 두드리는 부분 점검에 그치고, 가장 위험한 '보이지 않는 종속성'은 점검 범위 밖에 남는다. AIBOM이 점검 대상을 규정하지 못하면 능동적 방어의 범위 자체가 정해지지 않는다.

둘째, 능동적 방어의 결과는 AIBOM으로 환류되어 명세의 동적 갱신을 완성한다. 레드티밍과 취약점 사냥이 산출하는 것은 개별 취약점 보고서가 아니라, 해당 구성요소가 현재 운영 환경에서 실제로 익스플로잇 가능한지에 대한 상태 판정이다. 이 판정이 명세에 반영되면, AIBOM은 빌드 시점의 정적 스냅샷에 머물지 않고 평가 결과를 지속적으로 수용하는 명세로 기능한다. 이때 평가 결과를 명세에 결

합·갱신하는 구체적 방식은 본고의 범위를 넘어서며, 각주 2에 밝힌 출원 및 후속 연구에서 다룬다. 그 정책적 효과만 짚으면, 운영 환경의 통제 상태가 반영됨에 따라 보안 관제의 경고 피로가 완화되고, 한정된 평가 자원을 위험도에 따라 배분할 근거가 마련된다.

〈그림 3〉 동적 AIBOM과 능동적 사이버 방어의 환류 구조



출처: 저자 작성

이 두 방향을 결합하면 〈그림 3〉과 같이 가시화와 능동적 방어가 맞물리는 순환적 관계가 형성된다. 가시화된 명세가 방어 활동의 대상을 규정하고, 방어 활동의 결과가 다시 명세에 반영되어 다음 점검의 토대가 되는 구조이다. 이러한 순환이 사람의 개입 속도를 넘어 동작할 수 있을 때 Ⅱ장 2절에서 제기한 "에이전트가 쏟아내는 취약점 보고를 인간 패치 담당자가 따라잡지 못하는" 비대칭에 대응할 수 있

다. 공격자가 AI 군집으로 탐색 속도를 높이는 만큼, 방어자도 가시화—발굴—환류의 주기를 기계 속도로 단축해야 하기 때문이다.

이 순환 구조는 Ⅳ장에서 살펴볼 글로벌 정책 동향을 해석하는 분석 틀이 되며, Ⅴ장의 정책 제언, 특히 상시 레드티밍 체제와 평가 결과의 AIBOM 환류 제언(Ⅴ장 1절 3)의 이론적 근거가 된다.

Ⅳ. 글로벌 AIBOM 활성화 및 표준화 정책

AI 공급망 보안에 대한 정책적 대응은 2024~2026년 사이 다자 규범의 형성(G7·ISO), 주요국의 입법·조달 조치(미국·EU·일본·중국), 그리고 평가·검증 인프라의 분화(영국·싱가포르)라는 세 층위에서 동시에 진행되었다. 〈표 2〉는 이러한 흐름의 주요 정책·표준을 시점별로 정리한 것으로, 산업 분석 진영의 권고(가트너)와 국제 표준화 작업(ISO/IEC 42119-7)을 함께 포함한다. 본 장은 이러한 흐름이 어떻게 AIBOM을 산업계의 자율 권고에서 시장 진입 요건으로 격상시키고 있는지에 초점을 맞추어 다자 규범, 미국, EU 등 주요국의 순서로 정리한다. 이하의 분석은 Ⅲ장에서 정의한 동적 AIBOM의 세 보완재가 각국 제도에서 어떤 형태로 요구되고 있는지를 확인하는 작업이기도 하다.

〈표 2〉 주요국 AIBOM 관련 정책·표준 동향 (2023~2026)

Policy / Standard	Remark (주요 내용)
G7 SBOM for AI 공유 비전→최소 요소 (2025.06/2026.05)	2025년 6월 G7 사이버보안 워킹그룹이 '공유 비전' 채택(이탈리아 ACN·독일BSI 공동 작성, 캐나다 의장국). 2026년 5월 현재 의장국인 프랑스 주관 하에 '최소 요소' 발표, 7개 클러스터로 구조화

Policy / Standard	Remark (주요 내용)
US FY26 NDAA (2025.12)	국방부 조달 AI 시스템에 대한 생애주기 보안 의무화. 섹션 1513(위험 기반 사이버·물리 보안 프레임워크), 섹션 1533(AI 모델 윤리 검증 범부처 협업체), AI 시스템 단위 명세(AIBOM) 제출을 의회 차원에서 최초 명문화
OMB M-26-04 (2025.12)	행정명령(EO) 14319 이행 메모. 신규 LLM 조달 시 model card·system card·AUP 등 투명성 자료 제출을 계약상 의무화. 각 부처 2026.3.11.까지 조달 정책 갱신, 사실상의 AIBOM 조달 요건
EU AI Act 본법 (2024.08~)	고위험 AI 모델 제공자에 대한 기술문서 작성·제출 의무화. 위반 시 최대 3,500만 유로 또는 전 세계 매출 7% 벌금, 기술문서·추적성 의무가 AIBOM의 법적 수요를 창출
EU Digital Omnibus on AI (2026.05)	잠정 합의. 고위험 AI 의무 적용 시점을 2026.8 → 2027.12.2(Annex III) / 2028.8.2(Annex I)로 연기. 워터마킹 의무는 2026.12.2로 확정. 범용 인공지능 모델(GPAI) 의무 일정은 변동 없음
Japan AI Promotion Act (2025.05)	일본 최초의 AI 입법. AI를 전략 자산으로 규정하고 투명성·안전성·국제 정합성 원칙 천명. 처벌 조항 없는 소프트법 접근. METI·MIC 'AI Guidelines for Business v1.1'(2025.3.28.)과 연계, AIBOM·SBOM 산업 침투 모델
China Generative AI Interim Measures (2023.08)	중국 사이버공간관리국(CAC) 주도 첫 생성형 AI 행정 규정. 안전성 평가·알고리즘 등록 의무화. 2025년 말 누적 748건 서비스·5,000+ 알고리즘 등록. 포괄적 AI 단일법은 2025 입법 일정에서 제외, AIBOM 기능의 영역별 분산 구현
China TC260 AI 안전 표준 3종 (2025.11)	(i) 데이터 라벨링 안전, (ii) 사전학습·미세조정 데이터 안전, (iii) 생성형 AI 서비스 기본 보안 요구사항 국가표준. 중국 사이버보안법 개정으로 AI 조항 첫 국가법 편입, 데이터·학습 명세 요구가 AIBOM 데이터셋 속성에 대응
UK AI Security Institute (2023.11~)	30+ 프런티어 모델 평가 수행. 사이버, 화학, 생물, 자율 에이전트, 정렬 영역. '보충적 감독 계층'으로 자리매김, AIBOM 명세 정보(모델 카드 등)의 외부 검증 토대
Singapore AI Verify Foundation (2023.06~)	AI Verify 테스트 도구(2022)·Project Moonshot(2025)·Global AI Assurance Pilot(2025.2)·Singapore AISI(2025) AIBOM 평가 결과 항목의 다층 평가 체제 구축

Policy / Standard	Remark (주요 내용)
ISO/IEC AWI TS 42119-7 (2026, 초안)	인공지능 테스팅 제7부(Red Teaming) 국제 표준 초안. 산발적인 모의 해킹을 넘어, AI 시스템 수명 주기에 맞춘 표준화되고 감사 가능한(Auditable) 레드팀 평가 방법론 규격화, AIBOM 환류형 레드티밍의 감사 가능 방법론
Gartner 5 Steps to Secure AI Supply Chain (2026.03)	기존 SBOM의 한계 선언. 1~2단계: AIBOM 제출 의무화. 3~4단계: 암호화 해싱 및 에이전트 JIT 액세스 등 행동 제어. 5단계: 'AI 중심 레드티밍' 도입, AIBOM 제출 의무화(1~2단계)를 명시한 최초의 산업 권고

출처: 공개된 정책, 표준 자료를 저자가 종합[3)]

1. G7과 ISO: 다자 규범의 형성

다자 규범의 형성은 2025년 6월 G7 사이버보안 워킹그룹이 'AI를 위한 SBOM(SBOM for AI) 공유 비전'을 채택하면서 시작되었고, 2026년 5월, G7 의장국인 프랑스 주관 하에 'SBOM for AI: Minimum Elements'로 한 단계 구체화되었다(G7 Cybersecurity Working Group 2026). 후자는 이탈리아 ACN과 독일 BSI가 공동 주관하고 프랑스 ANSSI·캐나다 CSE·미국 CISA·영국 NCSC·일본 NCO 및 유럽위원회가 참여해 2025년 8월부터 2026년 2월까지 작성된 G7 최초의 AIBOM 가이던스로, AIBOM의 최소 요소를 메타데이터(Metadata), 시스템 수준 속성(SLP), 모델(Models), 데이터셋 속성(DP), 인프라(Infrastructure), 보안 속성(SP), 핵심성과지표(KPI)의 일곱 클러스터로 구조

3) G7 사이버보안 WG(2025; 2026), US FY26 NDAA, OMB M-26-04, EU Council 2026.5 보도, 日M ETI·MIC 가이드, 중국 CAC·TC260 자료, UK AISI·Singapore IMDA 공개 자료, ISO/IEC 공식 자료, Gartner Research 자료 등

화한다.

권고 형식임에도 이 문서의 클러스터 구성은 이후 소개할 미국 OMB Memorandum M-26-04(이하 OMB M-26-04)의 model card·system card 항목, EU AI Act가 요구하는 기술문서 구성과 상당 부분 중첩되며, 이러한 점에서 G7 다자 규범을 AIBOM의 '국제 공통 어휘'를 형성하는 작업으로 평가한다. 이 좌표축의 실증 도구로 Red Teaming과 가트너 5단계 로드맵이 보완재로 자리잡게 된다.

2. 미국: 입법·조달의 동시 진행

미국은 2025년 말부터 2026년 초까지 입법과 조달의 두 경로에서 거의 동시적으로 AIBOM 관련 조치를 내놓았다. 입법 측면에서는 2025년 12월 의회를 통과한 FY26 국방수권법이 국방부 조달 AI 시스템에 대한 생애주기 보안 정책 수립과 모델·시스템 단위 명세 제출을 명시적으로 의무화함으로써, AI 시스템에 대한 SBOM·AIBOM 적용을 의회 차원에서 처음으로 명문화하였다(U.S. Congress 2025). 산업 분석에 따르면, FY26 NDAA는 기존 SBOM이 가진 '투명성 공백(Transparency Gap)'을 AI 영역으로 메우는 입법적 시도로 평가된다(Manifest Cyber 2026). 행정 측면에서는 2025년 12월 11일 OMB M-26-04가 모든 신규 LLM 조달 시 model·system·data card와 사용자 피드백 채널 등 투명성 자료의 계약상 제출을 의무화하고 미준수 시 계약 해지를 가능케 함으로써, AIBOM 제출 의무를 행정 절차로 확립하였다(OMB 2025).

두 조치는 별개의 경로를 따르지만 결과적으로 동일한 방향을 가리킨다. AI를 조달하는 연방 기관이 모델·시스템 단위의 명세 자료를

계약 단계에서 확보·검증하지 못하면 거래 자체가 성립하지 않도록 만든 것으로, 이는 AIBOM이 산업 자율 권고에서 사실상의 시장 진입 요건으로 격상되었음을 의미한다.

3. EU: 시행 유예의 정치학과 그 함의

EU AI Act는 2024년 8월 발효 이후 단계적으로 적용되어 왔으며, Annex III 고위험 AI 시스템에 대한 핵심 의무는 본래 2026년 8월부터 적용될 예정이었다. 그러나 산업계와 회원국의 우려를 받아들여 집행위원회가 2025년 11월 제출한 '디지털 옴니버스(Digital Omnibus on AI)' 입법안(European Commission 2025)이 2026년 5월 7일 EU 의회·이사회·집행위 간 잠정 합의에 도달함으로써, Annex III 의무는 2027년 12월로, Annex I 의무는 2028년 8월로 각각 약 16개월 및 12개월 연기되었다(Council of the EU 2026). 동시에 워터마킹 의무는 오히려 3개월 단축되어 2026년 12월부터 적용되며, 범용 AI 모델 의무는 일정에 변동이 없다.

EU의 이번 조치가 강조하고자 하는 바는 명확하다. 이번 연기는 규제의 후퇴가 아니라, 회원국·산업계와 함께 기술 표준과 적합성 평가 인프라를 정비하기 위한 준비 기간(grace period)에 가깝다. 고위험 AI에 대한 실체적 의무는 그대로 유지되고 워터마킹 의무는 오히려 빨라졌으며, 새 일정은 명문 고정일로 못박혔다. 이 시기를 한국 산업이 '뒤늦은 따라잡기'가 아니라 G7·미국·EU 규범에 동시에 정합하는 AIBOM 체계를 선제적으로 설계할 수 있는 한정된 기회의 시기로 활용해야 한다.

4. 일본: 소프트법 중심의 통합 거버넌스

일본은 2025년 5월 일본 최초의 AI 입법인 'AI 추진법(AI Promotion Act)'을 통과시키면서도 처벌 조항 없는 소프트법(soft law) 접근을 채택했다(METI 2025a). 공급망 보안 영역에서는 2025년 9월 일본이 미국·EU와 공동으로 발표한 '사이버보안을 위한 SBOM 공유 비전 국제 가이드'(METI 2025b)와 2026년 3월 METI·NCO가 발표한 'Cyber Infrastructure Providers 책임 가이드'가 G7의 'AI를 위한 SBOM 공유 비전'(본 장 1절 참조)과 직접 연결되는 후속 조치를 형성한다(METI 2026). 일본 모델의 함의는 강제력보다 광범위한 가이드라인과 국제 협력을 통해 AIBOM·SBOM 체계를 산업에 침투시키는 접근으로, 한국이 AI 기본법 후속 입법을 설계할 때 비교 사례를 제공한다.

5. 중국: 수직 규제와 알고리즘 등록제

중국은 EU의 수평적(horizontal) AI 단일법과 정반대로, AI의 개별 응용 영역마다 독립된 규제를 도입하는 수직(sectoral) 규제 전략을 일관되게 유지해 왔다. 2023년 8월 발효된 사이버공간관리국(CAC)의 '생성형 AI 서비스 관리 잠정 방법'은 공개 출시 전 안전성 평가와 알고리즘 등록을 의무화하였고(CAC 2023), 2025년 11월부터는 TC260이 채택한 데이터 라벨링·사전학습·서비스 기본 보안에 관한 세 건의 국가표준이 시행되었다(TC260 2025). AIBOM 관점에서 보면 중국 모델은 알고리즘 등록·안전성 평가·모델 카드 의무가 영역별로 분산된 형태로 사실상 작동하고 있으며, 일본의 소프트법 모델과 함께 한국이 '기본법 + 영역별 후속 입법' 전략을 설계할 때의 양극단 참조점을 제공한다.

6. 영국과 싱가포르: 국가별 평가·검증 인프라의 분화

AIBOM이 단순한 문서가 아니라 실제 보안 보증으로 기능하려면 모델·시스템 단위의 평가·검증 인프라가 함께 갖춰져야 한다. 영국 AI Security Institute(AISI)는 2023년 출범 이후 프런티어 AI 모델 30종 이상을 평가하면서 자신을 규제 기관이 아닌 '보충적 감독 계층(supplementary layer of oversight)'으로 자리매김했다(UK AISI 2025). 싱가포르는 정보통신미디어개발청(IMDA)의 'AI Verify' 테스트 도구와 'AI Verify Foundation'을 중심으로 2025년에는 'Project Moonshot', 'Global AI Assurance Pilot', 별도의 싱가포르 AISI를 잇따라 신설하여 정부·산업·국제 표준 기구를 잇는 다층 평가 체제를 구축했다. 두 모델 모두 AIBOM이 명세하는 정보(모델 카드, 데이터 카드, 평가 결과)에 대한 외부 검증 토대를 제공함으로써, AIBOM 체계 자체의 신뢰성을 결정짓는 핵심 보완재가 된다.

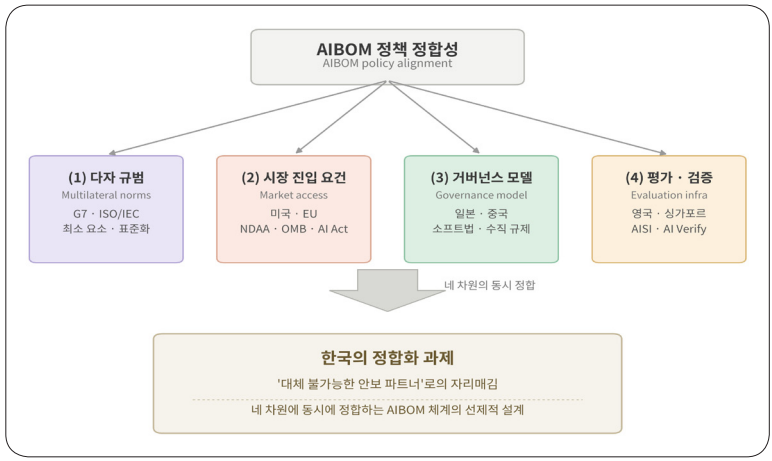
이러한 흐름의 공통된 방향은, AI 공급망 보안의 무게중심이 정적 문서 검토에서 능동적 평가·검증으로 빠르게 옮겨가고 있다는 것이다. 가트너가 2026년 3월 발표한 'AI 공급망 보안 강화 5단계 전략'과 ISO/IEC 42119-7은 모두 AIBOM 제출 의무화에 이어 암호화 해싱, JIT 액세스, AI 중심 레드티밍이 결합되어야 함을 공통으로 지적한다.

이러한 방향은 정책 영역에서도 가시화되었는데, 『New York Times』 2026년 5월 4일 보도에 따르면 미 백악관은 일정 능력 임계 이상의 AI 모델에 대해 공개 출시 이전 연방 정부 차원의 사전 검증을 요구하는 행정명령을 검토 중이며, 그 직접적 계기는 Anthropic이 자사 모델 'Mythos'의 출시를 자발 보류한 'Project Glasswing' 사례로 알려져 있다(New York Times 2026). 이 검토가 확정되면 AIBOM은

단순한 조달 문서가 아니라 모델 출시 가부 자체를 좌우하는 사전 검증의 입력값으로 격상된다.

2026년 시점에서의 AIBOM 정책은 더 이상 '도입 여부'의 문제가 아니라 '어떤 속도와 어떤 정합성으로 도입할 것인가'의 문제로 옮겨와 있으며, 그 정합성의 축은 (i) 다자 규범(G7·ISO), (ii) 시장 진입 요건(미국·EU), (iii) 거버넌스 모델(일본·중국), (iv) 평가·검증 인프라(영국·싱가포르)의 네 차원으로 정리된다. 〈그림 4〉는 이러한 네 차원의 정합성 구조와 한국이 직면한 정합화 과제를 시각화한 것이다.

〈그림 4〉 AIBOM 정책 정합성의 네 차원



출처: 저자 작성

Ⅴ. 결론 및 정책 제언

1. AIBOM의 재정의와 한국의 전략적 좌표

AI 혁신을 가능하게 하는 것이 기술이라면, AIBOM과 투명성, 그리고 능동적 사이버 방어는 그 혁신이 살아남도록 만드는 안보의 조건이다. 2025~2026년 사이 잇따른 에이전틱 공급망 침해 사태는, AIBOM을 둘러싼 정책 지형을 결정적으로 바꾸어 놓았다. 미국 FY26 NDAA, OMB M-26-04, 그리고 EU AI Act와 그 디지털 옴니버스로 이어지는 일련의 흐름은, Ⅳ장에서 정리한 시장 진입 요건으로의 격상을 입법·조달의 양 경로에서 동시에 굳히고 있다. 본고는 이러한 변화를 두 가지 명제로 압축한다. 첫째, AIBOM은 글로벌 시장 진출의 수출 장벽으로 작동하기 시작했다. 둘째, 그 작동 방식이 사이버 영역에서의 비대칭 능력 확보 경쟁과 분리될 수 없다는 점에서, AIBOM은 광의의 '사이버 무기 체계'의 일부로 이해되어야 한다.

이러한 재정의는 미국이 표방하는 '힘에 의한 평화(Peace through Strength)' 기조 아래에서 한층 가속되고 있다. 미국 주도의 보안 스택(security stack)이 동맹·파트너 국가에 사실상의 표준으로 확산되는 지정학적 환경에서, 산업 진흥 논리에 무게를 둔 한국의 AI 기본법만으로는 이 흐름의 수범자 위치를 벗어나기 어렵다. 한국이 표방하는 '소버린 AI(Sovereign AI)'가 자국 우선주의의 구호에 그치지 않고 국제 질서 안에서의 실질적 자율성을 의미하려면, 미국 주도 연합 내에서 한국이 자체 기여 능력으로 대체 불가능한 위치를 확보하는 전략이 필요하다. 필자는 그 위치를 사이버 억지력과 방어 무결성을 함께 제공

하는 '대체 불가능한 안보 파트너'로 정의한다. 이는 미국의 정책을 뒤따르는 데 그치지 않고, 한국이 자체적으로 검증과 평가, 방어 역량을 축적하여 동맹의 사이버 회복력에 실질적으로 기여할 때 비로소 성립한다.

다만 이러한 전략적 좌표 설정에 앞서, 한국의 현재 위치를 Ⅳ장에서 정리한 네 차원의 정합성 축 위에서 냉정하게 진단할 필요가 있다.

첫째, 거버넌스 모델 차원에서 한국은 2026년 1월 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 AI 기본법)을 전면 시행함으로써, EU에 이어 두 번째로 포괄적 AI 법제를 갖춘 국가가 되었다. EU가 디지털 옴니버스를 통해 고위험 AI 의무를 2027년 말 이후로 연기한 것을 고려하면, 포괄 규범을 실제로 전면 적용한 사실상 최초의 국가이기도 하다. 그러나 AI 기본법의 의무 구조는 고영향 AI에 대한 투명성·안전성 확보와 영향평가를 중심으로 설계되어 있어, 본고가 다루는 공급망 차원의 구성요소 명세에 대한 의무는 법령과 하위 고시 어디에도 존재하지 않는다. 미국이 FY26 NDAA와 OMB M-26-04로 조달 단계의 AIBOM 제출을 사실상 의무화한 것과 대비하면, 한국의 법제는 '시스템이 안전한가'를 묻되 '시스템이 무엇으로 만들어졌는가'는 묻지 않는 구조이다.

둘째, 시장 진입 요건 차원의 격차는 시간표로 드러난다. 한국의 SW 공급망 보안 정책은 2024년 5월 과기정통부·국정원·디지털플랫폼정부위원회가 합동 발표한 'SW 공급망 보안 가이드라인 1.0'을 기점으로 하며(과학기술정보통신부·국가정보원·디지털플랫폼정부위원회 2024), 관계부처 로드맵상 공공 납품 IT 제품에 대한 SBOM 제도는 2027년 도입, 2028년 본격 운영이 예정되어 있다. 이 일정 자체도 미국 연방 조달의 2026년 3월 갱신 시한보다 늦지만, 더 근본적인 공백

은 이 로드맵이 전통적 SW의 SBOM을 전제하고 있어 AI 시스템 고유의 비(非)코드 자산을 포괄하지 못한다는 점이다. 즉 한국 기업이 2027년 국내 SBOM 제도에 성실히 대응하더라도, 미국·EU 시장이 요구하는 AIBOM 수준의 명세와는 별도의 격차가 남는다.

셋째, 평가·검증 인프라 차원에서 한국은 2024년 11월 AI안전연구소를 출범시켜 영국·싱가포르형 평가 기관의 외형을 갖추었으나, 그 평가는 프런티어 모델의 안전성 평가에 무게가 실려 있고, 평가 결과가 공급망 명세(AIBOM)의 상태 정보로 환류되는 Ⅲ장 3절의 순환 구조는 제도적으로 설계되어 있지 않다. 다자 규범 차원에서도 한국은 G7 'SBOM for AI' 작성에 참여한 미국·영국·일본 등과 달리 규범 형성 과정의 바깥에 있었다.

종합하면 한국의 현재 좌표는 "포괄 법제는 가장 빠르나, 공급망 명세·평가 환류·규범 형성의 세 측면에서는 공백"으로 요약된다. 이 진단은 역설적으로 기회이기도 하다. AI 기본법이라는 수평적 법제가 이미 가동 중이므로, 주요국처럼 입법부터 새로 시작할 필요 없이 하위 고시·조달 기준의 보강만으로 AIBOM 체계를 빠르게 접목할 수 있는 제도적 접점이 존재하기 때문이다. 이하의 제언은 이 접점을 전제로 한다.

이러한 결론이 가리키는 공통의 방향은, 한국이 AIBOM을 단순히 외부 규범에 맞춰 도입해야 하는 행정 부담이 아니라, 자국의 산업 경쟁력과 안보 자율성을 동시에 떠받치는 전략 자산으로 재설계해야 한다. 이러한 인식 위에서 미국·EU·G7의 선행 입법과 그 시행 과정에서 드러난 한계를 함께 참조하여, 한국이 취해야 할 정책 방향을 다음과 같이 제언한다.

1) 보안 등급과 조직 성숙도를 함께 고려한 차등화된 AIBOM 체계의 설계

모든 자산과 조직에 동일한 잣대를 들이대는 'One-size-fits-all' 정책은 필연적으로 실패한다는 것이 미국·EU 시행 과정의 일관된 교훈이다. Ⅲ장에서 제시한 '동적 AIBOM'은 빌드·배포·런타임을 잇는 통합 신뢰 명세로서 그 자체가 정교한 체계이며, 모든 조직이 이를 동시에 갖추기를 기대하는 것은 비현실적이다. 따라서 다음과 같은 두 축의 차등화가 필요하다.

(1) 자산의 보안 등급에 따른 기술 차등

국방·의료·국가 핵심 인프라에 적용되는 1급 자산에는 동적 AIBOM의 전체 요소를 의무화하되, 일반 공공 데이터·서비스에는 정적 AIBOM의 핵심 요소만을 요구하여 활용도와 안전성 사이의 균형을 확보해야 한다.

(2) 조직 성숙도에 따른 의무 차등

글로벌 시장에 직접 진출하는 대기업·금융기관에는 G7 SBOM for AI 공유 비전과 EU 디지털 옴니버스의 새 일정에 정합하는 동적 AIBOM 체계를 우선 적용하되, 자체 보안 역량이 부족한 중소기업(SME)에는 의무 부과에 앞서 제로 트러스트(Zero Trust) 아키텍처 도입과 SBOM·AIBOM 자동 생성 도구 활용을 위한 바우처·인프라 지원이 선행되어야 한다. 이 단계적 접근은 OMB M-26-04가 미 연방기관에 부여한 단계적 갱신 일정(2026년 3월 11일까지의 조달 정책 갱신)과 EU 옴니버스가 확보한 준비 기간을 한국적 맥락에서 동시에 반영하는 방식이기도 하다.

2) Security by Design의 무기화와 Compliance-as-Code 생태계 구축

글로벌 무대에서 AI 규제는 점점 비관세 수출 장벽으로 작동하고 있다. EU AI Act, FY26 NDAA, OMB M-26-04는 모두 자국 또는 역내 시장 진입을 위한 사실상의 통과 요건을 부과한다. 한국이 이 흐름의 수동적 수범자에 머무르지 않으려면, 설계 단계부터 안전성이 확보된 'Security by Design(SbD)'을 국가 기본 표준으로 채택하고, 그 결과물을 G7 SBOM for AI 최소 요소와 호환되는 형태로 자동 생성·검증할 수 있는 '컴플라이언스 코드화(Compliance-as-Code)' 생태계를 구축해야 한다.

Ⅲ장에서 제시한 '동적 AIBOM'의 빌드 사슬 증명이 이러한 자동화된 검증 체계의 기술적 토대를 이룬다. CI/CD 파이프라인 상에서 빌드의 모든 단계가 암호학적 서명과 함께 기록되고, 그 서명 사슬이 AIBOM의 일부로 유통되며, 도착지에서 기계적으로 검증된다면, 이는 곧 한국 기업이 글로벌 시장에 제출할 수 있는 '안전 통행증'으로 작동한다. 앞 절에서 제안한 성숙도 차등 지원은 중소기업까지 이 체계에 편입시키기 위한 필수 조건이다.

3) 수동적 방어를 넘어선 '능동적 방어·억지' 인프라의 제도화

Ⅱ장 2절에서 살펴본 바와 같이, 인간 방어자가 기계의 속도를 따라잡지 못하는 비대칭이 정책적으로 대비해야 할 시나리오로 부상하고 있다. Ⅳ장 6절에서 다룬 미 백악관의 사전 검증 행정명령 검토와 그 직접 계기였던 Anthropic의 'Project Glasswing' 사례는, 표준화 단계의 권고(가트너 5단계, ISO/IEC 42119-7)가 실제 정책으로 결정화되어 가는 과정을 보여준다. 비록 이 단일 사례가 곧 글로벌 표준화의 완성을 의미하지는 않으나, 이 흐름은 한국 정책의 설계 변수에 포함되어야 한다.

(1) 능동적 방어: 국가 핵심 AI 인프라에 대한 상시적이고 표준화된 레드티밍 체제 도입

단발적 모의 해킹이 아니라 ISO/IEC 42119-7 초안에 부합하는 감사 가능한 평가 방법론을 기반으로, 영국 AISI가 보여준 '보충적 감독 계층' 모델과 싱가포르 AI Verify Foundation이 보여준 산업·정부·국제 표준 기구를 잇는 다층 평가 체제를 한국적 맥락에서 결합한 평가 인프라를 구축해야 한다. 이때 평가의 대상 범위는 Ⅲ장 3절에서 제시한 바와 같이 동적 AIBOM이 가시화한 종속성 그래프 전체로 정의되어야 하며, 이것이 AIBOM 의무화와 레드티밍 제도화가 별개 정책이 아니라 한 패키지로 추진되어야 하는 이유이다.

(2) 사이버 억지력: 고급 취약점 분석·대응 전문인력 양성으로의 인재 정책 전환

기존의 BoB(Best of the Best) 등 최고급 보안 인재 양성 패러다임을 LLM 기반 취약점 발굴과 검증, 패치 자동화에 특화된 인재 육성 체계로 신속히 전환해야 한다. 주의할 점은 이 축의 목표가 (1)의 능동적 방어와 같지 않다는 것이다. (1)이 자기 인프라의 결함을 줄이는 활동이라면, 인재 양성의 전략적 효과는 "한국을 표적으로 삼는 비용이 기대 이익을 초과한다"는 신호를 적대 세력에게 보내는 데 있다. 즉 능동적 방어 역량의 인적 토대가 충분히 두터워질 때, 그 역량은 방어를 넘어 실질적 사이버 억지력(Cyber Deterrence)으로 전환된다. (1)의 상시 레드티밍 체제는 이 인력이 실전 역량을 축적하는 훈련장이자 경력 경로로 기능하므로, 두 축은 제도 설계 단계에서부터 연동되어야 한다.

(3) 환류·국제화: 평가 결과의 AIBOM 환류와 국제 상호인정

위 평가 결과가 Ⅲ장 3절에서 제시한 환류 구조에 따라 AI-VEX 상태 정보로 AIBOM에 기록되어 보안 관제(SOC)의 경고 피로를 줄이고 차기 평가의 우선순위를 결정한다. 제도화 관점에서 남는 과제는 이 환류를 국가 단위에서 표준화하는 것과, 그 결과를 글로벌 평가 기관(영국 AISI, 싱가포르 AISI 등)과의 평가 결과 상호인정(mutual recognition)으로 연결하는 것이다. 한국의 평가 결과가 주요국 인프라에서 재평가 없이 수용될 때, 한국은 추격자의 위치에서 벗어나 글로벌 보안 조정자로 자리매김할 수 있다.

2. 국가 AI 조달·활용 프레임워크 정립: 가칭 'Safe Pass Gateway'

설계 단계부터 안전이 보증되고 적의 행동을 억제할 수 있는 AI만이 글로벌 무대에서 진흥될 자격을 얻는다. 대한민국의 AI 전략은 더 이상 '규제냐 진흥이냐'라는 이분법으로 환원될 수 없다. 앞 절에서 제안한 차등화된 AIBOM 설계, Compliance-as-Code 생태계, 능동적 사이버 방어 인프라는 모두 그 이분법을 넘어서기 위한 비대칭 전력에 해당한다. G7과 미국·EU, 영국·싱가포르의 정책 흐름은 이미 이 명제 위에서 움직이고 있으며, 한국이 그 흐름에 정합하는 자기 모델을 어떤 속도로 설계해 내느냐가 향후 AI 시대의 국가 사이버안보 위상을 결정할 것이다.

위의 정책 제언이 분산적으로 시행될 경우 에이전틱 AI 공급망의 본질적 특성상 실효성은 제한적이다. 규제 중심과 완전 자율의 양극단 사이에서 '조건부 신뢰 기반 활용(Conditional Trust-based Utilization)'이라는 현실적 접근이 부상하고 있으며, 이는 검증 가능한 신뢰

증거가 충족되는 조건 하에서만 AI 구성요소의 활용을 허가하는 동적 거버넌스 모델이다. 이를 실효성 있게 구현하려면 ▲보안성 ▲데이터 주권 ▲모델 투명성 ▲윤리 기준 준수의 네 가지 축을 종합적으로 평가하는 국가 차원의 표준화된 체계가 필요하며, 본고는 그 구체적 실현 방안으로서 국가 차원의 AIBOM 사전 검증 인프라, 가칭 '안전 통행 관문(Safe Pass Gateway)'의 구축을 제언한다(최윤성 2026).

그 실행의 기본 방향은 다음과 같이 설정할 수 있다. 첫째, 거버넌스 측면에서는 기존 SW 공급망 보안 협력체계를 AI 영역으로 확장하는 방식이 신규 기구 설립보다 현실적이며, AI안전연구소가 본 장 1절 3)에서 제언한 평가·환류 기능의 기술 주체로 결합될 수 있다. 둘째, 기존 사업과의 연계 측면에서는 'SW 공급망 보안 가이드라인'과 판교 테스트베드의 SBOM 실증 인프라를 AIBOM 실증으로 확장하고, 2027년 예정된 공공 조달 SBOM 제도에 AI 시스템 트랙을 병행 설계함으로써 별도 제도의 난립을 피할 수 있다. 셋째, 단계적 시행 측면에서는 본 장 1절 1)의 차등화 원칙에 따라 1급 자산(국방·의료·핵심 인프라)의 조달부터 적용을 개시하고, G7 최소 요소와의 필드 수준 정합성 검증을 거쳐 일반 공공 조달로 확대하는 경로가 EU 디지털 옴니버스가 확보한 준비 기간과도 시간적으로 정합한다. 다만 구체적 데이터 모델, 연동 인터페이스, 운영 세칙은 별도의 기술 표준화·실증 연구의 과제로 남긴다.

이러한 프레임워크가 실현되어야 서론에서 지적한 정책적 공백(Policy Gap) 즉, 신뢰 체인 전반을 실시간으로 검증하는 동적 인프라의 부재가 메워진다. 더 나아가 한국은 규제와 진흥, 통제와 활용 사이의 이분법을 '조건부 신뢰'라는 제3의 길로 돌파함으로써, 단순한 국제 규범 수범자가 아니라 동맹의 사이버 회복력에 실질적으로 기

여하는 '대체 불가능한 안보 파트너'의 위상을 기술적으로 뒷받침할 수 있게 된다.

〈참고문헌〉

최윤성. 2022. “미국의 소프트웨어 공급망 보안 정책 동향: SBOM 사례를 중심으로.” 『정보보호학회지』 32권 5호.

최윤성. 2024. “디지털 공급망 보안 위험과 관리체계 시사점: 미국 EO-14028의 SW 조달 개선 흐름 연구.” 『사이버안보연구』 1집 2호.

최윤성. 2026. “에이전틱 인공지능 공급망 보호를 위한 동적 자재명세서 기반 사전 검증 시스템 및 그 방법.” 대한민국 특허출원 제10-2026-0107657호 (2026년 6월 12일 출원).

Anthropic Frontier Red Team. 2026. “500 Validated High-Severity Vulnerabilities Found by Claude Opus 4.6.” Anthropic, March.

Aqua Security. 2026. “Trivy Supply Chain Attack: Update on Ongoing Investigation.” Aqua Security, March.

Arctic Wolf. 2026. “TeamPCP Supply Chain Attack Campaign Targets Trivy, Checkmarx (KICS), and LiteLLM.” Arctic Wolf, March.

BSI·ACN (G7 Cybersecurity Working Group). 2025. A Shared G7 Vision on Software Bill of Materials for AI. Endorsed in Ottawa, May 12–13.

CAC (Cyberspace Administration of China). 2023. Interim Measures for the Management of Generative Artificial Intelligence Services. Effective August 15.

CNCF (Cloud Native Computing Foundation). 2024. in-toto Specification v1.0. Linux Foundation.

Collins. 2025. “Word of the Year 2025: Vibe Coding.” Collins Dictionary.

Council of the EU. 2026. “Artificial Intelligence: Council and Parliament Agree to Simplify and Streamline Rules.” Press release (May 7).

Endor Labs. 2026. “TeamPCP Strikes Again: telnyx Compromised Three Days After LiteLLM.” Endor Labs, March.

European Commission. 2024. Regulation (EU) 2024/1689 (AI Act). Official

Journal of the European Union, August.

European Commission. 2025. Proposal for a Regulation Simplifying the Implementation of the AI Act (Digital Omnibus on AI). European Commission, November 19.

G7 Cybersecurity Working Group. 2026. Software Bill of Materials for AI: Minimum Elements. Joint publication by Germany BSI, Italy ACN, France ANSSI, Canada CSE, US CISA, UK NCSC, Japan NCO, in collaboration with the EU Commission. France G7 Presidency (May).

Gartner. 2024. "Gartner Unveils Top Predictions for IT Organizations and Users in 2025 and Beyond." Press release (October 22).

Gartner. 2026. "5 Steps to Secure Your AI Supply Chain." Gartner Research note ID G00840789 (March 24), by Angela Zhao, Deepti Gopal, et al.

GitGuardian. 2026. "The Team PCP Snowball Effect: A Quantitative Analysis." GitGuardian, March.

ISO/IEC. 2026. AWI TS 42119-7: Artificial Intelligence — Testing of AI — Part 7: Red Teaming. Working draft.

Karpathy, Andrej. 2025. "There's a new kind of coding I call 'vibe coding'…" X (formerly Twitter) post, February 2. https://x.com/karpathy/status/1886192184808149383.

Linux Foundation. 2024. SPDX Specification 3.0. Linux Foundation Project.

Mandiant. 2026. M-Trends 2026: The Rise of Agentic Swarms and Sub-minute Exploitation. Mandiant Threat Intelligence.

Manifest Cyber. 2026. The Transparency Gap & AI Security in the FY26 NDAA. Manifest Cyber.

METI (Japan), CISA, BSI. 2025. Shared Vision on Software Bill of Materials in Cybersecurity. Joint publication, September 4.

METI (Ministry of Economy, Trade and Industry, Japan). 2025. AI Promotion Act. Passed by the Diet on May 28.

METI (Ministry of Economy, Trade and Industry, Japan). 2026. Guidelines on the Roles Expected of Cyber Infrastructure Providers. METI·NCO, March 31.

Microsoft Security Blog. 2026. "Detecting, Investigating, and Defending against the Trivy Supply Chain Compromise." Microsoft, March.

New York Times. 2026. "White House Considers Vetting AI Models Before Release." New York Times (May 4).

Ng, Andrew. 2025. "The term 'vibe coding' is being misused." X (formerly Twitter) post, April.

OMB (U.S. Office of Management and Budget). 2025. Memorandum M-26-04: Increasing Public Trust in Artificial Intelligence Through Unbiased AI Principles. White House (December 11).

OpenSSF. 2024. OSV-Scanner: Vulnerability Scanner for Open Source Software. Open Source Security Foundation.

OWASP. 2024. OWASP Dependency-Check & CycloneDX Specification. Open Worldwide Application Security Project.

Ptacek, Thomas. 2026. "Vulnerability Research Is Cooked." Sockpuppet (Quarrelsome) (March 30). https://sockpuppet.org/blog/2026/03/30/vulnerability-research-is-cooked/.

ReversingLabs. 2025a. "Malicious ML Models Discovered on Hugging Face Platform: The nullifAI Campaign." ReversingLabs, February.

ReversingLabs. 2025b. "Shai-Hulud npm Supply Chain Attack: What You Need to Know." ReversingLabs Blog, September.

Singapore IMDA and AI Verify Foundation. 2024. Model AI Governance Framework for Generative AI. Infocomm Media Development Authority.

Socket.dev. 2026. "Trivy Under Attack Again: Widespread GitHub Actions Tag Compromise." Socket.dev, March.

StepSecurity. 2026. "TeamPCP Plants WAV Steganography Credential Stealer in telnyx PyPI Package." StepSecurity, March.

Sutton, Richard. 2019. "The Bitter Lesson." Incomplete Ideas. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

TC260 (National Information Security Standardization Technical Committee, China). 2025. AI Safety Standards System V1.0 (Draft) and three national standards on generative AI data and security. CAC, effective November 1.

Trend Micro. 2026a. "Your AI Gateway Was a Backdoor: Inside the LiteLLM Supply Chain Compromise." Trend Micro, March 26.

Trend Micro. 2026b. "TeamPCP's Telnyx Attack Marks a Shift in Tactics Beyond LiteLLM." Trend Micro, March.

UK AISI (AI Security Institute). 2025. Frontier AI Trends Report 2025. UK Department for Science, Innovation and Technology.

Unit 42 (Palo Alto Networks). 2026a. "Namespace Hijacking on Hugging Face: Pivoting into Vertex AI and Azure AI Foundry." Palo Alto Networks.

Unit 42 (Palo Alto Networks). 2026b. "Monitoring npm Supply Chain Attacks: The Mini Shai-Hulud Campaign." Palo Alto Networks, May.

U.S. CISA. 2022. Vulnerability Exploitability eXchange (VEX) — Use Cases. Cybersecurity and Infrastructure Security Agency.

U.S. Congress. 2025. National Defense Authorization Act for Fiscal Year 2026 (FY26 NDAA), Sections 1513 and 1533. Public Law signed December 2025.

Willison, Simon. 2025. "Not all AI-assisted programming is vibe coding (but vibe coding rocks)." simonwillison.net, March 19. https://simonwillison.net/2025/Mar/19/vibe-coding/.

Agentic AI Supply Chain Threats and National Response Strategy

: A Proposal for an Active Cyber Defense System Based on Dynamic AIBOM

Yunseong Choi | Korea University

A series of supply chain incidents in the artificial intelligence (AI) ecosystem between 2025 and 2026, together with the empirically demonstrated capacity of LLM agents to mass-produce zero-day vulnerabilities, reveals that the structural asymmetry between human defenders and machine-speed attackers has already become reality. In response, the AI Bill of Materials (AIBOM) norms now rapidly crystallising across the United States, the European Union, and the G7 are elevating AIBOM from an industry self-regulatory recommendation to a de facto 'export barrier' determining market access and, more broadly, to a component of cyber weapon systems.

However, a static SBOM produced only once at build time cannot adequately address the constantly evolving agent environment and the invisible dependencies of the 'vibe coding' era; an integrated trust specification spanning the build, deployment, and runtime stages, together with active evaluation and verification infrastructure, is equally required. Building on this understanding, the author proposes three policy directions for Korea to move beyond the position of a mere norm-taker and instead establish itself as an Indispensable Partner that supplies cyber deterrence and defensive integrity: a tiered AIBOM design calibrated to asset sensitivity and organisational maturity, a Compliance-as-Code ecosystem, and the institutionalisation of an active cyber defense infrastructure.

Keyword Agentic AI, Supply Chain Security, Dynamic AIBOM, Active Cyber Defense, Indispensable Partner