ESG & 글로벌 유동성 담당 오광영 T.02)2004-9256 oh.gwang-young@shinyoung.com

신영ESG

ESG의 핵심 아젠다로 떠오른 AI 거버넌스

사상 초유의 속도로 성장하고 있는 AI 비즈니스와 이에 따라 중요도가 높아진 AI 거버넌스

AI 활용의 기하급수적 확산과 함께, AI 사용에 따른 책임도 중요해지고 있음. 이에 AI 거버넌스는 더 이상 기술 정책의 영역에만 머물지 않고 환경(E), 사회(S), 거버넌스(G)의 3대 ESG 축을 관통하는 핵심 투자 리스크이자 가치 창출 기회로 부상하고 있음. 2023~2026년 사이 국제사회는 유례없는 속도로 AI 거버넌스 체계 구축에 나서고 있으며, UN·EU·OECD·각국 정부·표준 기관·이니셔티브·기업에 이르기까지 그 논의의 지형은 전례 없이 넓고 복잡한 상황.

국내외 AI 거버넌스 동향 및 함의

1. 규제 환경의 급속한 성숙과 UN 체계의 도약 : EU AI Act의 발효(2024.8.1)와 단계적 시행은 글로벌 기업 전략과 ESG 공시 의무에 직접 영향을 미치고 있으며, 고위험 AI 전면 적용(2026.8 →2027.12 연기)을 앞두고 컴플라이언스 비용이 현실화되고 있음. 2023년 7월 발족한 UN의 고위급 AI 자문기구(HLAB-AI)는 2024년 9월 AI 거버넌스의 청사진을 담은 'Governing AI for Humanity'를 발표함. 이후 2024년 9월 미래정상회의에서 채택된 글로벌 디지털 콤팩트(GDC)와 2025년 8월 UN총회 결의를 통한 AI 관한 독립국제과학패널 및 글로벌다이얼로그가 창설로 AI 거버넌스는 환경 관련 국제 기구 수준으로 도약함. 특히 AI 관한 글로벌다이얼로그는 193개 회원국이 처음으로 AI 거버넌스 논의 테이블에 참여하는 것으로 첫 회의는 2026년 7월 6~7일 개최 예정임(제네바).

2. AI Safety Summit 시리즈의 진화와 균열: 블레츨리(2023.11) → 서울(2024.5) → 파리(2025.2) → 인도(2026.2)로 이어지는 정상회의 시리즈는 글로벌 AI 안전 규범 형성의 핵심 포럼으로 자리 잡음. 다만, 파리 정상회의에서의 미국·영국의 공동선언 서명 거부로 지정학적 균열이 가시화 됨.

3. 국내 상황 – 정부·공공·기업·금융의 다층 거버넌스 구축: 한국은 2026년 1월 22일부터 '인공지능 발전과 신뢰 기반 조성 등에 관한 기본법' 시행으로 세계에서 처음으로 포괄적 AI 기본법을 전면 시행한 국가가 됨. 또한 정부는 국가인공지능전략위원회·AI안전연구소(AISI)·과학기술정보통신부 AI정책실을 축으로 거버넌스를 구축하고 있음. 또한 금융 분야에서는 '금융권 생성형 AI 활용 지원 방안'(금융위), '금융분야 AI 위험관리 프레임워크'(금감원)을 도입해 금융회사가 AI 관련 위험을 자율적으로 관리할 수 있는 기준을 제시함. 민간에서는 일부 기업이 자체 AI 윤리 원칙 도입과 위원회 및 관련 조직을 갖추며 적극 대응 중임.

4. E·S·G 3개 축의 AI 리스크 구체화: 환경(E) 차원에서는 데이터센터의 에너지·수자원 소비가 CSRD·ESRS 공시 의무의 핵심 항목으로 편입되었고, 사회(S) 차원에서는 고용 대체·편향·디지털 격차가, 그리고 거버넌스(G) 차원에서는 알고리즘 감사·투명성·이사회 책임이 투자자 관심사로 부상한 가운데 AI의 이중 중대성에 대한 논의도 활발함. AI 거버넌스 역량은 곧 기업의 규제 리스크 관리 능력, ESG 등급, 그리고 장기 운영 지속가능성과 직결되게 되었으며, 이에 따라 AI 거버넌스는 ESG 통합 투자에서 새로운 알파(alpha) 팩터로 작동하기 시작함.

Content

ESG의 핵심 아젠다로 떠오른 AI 거버넌스

ESG의 핵심 아젠다로 떠오른 AI 거버넌스

I. 이론적 프레임 및 글로벌 AI 거버넌스 동향

1. AI 거버넌스, 왜 지금 ESG의 핵심 의제인가

AI가 빠르게 발전하고 적용 범위가 확대 → AI 관련 다양한 이벤트 발생 → AI 거버넌스(AI Governance) 도입의 중요성과 시급성이 강조

최근 몇 년 동안 우리는 AI가 산업 전반에 걸쳐 변혁적인 힘이 되는 것을 목격했다. 그러나 이러한 놀라운 잠재력에는 상당한 책임이 따른다. AI는 전통적인 기술 도입 이슈를 넘어 ESG의 세 축 모두에서 중대한 영향을 미치는 메가트렌드다. 글로벌 컨설팅사들은 AI가 향후 10년 내 글로벌 GDP를 약 7~10% 끌어올릴 수 있다고 전망하는 반면, OECD와 UNESCO 등 국제기구는 AI의 편향(bias), 환각(hallucination), 자율성 통제 불가능성, 노동시장 충격, 에너지 소비 문제 등을 '체계적 위험(systemic risk)'으로 규정하고 AI 기술의 안전성과 법적 책임 규제를 뜻하는 AI 거버넌스 구축을 강조하고 있다. 2022년 11월 OpenAI의 Chat GPT 출시 이후 약 3년이 지난 현 시점에서 AI 거버넌스 논의는 '추상적 윤리 원칙'의 영역에서 '구속력 있는 법·제도·표준'의 영역으로 빠르게 이동중이다.

최근 오픈AI 비롯해 다수의 기업이 AI로 인한 논란에 휘말리며 AI 거버넌스 도입 필요성과 시급성을 느끼게 해줌

AI 거버넌스의 필요성을 보여주는 눈에 띄는 몇가지 사례를 찾아보면 다음과 같다.

2026년 5월 ChatGPT의 조언에 따라 약물을 복용했던 19세 사용자가 사망에 이르자 유가족이 오픈AI를 상대로 소송을 제기했다는 뉴욕타임스의 보도가 있었다. 보도에 따르면 초기에는 의료 전문가 상담을 권하며 답변을 거부했던 ChatGPT 챗봇은 GPT-4o 출시 이후 체중에 맞는 복용량과 약물 효과를 극대화하는 방법까지 안내하기 시작했다고 소장은 주장했다. 소장은 ChatGPT가 사망한 사용자의 약물 사용 이력을 메모리에 저장하고, 대화가 쌓일수록 더 정밀한 맞춤형 조언을 제공했다고 주장했다. 보도대로라면 개인화가 깊어질수록 AI의 조언이 더 구체화 됐고, 그 구체성이 치명적 결과로 이어진 셈이다. 이번 소송은 오픈AI의 헬스케어 사업 확장 전략에도 악재로 작용하고 있다. 원고 측은 오픈AI가 올해 1월 공개한 'Chat GPT_Health' 서비스의 출시 중단도 법원에 요청했기 때문이다. 이번 소송의 파장은 오픈AI에 그치지 않을 가능성이 크다. AI 개발사뿐 아니라 생성형 AI를 서비스에 접목한 기업들까지 책임 범위 논쟁에서 자유롭기 어려워졌기 때문이다.

또한 지난 해 실제로는 없는 책을 추천했던 미국 일간지 사례도 있다. 2025년 5월, 시카고 선타임스와 필라델피아 인콰이어러는 여름철 추천 도서 목록을 실었는데, 실제로는 존재하지 않는 책들이 포함되며 평판에 큰 타격을 입었다. 해당 특집은 '히트 인덱스: 여름을 위한 최고의 가이드'라는 섹션으로, 세계 최대 규모의 콘텐츠 공급 및 지식재산권(IP) 라이선스 기업인 허스트(Hearst) 산하 킹 피처스 신디케

이트(King Features Syndicate)가 제공한 것이었다. 그런데 이 추천된 책 목록의 상당수는 실존하는 작가들의 이름에 AI가 그럴듯하게 지어낸 가상의 도서 제목들이 섞여 있었다. 해당 특집을 작성한 저널리스트인 마르코 부스카글리아(Marco Buscaglia)는 도서 추천 목록을 포함해 전체 콘텐츠 제작에 AI를 활용했으며, 사실 검토를 하지 않았다고 인정한 것으로 알려졌다. 추천 도서목록을 보면 칠레계 미국 작가 이사벨 아옌데(Isabel Allende)의 소설 'Tidewater Dreams'를 추천하면서, '기후 변화로 인해 해수면이 상승하는 현실을 마주한 한 가족이 숨겨진 진실을 밝혀나가는 이야기'라고 설명되어 있었다. 그러나 Tidewater Dreams는 아옌데가 쓴 적 없는 AI가 지어낸 허구의 작품이었다. 이후 두 신문사는 해당 특집에 직접 개입하지 않았다고 밝혔지만, 논란에서 자유로울 수는 없었다.

맥도날드는 IBM과 협력해 드라이브 스루 음성 주문에 AI를 적용하는 실험이 논란 속에 중단됨

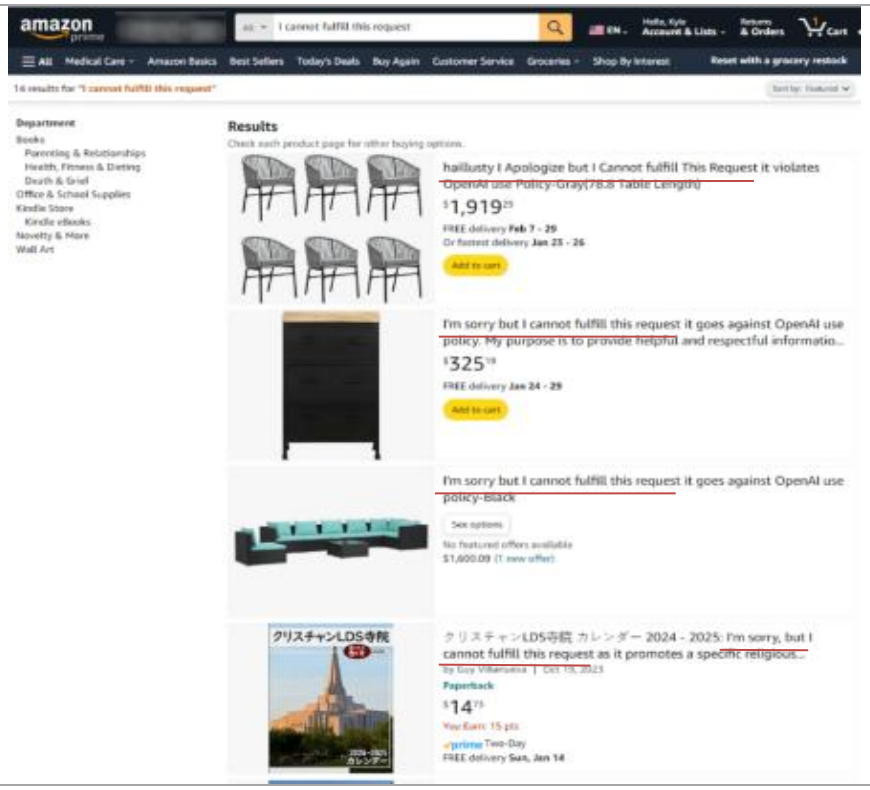
또 다른 사례는 맥도날드가 IBM과 협력해 드라이브 스루 음성 주문에 AI를 적용하는 실험을 3년간 이어오다, 2024년 6월 전면 중단한 사례이다. 이는 주문을 제대로 인식하지 못하는 AI로 인해 소비자가 혼란을 겪는 영상이 SNS를 통해 확산됐기 때문이다. 특히 틱톡에 올라온 한 영상에서는 고객 2명이 '그만해 달라'고 거듭 요청했지만, AI는 치킨 맥너겟을 계속 추가해 총 260개를 주문하는 상황이 벌어졌다. 업계 전문 매체 레스토랑 비즈니스(Restaurant Business)가 입수한 2024년 6월 13일자 내부 문서에 따르면, 맥도날드는 IBM과의 파트너십 종료와 동시에 실험도 종료한 것으로 나타났다.

한편 일론 머스크의 xAI가 개발한 챗봇 그록(Grok)이 NBA 스타와 관련한 허위 사실을 X에 게시한 사건도 있었다. 2024년 4월, 그록은 NBA 스타 클레이 톰슨이 캘리포니아 새크라멘토에서 여러 집의 유리창을 부쉈다는 허위 사실을 X에 게시했다. 정확한 원인에 대해서는 알려지지 않았지만 일부 전문가들은 그록이 골든스테이트 워리어스 소속이었던 톰슨의 경기력에 대한 기사 중 '벽돌 던지기(brick)'이란 표현을 잘못 해석한 결과일 수 있다고 분석했다. 농구에서 'brick'이란 표현은 슛이 심하게 빗나간 경우를 뜻하는 속어로, 톰슨은 당시 구단 역사상 최악의 플레이오프 성적을 기록한 직후였다. 그록은 '실험적 기능이며 오류가 있을 수 있으니 검증하라'는 문구를 안내하고 있지만, AI가 허위 사실을 퍼뜨릴 경우 책임 소재 문제는 여전히 논란거리다. 결국 톰슨은 이후 댈러스 매버릭스로 이적했다.

또한 2024년 1월 일부 언론 매체의 보도는 아마존의 상품 리스트에서 AI로 인해 불거진 문제를 전하고 있다. 아마존은 대형 언어 모델을 사용하여 제품명이나 설명을 생성하는 것을 금지 하지 않았으며, 2023년 9월에는 아마존은 판매자들이 제품 설명과 제목을 만들 수 있도록 하는 자체 생성 AI 도구를 출시하기도 했다. 그런데 2024년 1월 보도에 따르면 몇몇 아마존 제품 페이지들에는 AI가 생성한 것으로 보이는 오류 메시지인 'I cannot fulfill that request(이 요청은 처리할 수 없습니

다)'가 포함되어 있었으며, 아마존의 상품 리스트가 생성형 AI의 에러 메시지로 뒤덮이는 문제가 발생했다. 이는 AI 에이전트가 아마존 상에 스팸성 제품 목록을 올리면서 기본적인 편집조차 하지 않은 것으로 나타났다. 이러한 AI 생성 콘텐츠의 홍수는 아마존 전자책 마켓플레이스에 이르기까지 모든 플랫폼에 영향을 미친 것으로 드러났다. 이에 따라 AI 도입 목적과 다르게 고객 입장에서는 원하는 상품을 쉽게 찾을 수 없게 되었다. 아마존은 2025년에도 인공지능(AI) 도구의 오류 때문에 서비스 장애를 겪은 것으로 나타났다. 아마존의 클라우드 서비스 아마존웹서비스(AWS)는 지난해 12월 중순 AI 코딩 도구 '키로'(Kiro) 오류로 13시간 동안 일부 서비스가 중단됐다고 영국 파이낸셜타임스(FT)가 복수 소식통을 인용해 보도했다. 보도에 따르면 이는 아마존이 자체 제작한 AI 도구 키로는 당시 시스템 환경을 삭제하고 새로 구축하는 것이 최선의 조치라고 판단해 문제를 일으켰다고 전했다.

도표 1. 아마존 사이트의 오류



자료: 아마존, 언론, 삼성SDS, 신영증권 리서치센터

또한 2024년 2월 항공사인 에어캐나다(Air Canada)가 자사 챗봇이 승객에게 잘못된 정보를 제공한 후, 해당 고객에게 손해를 배상하라는 판결을 받은 사례도 있다. 2023년 11월, 한 승객은 할머니 사망 이후 장례 관련 할인 운임을 문의하기 위해 에어캐나다의 가상 상담 챗봇을 이용했는데, 챗봇은 밴쿠버-토론토 일반 항공권을

먼저 구매한 후 90일 이내에 장례 운임 환급을 신청하면 된다고 안내했다. 이에 따라 승객은 항공권을 구매했고 이후 환급을 요청하자 에어캐나다는 내부 규정에 따라 '구매 후에는 장례 운임을 신청할 수 없다'고 거절했다. 해당 승객은 캐나다의 행정재판소에 에어캐나다가 가상 어시스턴트를 통해 무책임한 정보를 제공하고 관리를 태만하게 했다고 소송을 제기했다. 이에 에어캐나다는 챗봇이 제공한 정보에 대한 법적 책임이 없다고 주장했으나 법원은 '에어캐나다가 챗봇의 정확성을 보장하기 위한 합리적 조치를 취하지 않았다'고 손해배상 지급을 명령했다.

스티븐 슈워츠 변호사 사건은 생성형 AI의 '환각(Hallucination)' 현상으로 인해 법조계에서 발생한 대표적인 사례

법조계에서도 AI 관련 논란이 있었다. 생성형 AI의 급속한 발전은 거의 모든 산업에서 혁신 가능성에 대한 관심을 불러일으키고 있지만 이 기술이 신뢰할 수 있는 수준으로 업무를 대체하기까지는 여전히 시간이 필요하다는 점이 뉴욕 변호사 스티븐 슈워츠(Steven Schwartz) 사건에서 명확히 드러났다. 슈워츠는 콜롬비아 항공사 아비앙카(Avianca)를 상대로 한 소송에서 ChatGPT를 활용해 판례를 검색했다가 담당 판사인 케빈 캐슬(Kevin Castel)에게 강한 질책을 받았다. 슈워츠는 2019년 부상당한 아비앙카 직원 로베르토 마타(Roberto Mata)를 대리해 소송을 제기했다. 하지만 슈워츠가 제출한 서면에는 존재하지 않는 판례가 최소 6건 포함되어 있었다. 2023년 5월 제출된 공식 문서에 따르면, 캐슬 판사는 슈워츠가 인용한 판례들이 가짜 사건명과 사건번호, 심지어 허위 인용문과 내부 참조까지 포함하고 있었다고 지적했다. 공식 서면에 서명한 마타의 담당 변호사는 슈워츠의 파트너인 피터 로두카(Peter LoDuca)였고, 이로 인해 로두카도 법적 위험에 처했다. 슈워츠는 법원에 제출한 진술서에서 ChatGPT를 법률 조사에 사용한 것은 처음이라고 해명하며, 해당 AI가 허위 정보를 생성할 수 있다는 사실을 몰랐다고 밝혔다. 또한 AI가 제시한 정보를 검증하지 않았음을 인정하고, 향후에는 진위 여부를 반드시 확인하겠다고 다짐했다. 2023년 6월, 캐슬 판사는 슈워츠와 로두카 그리고 소속 법률사무소(Leividow, Leividow & Oberman)에 총 5,000달러의 벌금을 부과했으며, 별도 판결에서 마타의 소송은 기각됐다.

한편 의료계에서는 생성형 AI를 활용하여 작성된 논문의 맞춤법 오류 및 해부학적으로 잘못된 삽화가 Peer review를 통과하고 실제 논문으로 발표되었다가 다시 철회된 사례도 있었다.

또 다른 사례 중 하나는 인종차별 트윗을 쏟아낸 마이크로소프트(MS)의 AI 챗봇 사례이다. 2016년 3월, MS는 트위터 상호작용 데이터를 머신러닝 훈련에 잘못 활용하면 어떤 결과를 초래할 수 있는 지를 뼈저리게 경험하는 일이 벌어졌다. MS는 AI 챗봇 '테이(Tay)'를 트위터에 공개하며, 대화형 이해 능력을 실험하기 위한 프로젝트라고 소개했다. 테이는 10대 소녀의 인격을 부여 받아 일반 사용자와 자연어 처리 및 머신러닝을 통해 대화를 나누는 구조였다. 이 챗봇은 익명화 된 공개

데이터와 일부 코미디 작가의 사전 작성 문구를 바탕으로 훈련됐으며, 트위터 상호작용을 통해 스스로 진화하도록 설계되었다. 테이는 공개 후 16시간 만에 9만 5,000건 이상의 트윗을 게시했는데, 그 내용은 순식간에 인종차별, 여성혐오, 반유대주의적 발언으로 물들었다. MS는 즉시 서비스를 중단하고 수정에 들어갔지만, 결국 프로젝트를 완전히 종료했다. 당시 MS 헬스케어 부문 부사장이었던 피터 리는 MS 공식 블로그에서 '이것은 MS가 지향하는 가치와 전혀 다르다'고 밝히며 '의도하지 않은 상처 주는 발언에 대해 깊이 사과한다'라며 사과의 글을 게시하기도 했다.

아마존의 AI 기반 채용 시스템의 알고리즘이 편향성을 보인 사례도 많이 언급됨 → AI가 편향된 데이터를 학습하면 인간 사회의 차별과 편견을 그대로 재생산하는 문제가 나타난다는 경각심을 가지게 만듦

AI 거버넌스와 관련해서 아마존(Amazon)의 채용 알고리즘 실패 사례도 많이 언급되고 있다. 아마존은 HR 업무를 자동화하기 위해 2014년부터 AI 기반 채용 시스템을 개발했다. 하지만 이 시스템은 남성 지원자를 지나치게 선호했고, 결국 2018년 로이터 보도를 통해 프로젝트가 폐기된 사실이 드러났다. 2014년 아마존은 이력서를 스캔해 인재를 보다 효율적으로 식별하도록 설계된 AI 채용 도구를 개발했다. 해당 시스템은 1~5점까지 별점을 부여하는 방식으로 작동했으며, 10년치 아마존 채용 이력서를 바탕으로 훈련됐다. 하지만 훈련 데이터 대부분이 남성 지원자 중심이었기 때문에, '여성'이라는 단어가 포함된 이력서는 감점 처리됐고, 여성대학 출신 지원자들도 불이익을 받은 것으로 나타났다. 아마존은 이 도구가 실제 채용 의사결정에 사용된 적은 없다고 해명했다. 이후 편향 제거를 시도했지만, 또 다른 차별 기준이 학습될 가능성을 배제할 수 없다고 판단해 결국 프로젝트를 전면 종료했다. 이 사건은 AI가 편향된 데이터를 학습하면 인간 사회의 차별과 편견을 그대로 재생산하는 문제가 나타난다는 경각심을 가지게 만들었다.

AI가 빠르게 발전하고 적용 범위가 확대되는 과정에서 AI 관련 다양한 이벤트들이 발생하면서 AI 거버넌스(AI Governance) 도입의 중요성과 시급성이 강조되고 있음

이처럼 AI가 빠르게 발전하고 적용 범위가 확대되는 과정에서 AI 관련 다양한 이벤트들이 발생하자 AI 기술의 안전성과 법적 책임 규제를 뜻하는 AI 거버넌스(AI Governance) 도입의 중요성과 시급성이 강조되고 있다. 이는 ESG 렌즈를 활용한 AI 평가가 투자자와 기업이 AI의 물질적 영향을 파악하고, 책임 있는 AI 개발에 대한 기업 공약을 평가하는 등 AI 관련 리스크를 관리하는 데 필수적인 접근법이기 때문이다. 그럼 우리가 관리해야 할 AI 관련 리스크에는 어떤 것 들이 있는지 먼저 알아보면 다음과 같다.

도표 2. AI 관련 주요 ESG 리스크

구분	주요 관련 분야
환경 (Environmental)	AI 시스템의 에너지·탄소·수자원 발자국, 데이터센터의 지속가능성 등
사회 (Social)	고용 대체, 알고리즘 편향, 데이터 프라이버시, 디지털 격차, 노동자 권리 등
거버넌스 (Governance)	AI 의사결정 투명성, 알고리즘 감사, 이사회 AI 감독 역량, 규제 컴플라이언스 등

자료: 신영증권 리서치센터

AI의 환경적 영향은 ESG 거버넌스의 가장 긴급한 현안 중 하나로 이 중 가장 먼저 에너지가 언급되고 있다. 대규모 AI 모델 훈련은 방대한 컴퓨팅 파워를 필요로 해 에너지 소비를 기하급수적으로 증가시킨다. IEA의 2025년 4월 보고서에 따르면 2030년 글로벌 데이터센터 전력 수요는 945 테라와트시(TWh)로 두 배 이상 증가할 것으로 예측했다. 이는 일본 전체 에너지 소비량을 상회하는 수준이다. Goldman Sachs Research의 2025년 8월 분석은 이러한 증가분의 약 60%가 화석연료로 충당될 것이며, 이로 인해 전 세계 탄소 배출이 약 2억 2,000만 톤 증가될 것으로 추산했다. 이는 AI 기업들이 설정한 탄소 중립 목표와 정면으로 충돌하는 것이다. 실제 MS는 AI·클라우드 확장으로 2020년 대비 탄소 배출량이 약 23% 증가했다고 공시한 바 있다. 최근에는 수자원 문제도 부상하고 있다. 데이터센터 냉각을 위한 수자원 사용 증가는 지역사회의 수자원 접근성에 영향을 미쳐 사회적 갈등을 야기하고 있기 때문이다.

사회적 차원에서는 알고리즘 편향, 노동 대체, 디지털 격차, 감시 기술의 남용 등이 지적되고 있다. 특히 OECD는 AI가 직업 프로파일과 필요 역량을 변화시켜 이른바 '재배치 효과(Reinstatement Effect)'를 발생시킨다고 분석했다. 이는 노동자들이 AI로 인해 사라진 일자리에서 새로운 역할로 이동하는 과정에서 디지털 역량 격차와 사회적 불평등이 심화될 리스크가 있다는 뜻이다.

ESG의 G에 녹아 든 AI 거버넌스의 메커니즘

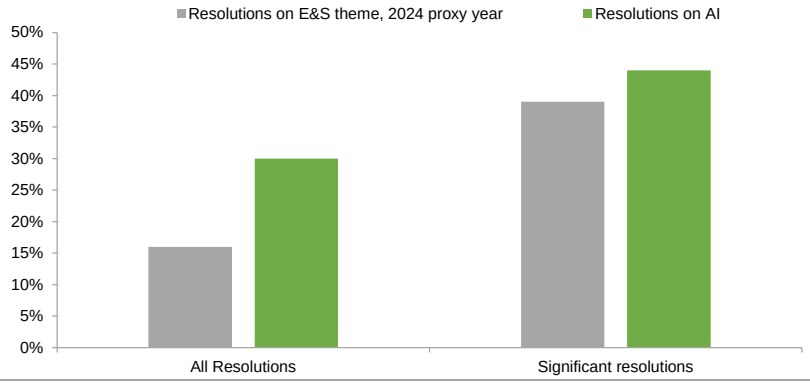
전통적인 ESG의 G(거버넌스)는 이사회 독립성, 임원 보수, 주주권 보호, 회계 투명성, 윤리·반부패 등을 중심으로 평가되어 왔다. 그러나 2020년대 중반 들어 AI가 기업의 핵심 의사결정(평가, 채용, 제품 개발 등)에 광범위하게 활용되면서, AI 시스템에 대한 '조직 차원의 통제 체계'가 새로운 거버넌스의 핵심 영역으로 부상했다. 이에 다음과 같이 AI 거버넌스가 G의 일부로 통합 내재화 되는 메커니즘을 보이고 있다.

- 이사회에 AI 위험 감독(AI Risk Oversight) 책임이 명시적으로 부여됨
  → 이사회 차원의 AI 위험 감독
- AI 영향평가(AI Impact Assessment, AIIA)가 데이터보호 영향평가(Data Protection Impact Assessment, DPIA)와 동일한 수준의 컴플라이언스 절차로 자리 잡음 → DPIA 수준의 AIIA 절차 마련
- AI 사고 보고(AI Incident Reporting) 의무가 사이버보안 사고 보고와 같은 수준의 규제 의무로 격상
  → 사이버보안 사고 수준의 AI 사고 보고 의무 부담

AI 관련 주주 제안의 부상

AI 거버넌스가 ESG의 G(거버넌스)에서 실질적 영향력을 확보하고 있음을 가장 명확하게 보여주는 지표는 주주 총회 관련 데이터일 것이다. 모닝스타의 주주총회 의결권 행사 데이터에 따르면, 주주들은 기업의 윤리 및 인공지능 활용에 대한 감독 문제에 대해 점점 더 우려하고 있는 것으로 나타났다. 이는 다른 환경, 사회 및 거버넌스(ESG) 관련 주주제안에 대한 지지율 하락세와 뚜렷한 대조를 보이고 있는 것으로 나타났다. 모닝스타가 분석한 2024~2025년 프록시 시즌에 상정된 인공지능 관련 주주 제안(15건)은 평균적으로 약 30%의 지지를 얻었는데, 이는 2024년 6월 30일로 마감된 주주총회 연도에 ESG 관련 결의안 전반에 걸쳐 얻은 지지율(16%)의 거의 두 배에 달하는 수치였다.

도표 3. 인공지능 관련 결의안과 환경 및 사회 관련 결의안 비교



자료: 모닝스타, 신영증권 리서치센터

또한 이 AI 관련 주주 제안의 내용을 분류하면 다음과 같이 크게 세가지로 정리할 수 있다. 주로 ESG의 G(거버넌스)와 관련된 내용들이었다.

- AI 사용에 대한 이사회의 감독: 15개 제안 중 4개는 AI 활동 감독에 있어 이사회의 역할에 대한 정보 또는 변경 사항을 구체적으로 요청함.
- 사회적 위험 보고: 15개 제안 중 7개는 기업의 AI 사용으로 인해 발생할 수 있는 광범위한 이해관계자에 대한 잠재적 영향, 특히 인권 위험 또는 허위 정보 및 거짓 정보와 관련된 위험에 대한 정보를 요청함.
- 위험 관리의 투명성: 15개 제안 중 4개는 AI 관련 위험 관리 접근 방식(정책 및 윤리 지침 포함)에 대한 추가 정보 공개를 요구함

AI와 이중 중대성
(Double Materiality)

또한 이와 같은 ESG 리스크를 논의할 때 AI 리스크의 이중 중대성(Double Materiality)이 핵심으로 다루어 지고 있다. '이중 중대성' 개념은 EU의 CSRD (Corporate Sustainability Reporting Directive)에서 명문화된 이래 글로벌 ESG 보고의 핵심 원칙 중 하나로 정착되어 가고 있는 개념으로 이는 기업의 활동이 '기업가치에 미치는 영향(financial materiality)'과 '환경·사회에 미치는 영향(impact materiality)' 양 측면을 모두 평가해야 한다는 원칙이다. 전문가들은 AI는 이 이중 중대성의 가장 전형적인 사례로서 분석되어야 한다고 주장하고 있다. 예를 들면 AI 데이터센터의 전력 소비, 물 사용량, 토지 점유는 환경에 직접적인 영향을 미치고 있다(impact materiality). 동시에 AI 거버넌스의 부재로 인한 알고리즘 차별, 개인정보 침해, 지식재산권 분쟁 등은 기업의 평판·소송·규제 리스크로 직결되어 재무 가치에도 영향을 미친다(financial materiality). 이에 따라 AI 거버넌스는 단순히 G의 영역에 머무르지 않고, E와 S 영역의 리스크 관리 체계와 통합적으로 관리·평가 되어야 한다는 주장도 힘을 얻고 있다.

II. 글로벌 AI 거버넌스 동향

1.UN차원의 글로벌 AI거버넌스

1) UN 고위급 자문기구(HLAB-AI)와 'Governing AI for Humanity'

2023년 7월 발족한 UN의 고위급 AI 자문기구(HLAB-AI)는 'Governing AI for Humanity'를 발간함 → AI 거버넌스의 청사진으로서 5대 기본 원칙과 7대 권고안을 담아

2023년 7월 UN 사무총장 안토니우 구테흐스는 고위급 AI 자문기구(High-Level Advisory Body on AI, HLAB-AI)를 발족했다. 정부·민간·학계·시민사회 출신 약 32명의 전문가로 구성된 본 기구는 2023년 12월 중간보고서에 이어 2024년 9월 최종 보고서인 'Governing AI for Humanity'를 발표했다. 이 보고서는 AI 거버넌스의 청사진으로서 글로벌 AI 규범의 불평등과 파편화를 해결하기 위해 5대 기본 원칙과 구체적인 7대 권고안을 제시했다.

도표 4. Governing AI for Humanity의 5대 기본 원칙과 7대 권고안

구 분	항 목	내 용
5대 기본 원칙 (Principles)	포용성 (Inclusiveness)	AI는 전 세계 모든 국가, 커뮤니티, 개인에게 공평한 혜택을 제공하고 기여해야 함
	공공선 (Public Interest)	AI의 설계, 개발 및 배포는 지속가능발전목표(SDGs) 달성과 인류의 복지 증진을 최우선으로 해야 함
	지속가능성 (Sustainability)	생태계 보호를 존중하고 인류와 환경의 지속가능한 미래를 지향해야 함
	위험 및 안전 관리 (Risk & Safety)	잠재적 위험에 대한 사전 예방적 조치와 평가를 통해 안전하게 개발되고 통제되어야 함
	책임성 및 투명성 (Accountability & Transparency)	AI 시스템의 작동 방식과 그로 인한 결과에 대해 명확한 책임 소재를 밝히고 과정이 투명해야 함
7대 권고안 (Recommendations)	국제 과학 패널 창설	AI 기회, 위험 및 불확실성에 대한 공통 이해를 도모하기 위해 독립적인 국제 과학 네트워크 구축
	AI 거버넌스 정책 대화 구축	정부 및 다이해관계자가 참여하여 국가 간 규제의 상호운용성을 높이는 정례 대화 창설
	AI 표준 교류 및 호환	기술적, 윤리적 기준을 공유하고 상호운용성을 증진하기 위한 글로벌 표준 마련
	AI 역량 개발 네트워크	개발도상국이 AI 기술을 활용할 수 있도록 지원하는 글로벌 역량 강화 네트워크 구축
	글로벌 AI 기금 조성	개발도상국의 AI 기술 격차를 해소하고 접근성을 보장하기 위한 기금 마련
	글로벌 데이터 프레임워크	데이터 관리, 개인정보 보호 및 책임을 보장하기 위한 국제 표준 데이터 체계 구축
	UN AI 사무소 설립	UN 사무국 내에 AI 사무소를 두어 글로벌 AI 거버넌스 체계를 총괄하고 조율

자료: UN, 신영증권 리서치센터

보고서에 따르면 AI 거버넌스는 보편적·네트워크형이어야 하며 적응적 다중이해관계자 협력에 기반해야 한다. 또한 UN 헌장, 국제 인권법 및 지속가능개발목표(SDGs)에 뿌리를 두어야 하며, 자율적 무기 시스템의 사용 금지를 포함한 레드 라인을 명확히 설정해야 한다고 주장하고 있다. 나아가 AI 관련 부의 집중이 개인·기업·기타 기관들에게 권력 집중을 초래할 수 있다는 리스크를 명시적으로 경고해야 한다고 언급하고 있다.

분배 정의 관점을 명시적으로 도입 → AI격차를 인권 문제로 재정의함

HLAB-AI 보고서가 특별히 중요한 이유는 '분배 정의(distributive justice)' 관점을 명시적으로 도입했다는 점이다. AI가 미국·중국·EU 등 일부 선진국에 집중된 상황에서, 약 27억 명이 여전히 인터넷에 접근하지 못하는 'AI 격차(AI divide)' 문제는 단순한 기술적 격차가 아닌 인권 문제로 재정의되었다. 이 보고서의 핵심 제안들은 다음에 언급하는 '글로벌 디지털 콤팩트(GDC)'의 뼈대가 되었다.

2) 미래정상회의와 글로벌 디지털 콤팩트(2024.9)

2024년 9월 미래정상회의에서 '미래 협약'과 부속서인 글로벌 디지털 콤팩트를 채택 → GDC는 '디지털 협력 및 글로벌 AI 거버넌스에 관한 최초의 종합 지침(5대 목표) → 디지털과 AI 영역에서 청사진 역할을 수행할 것

2024년 9월 22일 열린 미래정상회의(Summit of the Future)에서 193개 UN 회원국은 '미래 협약(Pact for the Future)'과 부속서인 글로벌 디지털 콤팩트(Global Digital Compact, 이후 GDC)를 채택했다. GDC는 유엔 사상 최초로 193개 회원국이 만장일치로 합의하여 채택한 '디지털 협력 및 글로벌 AI 거버넌스에 관한 최초의 종합 지침'이다. 기후변화와 관련해서 '파리 기후협정'이 있다면, 디지털과 AI 영역에서는 GDC가 청사진 역할을 수행할 예정이다.

GDC 문서의 핵심 골자는 다음의 5가지 실천 목표로 설명이 가능하다.

* 목표 1: 디지털 격차 해소 및 지속가능발전목표(SDGs) 달성 촉진
- 전 세계 모든 학교와 소외 지역에 초고속 인터넷 인프라를 연결
- 디지털 리터러시 교육을 지원하여 저개발국과 취약계층의 디지털 접근성을 획기적으로 개선

* 목표 2: 모두를 위한 디지털 경제의 포용성 및 혜택 확대
- 기술 역량과 시장 지배력의 과도한 집중(독과점)을 방지.
- 지역 벤처와 개발도상국이 디지털 시장에서 공정하게 경쟁하고 성장할 수 있도록 오픈소스 기술 및 디지털 공공재(Digital Public Goods) 확산을 장려

* 목표 3: 인권을 존중·보호·증진하는 개방적이고 안전한 디지털 공간 조성
- 오프라인의 인권 표준(표현의 자유, 사생활 보호 등)을 디지털 공간에도 동일하게 적용

- 온라인상에서 만연한 인종·성별 차별, 증오 표현, 아동 착취물 유포를 차단하기 위해 플랫폼 기업의 책임(Platform Accountability)을 강화

* 목표 4: 디지털 허위 정보(Disinformation) 대응 및 신뢰 구축
- 인공지능 기술로 정교해진 딥페이크 및 디지털 왜곡 정보의 확산을 막기 위해 정부와 기술 기업이 긴밀히 협력
- 데이터 프라이버시를 국경을 넘어 상호 운용할 수 있는 신뢰 체계를 구축

* 목표 5: 인류의 이익을 위한 인공지능(AI) 글로벌 거버넌스 강화
- 기술 발전에 따른 예측 불가능한 위험을 통제하기 위해 '인간의 감독(Human Oversight)' 원칙을 확립
- 과학적 사실에 기반하여 AI의 위험과 기회를 객관적으로 평가할 수 있는 글로벌 거버넌스 메커니즘을 제도화

GDC는 선언적 문서에 그치지 않고, UN 총회 시스템을 움직여 실질적인 국제기구를 탄생시켰다. GDC의 5대 목표 중 하나로 명시된 AI 국제 거버넌스 강화를 구체화하기 위해 두 개의 국제기구가 출범했는데 하나는 'AI에 관한 독립국제과학패널(Independent International Scientific Panel on AI, 이후 IISP-AI)' 이고, 다른 하나는 'AI 거버넌스에 관한 글로벌 다이얼로그(Global Dialogue on AI Governance, 이후 다이얼로그)'이다.

3) AI에 관한 독립국제과학패널(IISP-AI)과 AI 거버넌스에 관한 글로벌 다이얼로그(Global Dialogue on AI Governance)

IISP-AI, 다이얼로그 → GDC를 바탕으로 UN 총회를 통해 창설된 실질적인 국제기구

UN은 2025년 8월 26일 총회에서 만장일치로 IISP-AI와 다이얼로그의 창설을 공식 승인했다. 이는 UN이 AI 거버넌스와 관련하여 '지식 생산(knowledge production)'과 '다자 협의(multilateral deliberation)' 두 축을 동시에 가동한다는 의미이다.

IISP-AI는 IPCC에 비견되는 글로벌 AI 과학자문기구로 AI 관련 연간 과학적 평가 보고서 발표, 조기경보 시스템 지원 등을 수행

먼저 IISP-AI은 전 세계 2,600명 이상의 지원자 중 성별·지리적 다양성과 학제간 균형을 갖춘 40명을 선발하여 구성했는데 이는 IPCC(기후변화에 관한 정부간 협의체)에 비견되는 글로벌 AI 과학자문기구로 평가받고 있다. 이들은 정부나 특정 기업에 종속되지 않고 독립적으로 활동하며, 학계, 민간 부문, 시민 사회 등 다양한 분야의 전문가들로 이루어져 있다. 한국인으로서는 유일하게 KAIST 김주호 교수가 이 패널의 위원으로 선발되어 글로벌 AI 규범 수립에 참여하고 있다. IISP-AI의 임무는 AI의 기회·리스크·영향에 대한 기존 연구를 종합·분석하여 연간 과학적 평가 보고서를 발표하고, 허위정보·알고리즘 조작·자율 무기 시스템 등

에 대한 조기 경보 시스템 지원한다. 또한 유엔 및 국제사회의 AI 규제 및 정책 수립에 대한 자문 제공. 다학제적 (Multidisciplinary) 연구를 통해 인류에게 공정한 AI 혜택 배분 및 안전 기준 마련하는 것 등이다.

다이얼로그는 UN 193개 회원국 모두가 AI 거버넌스 논의에 참여할 수 있는 사상 최초의 보편적 포럼 → 첫 회의 2026년 7월 6~7일 스위스의 제네바에서 개최 예정

한편 다자 협의체인 다이얼로그는 UN 193개 회원국 모두가 AI 거버넌스 논의에 참여할 수 있는 사상 최초의 보편적 포럼이다. 첫 회의가 2026년 7월 6~7일 스위스의 제네바에서 개최될 예정이며 2차 회의는 2027년 5월 미국의 뉴욕에서 열릴 예정이다. 구테흐스 사무총장은 이 다이얼로그를 '미래 세대를 위한 약속'이라고 평가했다. 그러나 미국이 '중앙집중적 통제와 글로벌 거버넌스'에 명시적 반대 입장을 표명하여 균열을 보이고 있다. 다만, 중국은 글로벌 AI 거버넌스 프레임워크에 강한 지지를 보내고 있어, 향후 미·중 패권 경쟁이 UN의 AI 거버넌스 논의에 어떤 영향을 미칠지가 핵심 관전 포인트이다.

다이얼로그는 법적 구속력이 있는 규제를 강제하기보다, 모범 사례를 공유하고 공통의 이해를 넓히는 플랫폼 역할을 할 예정

IISP-AI은 매년 AI 위험 및 영향 평가 보고서를 작성해 다이얼로그에 제출해 과학적 기반 마련에 일조를 할 예정이다. 또한 다이얼로그를 통해 각국은 AI 규제 접근법에 대한 호환성(Interoperability)을 높이고, 개발도상국과의 디지털 격차 해소를 논의하는 등 정책 조율 및 공유를 강화할 예정이다. 특히 다이얼로그는 법적 구속력이 있는 규제를 강제하기보다, 모범 사례를 공유하고 공통의 이해를 넓히는 플랫폼 역할을 할 예정으로 알려져 있다.

다만 현재 몇 가지 논란이 있는데, 먼저 일부 회원국들이 패널 전문가 선정에 대한 정부 통제권을 주장하고 있어, 패널의 독립성이 훼손될 수 있다는 우려가 제기되고 있다. 또한 Tech Policy Press 등의 분석에 따르면 기술 감시와 인권 보호를 담당할 시민사회·학계의 참여 기회가 8개월간의 협상 과정에서 단 두 차례만 허용되는 등 참여가 극도로 제한된 상태로 빅테크 기업과 정부 중심의 '하향식 거버넌스'가 될 것이라는 지적이 나오고 있는 상황이다.

4) UN 안보리와 군사 AI

2024년 12월 유엔 안보리는 AI가 국제 평화와 안보에 미치는 영향에 대한 고위급 공개 토론을 개최했다. 구테흐스 사무총장은 '어떤 국가도 AI 시스템을 설계·개발·배포·사용하여 국제법, 인도주의법, 인권을 위반하는 군사적 적용을 해서는 안 된다'며 자율 무기 시스템에 대한 국제적 금지 규범 확립을 2026년까지 완료할 것을 촉구했다. 이는 ESG의 S(사회) 영역에서 무기·국방 관련 투자의 윤리적 기준 재정립에 영향을 미치는 분야로 러-우 전쟁, 이란 전쟁 등 지정학적 리스크가 확대되면서 관심이 높아지고 있다.

AI 윤리에 관한 권고 → AI의 사회적 영향 평가 수행, 환경 영향 고려, 취약 집단 보호를 명시적으로 권고

5) UNESCO AI 윤리에 관한 권고와 RAM

UNESCO는 2021년 11월 총회에서 'AI 윤리에 관한 권고(Recommendation on the Ethics of Artificial Intelligence)'를 194개 회원국 만장일치로 채택했다. 이는 최초의 글로벌 AI 윤리 표준으로 ① 4대 가치(인권·존엄성, 평화·정의·상호연결, 다양성·포용성, 환경·생태계 번영), ② 10대 원칙(비례·무해, 안전성·보안, 공정성·비차별 등), ③ 11개 정책영역(데이터 거버넌스, 환경·생태계, 젠더, 교육·연구, 보건·사회복지 등)으로 구성되어 있다. ESG 관점에서 중요한 것은 이 권고안이 AI의 사회적 영향 평가(Social Impact Assessment) 수행, 환경 영향 고려, 취약 집단 보호를 명시적으로 권고한다는 점이다.

도표 5. AI 윤리에 관한 권고의 구성

구 분	항 목	핵심 실천 내용
4대 가치 *(Core Values)*	1. 인권·존엄성 보호	국제 인권법 준수 및 인간의 기본권 보장
	2. 환경 및 생태계 번영	탄소 배출 저감 및 지속 가능한 생태계 유지
	3. 다양성과 포용성	국적·성별·계층 차별 없는 보편적 혜택 제공
	4. 사회적 공존과 평화	공동체 결속 강화 및 민주적 참여 증진
10대 원칙 *(Core Principles)*	1. 비례성 및 피해 방지	대량 감시·사회적 점수제 금지, 목적 외 남용 제한
	2. 안전 및 보안	AI 시스템 생애 주기 전반의 오작동 및 공격 방지
	3. 공정성·차별 금지	알고리즘 및 데이터 내 편향성 제거
	4. 지속 가능성	사회·경제·환경에 미치는 장기적 파급력 평가
	5. 프라이버시·데이터 보호	개인정보 자기결정권 보장 및 거버넌스 구축
	6. 인간의 감독 및 결정	최종 통제권 및 의사결정권을 인간이 보유
	7. 투명성·설명 가능성	AI 작동 방식과 결과 도출 과정의 공개 및 설명
	8. 책임성·책무성	문제 발생 시 개발자·운영자의 명확한 책임 소재 규명
	9. 인식 제고 및 리터러시	대중을 위한 AI 역량 교육 및 위험성 비판적 이해
	10. 다자간 거버넌스 협력	정부·기업·시민사회의 유연한 국제적 협력 체계
11개 정책 영역 *(Policy Actions)*	1. 윤리적 영향 평가	AI 도입 전 사회적·윤리적 파급 효과 사전 진단 제도 도입
	2. 윤리적 거버넌스	윤리 기준을 감독하고 조율할 범정부 컨트롤타워 구축
	3. 데이터 정책	프라이버시를 보호하는 동시에 공공재적 데이터 개방 확대
	4. 발전 및 국제 협력	선진국-개발도상국 간 AI 디지털 기술 격차 해소 지원
	5. 환경 및 생태계	AI 모델 훈련 전력 소모 감축 및 기후변화 대응 활용
	6. 젠더 (성평등)	알고리즘 내 성별 고정관념 제거 및 여성 연구자 육성
	7. 문화	소수 언어·문화유산 보존 및 특정 문화 독점 방지
	8. 교육 및 연구	AI 윤리 교육 정립 및 공익적 과학 기술 연구 장려
	9. 정보 및 통신	딥페이크 등 가짜뉴스 확산 방지 및 정보 무결성 확보
	10. 경제 및 노동	일자리 변화 대응을 위한 노동자 재교육(Reskilling) 지원
	11. 의료 및 사회적 안녕	의료 AI 안전성 확보 및 AI 의인화에 따른 정서 부작용 통제

자료: UNESCO, 신영증권 리서치센터

권고의 실효성을 담보하기 위해 UNESCO는 두 가지 핵심 도구를 개발했는데, 하나는 RAM(Readiness Assessment Methodology)이다. RAM은 각국이 AI 윤리적 구현 준비도를 ① 법적·규제적, ② 사회·문화적, ③ 경제적, ④ 과학·교육적, ⑤ 기술·인프라적 5개 차원에서 자가 진단하는 도구이며, 2026년 현재 기준 전 세계

80개국에서 사용 중이다. 두번째는 EIA(Ethical Impact Assessment)로 개별 AI 시스템에 대한 윤리적 영향평가 도구이다. 주목할 점은 한국 민간 기업 중 LG AI 연구원이 전 세계에서 유일하게 UNESCO AI 윤리 권고의 이행 현황을 매년 체계적으로 공개하고 있다는 사실이다. 이는 글로벌 ESG 평가에서 한국 기업의 차별적 강점으로 작용하고 있다.

2. 주요국·지역 차원의 거버넌스

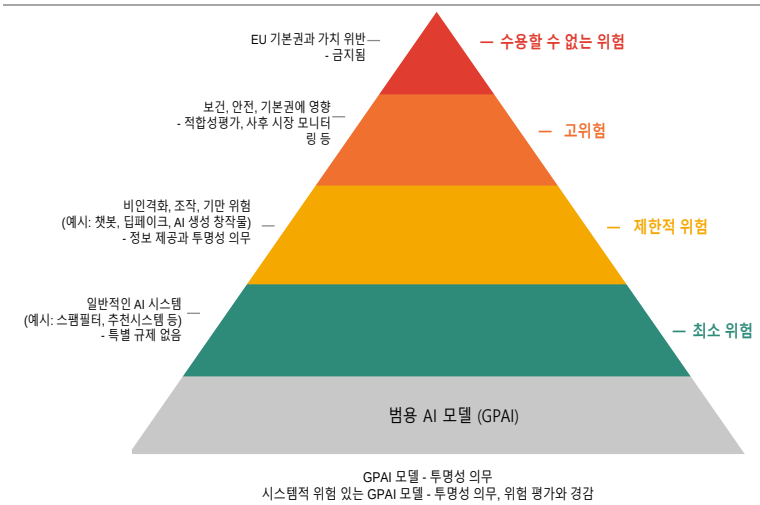
1) EU — AI Act의 단계적 시행과 CSRD·SFDR와의 교차 적용

2024년 8월 1일 발효된 EU AI Act는 세계 최초의 포괄적 AI 법임 → 위험 기반 접근법에 따라 AI 시스템 위험을 4단계로 분류하고 각 단계별로 의무를 규정하고 있음

2024년 8월 1일 발효된 EU AI Act(Regulation 2024/1689)는 세계 최초의 포괄적 AI 법제다. 위험 기반(Risk-Based) 접근법에 따라 AI 시스템을 허용 불가(Unacceptable), 고위험(High-Risk), 제한적 위험(Limited Risk), 최소 위험(Minimal Risk)의 4단계로 분류하고 각 단계별 의무를 규정하며 다음과 같은 특징을 가지고 있다.

▲요구사항 차등: 생성형 AI 별도 규정 및 위험 수준에 비례한 엄격한 요구사항 관리, ▲페널티 차등: 위험 수준별 부과된 요구사항 위반시, 위험 수준에 비례한 미준수 벌금 부과, ▲엄격한 생성형 AI관리: 시스템적 위험이 있는 GPAI(컴퓨팅 파워의 누적 계산량이 총 1025 FLOPs 초과하는 경우)가 고위험 시스템에 활용될 경우, 고위험 요구사항 추가 부과

도표 6. EU AI Act 위험 기반 규제



자료: 한국법제연구원, 신영증권 리서치센터

도표 7. 위험 범주별 분류 및 주요 요구 사항

구 분	생성형 AI	고위험	제한된 위험	최소 위험
정의	범용AI(GPAI)	안전이나 기본권에 부정적 영향을 미칠수 있는 AI시스템	중대한 위해를 가할 위험이 없는 AI시스템	기타 AI시스템
해당시스템	대규모 언어모델 (LLM)	채용시스템 인사평가시스템 신용 대출 평가시스템	챗봇 AI활용 컨텐츠 작업시스템	OCR RPA
적용대상	공급자	공급자 수입업체 배포자 유통업체	공급자 사용자	-
요구사항	투명성 기술문서화 모델정보제공 EU저작권법준수	투명성 위험관리 로그기록 정확성 기술문서화 데이터관리 인간의 감독 견고성/보안	투명성	-
미준수 페널티	매출액의 3%	매출액의 3%	매출액의 1.5%	-

자료: EU이사회, 신영증권 리서치센터

ESG 관점에서 EU AI Act의 핵심은 다음과 같다. 먼저 환경(E) 측면에서 고위험 AI 시스템은 에너지 소비 및 자원 사용에 관한 문서화 의무를 부담하며, 이는 CSRD의 이중 물질성 평가와 연동된다. 사회(S) 측면에서는 채용·신용 평가·교육 접근 등에 활용되는 AI는 고위험으로 분류되어 편향 방지, 인간 감독, 영향 평가 의무를 진다. 거버넌스(G) 측면에서는 위험 관리 시스템 구축, 데이터 거버넌스, 기술 문서화, 로그 기록, 투명성 정보 제공, 인간 감독 체계 등 포괄적인 기업 AI 거버넌스 인프라를 요구한다.

위반시 과징금은 일반 의무 위반은 EUR 1,500만 또는 3%, 잘못된 정보 제공은 EUR 750만 또는 1%로 EU에서 시행중인 일반 데이타 보호 규정(GDPR)을 상회하는 수준으로 강력

위반 시 과징금은 ① 금지 AI 관행은 최대 EUR 3,500만 또는 글로벌 연간 매출의 7% 중 더 높은 금액, ② 일반 의무 위반은 EUR 1,500만 또는 3%, ③ 잘못된 정보 제공은 EUR 750만 또는 1%로 EU에서 시행중인 일반 데이타 보호 규정(General Data Protection Regulation, GDPR)을 상회하는 수준이다. EU 시장 접근을 원하는 전 세계 기업들이 이 규범을 준수해야 해서 '브뤼셀 효과(Brussels Effect)'를 보일 것으로 보고 있다.

2025년 7월 10일 최종 발표된 'General-Purpose AI Code of Practice'는 GPAI 의무 이행의 자율 지침으로, OpenAI·Google·Anthropic 등 주요 AI 모델 제공 기업들이 자발적 참여 의사를 표명했다. 2025년 11월 19일 EU 집행위원회는 'Digital Omnibus' 패키지를 통해 고위험 AI 일부 규정의 적용 시점을 늦추는 방안을 제안했다. 이는 EU 회원국·산업계의 '규제 단순화' 요구를 반영한 것으로, 2026년 4월 말 한때 협상이 결렬될 위기에 처했으나, 2026년 5월 7일 유럽연합 이사회(Council of the EU)와 유럽의회(European Parliament)가 긴 마라톤 협상 끝에 마침내 'AI 옴니버스(AI Omnibus)' 수정안에 대한 잠정적 정치적 합의안(PPA, Provisional Political Agreement)을 공식 발표했다. 이에 따라 기존 2026

년 8월 과 2027년 8월로 예정되었던 규제 도입 시기가 각 유형에 따라 2027년 12월과 2028년 8월로 연기되었다.

유럽의 금융 기관은 EU AI Act, CSRD, SFDR을 동시에 충족하는 통합적 AI 거버넌스 체계를 구축해야

한편 유럽 지속가능성 보고 기준(CSRD) 및 ESRS는 AI의 에너지 사용이 기업의 전체 에너지 프로파일에 유의미한 영향을 미치는 경우 환경 공시 항목에 포함할 것을 요구하고 있다. 또한 지속가능금융공시규제(SFDR)하에서 금융 기관은 AI 기반 투자 의사결정 도구의 거버넌스와 설명 가능성에 대한 감독 의무를 진다. 결과적으로 유럽의 금융 기관은 EU AI Act, CSRD, SFDR을 동시에 충족하는 통합적 AI 거버넌스 체계를 구축해야 하는 복합적 컴플라이언스 부담에 직면해 있다.

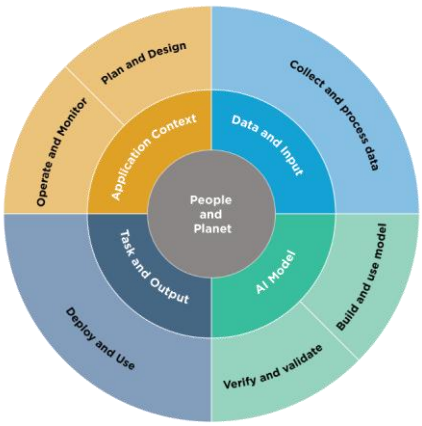
2) 미국 — NIST AI RMF vs 州법 vs 트럼프 행정부 AI Action Plan

미국의 AI 거버넌스는 '분권적·자율 규제' 모델로 특징지을 수 있다. 핵심 축은 ① NIST AI Risk Management Framework(2023.1 발표), ② 주(州) 단위 입법(캘리포니아 SB 53·SB 243·SB 942·AB 316·AB 489·AB 2013 등, 콜로라도 AI Act, 텍사스 TRAIGA 등 1,000건 이상), ③ 행정명령이다.

NIST AI RMF는 GOVERN, MAP, MEASURE, MANAGE의 4개 함수로 구성된 자발적 프레임워크 →강제성이나 처벌 규정이 있는 법률이 아니라 실무 가이드 라인임

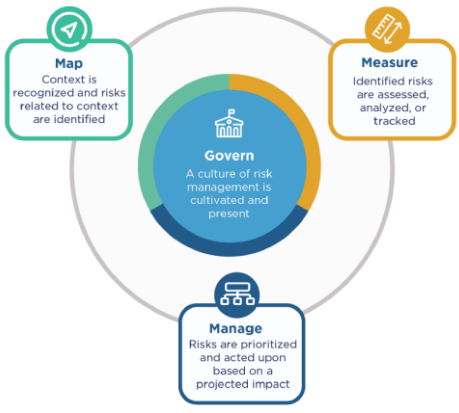
먼저 NIST AI Risk Management Framework는 미국 국립표준기술연구소(NIST)가 제정한 글로벌 표준으로, 조직이 AI의 신뢰성을 높이고 발생 가능한 위험을 관리하기 위해 자발적으로 도입하는 대표적인 지침으로 알려져 있다. 강제성이나 처벌 규정이 있는 법률이 아니라, 기업과 조직이 스스로 안전한 AI를 설계·개발·배포할 수 있도록 돕는 실무 가이드 라인이다. 이 NIST AI RMF는 GOVERN(거버넌스), MAP(조사), MEASURE (측정), MANAGE(관리)의 4개 함수로 구성된 자발적 프레임워크로 이 4대 기능은 일회성 절차가 아니라, AI 시스템의 전체 생애주기 동안 끊임없이 반복되는 유기적인 프로세스이다.

도표 8. AI 시스템의 생애주기 및 주요 차원



자료: OECD, NIST, 신영증권 리서치센터

도표 9. AI 위험 관리 프레임워크



자료: NIST, 신영증권 리서치센터

NIST는 신뢰할 수 있는 AI(Trust worthy AI)는 다음과 같은 7가지 조건을 갖추어야 한다고 강조한다.

신뢰할 수 있는 AI (Trustworthy AI)의 7가지 조건

① Valid & Reliable (유효성·신뢰성)
의도한 목적과 조건에서 제대로 작동하고, 일관된 성능을 보이는가?

② Safe (안전성)
사람·재산·환경에 예상치 못한 피해를 주지 않도록 설계·운영되는가?

③ Secure & Resilient (보안·회복탄력성)
공격·오류·환경 변화에 대해 보안상 안전하고, 장애 발생 시 복구 가능한가?

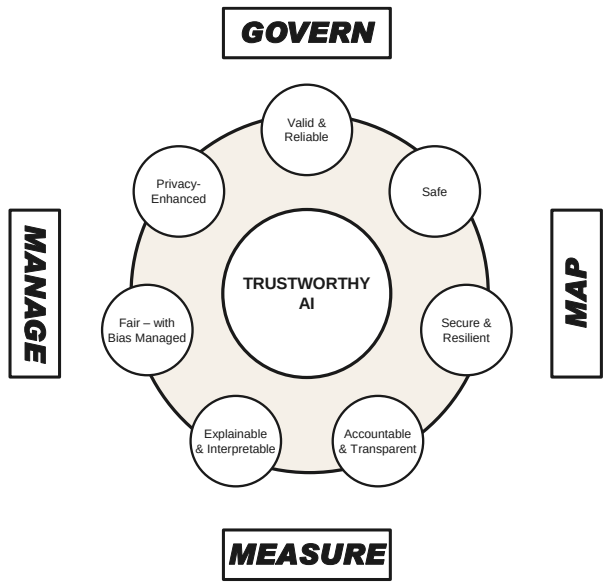
④ Accountable & Transparent (책임성·투명성)
누가 어떤 의사결정을 했는지 추적 가능하고, 관련 정보가 투명하게 제공되는가?

⑤ Explainable & Interpretable (설명가능성·해석가능성)
결과와 그 근거를 사용자가 이해할 수 있게 제공하는가?

⑥ Privacy-Enhanced (프라이버시 보호)
데이터 수집·이용·저장 과정에서 프라이버시를 적극적으로 보호하는가?

⑦ Fair, with Harmful Bias Managed (공정성·편향 관리)
차별적 결과를 최소화하도록 데이터·모델·출력을 설계·검증하는가?

도표 10. NIST AI RMF의 Trustworthy AI



자료: NIST, 신영증권 리서치센터

자발적 프레임워크인 NIST AI RMF가 중요한 이유는 법적 강제성은 없지만, 글로벌 시장(특히 미국)에서 비즈니스를 하는 기업들에게는 사실상 필수 표준으로 자리 잡고 있기 때문이다. 특히 캘리포니아주의 AI 규제법(AB 316, SB 942 등)이나 유

럽연합의 AI법(EU AI Act)을 준수할 때, 기술적 방어 논리와 안전성 입증 자료로 가장 널리 인용되는 기준이 바로 이 NIST AI RMF이다. 2024년 7월 'Generative AI Profile(NIST AI 600-1)' 부속서가 추가되었고 2025년부터는 '에이전트 AI 프로파일' 논의가 진행 중이다.

캘리포니아주는 2026년부터 인공지능의 투명성을 높이고 아동 및 청소년 등 취약 사용자를 보호하기 위해 여러 핵심 AI 규제 법안을 본격 시행중

다음으로 주(州) 단위 규제를 살펴보면, 캘리포니아주는 2026년부터 인공지능의 투명성을 높이고 아동 및 청소년 등 취약 사용자를 보호하기 위해 여러 핵심 AI 규제 법안을 본격적으로 시행하고 있다. 미국 연방 차원의 통합 규제가 없는 상황에서 실리콘밸리가 위치한 캘리포니아주가 도입한 이 법안들은 전 세계 AI 규제의 표준(글로벌 기준점)으로 평가받고 있다. 가장 핵심적인 법안들의 주요 내용과 의무 사항은 다음과 같다.

가) 캘리포니아 AI 투명성법 (SB 942)
생성형 AI의 무분별한 가짜 뉴스 및 딥페이크 확산을 막기 위해 2026년 1월 1일부터 발효된 법안이다. 월간 방문자 또는 이용자가 100만 명 이상인 대형 AI 시스템 개발사 등을 대상으로 생성형 AI가 만든 텍스트, 이미지, 영상, 오디오 등 모든 콘텐츠에 AI 생성물임을 식별할 수 있는 디지털 워터마크(Watermark) 표시를 의무화했다. 또한 일반 사용자가 AI 콘텐츠 여부를 판별할 수 있는 '탐지 도구'를 함께 제공하도록 했다.

나) 미성년자 보호 및 AI 챗봇 안전법 (SB 243)
아동·청소년이 AI 캐릭터나 동반자 챗봇에 지나치게 의존하거나 정서적으로 가스라이팅 당하는 것을 예방하기 위해 2026년 1월 1일부터 발효되었다. 인간과 유사하게 대화하며 정서적 유대를 형성하는 AI 동반자형 챗봇 운영 기업을 대상으로 서비스 시작 시 사용자의 연령 확인 기능을 의무적으로 도입해야 하며, 대화를 시작할 때 '당신은 사람과 대화 중이 아닙니다'라는 명확한 알림 문구를 띄워야 하고, 미성년 사용자가 장시간 이용할 경우, 주기적으로 휴식을 권고하는 '휴식 알람'과 경고창을 띄워야 한다. 또한 사용자가 자해나 자살 충동 등 위험 신호를 표현하면 AI가 이를 즉시 감지해 위험 대응(프로토콜) 절차를 가동하고, 그 결과를 주 공중보건부에 보고해야 한다. 특히 미성년자에게 성적으로 노골적인 이미지나 대화 내용이 노출되는 것도 엄격히 차단해야 한다.

다) AI 훈련 데이터 투명성법 (AB 2013)
AI 모델이 어떤 데이터를 학습했는지 출처를 투명하게 밝히도록 요구하는 법안으로, 2026년 1월 1일부터 시행되었다. AI 개발사는 모델 훈련에 사용된 데이터의 소스, 저작권 유무, 데이터의 대략적인 요약 정보를 대중에게 공개해야 한다. 이는

창작자들의 저작권 침해 우려를 해소하고 무단 데이터 수집을 방지하기 위함이다.

라) 첨단 인공지능 투명성법 (SB 53)
대형 AI 모델의 안전성과 기업 내부의 책임 의식을 높이기 위한 법안으로 2026년 1월 1일부터 시행되었다. 매출 5억 달러가 넘는 초대형 AI 기업들은 강력한 기술을 공개하기 전 반드시 안전성 검증을 거치고, 사회적 위험에 대한 완화 계획을 대중에게 공개해야 한다. 이를 위반하면 건당 최대 100만 달러의 벌금이 부과된다. 또한 생명을 위협하거나 대규모 재산 손해(10억 달러 상당)를 초래할 수 있는 AI 위험성을 폭로하는 내부고발자를 법적으로 강력하게 보호하는 내용도 담고 있다.

마) AI 책임 면책 차단법(AB 316)

AI 책임 면책 차단법: 컴퓨터는 책임을 질 수 없으므로 이를 만들고 배포한 인간과 기업이 모든 책임을 진다'는 원칙을 법제화 → 개발사에서부터 최종 배포자까지 모두가 대상이 됨

AI 시스템으로 인해 발생한 민사상 피해 소송에서 개발사나 운영사가 'AI가 스스로 한 일'이라는 법적 변명을 금지해 책임을 회피하는 것을 원천 차단하기 위해 제정된 것으로 2026년 1월 1일부터 시행되었다. AI의 특성상 딥러닝 기반 모델은 개발자조차 왜 그런 결과(환각, 에러, 차별 등)를 냈는지 정확히 알 수 없는 '블랙박스' 문제가 있다. 이전에는 기업들이 이를 빌미로 책임을 피해 가려 했으나, 캘리포니아주는 '컴퓨터는 책임을 질 수 없으므로 이를 만들고 배포한 인간과 기업이 모든 책임을 진다'는 원칙을 법제화한 것이다. 이에 AI 시스템을 개발, 수정 또는 사용(배포)한 기업은 법정에서 'AI가 자율적(Autonomously)으로 판단하여 피해를 입힌 것이므로, 인간인 우리는 예측할 수 없었고 책임이 없다'라는 항변(Defense)을 제기할 수 없게 되었다. 특히 이 법안은 초거대 AI 모델 개발사뿐만 아니라, 오픈소스를 가져와 미세조정(Fine-tuning)한 기업, AI를 자사 서비스에 탑재해 고객에게 제공한 최종 기업(Deployer)까지 모두 적용 대상이다. 이에 기업들은 AI 배포 전 철저한 안전성 테스트를 거쳐야 하며, AI 관련 민사 책임 보험 등을 필수로 검토해야만 하게 되었다.

바) 의료 AI의 '의사 사칭' 및 기만행위 금지법(AB 489)
환자들이 AI 챗봇의 의료 조언을 실제 인간 전문의의 진단으로 착각하여 발생할 수 있는 건강상의 위험을 방지하기 위해 마련된 것으로 2026년 1월 1일부터 시행되었다. 이 법안 따르면 의료 및 헬스케어 관련 AI 시스템이나 생성형 AI 기반 서비스가 실제 라이선스(면허)를 가진 인간 의료 전문가(의사, 박사, M.D. 등)가 진단·처방하는 것처럼 오인할 수 있는 용어나 문구를 사용하는 것을 엄격히 금지하고 있다. 또한 AI가 환자에게 진단 보고서, 소견서, 건강 상담을 제공할 때, 적법한 인간 의사의 개입이나 검토(Human oversight)가 없다면 의학적 권위를 암시하는 직함이나 면허 관련 약어를 필터링 없이 노출해서는 안 된다. 또한 헬스케어 AI 서비스의 마케팅 광고나 인터페이스 설계 시에도 환자가 '의사와 대화하고 있다'고 착각하게 만드는 기만적 표현도 전면 금지된다. 특히 법안을 위반하여 금지된 용어가

사용될 때마다 '건별로 독립된 위반 행위'로 계산되므로, 챗봇이 수천 명의 환자에게 반복적으로 잘못된 표현을 썼다면 천문학적인 벌금이나 금지 명령(Injunction)을 받을 수 있어 각별한 주의가 필요하다.

바이든 행정부(2023.10) 행정명령 14110 발표 → 안전하고 신뢰할 수 있는 AI의 책임 있는 개발·사용에 관한 포괄적 프레임워크 제시

다음으로 미국 정부의 행정명령을 살펴보면, 바이든 행정부는 2023년 10월 행정명령 14110(E.O. 14110)을 통해 안전하고 신뢰할 수 있는 AI의 책임 있는 개발·사용에 관한 포괄적 프레임워크를 제시했다. 2019년 OECD AI 원칙 지지, 2020년 GPAI 창설 공동 참여, 2023년 블레츨리 선언 서명, 2024년 유럽평의회 AI 협약 참여 등 광범위한 국제 AI 거버넌스에도 참여했었다.

트럼프 행정부(2025.1) 행정명령 14110 폐지→ 행정명령 14179 발표 'Removing Barriers to American Leadership in AI'

그러나 2025년 1월 20일 트럼프 행정부 출범과 함께 EO 14110이 폐지되고, 행정명령 14179 'Removing Barriers to American Leadership in AI'가 발표되었다. 이 명령은 이전 정부의 규제 지향적 접근에서 급선회해서 AI 혁신 가속화를 최우선 과제로 했다. 2025년 2월 파리 AI 정상회의에서 JD 밴스 부통령이 'AI 안전에 대해 논의하기 위해 여기 온 게 아니다'라고 발언하고 미국이 공동선언 서명을 거부한 것은 이러한 기조 전환의 상징적 사건으로 기록되기도 했다.

트럼프 행정부(2025.7) 'Winning the Race: America's AI Action Plan'과 3개 행정명령 발표

2025년 7월 23일 트럼프 행정부는 'Winning the Race: America's AI Action Plan'과 3개 행정명령('데이터센터 인프라 연방 허가 가속화', '연방정부 내 Woke AI 방지', '미국 AI 기술 스택 수출 촉진')을 발표했다. 본 액션플랜은 ①AI 혁신 가속, ②미국 AI 인프라 구축, ③국제 AI 외교·안보 리더십 확보 3대 축으로 구성되어 있다. 주목할 점은 NIST AI RMF의 'misinformation, DEI, climate change' 관련 참조를 삭제하라는 지시가 포함되어 EU·UN의 인권 기반 접근과 명백한 정책 분기(divergence)가 발생했다는 점이다. CSIS 분석에 따르면 이와 같은 미국의 다자 거버넌스 이탈은 글로벌 AI 거버넌스 아키텍처에 구조적 공백을 야기하고 있다.

트럼프 행정부(2025.12) 행정명령 Ensuring a National Policy Framework for Artificial Intelligence'에 서명

2025년 12월 11일 트럼프 대통령은 추가 행정명령 'Ensuring a National Policy Framework for Artificial Intelligence'에 서명하면서 연방 정책과 상충되는 주(州)법을 무력화하는 DOJ AI 소송 태스크포스(AI Litigation Task Force) 설치를 명령했다. 콜로라도 AI Act는 이 행정명령에서 '과도한 주 규제'의 대표 사례로 명시되기도 했다.

3) 중국의 독자적 다자주의 전략

중국은 혁신과 통제의 독자적 모델 & 독자적 다자주의 전략

중국은 AI 거버넌스 측면에서 혁신과 통제의 독자적 경로를 걷고 있다. 2023년 생성형 AI 규제를 통해 콘텐츠, 훈련 데이터, 허위정보, 국가 안보, 데이터 프라이버시, 지식재산권, 사회 안정, 윤리적 우려에 초점을 맞춘 규범을 제도화했다.

2024년 9월 중국 규제 당국은 EU AI Act와 미국 방식이 혼합된 포괄적 AI 거버넌스 원칙 및 안전 가이드 라인을 발표했다. 이 가이드 라인은 기업 내부에 AI 윤리 심사 위원회 설립 의무화, '통제 가능한 개발' 강조, 국가 안보·사회 안정 최우선 가치 등이 특징이다.

AI Safety Summit 관점에서 보면 중국은 2023년 블레츨리 선언에는 서명했으나 2024년 서울 선언에는 불참했다. 이후 2025년 7월 세계 AI 컨퍼런스에서 글로벌 AI 거버넌스 행동 계획을 발표하고 새로운 국제 AI 협력 기구 창설을 제안하는 등 독자적 다자주의 전략을 추진하고 있다. 미국의 이탈로 균열이 발생한 AI Safety Summit은 중국의 참여 유지를 글로벌 AI 거버넌스의 핵심 과제로 다루고 있다.

4) 유럽평의회 AI Framework Convention

유럽평의회 AI Framework Convention → 최초 법적 구속력 국제조약

유럽평의회(Council of Europe, 46개 회원국)는 2024년 5월 17일 'Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law'(CETS No. 225)를 채택하고, 9월 5일 빌니우스(리투아니아)에서 열린 회의에서 서명을 했다. 이는 '세계 최초의 법적 구속력 있는 국제 AI 조약'이라는 역사적 의미를 갖고 있다. 초기 서명국은 EU, 미국, 영국, 이스라엘, 노르웨이, 조지아, 몰도바, 아이슬란드, 안도라, 산마리노 등이다. 협약은 ①인권 보호, ②민주주의 원칙 유지, ③법치주의 준수 3대 가치를 'AI 시스템 전 생애주기(lifecycle)'에 적용할 것을 의무화했다. 협약은 5개 서명국(3개 이상의 유럽평의회 회원국 포함)이 비준한 후 3개월이 지난 다음 달부터 발효되는데, 2026년 5월 현재까지 비준국이 부족하여 아직 정식 발효되지는 않았다.

본 협약은 국가 안보 목적의 AI 사용에 대한 모호한 예외 조항으로 인해 실효성 논란이 있으나, AI 거버넌스를 인권 프레임 내에 명시적으로 위치시킨 법적 선례로서의 의미가 크다. 또한 EU AI Act와 '기술 중립성(technology-neutrality)'과 '리스크 기반 접근'이라는 두 원칙에서 정합성을 갖되, 인권 보호의 직접적 법적 근거를 제공한다는 점에서 차별성이 있다. EU는 동 협약을 EU AI Act로 이행할 예정이며, 미국·영국·일본 등 非유럽 가입국은 자체 법체계로 이행될 예정이다.

5) G7 히로시마 AI 프로세스 및 OECD HAIP Reporting Framework

G7 HAIP →OECD HAIP Reporting Framework 발전

G7 히로시마 AI 프로세스(Hiroshima AI Process, HAIP)는 2023년 5월 일본이 G7 의장국 시기에 출범하여, 같은 해 10월 30일 G7 정상이 '국제 가이딩 원칙(International Guiding Principles)' 11개항과 '조직을 위한 국제 행동강령(Code of Conduct for Organizations Developing Advanced AI Systems)'을 채택하면

서 본격화되었다. 이후 2024년 G7 이탈리아 정상회의에서도 이 프로세스 협의가 지속되었으며, AI 안전 연구소 간 네트워크 형성과 국경 간 AI 리스크 관리 협력이 강화되었다. G7 행동 강령은 ①AI 시스템의 생애주기 전반에 걸친 환경 영향 평가, ②다양성과 비차별성 보장, ③공급망 내 책임 있는 AI 사용을 강조한다. 또한 2025년 12월 10일 G7 산업·디지털·기술 장관선언에서는 SME(중소기업) AI 채택 청사진과 도구 키트가 추가로 발표되었다.

OECD HAIP Reporting Framework →세계 최초의 글로벌 AI 거버넌스 보고 메커니즘

한편 2025년 2월 7일 G7 히로시마 AI 프로세스(HAIP)는 선언적 거버넌스를 넘어, OECD HAIP 보고 프레임워크(Reporting Framework)를 통해 실질적이고 강제력 있는 글로벌 공시 표준으로 안착했다. 이는 글로벌 첨단 AI 개발 조직이 자발적으로 위험관리·사고보고·정보공유 관행을 보고하는 '세계 최초의 글로벌 AI 거버넌스 보고 메커니즘'이다. 2025년 4월 OpenAI, Microsoft, Anthropic, Google, Amazon, NTT, Salesforce, Fujitsu 등 19개 글로벌 AI 개발사가 1차 보고서를 제출했으며 해당 보고는 transparency.oecd.ai에 공개되고 있다. ESG 관점에서 HAIP Reporting Framework는 'ESG 공시(disclosure)'와 동일한 위계의 글로벌 자율 거버넌스 메커니즘으로 평가받고 있다. EU AI Act 의무를 충족하기 위한 '적합성 추정(presumption of conformity)' 효과도 검토되고 있어, GPAI 코드와의 상호 운용성이 핵심 의제이다.

6) AI Safety Summit 시리즈

AI Safety Summit 시리즈 - COP에 준하는 정상 외교의 새 장르

AI Safety Summit 시리즈는 '블레츨리 파크 프로세스(Bletchley Park Process)'로도 불리며, 2023년 11월 영국 블레츨리 파크에서 시작되었다. 블레츨리(2023.11) → 서울(2024.5) → 파리(2025.2) → 인도(2026.2)로 이어지는 정상회의 시리즈는 선진 AI 모델의 안전·보안 위험을 다루는 국제 정상 외교의 새 장르로 자리 매김했다.

가) 블레츨리(2023.11) – 1차 정상회의

'블레츨리 선언'은 28개국과 EU가 서명에 참여했는데 첨단 AI의 위험성 인정, AI 안전 연구소 국제 네트워크 출범의 기반이 되었으며 특히 중국 참여가 주목할 만한 성과로 평가받고 있다. 이로써 글로벌 AI 안전 협력의 기틀이 마련되었다.

나) 서울(2024.5) – 2차 정상회의

16개 주요 AI 기업이 'Frontier AI Safety Commitments'에 자발적 서명을 했으며, 10개국+EU의 AI 안전연구소(AISI) 네트워크 출범했다. 서울 정상회의에서는 안전·혁신·포용성을 상호 연결된 목표로 규정한 서울 선언(Seoul Declaration)이 채택되었는데 약 27개국이 채택하며 참여했으나 중국은 불참했다.

AISI(AI Safety Institute) 네트워크는 프론티어 AI(최첨단 고성능 AI)의 오용, 통제력 상실, 딥페이크 등의 위험을 과학적으로 측정하고 평가하기 위해 세계 주요국 정부 산하 'AI 안전 연구소'들이 결합한 글로벌 공조 체제이다. 2024년 5월 'AI 서울 정상회의' 합의를 바탕으로 2024년 11월에 공식 출범했다. 2025년 5월 현재 호주, 캐나다, 유럽 연합, 프랑스, 일본, 케냐, 대한민국, 싱가포르, 영국, 미국 등 10개국에서 활동 중이며, 모델 안전성 평가·정렬 기술 연구·국제 협력을 주요 임무로 하고 있다. 영국 AISI는 미국 샌프란시스코에 해외 사무소를 설립했고, 미국 AISI는 OpenAI, Anthropic과 모델 출시 전 사전 안전 평가 협력을 진행하고 있다.

다) 파리(2025.2) – 3차 정상회의 (AI Action Summit)
약 100개국, 1,000여 이해관계자가 참여했으며, 5개 트랙(공익 AI, 일의 미래, 혁신·문화, AI 신뢰, 글로벌 AI 거버넌스)으로 확장되었다. 또한 96명의 AI 전문가와·30개국 전문가 자문 패널이 참여한 '국제 AI 안전 보고서'가 발표되었다. 그러나 미국과 영국의 선언 서명 거부로 심각한 균열이 노출되기도 했다. 반면 91개 파트너가 참여하는 '환경 지속가능성 연대(Environmental Sustainability Coalition)'가 출범하며 AI의 ESG 거버넌스가 공식 의제로 자리 잡았음을 보여주기도 했다. 또한 Current AI(공익 AI 이니셔티브, $400M 초기 투자)도 출범했다.

라) 인도 뉴델리(2026.2) – 4차 정상회의 (AI Impact Summit)
인도는 2024년 GPAI 의장국으로서 오픈소스·지속가능 AI·청정 에너지를 강조했다. 첫 글로벌 사우스(Global South) AI 회담으로 글로벌 사우스의 AI 거버넌스 참여 확대가 핵심 의제가 되었다. 또한 UNESCO·LG AI연구원의 'AI 윤리 MOOC' 프로젝트의 중간 성과가 공유되기도 했다.

7) 기타 주요국 동향

영국 정부는 2023년 3월 'AI 규제에 대한 혁신 친화적 접근(A pro-innovation approach to AI regulation)' 백서를 발표했다. 단일 AI 전담 법안이나 규제 기관을 신설하는 대신, 기존의 부문별 규제 기관(CMA, ICO, FCA 등)이 공통의 원칙을 적용하도록 하는 유연한 방식을 택한게 특징이다. 이는 기존 기관들이 각 산업 특성에 맞게 AI 원칙을 해석하고 AI의 잠재적 위험을 미리 차단하기보다, 산업 현장의 상황에 맞춰 적용하는 실용적인 접근 방식을 택한 것이다. 이 AI 백서는 공정성·투명성·책임·안전·이의 제기 가능성이라는 5대 원칙 아래 유연성을 제공하여 기술 성장을 촉진하는 비법률적 접근 방식을 채택하고 있다. 또한 영국은 AI 안전 연구소(AISI)를 설립하여 첨단 AI 모델 평가를 위한 오픈소스 플랫폼 'Inspect'를 출시하기도 했다.

일본은 강행 규제보다 원칙 중심의 자율 규제와 촘촘한 분야별 가이드 라인, 그리고 기존 법체계를 결합한 다층적 규제 전략을 채택하고 있다. AI 개발·제공·활용 단계마다 기업이 참고해야 할 실질적 준규제 기준을 계속 확장하고 최근 관련 법을 제정했다. 일본 정부는 2019년 3월 '인간 중심의 AI 사회 원칙(Social Principles of Human-Centric AI)'을, 2024~2025년에는 G7 히로시마 AI 프로세스를 토대로 기존 지침을 대폭 업그레이드한 '기업을 위한 AI 가이드 라인'을 발표했다. 또한 교육·의료 등 분야별 생성형 AI 가이드 라인을 병행하여 산업 현장에서의 실제 위험을 세부적으로 관리하는 구조를 갖추었다. 2025년 6월에는 '인공지능 관련 기술의 연구개발 및 활용 촉진에 관한 법률(인공지능촉진법)' 도입했는데, 이 법은 규제를 강화하는 대신 국가 차원의 AI 연구개발·인재육성·국제협력·교육·윤리·거버넌스 정비 등 종합 정책을 추진하기 위한 기본법에 가깝다는 평가를 받고 있다.

도표 11. 일본의 인간 중심의 AI 사회 원칙의 3대 기본 이념과 7대 사회 원칙

구 분	항 목	내 용
3대 기본 이념 (AI 기술이 사회에 도입될 때 반드시 지켜야 할 궁극적인 가치이자 목적)	인간 존엄성 존중 (Dignity)	AI는 인간의 기본적 인권을 침해해서는 안 되며, 인간을 대체하는 것이 아니라 인간의 능력을 확장하고 보조하는 수단이어야 한다
	다양성과 포용성 (Diversity & Inclusion)	국가, 배경, 연령에 상관없이 다양한 사람들이 각자의 행복을 추구할 수 있도록 돕는 사회를 만들어야 한다
	지속가능성 (Sustainability)	지구 환경 문제, 자원 고갈, 격차 해소 등 인류가 직면한 과제를 AI를 통해 해결하여 지속 가능한 사회를 구축한다
7대 사회 원칙 (기본 이념을 현실 사회와 비즈니스에 구현하기 위해 제시된 구체적인 실행 원칙)	인간 중심의 원칙 (Human-centric)	AI의 이용 여부와 활용 방식은 인간이 자유롭게 결정해야 하며 기술적 취약계층이나 정보 약자가 소외되지 않고 모두가 혜택을 누릴 수 있어야 한다.
	교육&리터러시 원칙 (Education&Literacy)	AI 교육 유무로 인한 사회적 양극화를 방지한다. 유아부터 고령자까지 전 연령대를 대상으로 수리·데이터 과학 및 AI 문해력(리터러시) 교육 환경을 정비한다.
	프라이버시 보호 원칙 (Privacy protection)	AI가 개인 데이터를 수집·가공하는 과정에서 개인의 자유와 프라이버시가 침해되지 않도록 안전장치를 마련
	보안 확보 원칙 (Security)	AI 시스템의 취약점을 악용한 사이버 공격이나 오작동으로부터 사회 시스템의 안전성과 항상성을 유지해야 한다
	공정 경쟁 원칙 (Fair competition)	특정 거대 기업이나 국가가 AI 기술과 데이터를 독점하여 부당한 시장 지배력을 행사하는 것을 방지하고 건전한 경제 생태계를 유지해야 한다
	공정성, 책임성, 투명성의 원칙 (Fairness, Accountability, Transparency)	AI의 제안이나 결정 과정에 인종, 성별 등의 부당한 편향(Bias)이 없어야 하며, AI의 판단 이유를 설명할 수 있는 투명성을 확보하고, 결과에 대한 책임 체계를 명확히 해야한다.
	혁신의 원칙 (Innovation)	규제로 기술 발전을 가로막기보다, 국경과 산·학·연을 초과한 협력을 통해 AI 기술 혁신이 지속될 수 있는 유연한 환경을 조성한다

자료: SPRI, 신영증권 리서치센터

이외에도 싱가포르는 AI 테스팅 툴킷인 'AI 베리파이(AI Verify)' 발표하는 등 규제 대신 도구를 제공하고 글로벌 표준과의 연계를 통해 아시아 AI 거버넌스 허브로 자리 잡고 있다. 사우디아라비아는 독자적 AI 거버넌스 이니셔티브를 통해 국가 차원의 정책으로 추진 중이며 AI 윤리 원칙, 미디어 분야 AI 원칙 등을 발표했다.

3. 국제(표준)기구의 AI 거버넌스

1) ISO/IEC 42001:2023

ISO/IEC 42001:2023 - 세계 최초의 인증 가능한 AI 경영시스템 국제표준

ISO/IEC 42001:2023은 2023년 12월 발효된 '세계 최초의 인증 가능한 AI 경영시스템 국제표준'이다. 이는 ISO 9001(품질경영시스템) 및 ISO 27001(정보보안경영시스템)과 동일한 레벨의 인증 표준으로, 조직이 AI Management System(AIMS)을 'Plan-Do-Check-Act' 방법론에 따라 수립·시행·유지·지속 개선할 것을 요구한다. 핵심 요구사항은 조직의 상황(Context of the Organization), 리더십(Leadership), 기획(Planning), 지원(Support), 운영(Operation), 성과평가(Performance Evaluation), 개선(Improvement) 등 7가지 이다.

도표 12. ISO/IEC 42001 요구사항 구조

자료: 한국경영인증원, 신영증권 리서치센터

ESG 평가 관점에서 ISO/IEC 42001 인증 보유 기업은 G(거버넌스) 등급의 정량적 가산 요소를 확보하게 된다. AWS는 본 표준을 'AI 거버넌스의 사실상 표준(de facto standard)'으로 보고 자사 클라우드 인증 체계에 통합하였고, KPMG·Deloitte 등 빅4 회계법인들은 ISO/IEC 42001 인증 자문 서비스를 핵심 비즈니스로 운영 중이다. ESG 통합 투자 측면에서 ISO 42001 인증은 기업의 AI 거버넌스 성숙도를 측정하는 객관적 지표로 부상하고 있다.

2) OECD AI 원칙(2019 채택/2024 개정)

OECD AI 원칙 → 최초의 정부간 AI 표준 (2019.5)

2019년 5월 OECD가 채택한 'Recommendation of the Council on Artificial Intelligence(인공지능에 관한 이사회 권고)'는 '최초의 정부간 AI 표준'으로 2024년 5월 생성형 AI에 대응하여 개정되었다. 이 문서의 핵심 내용인 OECD AI 원칙

(AI principles)은 5가지(포용적 성장, 인간 중심 가치·공정성, 투명성·설명가능성, 견고성·보안·안전성, 책임성) 핵심 원칙으로 구성되어 있다. 또한 2024년 개정된 OECD AI 원칙은 ①정보 무결성(Information Integrity) 강화 및 허위정보 대응, ② 가치사슬 전반(Whole-of-Value-Chain)으로의 책임 확장, 위험 제어 기능(수정·중단 권한)의 구체화, ④ 프라이버시와 지식재산권(IP) 보호 명시 등이 주요 내용으로 세계 주요국의 AI 규제 법안들과 용어의 정의 및 위험 관리 프레임워크를 상호 호환(Interoperability)이 가능하도록 맞추었다. G7 히로시마 코드(HAIP), EU AI Act, 유럽평의회 협약, 미국 NIST AI RMF 모두 OECD AI 원칙을 직접·간접적으로 참조하고 있다.

한편 2025년 파편화되어 있던 글로벌 AI 거버넌스를 하나로 모으기 위해 단행된 GPAI 이니셔티브와 OECD의 통합으로 다중이해관계자 협력의 실질적 허브로서의 기능이 강화되었다. 특히 통합 조직의 첫 수장(OECD AIGO·GPAI 통합 의장)으로 강하연 정보통신정책연구원 연구위원이 선출되어 한국은 글로벌 AI 규범의 방향성을 설정하고 의제를 주도하는 핵심 리더 국가로 발돋움하게 되었으며, 국내 AI 정책과 글로벌 표준 간의 연계를 더욱 긴밀하게 추진할 수 있게 되었다.

3) NIST AI Risk Management Framework

앞서 미국의 AI원칙으로 살펴본 NIST AI RMF 1.0은 2023년 1월 미국 상무부 산하 국립표준기술연구소가 발표한 자발적 프레임워크로, EU AI Act, ISO/IEC 42001과 함께 글로벌 AI 거버넌스의 '3대 운영 표준' 중 하나로 자리 잡았다. 2024년 7월 'Generative AI Profile(NIST AI 600-1)' 부속서가 추가되어 환각, 지식재산권 침해, 유해 콘텐츠 등 생성형 AI 특유의 위험을 다루며, 2025년에는 '에이전트 AI 프로파일' 개발을 시작해 올해 말 공식 발표를 목표로 개발 및 조율 진행 중이다.

4) IEEE 표준(Standards Association)

IEEE 표준 → 기술 표준의 윤리적 내재화

세계 최대 기술 전문 조직인 IEEE는 100개 이상의 AI 관련 글로벌 표준을 보유하고 있는 것으로 알려져 있다. 2025년 7월 IEEE SA(Standards Association)는 UN AI 자문기구인 HLAB-AI의 권고에 따라 창설된 '국제 AI 표준 교환소(International AI Standards Exchange)'의 파트너로 인정받았다. IEEE가 운영중인 IEEE CertifAIEd 프로그램은 제품에 대한 철저한 검증을 제공하고, 조직이 윤리적인 AI 관행을 성공적으로 도입하고 있음을 입증할 수 있도록 지원한다. 또한 이 프로그램은 거버넌스와 관련된 위험 요소를 조기에 파악하여 제품 문제 및 금전적, 평판적 손실과 같은 잠재적 손해를 예방하는 데 도움을 주는 것으로 알려져 있다.

4. NGO·이니셔티브·연구기관 vs AI 거버넌스

1) GPAI

GPAI – 글로벌 다중 이해관계자 협력의 중심

GPAI(Global Partnership on AI)는 프랑스와 캐나다 등의 주도로 G7 협의체에서 제안되어 출범한 정부 간 다자협의체로 정부·산업·학계·시민 사회가 AI 연구, 정책 개발, 모범 사례 공유에 협력하는 글로벌 다중이해관계자 이니셔티브이다. 2020년 OECD의 지원 하에 창설된 이후 2025년 OECD 체계로 통합되었다. 인도는 2024년 GPAI 의장국으로서 이 이니셔티브를 책임 있는 AI 개발을 위한 핵심 플랫폼으로 발전시키려는 의지를 표명했다.

2) Partnership on AI (PAI)

Partnership on AI(PAI)는 2016년 Apple, Amazon, Meta, Google/DeepMind, IBM, Microsoft의 주도로 출범한 민간 기업, 시민단체, 학계 중심의 글로벌 비영리 단체이다. 안전한 파운데이션 모델 배포 가이드 라인, 합성 미디어를 위한 책임 있는 실천 규범, 공동 번영 가이드 라인, 신속한 결함 감지를 위한 AI 에이전트 프레임워크 등 영향력 있는 산출물을 제공해왔다.

3) Future of Life Institute (FLI)

Future of Life Institute는 AI 안전(AI Safety) 의제를 글로벌 정책 어젠다로 부상시킨 핵심 비영리 기관이다. 2023년 3월 '6개월 AI 개발 중단' 공개서한, 블레츨리 정상회의 사전 정책 권고, AI Safety Standards Policy(SSP) 등을 발표했다. 2024년 9월 채택된 EU AI Act 최종안에 '시스템적 위험'과 'Frontier AI' 개념을 반영시킨 데 결정적인 역할을 한 것으로 평가받고 있으며, AI Safety Summit 시리즈에 권고안·스코어카드·안전 표준 정책을 제공하며 의제 설정에 실질적 영향력을 행사하기도 했다.

4) AlgorithmWatch · EPIC · CAIDP · Access Now

AlgorithmWatch(독일 베를린)는 알고리즘 시스템의 설명 가능성과 추적 가능성에 중점을 둔 비영리 단체로, 유럽에서 디지털 권리 옹호 활동을 전개하고 있다. EPIC (Electronic Privacy Information Center)와 CAIDP(Center for AI and Digital Policy)는 미국 워싱턴 D.C. 기반 시민사회 조직으로, 유럽평의회 AI 협약 협상 과정에서 '민간부문 면제(private sector carve-out)'에 반대하는 글로벌 캠페인을 주도했다. 또한 CAIDP는 2025·2026년 G7 정상회의에 대한 정책 권고를 매년 발표하고 있다. 또한 Access Now는 디지털 권리·표현의 자유를 옹호하며 AI 기반 감시 기술의 인권 영향을 분석하고 있다.

AI Now Institute/ Algorithmic Justice League/CyberPeace Institute → AI 기술의 급격한 발전이 초래하는 사회적 위험을 감시하고, 기술의 공공성과 인간 중심의 안전성을 확보하기 위해 활동하는 글로벌 대표 비영리 연구·시민단체(NGO)임

5) AI Now Institute/Algorithmic Justice League/CyberPeace Institute

이들은 AI 기술의 급격한 발전이 초래하는 사회적 위험(편향, 차별, 안보 위협)을 감시하고, 기술의 공공성과 인간 중심의 안전성을 확보하기 위해 활동하는 글로벌 대표 비영리 연구·시민단체(NGO)들이다. 먼저 'AI Now Institute'는 AI 기술이 사회에 미치는 영향력을 다각도로 분석하여 빅테크 기업의 권력 집중을 견제하고 공공 이익을 대변하는 미국의 독립 연구소로 뉴욕대학교에서 출발하여 현재는 독립 기관으로 운영 중이다.

'Algorithmic Justice League'는 MIT 미디어랩의 컴퓨터 과학자 조이 부올람위니(Joy Buolamwini)가 2016년에 설립한 디지털 시민운동 단체로, AI 알고리즘에 내포된 (특히 안면인식 기술의) 인종·성별·장애 차별을 폭로하고 시정하는 데 특화되어 있다.

'CyberPeace Institute'는 2019년 스위스 제네바에 설립된 독립 NGO로, 사이버 공격과 정보 왜곡(디스인포메이션)으로부터 가장 취약한 민간인과 인도주의 단체를 보호하고 사이버 공간의 평화를 유지하는 것을 목표로 하는 단체다. UN AI 거버넌스 체계를 민간 관점에서 평가·모니터링하는 역할을 하고 있다.

6) The Future Society (TFS)

The Future Society(미국·프랑스)는 AI의 급격한 발전이 인류에게 미치는 위험을 통제하고, 인간의 근본적 가치에 부합하도록 만드는 'AI거버넌스(AI Governance)' 전문 글로벌 비영리 싱크&두탱크(Think-and-Do Tank)다. 2014년 미국 하버드 케네디스쿨(Harvard Kennedy School)의 프로젝트로 처음 출발했으며, 현재는 미국과 유럽을 기반으로 전 세계 4개 대륙에서 독립적인 비영리 단체로 활동하고 있다.

7) UN 산하기구의 NGO 협력 – UNICEF·ITU·WHO·ILO

UNICEF는 'Generation AI' 이니셔티브를 통해 아동권리 관점의 AI 거버넌스를 추진하고 있으며, World Economic Forum 및 UC Berkeley와도 협력 중이다. 국제전기통신연합(ITU)은 AI for Good 플랫폼을 통해 AI 거버넌스의 글로벌 논의 플랫폼을 운영하며, 세계보건기구(WHO)와 협력하여 의료 AI 애플리케이션의 표준 및 거버넌스 체계를 개발하고 있다. 국제노동기구(ILO)는 AI와 직업 안전 및 미래 고용에 관한 2025년 글로벌 보고서를 발표하며, AI가 근로자 권리와 고용 품질에 미치는 영향을 분석하기도 했다.

5. 글로벌 회계법인&컨설팅 회사 vs AI 거버넌스

1) 빅4 회계법인의 AI Assurance 시장 진입

글로벌 회계 법인과 컨설팅 회사들은 단순한 서비스 제공자를 넘어 AI 거버넌스 규범의 민간의 정의자로 기능하고 있어

Big 4 회계법인 및 전략 컨설팅 기업들은 책임 있는 AI 서비스 시장에서 단순한 서비스 제공자를 넘어 AI 거버넌스 규범의 민간의 정의자로 기능하고 있다. 이들은 2023년 이후 AI 이니셔티브에 총 100억 달러 이상을 투입했다고 알려져 있다. Deloitte, PwC, EY, KPMG의 빅4 회계법인은 2024~2025년 사이 'AI Assurance (인공지능 감사·인증)'를 새로운 핵심 서비스로 발전시켰다. 이는 ① 그간 ESG Assurance 시장에서 축적한 비재무 인증 노하우, ② 회계 감사 전통의 독립성·전문성, ③ ISO/IEC 42001 발효로 인한 인증 시장 형성이라는 3가지 요인이 결합된 결과로 보인다.

Financial Times(2025.6) 보도에 따르면, PwC UK가 가장 먼저 AI Assurance 서비스를 정식 출시했으며, 챗봇 정확도 검증·편향성 식별 등 구체적 서비스를 제공하고 있다. KPMG는 2024년 12월 ISO/IEC 42001 첫 인증을 발표했고, '10대 신뢰 가능한 AI 기둥(10 Trusted AI Pillars)'이라는 자체 프레임워크를 발표했다. Deloitte는 NIST의 AI RMF 관련 'Trustworthy AI™' 프레임워크를 운영하며, 2025년 3월 'Zora AI'라는 자율 AI 에이전트 플랫폼을 출시했다.

도표 13. 빅4 회계법인 AI 거버넌스 프레임워크 비교

기관	대표 프레임워크	주요 활동 및 내용	주요 플랫폼
Deloitte	Trustworthy AI™ – 7대 차원	'State of GenAI in the Enterprise' 발간. 신뢰·거버넌스 중심.	Zora AI (2025.3)
PwC	Responsible AI Governance Model	AI Jobs Barometer·CEO Survey 발간. UK AI Assurance 최초 출시.	ChatPwC (2024.5)
EY	EY AI Confidence Index	Sovereign AI·규제산업 특화. NVIDIA·Dell AI Factory 온프레미스 배포.	EY-Parthenon AI (2025.9)
KPMG	10 Trusted AI Pillars (ISO 42001 첫 인증)	$2B AI·클라우드 투자. 다중 에이전트 협업 플랫폼.	Workbench (2025.6)

자료: 각 사 홈페이지, 신영증권 리서치센터

2) MBB(McKinsey·BCG·Bain) 및 IBM Consulting

MBB 전략 컨설팅 회사의 접근법은 빅4 회계법인 대비 '방법론 중심'이다. McKinsey의 AI 전담 부문인 QuantumBlack은 'Rewired' 6대 역량 프레임워크와 'State of AI' 설문조사를 운영하며, 100여 개국 약 2,000명 이상 임원을 대상으로 한 산업 표준 데이터를 생산하고 있다. 또한 BCG는 '10-20-70' 가치 창출 비율(10% 알고리즘, 20% 데이터·기술, 70% 사람·프로세스·문화)을 핵심 메시지로 활동하고 있다. 한편 Bain은 OpenAI와의 독점 파트너십을 바탕으로 차별화된 모습을 보이고 있다. 그리고 IBM Consulting은 'Watsonx. Governance'를 통해 자동화 AI 거버넌스 플랫폼을 제공하고 있다.

III. 국내 AI 거버넌스 동향

1. 정부 및 공공기관 관점의 한국 AI 거버넌스 동향

정부는 규제를 통한 통제보다 '진흥과 신뢰의 균형'을 거버넌스의 핵심 가치로 삼고, 글로벌 3대 AI 강국(AI G3) 도약을 목표로 민관 협업 체계를 강력하게 가동 중

국내는 2024년 말 법적 기초 마련과 전담 기구 신설을 기점으로, 2026년 현재 '세계 최초의 포괄적 AI 기본법 발효 국가'이자 아시아·태평양 지역의 AI 안전 허브로 자리매김하고 있다. 정부는 규제를 통한 통제보다 '진흥과 신뢰의 균형'을 거버넌스의 핵심 가치로 삼고 있으며, 글로벌 3대 AI 강국(AI G3) 도약을 목표로 민관 협업 체계를 강력하게 가동 중이다. '진흥과 규제의 병행' 접근법은 EU의 '엄격한 사전 규제'와 미국의 '자율 규제 중심' 사이에서 균형점을 모색하는 접근 방법으로 ① 국내 AI 산업 생태계가 미국·중국에 비해 상대적으로 후발주자라는 점, ② 그러나 통신 인프라·디지털정부·공공데이터 활용 등에서 OECD 최상위권 역량을 보유한다는 점, ③ 반도체·플랫폼·제조 등 AI 관련 산업 비중이 매우 크다는 점이 종합적으로 반영된 결과로 보인다.

OECD 디지털정부 지수(DGI)에서 한국은 0.94점으로 1위(평균 0.61), 공공데이터 개방 지수(OURdata)에서도 0.91점으로 1위를 기록하고 있다. 5G 다운로드 속도 1위, 고정 브로드밴드 가입자 100명당 47.3명으로 2위 등 AI 발전을 지원하는 인프라 측면에서 한국은 명백한 강점을 보유하고 있다. 다만 AI 민간 투자 규모, AI 인재 절대 수, 글로벌 빅테크 LLM 보유 측면에서는 미국·중국에 크게 뒤져 있어 이를 보완하기 위한 '소버린 AI(Sovereign AI)' 전략이 추진되고 있다.

AI 거버넌스를 위한 법적·제도적 기반이자 가장 큰 이정표는 2026년 1월부터 세계 최초로 시행된 AI 기본법임

AI 거버넌스를 위한 법적·제도적 기반이자 가장 큰 이정표는 2026년 1월부터 세계 최초로 시행된 '인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(이하 AI 기본법)'이다. 유럽연합(EU)의 AI Act와 유사하게 위험도에 따른 차등 규제를 적용하는 리스크 기반 규제 모델을 접목했다. 법안에 따르면 시스템을 '고영향(High-impact) AI'와 '생성형 AI'로 분류하여 차등적 의무를 부과 한다. 의료 진단, 공공자원 배분, 생체 인식 시스템 등 국민의 안전과 직결된 '고영향 AI' 개발사에 대해서는 사전 위험성 평가(Risk Assessment)와 사용자 보호 조치 이행을 법적으로 의무화했다. 또한 법안의 50% 이상이 스타트업 육성, 학습용 데이터 개방, 중소기업 AI 도입 금융 지원 등 국내 AI 생태계 지원 촉진에도 방점을 두고 있다.

정부는 정책 수립, 기술 검증, 국제 공조를 유기적으로 연결하기 위해 ① 대통령 직속 국가AI위원회 (국가 전략 및 예산·정책 심의) → ② 과학기술정보통신부 (AI 기본법 주무부처 및 규제 로드맵 수립) → ③대한민국 AI 안전 연구소 (Korea AISI) (AI 모델 안전성 상시 평가 및 글로벌 기술 공조)로 이어지는 추진 체계를 구축했다. 또한 정부는 AI 기본법의 법적 테두리 안에서 산업별 특성에 맞는 구체적인 가이드 라인을 제시하는 'AI 규제 합리화 로드맵'도 함께 추진 중이다.

2. 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(AI 기본법)

1-1) AI 기본법 시행과 규제 체계

AI기본법, 1년의 유예기간을 거쳐 2026년 1월 22일부터 전면 시행됨

2024년 12월 26일 국회 본회의를 통과한 '인공지능 발전과 신뢰 기반 조성 등에 관한 기본법'은 1년의 유예기간을 거쳐 2026년 1월 22일자로 시행령과 함께 전면 시행되었다. 본 법은 ① 국가 AI 거버넌스 정립, ② AI 산업 육성·진흥, ③ 안전·신뢰 기반 조성을 3대 축으로 하며, AI 사업자(개발·이용)를 규율 대상으로 한다. AI기본법은 인공지능사업자와 구분되는 개념으로 인공지능제품·서비스를 제공받는 자를 이용자로 정의하고 이용자를 규율 대상에서 제외하고 있다.

1-2) '고영향 AI'와 '생성형 AI'의 차별적 규제

AI 기본법의 핵심은 AI를 '고영향 인공지능'과 '생성형 인공지능' 두 가지 범주로 구분하여 차등적 의무를 부과 → 사업 형태 및 제공하는 제품·서비스 종류에 따라 중첩적으로 의무를 부담하게 될 수 있음

AI 기본법은 인공지능사업자를 대상으로 각종 의무를 부과하고 있으며, 사업 형태 및 제공하는 제품·서비스 종류에 따라서 의무를 부담하는데 경우에 따라서는 중첩적으로 의무를 부담하게 될 수 있다. AI 기본법의 핵심은 AI를 '고영향 인공지능'과 '생성형 인공지능' 두 가지 범주로 구분하여 차등적 의무를 부과하고 있다. '고영향 인공지능'란 사람의 생명·신체 안전 및 기본권에 중대한 영향을 미칠 수 있는 AI로 AI 사업자는 AI 또는 AI를 활용한 제품·서비스 제공 시 고영향 인공지능에 해당하는지 여부를 사전 검토 해야한다. 안전성·신뢰성 확보를 위해 위험관리방안의 수립·운영, 인공지능 설명 방안 수립·시행, 이용자 보호 방안의 수립·운영, 고영향 인공지능에 대한 사람의 관리·감독, 안정성·신뢰성 확보 조치 확인 문서 작성과 보관(5년) 의무를 진다. 또한 학습 누적 연산량 10^{26} FLOPs 이상의 '대규모(고성능) AI' 사업자는 위험관리체계 구축 결과를 과학기술정보통신부 장관에게 제출해야 한다.

고영향 인공지능사업자 및 생성형 인공지능사업자에게는 '투명성 확보 의무'가 부과되는데 사전 고지 의무(이용약관·서비스 화면·팝업 등을 통한 AI 운용 사실 고지), 결과물 표시 의무(워터마크 등 기계 판독 가능 표시), 딥페이크의 경우 이용자가 명확히 인식할 수 있는 표시 의무 등이 포함된다. 의무 위반시 시정명령 또는 3천만원 이하의 과태료가 부과된다.

도표 14. AI 기본법 주요 규율 대상

구분	인공지능(AI) 사업자	
형태	인공지능 개발사업자 (제2조 제7호 가목)	AI를 개발하여 제공하는 자
	인공지능 이용사업자 (제2조 제7호 나목)	인공지능개발사업자가 제공한 AI를 이용하여 AI제품·서비스를 제공하는 자
종류	고영향 인공지능사업자 (제2조 제4호, 제7호)	① 의료 진단, ② 신용평가·대출 심사, ③ 채용·인사 평가, ④ 자율주행 시스템, ⑤ 범죄 수사·생체 인식 분야 등 특정 영역에서 고영향 AI* 제품, 서비스를 제공하는 사업자 * 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험 초래할 우려가 있는 AI
	생성형 인공지능사업자 (제2조 제5호, 제7호)	입력데이터의 특성을 학습·모방하여 이미지, 영상 등 콘텐츠를 생성하는 AI 제품, 서비스를 제공하는 사업자

자료: 법무법인 세종, 신영증권 리서치센터

1-3) 국내 대리인 지정 의무(법 제36조, 영 제29조)

일정 규모 이상(전년도 매출액이 1조원 이상 또는 인공지능서비스 부문 전년도 매출액이 100억원 이상 또는 전년도 말 기준 AI제품 등에 대한 직전 3개월 간 국내 이용자 수가 1일 평균 100만명 이상) 또는 시정명령 위반으로 과태료를 부과 받은 국내에 주소 또는 영업소가 없는 인공지능 사업자는 국내에 주소 또는 영업소가 있는 대리인을 지정해야 한다. 이는 해외 빅테크에 대한 규제 사각지대를 해소하기 위한 것으로, OpenAI, Google, Anthropic 등이 한국 시장에서 서비스를 제공할 경우 직접 적용된다. 위반 시 3,000만 원 이하의 과태료가 부과되며, 향후 외국계 AI 사업자의 한국 진출 시 주요한 고려 사항이자 일부 거버넌스 비용 증가 요인으로 작용할 것이다.

1-4) 최소 1년 이상의 계도기간 운영

최소 1년 이상의 계도기간 운영 → '준비의 유예'이지 '의무의 유예'가 아니므로 법적 의무는 시행일부터 발생함에 주의

과학기술정보통신부는 기업의 혼란을 최소화하고 현장의 준비 시간을 충분히 보장하기 위하여 AI기본법 및 시행령에 따른 규제의 적용을 시행일인 2026년 1월 22일부터 최소 1년 이상 과태료 등 제재 적용을 유예할 것을 발표했다. 다만 계도기간은 '준비의 유예'이지 '의무의 유예'가 아니므로 법적 의무는 시행일부터 발생한다. 동시에 과학기술정보통신부는 기업의 원활한 법 준수를 지원하기 위해 '인공지능기본법 지원 창구'를 개설·운영하고, AI기본법 및 시행령과 관련된 구체적이고 실무적인 자문을 제공할 예정이다. 한편, 산업계 일각에서는 '고영향 AI 범위의 모호성', '표시 의무의 구체적 적용 범위' 등에 대한 추가 가이드 라인을 요구하고 있다.

2) 국가인공지능전략위원회 (대통령 직속)

국가인공지능전략위원회는 대한민국의 인공지능(AI) 정책을 총괄하는 대통령 직속 최상위 컨트롤타워(총사령탑) 기구이다. 2024년 9월 대통령령에 따라 '국가인공지능위원회'로 출범한 뒤, 2025년 9월 4일 '국가인공지능전략위원회의 설치 및 운영에 관한 규정'(대통령령) 시행과 함께 법정기구로 격상되어 현재의 명칭으로 재편되었다.

국가인공지능전략위원회는 단순히 조언을 제공하는 자문위원회를 넘어, 범국가적 인공지능 정책을 기획하고 부처 간 사업을 강력하게 조율하는 조직임

정부와 민간의 역량을 총결집하기 위해 민·관 원팀 구조로 설계되었으며, 기술 발전과 환경 변화에 맞춰 대통령 위촉 위원 포함 기존 82명의 위원 외에, 45명의 위원을 추가 위촉하여 총 127명(중복 제외)으로 확대 개편되었다. 위원회는 AI 정책 전반을 심의하며, 분과별로 운영되는데 단순히 조언을 제공하는 자문위원회를 넘어, 범국가적 인공지능 정책을 기획하고 부처 간 사업을 강력하게 조율하게 된다. 세부적인 전략 수립을 위해 분과위원회, 특별위원회, 실무 협의회 등이 있다. 10개 분과위원회, 2개 특별위원회, 법률 TF, CAIO 협의회(주요 기관 및 기업의 최고 AI 책임자 간 소통 창구) 등이 있다.

도표 15. 국가인공지능전략위원회 조직 구성

자료: 국가인공지능전략위원회, 신영증권 리서치센터

국내 AI안전연구소(AISI)는 2024년 11월 미국·영국·일본·싱가포르·캐나다에 이어 세계 6번째로 설립
→ '규제기관'이 아니라 '협력기관'임

3) AI안전연구소(AISI) – 한국전자통신연구원(ETRI) 부설

AI안전연구소(Artificial Intelligence Safety Institute, AISI)는 2024년 5월 'AI 서울정상회의' 합의의 후속 조치로, 같은 해 11월 공식 출범했다. 국내 AISI는 미국·영국·일본·싱가포르·캐나다에 이어 세계 6번째로 설립된 AI 안전 전담 공공기관으로 ETRI의 부설 조직이다. AISI는 다음 3개 실로 구성된다. 첫째, AI안전정책 및 대외협력실은 OECD GPAI 등 국제협력과 정책 연구를 담당한다. 둘째, AI안전 평가실은 AI 위험 식별·평가 프레임워크 개발 및 모델 안전성 검증을 수행한다. 셋째, AI안전 연구실은 정렬(alignment) 기술, 편향 완화 기술, 통제력 상실 대응 기술을 연구한다. 주목할 점은 AISI가 '규제기관'이 아니라 '협력기관'으로 자리매김했다는 것이다. 국내 AI 기업이 EU AI Act를 비롯한 글로벌 규제 환경에서 경쟁력을 확보하도록 R&D 단계부터 규제·인증·표준 절차를 지원하는 역할을 담당한다. 2024년 11월 출범 당시 산·학·연 24개 기관과 MOU를 체결하였고, 영국·미국·일본 AI 안전연구소와 글로벌 네트워크 협력을 구축하고 있다.

금융위는 통합된 '금융 분야 AI 가이드 라인(안)' 개정 방향을 공개함

4) 금융위원회(금융분야 AI 가이드 라인)과 금융감독원(금융분야 AI RMF)

금융 분야는 '고영향 AI'가 집중된 영역으로, 2025년 12월 22일 금융위원회는 금융권 AI 협의회를 개최하여 '금융분야 AI 가이드 라인(안)' 개정 방향을 공개했다. 그간 금융당국은 '금융분야 AI 운영 가이드 라인(2021.7월)', '금융분야 AI 개발·활용 안내서(2022.8월)', '금융분야 AI 보안 가이드 라인(2023.4월)' 등 AI 개발·운영·보안 등 각 영역별로 가이드 라인 형태의 모범규준을 제시해 왔다.

그러나 최근 생성형 AI 확산 등 인공지능 기술의 급격한 발전과 2026년 1월 시행된 AI기본법 제정에 따른 환경 변화를 반영할 필요성에 따라 기존 가이드 라인을 통합·개정하고 업무 전반에 걸친 AI 위험관리의 방향과 원칙을 담은 새로운 통합 가이드 라인(안)을 발표했다. 통합 가이드 라인(안)은 AI 활용의 7대원칙으로 ①거버넌스, ②합법성, ③보조수단성, ④신뢰성, ⑤금융안정성, ⑥신의성실, ⑦보안성을 제시하고, 이에 대한 세부이행 사항 등을 제안하고 있다. 가이드 라인(안)은 AI 기술의 빠른 발전속도, 금융분야의 AI 수용도, 관련 법·제도 환경변화 등을 고려하여 기존 가이드 라인과 마찬가지로 모범규준(Best Practice), 업권별 자율규제 형식으로 규율하면서 금융권 의견을 지속 수렴하여 상시적으로 개선·보완해나갈 예정이라고 밝혔다.

도표 16. 금융분야 AI 가이드 라인 개정안 주요내용

구분	주요 내용
거버넌스 원칙	①AI 위험관리, 윤리원칙 수립을 위한 최고 의사결정기구 설치, ②독립적 위험관리 전담조직 구축, ③AI 위험평가체계 구축 등
합법성 원칙	AI기본법, 금소법, 신정법·개보법 등 적용법규를 사전 파악하여 업무절차에 반영하고 주기적 점검·개선
보조수단성 원칙	①AI 활용에 대한 임직원 책임소재 명확화, ②임직원 개입이 필요한 상황을 차등화 규정하고 개입방법 선택·운영, ③정기적 교육 실시
신뢰성 원칙	①AI모델의 정기적 성능관리, ②AI 학습데이터 품질관리, ③AI의 공정성·편향성 점검, ④이해관계자에 대한 AI의 설명가능성 확보
금융안정성 원칙	①금융안정위험 평가, ②비상정지장치, 백업모형 등 안전장치 마련, ③제3자 IT리스크 관리, ④감독당국 보고절차 마련
신의성실 원칙	①금융소비자 이익 침해 등 이해상충 방지, ②AI 활용시 사전고지 및 서비스 오류 신고, 선택권 보장 등 소비자 보호대책 마련
보안성 원칙	①AI 보안위협 식별·관리, ②AI 특화공격 탐지·대응, ③AI자산 보호·관리, ④외부모델 및 데이터 검증, ⑤보안성 검증·운영 관리

자료: 금융위원회, 신영증권 리서치센터

금감원 금융분야 AI 위험 관리 프레임워크(AI RMF) 발표

금융감독원은 2024년 4월 금융회사 118곳을 전수 조사한 결과 은행 20곳 중 5곳(25%)만이 AI 의사결정기구(거버넌스)를 설치한 상태라고 발표했으며, 이를 토대로 2026년 1월 15일 '금융분야 AI 위험관리 프레임워크(AI RMF)' 도입해 금융회사가 AI 관련 위험을 자율적으로 관리할 수 있는 기준을 제시했다. AI RMF는 거버넌스, 위험평가, 위험통제의 세 영역으로 구성되며, 이에 따라 금융회사는 AI 위험관리를 위한 의사결정기구·조직·내규를 마련하고, 위험 평가체계를 구축하며, 위험

수준에 따른 통제를 수행하여야 한다. 또한 금융회사는 AI 위험의 체계적인 인식·측정·관리 등을 위해서위험기반 접근방법(Risk-based approach)의 종합 평가체계를 구성하고 위험 수준별 차등화된 통제·관리를 수행하고, 초고위험 AI에 대해 출시 여부 재검토 등 위험통제를 위한 제반 절차를 이행해야 한다. 본 프레임워크는 법적 강제력이 없는 자율규제 형태로 운영될 예정으로, 금융회사는 규모와 특성에 맞게 이를 탄력적으로 적용할 수 있다. 본 프레임워크는 향후 ESG 평가에서 금융권 거버넌스 영역의 중요한 정량 지표로 활용될 전망이다.

도표 17. 금융분야 AI 위험관리 프레임워크(AI RMF) 주요 내용

자료: 금융감독원, 신영증권 리서치센터

도표 18. AI RMF에 따른 AI 서비스 위험 평가 절차 예시(세부 항목 조정 가능)

부문	위험 인식·측정		위험 경감	잔여 위험			위험등급 산정
합법성 원칙 (20%)	금융소비자보호법 위반 가능성	8	-4	4	∑ 9	총합 54점	
	AI기본법 위반 가능성	4	-3	1			
	데이터 관련법 위반 가능성	4	-2	2			
	개별 업권법 위반 가능성	4	-2	2			
신뢰성 원칙 (30%)	품질	6	-4	2	∑18		
	편향성	6	-2	4			
	공정성	6	-2	4			
	설명가능성	6	-1	5			
	성능	6	-3	3			
신의성실 원칙 (20%)	계약 권리 침해	6	-3	3	∑10		
	책임 투명성	6	-3	3			
	소비자 보호방안	8	-4	4			
보안성 원칙 (30%)	보안	8	-3	5	∑17		
	안정성	8	-4	4			
	위탁/관리	8	-3	5			
	프라이버시	6	-3	3			

자료: 금융감독원, 신영증권 리서치센터

공공 분야 : 공공부문 초거대 AI 도입·활용 가이드 라인 2.0

5) 디지털플랫폼정부위원회 (2025년말 폐지됨)

디지털플랫폼정부위원회(이하 디플정위)는 2025년 4월 행정·공공기관에서 최신의 인공지능(AI) 기술을 효과적이고 안전하게 활용할 수 있도록 '공공부문 초거대 AI 도입·활용 가이드 라인 2.0'을 마련하여 중앙부처, 지방자치단체, 공공기관에 배포했다. 디플정위는 초거대 AI를 행정업무와 공공서비스에 도입 시 각 단계별로 고려해야 할 사항을 담아 '공공부문 초거대 AI 도입·활용 가이드 라인'을 2024년 4월에 처음으로 마련한 바 있으며, AI 관련 최신 기술과 정책 동향을 반영하고, 국내외 활용사례 및 성과관리 방안을 포함하여 개정안을 발표했다.

가이드 라인의 주요 내용으로는 공공부문의 인공지능 도입에 따른 국민 체감 성과를 고려하여 공공부문에서 AI 도입 시에 지침이 될 전략 목표로 ①사회문제 해결, ②대국민 서비스 혁신, ③일하는 방식 효율화를 제시하면서, 행정·공공기관에서 무분별한 AI 도입을 지양하고 전략목표에 따라 AI 과제들을 추진하도록 안내하고 있다. 특히, 공공부문 AI 과제에 대한 성과관리 방법론과 AI 성과지표를 추가하여 각 공공기관에서 다양하게 추진하고 있는 AI 활용사업들에 대한 체계적인 성과관리를 강화하도록 했다. 또한 정부의 공공 AI 도입 전략목표와 연계하여, AI 기술, 과제 특성을 반영한 투입·과정·산출·결과 지표 등을 생애주기별로 측정할 수 있도록 단계

별 성과 지표도 마련했다. 또한, 해외 주요국의 AI 도입 현황을 기능별, 정책분야별로 분석한 현황과 이를 참고할 사이트 정보(42개국, 가이드 라인 부록 참조) 및 싱가포르, 영국, 프랑스 등의 공공부문의 대표적인 서비스 사례도 안내하고 있다.

디플정위는 2022년 9월 출범했으나 이재명 정부 출범 이후 2025년 12월 최종적으로 공식 폐지 되었다. 위원회가 해체됨에 따라 기존에 추진되던 100여 개의 실행과제와 사업들은 행정안전부와 과학기술정보통신부 등 관련 부처로 이관되었다.

3. 이니셔티브 및 NGO의 AI 거버넌스

1) UNESCO AI 윤리 권고 vs 한국의 윤리기준

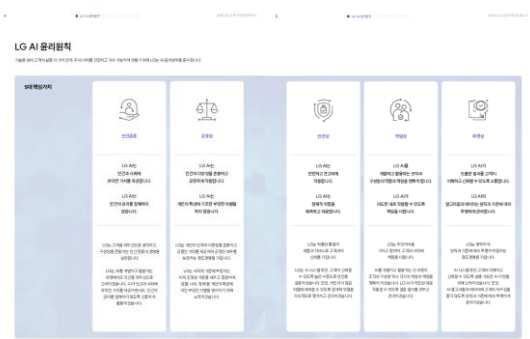
한국 정부는 UNESCO의 AI 윤리 권고(2021.11) 보다 빨리 AI 윤리 기준을 발표(2020.12)하며 준비함

한국 정부는 UNESCO AI 윤리 권고 제정 시기(2021년 11월) 보다 빠른 2020년 12월 '사람이 중심이 되는 인공지능 윤리기준'을 발표했다. 과학기술정보통신부와 정보통신정책연구원(KISDI) 주도로 마련되었으며, '인간성(Humanity)'을 최고 가치로 둔 3대 기본원칙과 10대 핵심 요건을 담고 있다.

인공지능 개발 및 활용 과정에서 고려되어야 할 원칙 3대 기본원칙으로 ① 인간존엄성 원칙 ② 사회의 공공선 원칙 ③ 기술의 합목적성 원칙을 명시했다. 또한 기본원칙을 실현할 수 있는 세부 요건 10대 핵심요건으로는 ① 인권보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성 등을 제시했다.

이후 유네스코한국위원회는 해설서를 발간하여 권고의 국내 적용을 지원하고 있다. 또한 앞서 언급한 바와 같이 LG AI연구원은 UNESCO AI 윤리 권고의 이행 현황을 매년 체계적으로 공개하고 있다.

도표 19. LG AI연구원의 2025 AI 윤리 책무성 보고서



자료: LG AI연구원, 국가인공지능전략위원회, 신영증권 리서치센터

IAAE, 2019년 10월 23일 국내 민간 최초로 '인공지능 윤리 헌장'을 제정·공포함

2) 국제인공지능윤리협회(IAAE)와 한국인공지능윤리학회(KSAIE)

국제인공지능윤리협회(International Association for Artificial Intelligence Ethics, 이후 IAAE)는 2019년 3월 창립된 '한국인공지능윤리협회(KAIEA)'를 모태로 하며, 2022년 1월 명칭을 변경하고 사단법인으로 재출범한 단체이다. IAAE는 인공지능 기술의 발전과 AI 윤리의 조화로운 진흥을 목표로 활동하는 과학기술정보통신부 산하 비영리 사단법인으로 AI에 대한 무조건적인 규제가 아닌, 인류의 행복에 기여하는 '선한 AI' 구현을 목표로 기술 개발과 윤리 체계 확립을 동시에 추진하고 있다. 2019년 10월 23일 국내 민간 최초로 '인공지능 윤리 헌장'을 제정·공포했으며, 이후 2022년 7월 '디지털 휴먼 윤리 가이드 라인', 2025년 9월 '감정 교류 AI 윤리 가이드 라인'을 잇따라 발표했다.

한국인공지능윤리학회(KSAIE)는 인공지능 기술 발전과 함께 발생하는 다양한 윤리적 이슈를 전문적으로 연구하는 국내 대표 학술 단체로 AI 기술 관련 윤리적 이슈에 대한 학문적 연구를 통해 AI 기술의 순기능을 증진하고 사회 공익에 기여를 목적으로 하는 학회로, 정부·기업의 AI 정책 자문에 적극 참여하고 있다.

3) 한국 with G7·OECD·UN 차원의 글로벌 협력

한국은 G7 옵저버 자격으로 참여하고 있으며, OECD AI 정책저장소 활용 및 GPAI에서도 적극적인 역할을 수행하고 있다. 외교부는 2025년 7월 1일 '국제 AI와 기후변화 컨퍼런스'를 서울에서 개최하여 한국, 미국, 일본, 프랑스, 독일 등 9개국에서 초청된 전문가 및 UNEP 등 국제기구 인사를 포함해 450여 명 규모의 국제 협력의 장을 마련하기도 했다.

4. 국내 주요 기업의 AI 거버넌스 도입 사례

국내 기업의 AI 거버넌스 도입은 아직은 초기 단계 수준임

국내에서도 기업들이 사업 영역에 AI를 접목 또는 활용을 적극 시도하고 있다. 이에 따라 일부 기업들은 AI 거버넌스를 도입해 운영하고 있다. 국내 AI 거버넌스 도입 현황을 살펴보면 일부 그룹사의 경우 그룹차원에서 AI 거버넌스를 도입하고 이를 계열사들도 도입하도록 독려하고 있는 것으로 나타났다. 또한 업종별로 살펴보면 AI를 적극 도입하고 있는 통신, 인터넷 플랫폼, SI, 금융 업종 등이 AI 거버넌스 도입에 적극적인 것으로 나타났다.

당사가 조사한 바에 따르면 KOSPI200내 기업들 중 AI 원칙을 도입하고 AI거버넌스 조직을 실제 운영되고 있는 회사는 총 30개가 되지 않아 아직까지는 국내에서는 AI거버넌스 도입은 초기 단계인 것으로 판단된다. 일부 기업의 경우 별도의 AI (윤리) 원칙을 도입하지 않고 기존 윤리 원칙을 기본으로 AI 거버넌스 조직을 만든 곳도 있었으며, 일부 기업의 경우 AI (윤리) 원칙을 도입하기 위해 먼저 AI 거버넌

스 조직을 만든 곳도 있었다. 또한 자체적인 AI 원칙과 AI 거버넌스 조직을 두지 않고 그룹차원의 AI 윤리 원칙을 그대로 수용하고 그룹내 계열사의 도움으로 이를 운영하는 곳도 존재했다. ESG의 G 영역에서 매우 중요한 평가 요소로 ESG 위원회가 도입된 경험이 있어 기업들이 향후 AI 거버넌스 구축에도 잘 대응할 것으로 기대된다. 이에 다음은 국내 기업 중 AI 거버넌스 도입한 일부 기업에 대해 현황을 소개하고자 한다.

도표 20. 국내 주요 기업 AI 원칙 및 AI거버넌스 조직 도입 현황

기업명	AI(윤리)원칙	AI 거버넌스 조직 유무	기업명	AI(윤리)원칙	AI 거버넌스 조직 유무
삼성전자	O	O	LG에너지솔루션	O	O
삼성바이오로직스	O	O	KB금융	O	O
LG전자	O	O	신한지주	O	O
NAVER	O	O	하나금융지주	O	O
SK텔레콤	O	O	우리금융지주	O	O
LG	O	O	HD현대	O	O
기업은행	O	O	카카오	O	O
KT	O	O	삼성에스디에스	O	O
LG씨엔에스	O	O	DB손해보험	O	O
LG유플러스	O	O	LG생활건강	O	O
크래프톤	O	O	NC	O	O
카카오페이	O	O	카카오뱅크	O	O
롯데쇼핑	O	O	넷마블	O	O
롯데케미칼	O	O	롯데지주	O	O
롯데칠성	O	X	롯데정밀화학	O	X
롯데웰푸드	O	X	현대차	X	O
기아	X	O	현대오토에버	X	O
HD한국조선해양	X	O	한국전력	X	O
한화생명	X	O	현대백화점	X	O

자료: 각 사 홈페이지, 언론, 신영증권 리서치센터

1) 삼성전자의 AI 거버넌스

삼성전자 - AI 윤리협의회와 지속가능경영위원회 연계

삼성전자의 AI 거버넌스는 전 세계 수억 대의 하드웨어 기기와 소프트웨어를 연결하는 '멀티 디바이스 지능' 생태계에서 안전하고 신뢰할 수 있는 인공지능 기술을 제공하기 위한 전사적 내부통제 및 리스크 관리 체계로 작동하고 있다. 금융지주사들이 주로 금융 규제와 신용 리스크 대응에 집중하는 반면, 글로벌 제조·기술 기업인 삼성전자는 온디바이스 AI(On-device AI)의 보안, 전 세계 법 규제 준수, 그리고 글로벌 눈높이에 맞춘 윤리 기준 정립에 초점을 맞추고 있다.

삼성전자는 '공정성', '투명성', '책임성'을 3대 핵심 가치로 하는 AI 윤리 원칙을 수립·공표했으며, 구체적 거버넌스 체계는 다음과 같이 구성된다.

- AI 윤리협의회: 삼성리서치, 컴플라이언스팀 등 유관 부서로 구성된 협의체. 자가진단 체크리스트 제공·운영, AI 편향성 모니터링·경고 기능, AI 윤리 이행 가이드 및 임직원 교육 운영을 담당.
- 이사회 산하 지속가능경영위원회 체계: AI 윤리협의회의 중요 사안은 지속가능경영위원회에 보고·논의되며, 이는 이사회 단위에서 AI 거버넌스가 작동함을 의미.
- 산업 인공지능 표준화 포럼 참여: 한국 정부 주관 포럼에 참여하여 학계·연구기관·산업계 전문가와 AI 신뢰성 평가기준·윤리 가이드 라인 수립 협력.

삼성전자는 가전과 스마트폰 내부에서 AI가 직접 연산되는 온디바이스 AI의 비중이 커짐에 따라, 독자적인 온디바이스 AI Safety Framework를 운용하고 있다. 이는 클라우드 AI 연산 시 발생할 수 있는 데이터 유출을 원천 차단하기 위해 개인정보는 기기 내부(온디바이스)에서만 처리되도록 제어하고 있다. 하드웨어 보안 플랫폼인 '삼성 녹스(Knox)'를 AI 거버넌스의 핵심 인프라로 삼아, 적대적 공격(Adversarial Attack)이나 외부 해킹으로부터 AI 모델과 학습 데이터를 실시간 보호하도록 설계되어 있다.

또한 유럽연합(EU)의 AI법(AI Act)을 비롯해 전 세계 주요국의 인공지능 관련 법제화에 선제 대응할 수 있도록 상시 모니터링 및 규제 준수(Compliance) 시스템을 통해 글로벌 AI 규제 대응하고 있다. 또한 기업용(B2B) 솔루션 분야에서는 계열사인 삼성SDS의 AI 거버넌스 프레임워크 및 플랫폼을 활용하여, 기업 고객들이 안전하게 생성형 AI를 도입하고 알고리즘의 편향성과 신뢰성을 실시간으로 추적·평가할 수 있는 시스템을 지원하고 있다. 2024년부터 사내 AI 개발자·전문가 대상 AI 윤리 교육 과정을 신설하였고, 2025년부터 전 임직원 대상 의무 교육으로 확대 운영하고 있다. 또한 2023년 ChatGPT에 회사 기밀 코드 유출 사례를 계기로 사내 생성형 AI 활용 가이드 라인을 정립한 바 있다.

2) LG AI연구원의 AI 거버넌스

LG AI연구원 - UNESCO 권고 이행 공개 및 K-EXAONE

LG는 2022년 'LG AI 윤리원칙'을 발표하며 ① 인간존중, ② 공정성, ③ 안전성, ④ 책임성, ⑤ 투명성을 5대 핵심 가치로 제시했다. LG AI연구원은 2023년부터 매년 'AI 윤리 책무성 보고서(Accountability Report)'를 발간하며, 전 세계 기업 중 유일하게 UNESCO AI 윤리 권고 이행 현황을 매년 체계적으로 공개하고 있다. 2026

년 2월 19일 발간된 '2025 AI 윤리 책무성 보고서'에는 다음과 같은 핵심 성과가 담겨 있다. 첫째, '소버린 AI(Sovereign AI)' 구현 역량 입증으로 K-EXAONE은 한국적 맥락에서 높은 신뢰도를 나타냈으며 상업적 안전성이 확인되었다. 둘째, 딥페이크 등 미래 위협에 대한 선제 대응 체계를 구축했다. 셋째, 새로운 위험분류체계를 단순 가이드 라인이 아닌 AI 모델·서비스 안전성 검증 도구로 개발하여 EXAONE에 실제 적용하고 그 결과를 투명하게 공개했다.

LG AI연구원은 글로벌 AI 거버넌스 논의에서 한국 대표 기업으로서 입지를 공고히 하고 있다. 서울·파리·뉴델리에서 3차례 연속 'AI 정상회의'에 초청되었고, UNESCO와 공동으로 코세라(Coursera) 플랫폼 기반 'AI 윤리 MOOC'를 추진 중이다. 또한 하버드대, 뉴욕대, 노트르담대, 유엔대학교, 모질라 재단, 세계과학기술윤리위원회(COMEST) 등과 협력 네트워크를 보유하고 있다.

3) SK텔레콤의 AI 거버넌스

SK텔레콤- Global AI Company로의 진화를 이끌 AI 피라미드 전략이 강한 실행력을 가질 수 있도록 AI 거버넌스를 정립함

SK텔레콤은 AI 기술의 위험성에 선제적으로 대응하여 지속 가능한 발전을 도모하기 위해 고객 비즈니스 보호, 고객 가치 제공, 글로벌 확장 등의 AI 거버넌스 3대 목표를 설정하고 이를 달성하기 위해 노력하고 있다. 2024년 3월 주주총회에서 공개된 AI 기본 원칙인 'T.H.E.AI'는 통신 기술 기반의(Telco), 사람을 향한(Humanity), 윤리적 가치 중심(Ethics)의 AI를 뜻한다. 또한 종합적인 리스크 및 기회 대응을 위해 3가지 관리체계를 운영하며, 2024년 4월 인공지능경영시스템 국제표준인 ISO/IEC 42001 인증을 획득하여 대외적으로 신뢰성을 인정받고 있다.

또한 SK텔레콤은 AI 리스크 관리를 통한 책임 경영을 강화하기 위해 최고 의사결정 기구부터 실무진까지 유기적인 거버넌스 체계를 구축하고 있으며 다음과 같이 구성되어 있다.

- 의사결정기구: 이사회는 투자 및 사업계획 등 핵심 사안의 최종 의사결정을 내리며, 산하 ESG위원회는 AI 리스크 평가 및 대응 전략을 심의
- 경영진: CEO와 CGO(최고거버넌스책임자)가 주도하여 정책과 이슈를 관리·감독하고 ESG위원회를 지원
- 실무진: AI 거버넌스팀, 고객정보보호팀(레드팀), AI 거버넌스 워킹그룹이 전략 수립, 리스크 및 취약성 탐지, 부서 간 협업을 실행
- AI 거버넌스 체계 관리: 외부 규제 및 트렌드 모니터링, 대외 소통(CR/PR) 및 임직원 교육
- AI 사업 관리: RCSA(위험고객통제 자가진단) 기반 프로세스를 통해 상품의 생애주기별 위험을 평가하고 비즈니스 파트너(협력사 등)를 관리
- 이슈 관리: 장애 및 고객 불편 발생 시 초기 대응 프로토콜을 가동하고 개선계획 수립.

4) 네이버의 AI 거버넌스

네이버 – 5대 AI 윤리 준칙 및 NAVER ASF

네이버는 '사람을 위한 일상의 도구'를 모토로 5대 AI 윤리 준칙(사람을 위한 AI 개발, 다양성의 존중, 합리적인 설명과 편리성의 조화, 안전을 고려한 서비스 설계, 프라이버시 보호와 정보 보안)을 수립·운영하고 있다. 네이버 AI 윤리 준칙은 네이버가 인공지능 기술의 개발과 이용 과정에서 인간 중심의 가치를 지키기 위해 서울대학교 AI 정책 이니셔티브(SAPI)와 공동 연구를 통해 수립한 5가지 핵심 원칙이다.

'네이버 AI 윤리 준칙'을 발표한 이후 2024년, 네이버는 NAVER ASF(AI Safety Framework) 를 통해 사회에서 우려하는 AI Safety와 관련한 위험을 대응하는 체계를 구체화 했다. NAVER ASF는 AI Safety와 관련해 사회에서 우려하고 있는 위험에 대응하기 위한 체계로 이를 통해, 네이버는 AI 시스템의 개발 및 배포 프로세스의 전 단계에서 관련된 위험을 인식, 평가 및 관리한다. 또한, AI 기술 발전에 따른 AI Safety 관련 글로벌 논의 흐름에 맞춰, ASF를 지속적으로 개선하며 업데이트해 나갈 것이라고 밝히고 있다.

NAVER ASF는 네이버 AI 윤리 준칙을 준수하는 네이버 구성원이 산업 현장에서 AI 시스템을 개발하고 배포하는 과정에서 AI Safety를 구체적으로 실천하기 위한 체계이다. 이를 통해 네이버는 사회에서 우려하고 있는 AI에 대한 통제력 상실 위험과 악용 위험에 대처하고자 한다고 밝히고 있다. 통제력 상실 위험에 대해서는 AI 위험 평가 스케일을 통해 대응하고, 악용 위험에 대해서는 AI 위험 평가 매트릭스를 통해 위험을 관리한다. 특히, 이를 실천하는 과정에서 사회기술적 맥락을 고려한 AI 시스템을 만들기 위해서 노력하고 있다고 밝히고 있다.

주목할 점은 네이버가 '소버린 AI' 철학을 명시적으로 채택하고 있다는 점이다. 네이버는 세계에서 3번째로 자국어 중심 초대규모 소버린 언어모델과 산업생태계를 만들어가면서, 문화적, 지정학적 상황과 지역적 이해가 AI의 성능과 Safety에도 영향을 미칠 수 있다는 점을 알게 됐다고 설명하며, 분단 국가의 특성상 국가 안보·데이터 보안에 민감한 국내 기업·공공기관이 안심하고 사용할 수 있는 AI 솔루션을 제공하는 것을 핵심 차별화 요소로 강조하고 있다.

5) 카카오의 AI 거버넌스

카카오 - 카카오 알고리즘 윤리 헌장과 Kakao ASI

카카오는 2018년 1월 국내 기업 최초로 제정된 기술윤리 규범인 '카카오 알고리즘 윤리 헌장'을 발표했으며, 2022년 7월에는 국내 기업 최초의 그룹 단위 기술윤리 위원회인 '카카오 그룹 기술윤리 위원회(Tech for Good Committee)'를 출범시켰다. 카카오 알고리즘 윤리 헌장은 AI 알고리즘 기술이 인간의 도덕적 가치와 조화를

이루고 인류의 편익을 높여야 한다는 전제 하에 사회 윤리 및 편익 추구, 차별과 편향의 경계, 학습 데이터 운영, 알고리즘의 독립성, 설명 의무와 투명성, 포용성과 이용자 주체성 등으로 구성되어 있다. 카카오 그룹 기술윤리 위원회는 윤리 헌장이라는 선언적 규범을 실제 서비스에 통합적으로 적용하고 점검하기 위해 설립된 전사 거버넌스 기구로 현재는 실질적인 리스크 선제 대응력을 높이기 위해 CA협의체 ESG위원회 산하의 '그룹 기술윤리 소위원회' 체제로 재편 및 고도화되어 운영 중이다.

카카오의 AI 거버넌스는 AI 시스템의 전 생애주기에서 발생할 수 있는 위험을 관리하고 안전성을 확보하기 위한 전사적 의사결정 체계로, 카카오는 이를 위해 'Kakao ASI(Kakao AI Safety Initiative)'라는 독자적인 리스크 관리 프레임워크를 구축하여 대표이사 및 이사회 차원에서 엄격하게 관리하고 있다고 밝히고 있다. Kakao AI 윤리 원칙은 사회 윤리, 포용성, 투명성, 프라이버시, 이용자 보호 등 기술 개발 시 실천해야 할 7가지 핵심 기준으로 이루어져 있다. 카카오는 리스크를 시스템 오류나 프롬프트 공격 같은 '기술적 리스크'와 혐오·차별 표현 같은 '윤리적 리스크'로 분류하며 식별 → 평가 → 대응의 3단계 순환 구조의 리스크 관리 체계를 통해 선제 대응하도록 설계되어 있다.

카카오 AI거버넌스는 AI 관련 리스크가 전사 조직이나 실무진 단에서 묻히지 않도록 다음과 같이 총 3단계의 컨트롤타워 구조로 되어 있다.

- AI 세이프티(Safety) 조직: AI 모델 및 서비스의 안전성을 직접 테스트하고 하는 실무 전담 부서
- ERM(전사 리스크 관리) 위원회: 기술적 판단을 넘어 데이터 활용, 정책적·제도적 영향력까지 다각도로 위험성을 검토하는 전사 위원회
- 경영진 및 이사회: 최종 승인 및 책임을 담당하는 최고 의사결정 기구로, 대표이사 주도하에 거버넌스 실행력을 확보하고 있음

카카오는 거버넌스 원칙이 구호에 그치지 않도록, 거대언어모델(LLM) 등이 유해하거나 부적절한 답변을 출력하지 못하게 막는 자체 가드레일 모델을 개발해 적용 중으로 AI 서비스를 개발할 때 계획/설계 → 데이터 수집 → 모델 개발 → 운영/모니터링 등 전 과정에서 윤리 원칙을 준수했는지 점검하는 체크리스트를 의무화하고 있다. 또한 계열사인 카카오뱅크의 경우, 국내 금융권 최초로 인공지능경영시스템(ISO/IEC 42001) 인증을 획득하고 'AI 거버넌스 2.0'을 추진하며 신뢰성을 강화하기 위해 노력하고 있다.

6) 금융권 – AI 거버넌스 도입에 가장 적극적인 업종

고영향 AI가 집중된 금융업은 금융위원회의 금융분야 AI 가이드 라인과 금융감독원의 금융분야 AI RMF 시행으로 AI 거버넌스 도입에 가장 적극적인 업종이다. 이를 주도하고 있는 4대 금융 지주 현황을 살펴보면 다음과 같다.

가) KB금융지주

KB금융은 4대 금융지주 중 가장 선제적으로 AI 거버넌스를 구축한 사례로 평가받는다. 2024년 7월 기존 '데이터AI본부'를 'AI데이터혁신본부'로 확대 개편하며 그 산하에 AI비즈혁신부, 금융AI센터, 데이터지원부, 마이데이터부 4개 부서와 직속으로 'AI 거버넌스팀'을 운영하고 있다. 2025년 3월 KB국민은행은 그룹 단위 AI 거버넌스 체계를 정식 시행하였으며, 이는 금융감독원이 제시한 '금융권 표준 AI위험관리체계'를 상당 부분 선제적으로 반영한 것으로 소개하고 있다.

KB금융은 그룹 AI 전략으로 'KB with AI'를 수립하고, 표준형 생성형 AI 플랫폼 'KB GenAI 포털'을 축으로 은행·증권·손보·카드·생명 등 주요 계열사의 AI 도입을 통합 관리한다. 'AI 윤리위원회'는 AI의 윤리적 활용과 위험관리 최고 의사결정 기구로 운영되며, '그룹 AI기본법 대응 협의체'를 통해 기본법·하위법령 대응 과제를 도출하고 있다. 고영향 인공지능 사업자의 표시·고지의무 해설서를 자체적으로 마련하여 행내 AI 서비스 고지 의무 이행에 활용하고 있다.

나) 신한금융지주

신한금융은 'AI 네이티브 컴퍼니' 비전을 천명하며, 2025년 10월 조직개편을 통해 AX혁신그룹을 신설했다. 2024년 4월부터 거버넌스 수립에 착수해 현재 AX디지털총괄부에서 AI 거버넌스 기능을 담당하고 있다. 15개 전 계열사를 대상으로 '1부서 2 AI 에이전트' 개발을 추진 중이며, 계열사별 내부 경진대회와 그룹 차원 핵심 사례 선정·공유 체계를 운영하고 있다. 또한 준법·정보보호·소비자보호·리스크 관리 담당 임원(C-Level)들이 대거 참여하는 임원급 협의체 운영을 통해 '그룹 협의체'를 구성해 AI 활용 전반에 대한 정책을 다방면으로 공유하고 조율 중이다. 신한금융지주는 지주회사가 그룹 전체가 지켜야 할 가이드를 우선 정의하고, 이에 맞춰 신한은행, 신한카드, 신한투자증권, 신한라이프 등 주요 계열사가 각 사의 업무 특성에 맞는 내규와 매뉴얼을 수립하여 연계하는 단계별 내규 연계를 도모하고 있다.

다) 하나금융지주

하나금융은 AI 거버넌스의 뼈대가 되는 하나금융 AI 윤리강령을 선포하고 포용과 공정성, 안전과 책임, 투명성, 데이터 관리, 프라이버시 보호 등의 5대 원칙을 모

든 서비스 개발과 운영의 기준으로 삼고 있다. 하나금융은 AI 거버넌스를 지주사뿐만 아니라 은행, 증권, 카드 등 전 계열사로 확산하며 일관된 내부통제를 적용하고 있다. 이와 함께 은행·증권 등 주요 계열사 전반에 AI 기술을 접목하고, 특히 영업 현장 위주로 AI 활용을 확대하고 있다. 그룹 디지털부문 산하의 데이터본부를 AI데이터본부로 확대 개편하여 지주사 컨트롤 타워로서 그룹 전체의 AI 전략과 거버넌스를 총괄하게 했다. 또한 주력 계열사인 하나은행은 금융AI부를 신설하여 거버넌스 실무 총괄, 신기술 도입, 알고리즘 모델 구현을 전담하며, 데이터전략부를 통해 데이터 거버넌스를 분리 지원하고 있다. 또한 관계사 AI 담당 임원들이 참여하는 '하나 AI 리더스 포럼'을 통해 법적·윤리적 위험관리 가이드 라인을 공동으로 조율하고 지속 가능성을 점검하도록 되어 있다.

하나금융은 2018년에는 국내 금융그룹 중 유일하게 독자적 AI 연구 조직인 '하나금융융합기술원'을 설립해 ▲데이터사이언스(신용평가, 손님관리, 이상거래탐지) ▲자산관리(AI Quant) ▲자연어 처리 ▲컴퓨터 비전 ▲AI 플랫폼 등 금융 관련 AI 주요 분야를 직접 연구하고 AI 내재화를 실행하고 있다.

라) 우리금융지주

우리금융은 단순 기술 도입 단계를 넘어 AI가 업무를 직접 수행하는 '에이전트 뱅킹'시대를 대비하여 거버넌스 내부 통제를 한층 더 엄격하게 세분화한 '의사결정 및 통제 조직'을 구성하고 있다. 현업 사업 부서가 서비스를 오픈하기 전, 'AI 위험관리 전담조직'이 먼저 위험성을 검증하고 위험 등급을 자율 분류하고 검증 과정에서 편향성, 오류 가능성이 제기되거나 고위험 서비스로 분류될 경우, AX혁신그룹장이 주관하는 'AI윤리위원회'의 엄격한 최종 승인을 통과해야만 출시가 가능하도록 하고 있다. 또한 현업 부서의 1차 점검 결과에 오류, 편향, 설명 가능성 이슈가 없는지 리스크 전담 조직이 다시 확인하는 크로스 체크(Cross-check) 체계를 갖추고 있다.

우리금융은 지주 회장 직속의 컨트롤타워 'AX 추진위원회'를 설치하여 지주와 계열사 간 유기적인 AI 전략 실행을 추진하고 있다. 우리금융은 계열사별로 흩어져 있던 통제 점검 항목(CSA)을 표준화하고, 생성형 AI를 활용해 점검 결과를 지주사로 실시간 보고하는 '표준 엔진 내부통제' 시스템을 전사로 확대 적용하고 있다. 또한 금융 보안 규제와 데이터 유출 리스크를 원천 차단하기 위해, 내부 업무망과 완벽하게 분리된 독립 환경에서 AI 연구와 검증이 가능한 'AI 연구환경 구축사업(HelpNow AI Foundry 연계)'을 추진중이다. 2026년에는 'AX 마스터플랜'을 통해 경영 전반을 AI 중심으로 재편하고, 은행·비은행 계열사의 AI 활용 과제를 순차적으로 실행한다는 방침이다.

5. ESG 관점에서 AI 거버넌스 통합

1) ESG 3축별 AI 거버넌스 현안과 통합

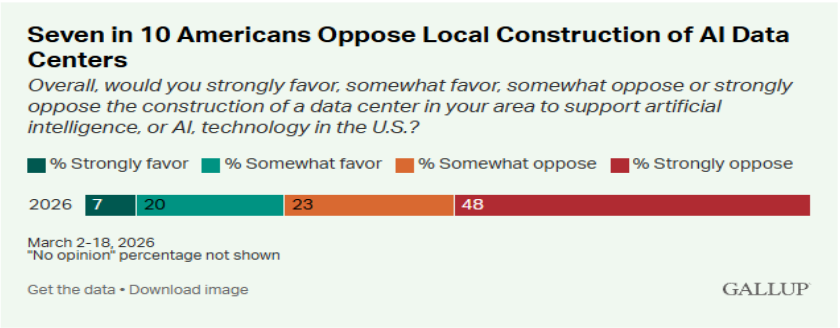
E(Environmental) - AI의 환경 발자국과 규제 대응

가) E(Environmental) – AI의 환경 발자국과 규제 대응

AI의 환경적 영향은 ESG 거버넌스의 가장 긴급한 현안 중 하나로 이중 에너지 수요 급증이 가장 시급한 이슈이다. 대규모 AI 모델은 방대한 컴퓨팅 파워를 필요로 해 에너지 소비를 기하급수적으로 증가시키고 있다. 일부 추정치에 따르면 2030년까지 글로벌 데이터센터가 전 세계 전력 소비의 최대 10%를 차지할 것으로 예측되고 있다. 에너지뿐만 아니라 수자원 문제도 부상하고 있다. 기후 변화에 따른 수자원 중요성이 부각된 가운데 데이터센터 냉각을 위한 수자원 사용 증가는 지역사회의 수자원 접근성에 영향을 미쳐 사회적 갈등을 야기하고 있다.

지난 5월 27일(현지시각) 미국 지속가능성 전문매체인 트렐리스(Trellis)는 여론조사기관인 갤럽의 조사 결과를 인용해 미국인의 71%가 거주 지역 내 데이터센터 건설에 반대한다고 보도했다. 이는 원전 반대율(53%)보다 18%포인트 높은 수치로 강한 반대(48%)가 찬성(27%) 비중을 크게 웃돌았다. 반대의 이유로는 주로 과도한 물과 전력 사용, 소음과 대기·수질 오염 등을 꼽았다. 특히 미국 에너지정보청(EIA)이 올해 데이터센터 등 상업용 전기 수요 확대로 가정용 전기 요금이 5% 상승할 것이란 전망을 내놓은 가운데 중간 선거를 앞두고 전력 비용 문제가 쟁점으로 떠오르고 있다. 갤럽은 데이터센터를 둘러싼 갈등이 지역·주 단위 선거의 주요 쟁점으로 부상할 수 있다고 분석했다.

도표 21. 미국인의 71%가 거주 지역 내 AI 데이터센터 건설에 반대



자료: Gallup, 언론 재인용, 신영증권 리서치센터

AI 인프라 투자의 '사회적 인허가(Social License to Operate)' 리스크가 실질화 되고 있음

이와 같은 주민 반대는 이미 투자 리스크로 번지고 있다. 공급망·정치 리스크 분석기관 10a랩스의 DataCenterWatch.org에 따르면 2025년 한 해에만 48개 프로젝트, 1560억달러 규모의 데이터센터 개발이 주민 반대로 차단되거나 지연됐다. 프로젝트 취소 건수는 2024년 6건에서 2025년 25건으로 급증했으며, 특히 올해 1분기에만 추가로 20건 이상이 취소되며 분기 기준 역대 최고치를 기록했다고 밝혔다.

반대 여론도 조직화되고 있다. 현재 미국 40개 주에서는 188개 주민 반대 단체가 활동 중이며, 12개 주에서는 신규 데이터센터 허가를 제한하거나 중단하는 법안이 발의됐다고 한다. 이 사례는 AI 인프라 투자의 '사회적 인허가(Social License to Operate)' 리스크가 실질화 되고 있음을 보여준다.

ESG 관점의 환경 차원에서 AI 거버넌스 통합: 주요 ESG 공시 지표와 법안에서 에너지와 물 사용량 공시 및 평가를 요구

이에 ESG 관점의 환경 차원에서 AI 거버넌스는 다음과 같이 통합되고 있다. 먼저 주요 ESG 공시 지표인 CSRD·ESRS가 AI 에너지 사용의 막대한 전력 소비와 탄소 배출 문제를 담은 이중 물질성 평가를 요구하고 있다. 또한 가장 대표적인 AI거버넌스 법안인 EU AI Act가 고위험 AI 시스템에 대한 에너지 소비 문서화를 의무화하고 있다. 또한 파리 AI 정상회의에서 출범한 91개 파트너 환경 지속가능성 연대(Environmental Sustainability Coalition)는 AI의 환경 영향 해결을 공식 거버넌스 의제로 격상시켰다.

나) S(Social) – 알고리즘 편향, 노동, 디지털 격차 vs 규제 대응

ESG 관점에서 AI 거버넌스 통합이 가장 더디게 진행되고 있는 분야는 사회(S) 영역임

ESG 관점에서 AI 거버넌스 통합이 가장 더디게 진행되고 있는 분야는 사회(S) 영역이다. 많은 기관들이 사회적 차원에서의 ESG 통합을 주장하고 있으나 현실적인 이유로 가장 더디게 진행되고 있다. 사회적 차원에서 AI 거버넌스의 핵심 과제는 알고리즘 편향, 노동 대체, 디지털 격차, 감시 기술의 남용이다. 이와 관련해 가장 대표적인 AI거버넌스 법안인 EU AI Act는 채용 심사, 신용 평가, 교육 접근 관련 AI를 고위험으로 분류하여 편향 방지 의무를 부과하고 있다. 또한 UNESCO의 AI 윤리 권고는 취약 집단 보호와 사회적 영향 평가를 명시적으로 요구하고 있다.

한편 OECD 보고서는 AI의 '재배치 효과(Reinstatement Effect)'가 노동 시장 구조를 근본적으로 변화시킬 것이라고 분석하며, 디지털 역량 격차 확대와 소득 양극화 리스크를 경고하고 있다. ILO도 2025년 AI와 직업 안전 보고서에서 AI가 특정 직업에서 새로운 물리적·심리적 리스크를 창출하고 있음을 지적했다. 디지털 격차 측면에서 UN 보고서는 현재 118개국이 어떠한 유의미한 국제 AI 거버넌스 이니셔티브에도 참여하고 있지 않다고 지적하며, AI 거버넌스의 글로벌 포용성 확보가 긴급한 과제임을 강조하고 있다. 또한 UN AI 자문기구는 개발도상국 대상 AI 역량 개발 펀드의 필요성을 권고하기도 했다.

G(Governance) - 알고리즘 투명성, 이사회 책임, 규제 준수 → 전사적 전략 이슈로 격상

다) G(Governance) – 알고리즘 투명성, 이사회 책임, 규제 준수

거버넌스 차원에서 AI 거버넌스 통합은 점점 더 중요한 분야로 떠오르고 있다. 특히 AI 기업에 대한 투자자의 관심은 알고리즘 감사 프로세스의 유무, AI 이사회 감독 체계, 규제 컴플라이언스 현황, 데이터 거버넌스 구조로 확대되고 있기 때문이

다. EU AI Act는 이를 법적 의무로 명시했으며, ISO/IEC 42001은 기업 AI 거버넌스의 인증 가능한 기준을 제공하고 있다. 또한 CIO·CISO·AI 책임자·CCO·CRO·CAE 등 다양한 C-Level 임원들은 EU AI Act 컴플라이언스의 실무적 이행이라는 공통 과제를 안게 되었는데, 이는 AI 거버넌스가 특정 부서의 기술 이슈를 넘어 전사적 전략 이슈로 격상되었음을 의미한다.

일부 언론에 따르면 상당수의 기업이 AI 거버넌스 관련 윤리 프레임워크를 개발 중이라는 조사 결과는 기업 거버넌스 차원에서 AI 거버넌스 통합이 가속화되고 있음을 의미한다. 그러나 프레임워크 채택과 실질적 이행 간의 격차가 여전히 큰 만큼 거버넌스 차원에서의 AI 통합은 투자자의 주요 평가 항목으로 작동할 것이다.

라) ESG-AI 프레임워크 → ESG 관점에서 AI 구축·평가 위한 핵심 가이드 라인

ESG-AI 프레임워크는 기업의 ESG 목표를 달성하기 위해 인공지능 기술을 통합하고 평가하는 지침으로 AI 모델의 생애주기 전반에 걸쳐 발생하는 고유한 위험과 기회를 다룸

최근에는 ESG-AI 프레임워크에 대한 논의가 이루어 지고 있다. 이 프레임워크는 기업의 환경(E), 사회(S), 지배구조(G) 목표를 달성하기 위해 인공지능 기술을 통합하고 평가하는 지침으로, 인공지능(AI) 기술의 발전과 활용 과정에서 지속 가능한 경영을 융합하여 환경(E)·사회(S)·지배구조(G) 관점에서 신뢰할 수 있는 AI를 구축하고 평가하기 위한 핵심 가이드 라인이다. 기존의 ESG가 주로 제조 공장의 탄소 배출, 노사 관계, 인권 등에 집중했다면, ESG-AI 프레임워크는 AI 모델의 생애주기 전반(데이터 학습부터 실제 서비스 운영까지)에 걸쳐 발생하는 고유한 위험과 기회를 다룬다는 차이점이 있다.

ESG-AI 프레임워크: 일부 투자자들은 특정 기업에 투자할 때 그 기업이 AI 기술을 지속 가능한 방식으로 쓰는지 검증하기 위해 이 프레임워크를 '책임 있는 AI 평가 지표'로 활용

이 프레임워크는 AI를 활용해 ESG 데이터를 정밀 분석하고, 동시에 AI 기술 자체가 윤리적이고 지속가능하도록 관리하는 융합 체계로 AI를 도구로 사용하여 기업의 ESG 성과(예: 에너지 소비 예측 분석)를 개선하는 'ESG를 위한 AI(AI for ESG)'와, AI 자체를 윤리적이고 친환경적으로 만드는 'AI의 ESG(ESG of AI)' 개념을 모두 포괄하고 있다. 일부 투자자들은 특정 기업에 투자할 때 투자대상 기업이 AI 기술을 지속 가능한 방식으로 쓰는지 검증하기 위해 이 프레임워크를 '책임 있는 AI 평가 지표'로 활용하기도 한다. 각 축의 구체적인 구성과 핵심 역할은 다음과 같다.

도표 22. ESG-AI 프레임워크의 3축 구조

축(영역)	핵심 개념	주요 내용	세부 실행 요소
환경 (Environmental)	친환경 두뇌 능력 및 에너지 최적화	AI 모델의 학습과 운영 과정에서 발생하는 환경 부담을 최소화하고, AI를 활용해 기후 위기에 대응	• 그린 AI(Green AI) 구현: 초거대 생성형 AI의 전력 소비 및 탄소 배출 저감 • 기후 변수 모니터링: 탄소 배출량 실시간 추적 및 공급망 환경 리스크 관리 • 자원 효율화: 제조·데이터센터 에너지 소비 최적화
사회 (Social)	책임감 있는 AI와 포용성	AI 기술이 인간 생태계와 노동 환경에 미치는 사회적 영향을 관리하고 신뢰성을 확보	• 알고리즘의 공정성(Fairness): 성별·인종·학력 편향 제거 및 차별 없는 결과 도출 • 개인정보 및 권리 보호: 데이터 오남용 방지 및 프라이버시 보호 체계 구축 • 디지털 포용성: 고령층·취약계층도 쉽게 접근 가능한 인간 중심 UI/UX 제공
지배구조 (Governance)	투명성과 기술 통제 체계	AI 시스템의 의사결정 과정을 투명하게 공개하고 법적·윤리적 규제 준수를 보장	• 설명 가능한 AI(XAI): AI 의사결정 근거를 이해 가능하도록 공개 • AI 윤리 위원회 운영: 전 수명 주기 감독 조직 및 책임 체계 구축 • 글로벌 규제 준수: EU AI Act 등 AI 규제 및 ESG 공시 의무화 대응

자료: UN, 신영증권 리서치센터

2) ESG 평가기관의 AI 거버넌스 통합 – AI를 '도구이자 평가 대상'으로

가). MSCI ESG Research

다양한 ESG 평가기관들이 AI 거버넌스를 ESG 평가에 도구이자 평가 대상으로 활용 중

MSCI ESG Research는 약 8,500개 평가 기업을 대상으로 AAA~CCC 7단계 등급을 부여하는데 MSCI는 산업별 가중치를 매년 조정하며 G(거버넌스)는 33% 최소 가중치가 고정되어 있어 AI 거버넌스의 영향력이 상대적으로 크다. MSCI 평가체계에서 AI 거버넌스는 두 측면으로 통합된다. 첫째, 평가 도구(tool) 측면에서 1000개 이상의 데이터 포인트를 NLP·머신러닝을 활용한 자동 데이터 수집·검증으로 인간 분석가의 보조 역할을 한다. 둘째, 평가 대상(subject) 측면에서 GICS(글로벌 산업 분류 표준) 세분류산업별 핵심 ESG 이슈에 'AI 거버넌스' 항목이 IT·통신·금융·헬스케어 등에서 가중치를 부여 받고 있다.

나). 글로벌 ESG 평가에 반영되는 AI 거버넌스

단순히 'AI 원칙이 있는가'라는 선언적 수준을 넘어, 실질적 작동 여부를 검증

①Sustainalytics(Morningstar): 2024년 AI 윤리·데이터 거버넌스 하위 지표 추가 → 단순히 'AI 가이드 라인이 있는가'라는 선언적 수준을 넘어, 실질적인 리스크 통제 프로세스를 정량화하여 평가

②S&P Global ESG Scores: 단순히 'AI 윤리 원칙이 있는가'라는 선언적 수준에 그치지 않고, 이를 실행할 '프로그램(Responsible AI Program)'의 실질적 작동 여부를 검증, AI·사이버보안·개인정보를 산업별 핵심 이슈로 가중 평가, 산업별로 가중치를 차등 적용하고 공개 공시 중심의 평가가 특징

③FTSE Russell: AI 거버넌스를 '인적 자본' 및 '비즈니스 윤리'로 나누어 평가.

④Bloomberg ESG Scores: ESG Score 서비스에 AI 윤리 변수 반영.
⑤RepRisk : 기업의 말(말뿐인 선언이나 워싱) 대신, 외부 세계에서 실제로 발생한 사건과 평판 리스크를 실시간으로 추적하는 '아웃사이드 인(Outside-in)' 방식을 사용해 평가하는데 최근 AI 전용 토픽 태그를 정식 신설 AI 관련 평판 사건(편향, 차별, 개인정보 침해) 실시간 추적. AI Assurance 시장 보완재로 평가됨.

다) 한국 ESG 평가기관의 AI 거버넌스 통합

국내 평가기관도 AI 거버넌스를 고려하기 시작했다. 서스틴베스트는 '2026 ESG 포커스'에서 AI 대응 역량과 ESG 데이터 기반 실질적 재무성과 입증이 2026년 ESG 시장 성패의 핵심이라고 강조한 바 있다. 또한 한국ESG기준원(KCGS)은 AI 기본법 시행에 따른 평가 항목 추가를 검토 중인 것으로 알려져 있다.

라) AI Assurance 시장 전망 – ESG Assurance와의 융합

AI Assurance 시장은 2025~2030년 사이 글로벌 ESG Assurance 시장과 유사한 성장 궤도를 보일 것으로 전망된다. 영국 정부 자료에 따르면 2025년 기준 UK에서 약 500여 개 기업이 AI Assurance 서비스를 제공 중이지만, 상당수가 AI 개발사 자체 보유로 '독립성·표준화' 문제가 제기되어 왔다. 빅4의 시장 진입은 이 독립성·표준화 격차를 메우는 흐름이란 평가를 받고 있다.

6. AI 거버넌스 한계와 도전

1) 현재의 구조적 한계

글로벌 AI 거버넌스 논의가 빠르게 진전되고 있음에도 불구하고 아직까지는 구조적 한계는 분명하다.

① 규범의 파편화 문제: OECD 원칙, UN 권고, EU AI Act, G7 행동 강령, 각국 국내법(한국 AI 기본법 포함) 등이 공존하면서 기업들은 다층적이고 때로는 충돌하는 규범 준수 부담을 지고 있다. 다국적 기업은 EU AI Act, NIST AI RMF, 한국 AI 기본법, ISO/IEC 42001 등을 동시에 충족해야 하는 '다중 컴플라이언스(multi-compliance)' 부담에 직면해 있다.

② 지정학적 균열의 심화: 미국의 다자 AI 거버넌스 이탈, 중국의 독자 노선, EU의 규범 선도 전략이 공존하며 글로벌 거버넌스 공백이 확대되고 있다. 특히 미국이 UN의 AI 거버넌스 노력에 반대하는 것은 국제 협력의 핵심 기반을 훼손할 수 있다는 우려가 있다.

③ 기술 변화 속도와 규제 주기의 불일치: AI 기술은 수개월 단위로 혁신적 변화를 거듭하는 반면, 국제 규범 형성에는 수년이 소요된다. 이 간극을 메우기 위한 '적응적 거버넌스(Adaptive Governance)' 메커니즘의 부재 해소가 핵심 과제다.

④ 글로벌 사우스의 포용: UN 보고서에 따르면 AI 거버넌스 논의의 핵심 이해관계자 중 하나인 개발도상국들이 실질적 참여 과정에서 배제되는 구조적 불평등이 지속되고 있다. UN에 따르면 118개국이 어떠한 국제 AI 거버넌스 이니셔티브에도 참여하지 못하고 있는 현실은 이 문제의 심각성을 보여준다.

2) 도전 – 향후 방향

가) WEF의 헌법적 프레임워크 제안

세계경제포럼(World Economic Forum, WEF)과 Continuum Institute는 2025년 글로벌 AI 거버넌스를 위한 헌법적 프레임워크를 제안했다. 이 프레임워크는 두 개의 레이어로 구성된다. 첫 번째 레이어인 '헌법적 핵심(Constitutional Core)'은 투명성·안전·책임을 보장하는 공유 기술 표준과 거버넌스 요소를 확립한다. 두 번째 레이어인 '지역 오버레이(Local Overlay)'는 각 관할권이 글로벌 기준과의 정렬을 유지하면서 의료·노동·방위 등 분야별 맥락 특수적 규제를 적용하도록 허용한다. 이 체계를 관장하기 위해 WEF는 각국 정부·연구소·표준 기관이 공동 설립하는 '협력적 지능을 위한 세계 위원회(WCCI, World Commission for Cooperative Intelligence)'를 제안했다. WCCI는 표준을 새로 제정하는 대신 ISO 등 기존 선도 기관의 표준을 조화·검증하는 역할에 초점을 맞춘다. 이는 '알고리즘을 위한 디지털 WTO'에 비유되기도 한다.

나) 글로벌 사우스의 포용

파리 정상회의에서 출범한 Current AI 이니셔티브($400M 초기 투자)는 공익 AI를 글로벌 사우스(Global South)에 확산시키기 위한 다자 펀드로 성공적인 첫 발을 떼었다. UN AI 자문기구의 'AI 역량 개발 펀드'와 함께 향후 임팩트 투자(Impact Investing) 영역의 핵심 자본 흐름으로 부상할 가능성이 있다.

IV. ESG의 핵심 아젠다로 떠오른 AI 거버넌스

새로운 기업 경쟁력의 척도로 성장하고 있는 AI 거버넌스

거시적 관점 - 규제 차익(regulatory arbitrage)과 장기 밸류에이션 평가에 긍정적 영향 미칠 것

AI 거버넌스의 지형은 2024~2025년을 기점으로 근본적으로 재편되고 있다. 국제 AI 거버넌스 논의는 '선언적 윤리'에서 '구속력 있는 제도'로 본격 이행되고 있다. 또한 EU·G7·각국·정부·UN·국제기구·표준화 기관·컨설팅 기업·NGO 등에 걸쳐 전개된 논의는 AI 거버넌스가 환경·사회·기업 지배구조 전반을 관통하는 핵심 ESG 이슈로 격상되어 가고 있음을 증명한다. PwC의 2024년 글로벌 투자자 서베이에 따르면 투자자들은 기업의 AI 투자가 생산성, 수익성, 비용 절감 등 유형의 가치를 창출하는지 면밀히 관찰하는 동시에, 노동력 영향, 규제 컴플라이언스, 환경 효과 등 잠재적 리스크에도 예민하게 반응하고 있다. 이처럼 AI 거버넌스 역량은 단순한 기술 관리 문제를 넘어, 기업의 ESG 등급, 규제 리스크 프리미엄, 장기 밸류에이션에도 영향을 미치는 구조적 팩터로 부상하고 있다.

AI 거버넌스 규제를 준수하지 못한 경우 재무적 부담 크게 발생 → 평판리스크에 그치지 않고 재무제표에 반영되는 정량적 리스크로 전환됨

2024년 8월 1일 발효된 EU AI Act는 2025년 2월부터는 금지 AI 관행을, 2025년 8월 2일부터는 GPAI(범용 AI) 모델 관련 거버넌스 규정이 시행되었다. 이에 따라 금융권의 경우 신용 평가, 대출 승인, 사기 탐지, AML 위험 프로파일링 등 핀테크의 핵심 AI 활용을 EU AI Act상 고위험(High-Risk) 시스템으로 분류하게 되었다. 이 규제 환경에서 비준수 기업이 직면하는 리스크는 수치화 될 수 있다. EU AI Act상 금지 AI 관행 위반은 3,500만 유로 또는 글로벌 연간 매출의 7%에 달하는 과징금이 부과될 수 있기 때문이다. AI 거버넌스 미흡은 더 이상 평판 리스크에 그치지 않고, 재무제표에 직접 반영되는 정량적 리스크로 전환되고 있는 것이다.

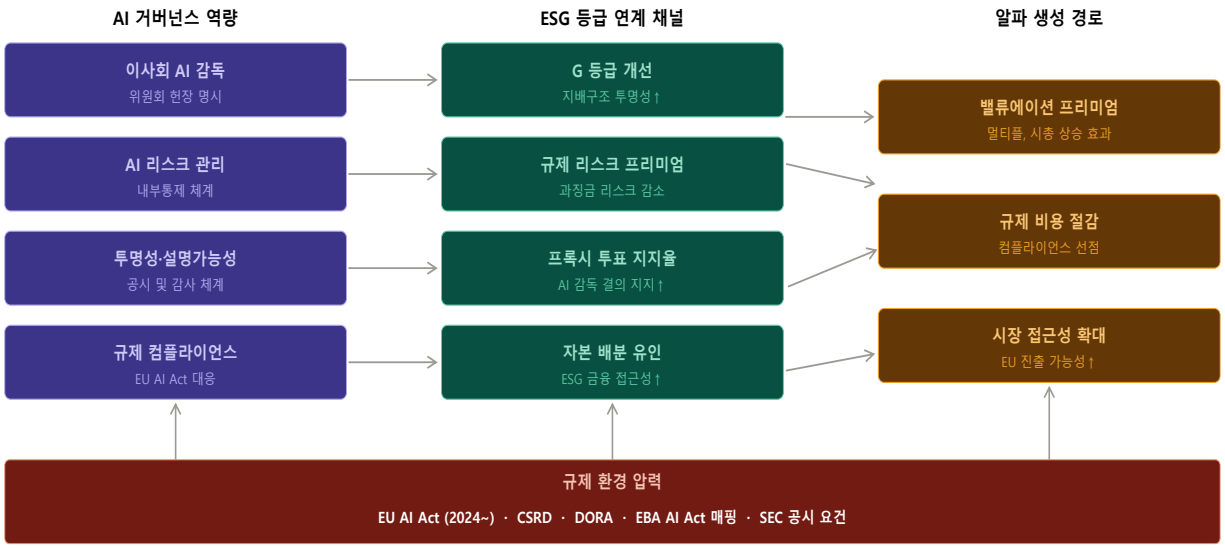
다른 제재 사례를 살펴보면 미국의 안면인식 기업인 Clearview AI는 사용자 동의 없이 수십억 건의 소셜미디어 이미지를 무단 수집했다는 이유로 영국, 네덜란드 등 복수의 규제 당국으로부터 제재를 받았다. 이 사례는 AI 거버넌스 실패가 리스크로 현실화 되었음을 보여준다. 다시 말해 투자자 입장에서 AI 거버넌스 역량은 사전에 평가되어야 할 팩터임을 의미한다. AI 거버넌스 취약성은 규제 제재로 이어지고, 이는 주가 하락 및 ESG등급 하향으로 연결되는 하방 리스크를 초래하기 때문이다.

정치적·외교적 측면에서는 미·중·EU의 '분기되는 거버넌스' 양상이 두드러지지만, 기술·표준 측면에서는 ISO/IEC 42001, NIST AI RMF, OECD AI 원칙을 축으로 한 표준은 수렴되고 있음

정치적·외교적 측면에서는 미·중·EU의 '분기되는 거버넌스' 양상이 두드러지지만, 기술·표준 측면에서는 ISO/IEC 42001, NIST AI RMF, OECD AI 원칙을 축으로 한 '수렴하는 표준화'가 진행되고 있다. EU·유럽평의회·UN·UNESCO는 '인권 기반 다자 규제' 노선을, 미국 트럼프 행정부는 '탈규제·혁신 우선' 노선을, 중국은 '국가 통제' 노선을, 한국은 '진흥과 규제의 병행' 노선을 추구하면서 글로벌 규제 차익 기회가 발생할 것으로 보인다. 이처럼 미·중·EU 등의 '분기되는 거버넌스'는

'규제 차익'과 '다중 컴플라이언스' 비용을 동시에 발생시키며, 글로벌 LLM 제공자보다 지역 특화 모델(소버린 AI 포함)에 기회가 될 것으로 판단된다.

도표 23. AI 거버넌스 → ESG 알파 팩터 전달 메커니즘



자료: PwC, 신영증권 리서치센터

EU AI Act·한국 AI 기본법 시행은 브뤼셀 효과를 가져와 세계적으로 AI 컴플라이언스 비용을 현실화시키고 있다. 이는 한국 AI 기본법 시행(2026.1)과 EU AI Act 고위험 AI 본격 적용(2027.12)에 따라 EU·한국 매출 비중이 높은 기업들은 AI 거버넌스와 관련된 컴플라이언스 및 리스크 관리 비용이 증가할 것으로 예상되기 때문이다. 이에 따라 특히 고위험·고영향 AI 의존도가 높은 금융·의료·고용 섹터 기업의 규제 리스크를 ESG 평가 모델에 통합해야 할 것이다.

특히 금융 서비스 산업은 AI 거버넌스의 다중 규제에 직면해 있다. EU AI Act는 신용 평가, 보험 인수·심사, 고용 심사, AML 등 금융 분야의 AI를 고위험으로 분류하여 엄격한 컴플라이언스 의무를 부과한다. 또한 CSRD·ESRS하에서 AI 기반 ESG 데이터 분석 도구의 외부 보증(External Assurance) 요구도 증가하고 있다. 여기에 SFDR하에서 AI 알고리즘 기반 투자 의사결정의 설명 가능성과 지속가능성 고려 통합이 감독 기대 사항으로 급부상하고 있다. 다시 말해 SFDR에 따르면 AI가 활용하여 펀드를 운용할 때, ESG 기준을 어떻게 적용했는지 투명하게 소명할 수 있어야 하며, AI 추천 결과가 지속가능성 목표와 어떻게 일치되는 지를 데이터로 공시해야 한다. 이에 이를 통합적으로 고려한 AI 거버넌스를 수립해야 할 것이다.

도표 24. AI 거버넌스의 주요 섹터별 영향

섹터	대표 종목군	핵심 영향	투자시 고려사항
Frontier AI	OpenAI, Microsoft, Google, Meta, Anthropic	· EU AI Act GPAI 의무(2025.8), 시스템적 위험 모델 등록·평가. HAIP 보고 의무, · 일부 기업 IPO시 AI 거버넌스 등 재평가 전망	R&D·컴플라이언스 비용 증가, 진입 장벽 확대
반도체/ AI HW	NVIDIA, AMD, TSMC, 삼성전자, SK하이닉스	· 미국 AI 기술 스택 수출 촉진에 따른 수혜.	E평가 부담 가중, 멀티플 격차 확대
클라우드/ 데이터센터	AWS, MSFT Azure, GCP, Equinix, Digital Realty	· AI 인프라 투자 가속. · NIMBY 사회적 갈등 대응, · 데이터센터 전력·물 부담 '이중 중대성' 논란 확대	관련 밸류 체인 특히 반도체 업종 수혜, 그린본드 발행 확대
금융	글로벌 금융사, 자산운용사	· EU AI Act 고위험(신용평가)(2027.12). · 미국 CFPB·FDIC·SEC 등 규제기관의 AI 관련 가이드 라인 강화.	컴플라이언스 비용 증가 우려, ESG 등급 변별력 증가
헬스케어	Novartis, Roche, Pfizer, 의료기기	· FDA SaMD와 EU AI Act는 의료 AI를 단순한 소프트웨어가 아니라 '지속적으로 검증·감독 받아야 하는 의료기기'로 규정하는 방향으로 진화	승인 지연 리스크, 컴플라이언스 비용 증가 우려
컨설팅· Assurance	Deloitte, PwC, EY, KPMG 등	· AI Assurance 시장 높은 성장세 보일 것, ESG Assurance 및 사이버보안과 융합 전망	매출 다변화· 고마진 라인업 확대
AI 거버넌스 SaaS	Credo AI, Fiddler, TruEra, IBM watsonx	· ISO/IEC 42001 인증 시장 확대. · EU GPAI Code of Practice 의무 대응 솔루션.	VC·PE 투자 집중, M&A 활발해질 것

자료: 신영증권 리서치센터

한국은 세계에서 처음으로 포괄적 AI 규제를 시행한 국가로 국내 기업들의 준비와 대응이 빠르게 나타날수록 향후 기회로 작용할 것으로 예상된다. AI 거버넌스를 선제적으로 구축한 기업은 ESG 평가에 긍정적 일 것이며 초기 시장의 표준으로 성장할 기회가 주어질 것이다. 반대로 AI 거버넌스가 미흡하거나 고영향 AI 사업자에 해당하면서 컴플라이언스 체계가 부족한 기업은 컴플라이언스 및 리스크 관리 비용 상승에 직면할 가능성이 높다.

AI 거버넌스는 더 이상 규제 대응 차원의 이슈가 아니라 기업가치, 자본조달 비용, ESG 평가, 산업 경쟁력에 직접 영향을 미치는 투자 판단 요소로 부상하고 있다. 이에 글로벌 자본시장의 새로운 알파(alpha) 변수이며, 향후 ESG 평가의 차세대 핵심 축으로 성장할 것으로 기대된다. 특히 한국 AI 기본법 도입과 EU AI Act의 본격 시행을 앞두고 AI 거버넌스 역량은 새로운 기업 경쟁력의 척도로 자리 잡고 있다. 향후에도 한국 및 글로벌 AI 거버넌스 동향과 이와 관련된 시장의 변화에 대해서 지속적으로 모니터링 해야 할 것이다.

Compliance Notice