Ver. 2026

인공지능 전환(AX) 과정에서의 보안규제 기반 신뢰성 확보



과학기술정보통신부
Ministry of Science and ICT

nipa 정보통신산업진흥원
National IT Industry Promotion Agency



인공지능 전환(AX) 과정에서의 보안규제 기반 신뢰성 확보

CONTENTS

1. 서론: 인공지능 전환(AX) 신뢰성은 모델 성능이 아니라 기업 책임의 문제

가. AI 업무 위임과 기업 책임의 이동

나. 보안과 규제 요구 확대의 배경

다. 도입 전 판단 기준

2. 주요국의 AI 보안과 규제 요구 수준 비교

가. 미국: 자율 기준, 조달 규율, 주법의 결합 구조

나. EU: 위험 등급과 공급망 역할별 의무 배분

다. 한국: AI 기본법과 개인정보 규율의 중첩

라. 주요국 제도의 공통 원리와 국내 시사점

3. 국내 인공지능 전환(AX) 통합 안전관리체계의 생명주기별 실행

가. 도입 판단: 업무 경계와 책임자 결정

나. 기획과 조달: 공급자 책임 계약화

다. 데이터와 설계: 읽는 정보와 실행 권한의 분리

라. 검증과 배포: 실패 조건 시험과 승인 전제의 기록

마. 운영과 변경: 변경 관리와 이행 기록

바. 실행 장애와 해법: 기업이 수행할 보안·안전 과업

사. 이용자 접점과 인간 감독: 고지와 설명 및 실질적 감독 권한

아. 폐기: 종료 계획과 잔여 책임의 관리

자. 생명주기별 의무와 실행 설계

차. 기업의 현실적 한계와 소결

4. 결론: 민관 협력을 통한 신뢰 기반 인공지능 전환(AX) 완성 방안

가. 기업 자체 대응의 한계와 정부 지원의 필요성

나. 정부 인공지능 전환(AX) 지원사업과의 연계 방향

다. 협력으로도 대신할 수 없는 기업의 몫

참고문헌

1 서론: 인공지능 전환(AX) 신뢰성은 모델 성능이 아니라 기업 책임의 문제

가. AI 업무 위임과 기업 책임의 이동

AI가 대신 내린 판단의 결과를 기업은 끝까지 설명하고 책임질 수 있는가.

인공지능 전환(AX)이 진행될수록 판단과 실행은 AI로 옮겨 가지만 그 결과에 대한 법적 책임과 고객의 신뢰는 기업에 남는다. 고객센터의 AI는 반품과 환불의 승인까지 맡고 인사팀의 AI는 서류 전형의 탈락자를 가려내며 생산 현장의 AI는 이상 징후를 감지해 라인을 세운다. 이때 경영진이 답할 질문은 어느 모델이 더 똑똑한가가 아니다. 잘못 승인된 환불과 탈락자의 이의 제기와 불필요한 라인 정지로 어긋난 납기를 누가 책임지는가이다. 즉 성능시험을 통과하는 것과 그 손해를 감당하는 것은 다른 문제이다.

신뢰성은 오류가 없다는 확인이 아니다. 허용한 범위 안에서 위험을 통제하고 그 상태를 입증하며 문제가 생겼을 때 회복하는 조직의 능력이다. 고객이 동의한 범위를 벗어난 개인정보로 내린 판단은 아무리 정확해도 정당성을 얻기 어렵고 승인 한도를 넘어 실행된 명령은 내용이 맞더라도 안전하지 않다. 신뢰성은 모델의 품질에서 출발해 데이터와 업무 절차를 거쳐 경영의 책임 구조에서 완성된다. 안전성만이 아니라 보안과 회복력 및 설명 가능성을 함께 관리하도록 요구하는 국제 위험관리 기준도 같은 관점에 서 있다[1].

나. 보안과 규제 요구 확대의 배경

보안과 규제의 요구가 함께 늘어나는 배경에는 AI의 역할 변화가 있다. 정보를 제시하던 도구가 행동을 실행하는 시스템으로 바뀌면서 협력사의 견적서에 숨겨 둔 지시문이 구매팀의 AI를 거쳐 발주 시스템에 전달되고 정상 계정으로 로그인한 직원의 질문에 열람 권한이 없는 인사 자료가 섞여 나온다. 공격자가 침입하지 않아도 잘못된 추론이 정상 업무처럼 반복되는 것이다. 프롬프트 인젝션과 권한 오남용 및 데이터 유출을 기존 정보보호 체계만으로 검증하기 어렵다는 진단이 정부 기술지침의 출발점이 된 것도 같은 맥락이다[2].

또 다른 배경은 책임의 귀속이 흐려졌다는 점이다. 기반 모델을 만든 해외 기업과 그 모델로 상담 솔루션을 만든 국내 개발사와 그 솔루션을 도입한 유통기업은 서로 다른 회사여서 고객 정보가 새어 나갔을 때 누구의 통제 실패였는지 가려내기 어렵다. 각국 제도는 이 공백을 사업자의 역할과 사용 목적 및 위험의 크기에 따라 의무를 배분하는 방식으로 메우고 있다. 즉 AI 규제의 확대는 AI 활용을 제한하는 작업이라기보다 책임의 경로를 다시 그리는 작업이라 할 수 있다.

다. 도입 전 판단 기준

기업이 먼저 정할 것은 어느 모델을 살 것인가보다 어떤 판단을 AI에 맡겨도 되는가이다. AI가 승인해도 되는 환불 금액의 상한, 서류 전형에서 AI 점수가 갖는 효력의 범위, 라인 정지에 사람의 확인이 필요한 조건을 정해 두어야 보안 수준과 개입 지점을 설계할 수 있다. 같은 업무를 놓고 기대 편익과 허용하지 않을 피해를 함께 정의하고 그 조건을 충족했을 때만 자동화 범위를 넓히는 방식이 안전하다. 끝으로 경계할 것은 공급자의 기술을 도입하면 이용기업의 법적 의무까지 해결된다는 오해이다.

2 주요국의 AI 보안과 규제 요구 수준 비교

가. 미국: 자율 기준, 조달 규율, 주법의 결합 구조

미국의AI 규제는 하나의 법률로 사전 허가를 요구하는 방식이 아니라 자율적 위험관리 기준과 연방 조달 규율 및 주법이 서로 다른 층에서 작동하는 분산형 구조이다. 위험관리 프레임워크는 어디까지나 자발적 기준이지만 연방기관에 납품하는 순간 계약이 그 내용을 요구 사항으로 바꾸어 놓고 캘리포니아에서 첨단 모델을 개발하는 기업에는 주법이 강제 의무를 부과한다. 어느 시장에서 누구에게 무엇을 공급하는지에 따라 적용되는 규율이 달라지는 셈이다[1][3][4].

첫 번째 층인 자율 기준은 거버넌스를 세우고 사용 맥락을 파악하며 위험을 측정하고 관리하는 순환을 요구한다. 생성형AI에 대해서는 허위 생성과 정보 무결성, 개인정보와 공급망의 위험을 추가로 다루도록 하는 별도의 지침이 붙어 있다. 기업에 필요한 것은 한 장의 평가표가 아니라 용도별로 위험의 책임자와 허용 수준을 정하고 모델이 바뀐 뒤에도 같은 기준으로 다시 평가하는 운영 구조이다[1][5].

두 번째 층인 연방 조달은 이 원칙을 계약의 언어로 바꾼다. 국민의 권리나 안전에 중대한 영향을 주는 용도로 AI를 쓰려는 연방기관은 영향평가와 배포 전 시험, 운영 중 감시와 인적 감독을 거쳐야 하고 조달 과정에서 데이터 권리와 특정 공급업체에 대한 종속을 관리해야 한다. 그 결과 납품기업은 평가 자료의 제출과 데이터 이용 조건 및 서비스 변경의 사전 통지를 계약 조건으로 요구받는다[3][6]. 연방 차원의 보안 규율은 첨단 모델의 국가안보 위험에 집중되어 배포 전 평가를 개발사와의 자발적 협력으로 확보하고 AI 사이버보안 정보 공유 창구를 민간과 함께 운영하되 의무적 허가나 사전 승인 제도는 두지 않는다[7].

세 번째 층인 주법의 강제성은 뚜렷하다. 학습 연산량이 법정 기준을 넘는 모델을 개발하는 일정 규모 이상의 개발자는 재난적 위험을 평가하고 완화하는 체계를 공개한 뒤 이를 지켜야 하며 중대한 안전사고는 정해진 기한 안에 보고해야 한다. 위반에는 건당 상한이 있는 민사 제재금이 따른다[4]. 독립 검증기관의 기준과 AI 감사인 등록제를 도입하는 법도 서명되었지만 단계적으로 발효되므로 지금 당장 적용되는 외부감사 의무로 보기는 어렵다[8]. 즉 미국에서 규제가 사라지고 있는 것이 아니라 위험의 종류와 시장 접점에 따라 책임의 강도가 달라지고 있는 것이다.

나. EU: 위험 등급과 공급망 역할별 의무 배분

EU의 규제는 위험의 크기와 공급망에서의 역할에 따라 의무를 배분하고 시장 진입 이후에도 이를 추적하는 법률 중심 구조이다. 금지되는 관행, 고위험 시스템, 이용자 접점의 투명성, 범용AI 모델이라는 네 갈래의 의무가 서로 구분되며 고위험 여부는 제품의 안전 기능에 해당하는지 또는 법이 열거한 사용 목적에 들어가는지로 결정된다. 채용 심사에 쓰는AI와 회의록을 정리하는AI는 같은 모델이라도 다른 등급으로 다루어진다는 뜻이다[9].

의무의 내용과 적용 시점은 나누어 보아야 한다. 최근의 간소화 규정으로 고위험 시스템의 의무는 조화표준과 지원 도구의 준비 상황에 맞추어 2027년 말과 2028년 여름으로 나뉘어 적용이 늦춰졌다. 그러나 AI와 대화하고 있다는 사실을 알리고 생성물임을 표시하는 투명성 의무는 2026년 8월부터 원래 일정대로 적용되고 있다. 고위험 의무의 연기를 AI 의무 전체의 유예로 받아들이면 지금 해야 할 조치를 놓치게 된다[10][11].

고위험 시스템의 공급자는 위험관리 체계와 데이터 품질 관리, 기술문서와 자동 기록, 배포자에 대한 정보 제공, 인간 감독 장치, 정확성과 견고성 및 사이버보안을 갖춰야 한다. 이를 도입해 쓰는 배포자는 사용 지침을 지키고 입력 데이터를 살피며 운영을 감시할 책임을 지고 자기 이름으로 시장에 내놓거나 사용 목적을 실질적으로 바꾸는 이용자는 공급자로 취급된다. 조화표준을 따르면 적합성이 추정되지만 법적 책임까지 표준기관으로 넘어가지는 않는다. 범용AI 모델의 공급자에게는 문서화와 하위 공급망에 대한 정보 제공이 요구되고 시스템적 위험을 가진 모델에는 추가 시험과 사고 보고가 붙는다[9].

보안 규율은AI 법 바깥에도 있다. 디지털 요소를 가진 제품을 EU 시장에 공급하는 제조자는 실제로 악용되는 취약점이나 중대한 보안사고를 알게 되면 24시간 안에 초기 경보를 보내고 72시간 안에 후속 보고를 해야 한다. 이 보고 의무는 2026년 9월부터 제품 전반의 요구보다 먼저 적용되므로 EU에 AI 내장 장비나 소프트웨어를 공급하는 국내 기업이라면 사고 대응 준비를 서둘러야 한다[12]. 제재는 금지 관행 위반에 가장 높은 상한이 적용되어 최대 3천500만 유로 또는 전 세계 연간 매출의 7%에 이르고 다른 주요 의무 위반은 최대 1천500만 유로 또는 3%이며 중소기업에는 정액과 매출 비율 중 낮은 쪽이 상한이 된다. 즉 EU가 기업에 주는 실질적 압력은 제재액의 크기보다 공급망 안에서 책임을 입증할 자료를 요구하는 시장 구조에 있다[9].

다. 한국: AI 기본법과 개인정보 규율의 중첩

한국의 규제는 산업 진흥과 신뢰 기반 조성을 한 법률에 담은 기본법 위에 개인정보와 분야별 법률의 책임을 겹쳐 놓은 중첩 구조이다. 2026년 1월부터 시행된 AI 기본법은 고영향AI와 생성형AI 및 초대규모AI에 서로 다른 의무를 붙이고 그 의무를AI를 개발하거나 이용해 제품과 서비스를 제공하는 인공지능사업자에게 지운다. 완성된 서비스를 쓰기만 하는 최종 이용자는 여기서 구분된다. 고영향AI는EU의 고위험 분류를 그대로 옮겨 온 용어가 아니다. 에너지와 보건의료, 채용과 대출 심사, 교통, 공공서비스, 학생 평가 등 법이 열거한 영역의 업무이면서 생명과 신체의 안전 또는 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있어야 성립한다[13].

규제 방식의 특징은 의무의 출발점을 사업자 스스로의 검토에 두었다는 점이다. 생성형AI로 결과물을 내놓거나 고영향AI를 서비스에 쓰는 사업자는 그 사실을 이용자에게 미리 알리고 생성물임을 표시해야 하며 자신의 서비스가 고영향에 해당하는지를 스스로 검토해야 한다. 정부에 확인을 요청하는 절차는 선택이다. 고영향 사업자로 판단되면 위험관리 방안과 설명 방안, 이용자 보호 조치, 사람의 관리와 감독, 관련 문서의 작성과 보관을 갖추어야 한다. 검증과 인증 및 기본권 영향평가는 노력의무이지만 고영향 사업자의 조치는 법정 의무이므로 권고 근거가 함께 있다는 이유로 자율 사항으로 다룰 수는 없다[13][14].

초대규모 AI의 안전성 확보 의무는 고영향 여부와 별개로 작동한다. 누적 학습 연산량이 법정 기준 이상이고 최첨단 기술을 활용하며 광범위하고 중대한 영향을 미칠 우려가 있다는 세 조건을 모두 충족하는 모델이 대상이며 위험의 식별과 평가 및 완화, 안전사고의 감시와 대응, 그 결과의 제출이 요구된다[13][14].

제재는 정밀해지는 방향으로 다듬어졌다. 사전 고지는 이용자가 AI의 관여를 알고 대응할 수 있게 하는 최소 조건이므로 이를 이행하지 않거나 국내 대리인을 두지 않거나 시정과 중지 명령을 어기면 3천만 원 이하의 과태료가 따른다. 시행 초기에 과학기술정보통신부가 밝힌 1년 이상의 계도기간은 법령상 유예가 아니라 사실조사와 과태료에 관한 행정 운영 방침이며 인명이나 인권에 관한 중대한 사안에는 예외가 있다[13][15].

개인정보 쪽에서는 대형 유출이 되풀이되면서 책임의 소재를 경영진으로 끌어올리는 방향으로 규율이 강화되었다. 사업주와 대표자가 최종 책임자로 명시되고 일정 규모 이상의 기업과 대학, 상급종합병원, 공공시스템 운영기관은 보호책임자의 지정·변경·해제에 이사회 의결과 신고를 거쳐야 한다. 반복적이거나 고의와 중과실에 의한 유출에는 매출액의 10% 이내에서 징벌적 과징금이 부과되고 통지 의무의 시작 시점도 유출 가능성이 높다고 판단되는 때로 앞당겨졌다[16][17][18]. 여기에 가명과 익명 처리만으로는 만들기 어려운 AI 학습 데이터를 개인정보보호위원회의 심의와 강화된 안전조치를 전제로 쓸 수 있게 하는 특례가 2027년 3월 시행(예정)을 앞두고 있다. 데이터 활용의 길과 감독의 틀이 함께 넓어지는 국면이라 할 수 있다[19].

라. 주요국 제도의 공통 원리와 국내 시사점

미국에서는 조달과 주법의 경로를, EU에서는 위험 분류와 사업자 역할 및 적용 일정을, 한국에서는 고영향 판단과 개별 법률의 중첩을 확인해야 한다. 〈표 1〉은 주요국이 무엇을 어느 수준으로 요구하는지를 핵심어로 정리한 것이다.

〈표 1〉 주요국(미국, EU, 한국) AI 보안과 신뢰성 규율 비교

비교항목	미국	EU	한국
핵심 문건	• NIST AI RMF 1.0과 생성형AI 프로파일 • OMB 연방AI 이용·조달 지침(M-25-21, M-25-22) • 캘리포니아SB 53, 행정명령14409, SB 813·AB 1405	• AI Act • 사이버복원력법(CRA) • AI 디지털 옴니버스 규정	• AI 기본법과 시행령 • 개인정보 보호법 개정 • AI 개발 개인정보 활용 특례, AI 보안 위협 대응 매뉴얼
규제 방식	• 자발적 위험관리 기준 • 조달 계약을 통한 요구 사항 전환 • 첨단 모델 개발자에 대한 주법상 강제 의무	• 위험 등급별 법정 의무 • 조화표준 준수 시 적합성 추정, 책임은 불이전 • 고위험 의무 적용 연기, 투명성 의무 유지	• 법정 의무인 사전 고지와 고영향 사업자 조치 • 노력의무인 검증·인증과 기본권 영향평가 • 비구속적 보안 매뉴얼과 개인정보 안내서
추진 체계	• OMB: 연방기관AI 이용·조달 관리 • NIST: 기술 기준 • 주 검찰: 주법 집행	• EU AI Office: 범용 모델 감독 • 회원국 시장감시당국: 고위험 시스템 감독 • 지정 적합성평가기관: 제3자 평가	• 과학기술정보통신부: 고영향 확인, 사실조사 • 개인정보보호위원회: 개인정보 감독 • 한국인터넷진흥원: 보안 지침, 검증 지원
규제 대상	• 권리·안전 관련 연방기관의AI 이용 • 연산량 기준 초과 첨단 모델 개발자, 대규모 개발자 강화 의무	• 금지 관행 • 사용 목적과 제품 안전에 따른 고위험 시스템 • 이용자 접점 시스템과 범용AI 모델	• 고영향AI • 생성형AI • 연산량·최첨단성·중대 위험 조건을 충족하는 초대규모AI
보안, 신뢰성 요구	• 맥락별 위험 측정과 관리의 순환 • 사전시험, 영향평가, 지속 감시, 인간 감독 • 위험 체계 공개, 가중치 보안, 사고 보고	• 위험관리, 데이터 품질, 기술문서, 기록, 인간 감독, 정확성·견고성·사이버보안 • 취약점·사고의24시간 경보와72시간 보고	• 사전 고지와 생성물 표시 • 위험관리, 설명, 이용자 보호, 인간 감독, 문서화 • 개인정보 안전조치와 보호책임자 체계
확인, 평가 방식	• 자율 프레임워크 활용과 연방기관의 영향평가 • 위험 체계 공개·준수와 사고 보고	• 내부 통제 절차 또는 제3자 적합성평가 • 출시 후 시장감시와 중대 사고 보고	• 고영향 사전 검토와 정부 확인 요청 • 고영향 이행 문서의5년 보관 • 공공기관 개인정보 영향평가
위반 시 주요 제재	• 자율 기준 자체의 법정 제재 부재 • 대규모 개발자의 특정 위반에 건당 최대100만 달러	• 금지 관행 위반 시 최대3천500만 유로 또는 매출7% • 다른 주요 의무 위반 시 최대1천500만 유로 또는3%	• 특정 위반에 대한3천만 원 이하 과태료 • 반복·고의·중과실 유출에 매출10% 이내 징벌적 과징금
정부 역할	• 기술 기준과 평가 역량 제공 • 공공 조달을 통한 책임 있는 도입 유도 • 자발적 사전평가 협력	• 규율과 감독의 조정 및 표준화 지원 • 규제 샌드박스 운영 • 사후 시장감시	• 고영향 확인과 법령 해석 지원 • 기술 지침과 검증 기반 제공 • AX 원스톱 바우처 등 실증지원

주요국 제도의 공통점은 모델의 이름이나 기업의 규모만으로 책임을 정하지 않는다는 데 있다. 제공자와 이용자의 역할 및 실제 사용 목적과 위험의 크기로 의무를 나누고 사람의 감독과 기록의 보존 및 사고의 보고를 요구한다. 차이는 강제력의 경로와 시점에서 나타난다. 미국은 계약과 주법이 자율 기준을 강제로 바꾸고 EU는 법률이 공급망 전체에 의무를 배분하며 한국은 기본법의 책무와 개인정보법의 제재가 서로 다른 축으로 작동한다.

즉 국가별 체크리스트를 따로 만드는 방식은 같은 위험을 중복 점검하면서도 실행 권한의 공백은 남기게 된다. 업무 목적과 데이터 출처, 평가 결과, 운영 변경의 기록을 한 벌로 정리해 여러 제도의 요구에 연결하고 법적 역할과 시행 시점이 다른 부분만 따로 관리하는 편이 낫다. 국내 기업의 통합 안전관리도 여기에서 출발할 필요가 있다.

3 국내 인공지능 전환(AX) 통합 안전관리체계의 생명주기별 실행

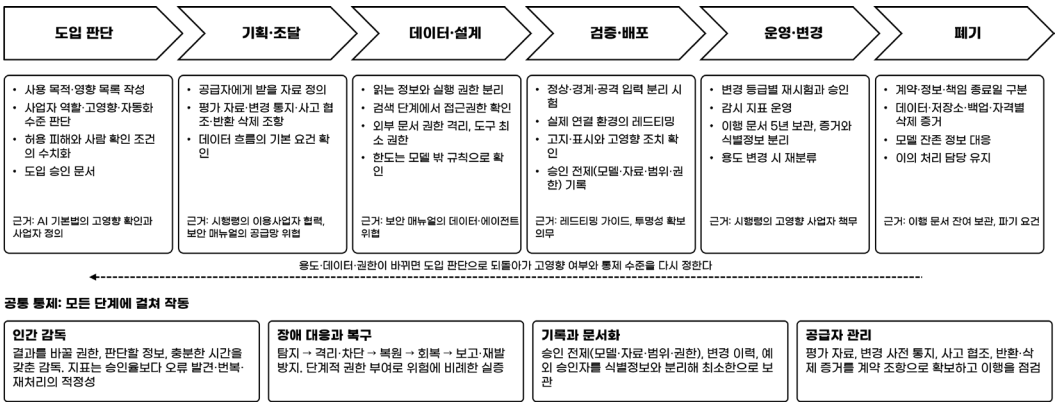
인공지능 전환(AX)에서 기업이 먼저 정할 것은 AI에 맡길 업무의 한계와 그 한계를 지킬 책임자이다. 영업본부가 효과를 정하고 IT팀이 기능을 붙이고 보안팀이 나중에 점검하는 순서로는 부족하다. 어떤 피해를 허용하지 않을 것인지와 어느 조건에서 사람의 판단으로 넘길 것인지를 도입 승인 전에 함께 정해 두어야 한다.

통합 안전관리의 단위는 개별 모델이 아니라 고객의 정보가 들어와 기업의 결정이나 행동으로 이어지는 하나의 업무 흐름이다. 규제 대상도 기업에 한 번 붙는 단일 등급이 아니어서 같은 유통기업이라도 채용 업무에서는 고영향 의무를, 상담 기록에서는 개인정보 처리 근거를, 해외 클라우드의 모델을 쓰는 업무에서는 국외 이전 요건을 따져 보아야 한다. [그림 1]은 업무 하나를 놓고 누가 제공하는지와 어디에 무엇을 위해 쓰는지 및 어떤 정보를 다루는지를 차례로 물어 참조할 규제 항목을 찾아가는 흐름이다[13][17].



[그림 1] 인공지능 전환(AX) 시에 고려할 규제 항목 판단 프로세스

인공지능 전환(AX)은 도입 판단부터 폐기까지 생명주기 전체로 접근할 필요가 있다. 도입 시점의 위험 분류와 계약 조건이 검증의 자료가 되고 그 검증이 다시 운영 중 변경 판단의 기준이 되기 때문이다. 장애 대응과 인간 감독도 마지막에 덧붙이는 단계가 아니라 전 과정에서 작동하는 공통 통제이다. [그림 2]는 이 연결을 단계별 과제와 근거로 정리한 것이다.



[그림 2] 인공지능 전환(AX) 통합 안전관리의 생명주기와 공통 통제

가. 도입 판단: 업무 경계와 책임자 결정

도입 판단에서 기업이 만들어야 할 산출물은 도입할 AI의 목록이 아니라 사용 목적과 영향의 목록이다. 답변 초안을 만들어 상담원이 확인하게 하는 업무와 보험금 청구를 AI가 자동으로 거절하는 업무는 같은 모델을 써도 전혀 다른 업무이다. 담당 부서는 적용 분야와 판단 대상, 오류가 났을 때의 피해와 회복 가능성을 적고 그 업무가 고영향에 해당하는지를 도입 전에 검토해야 한다. 판단이 어렵다면 정부에 확인을 요청하는 길이 열려 있지만 검토 자체를 생략할 수는 없다.

고려할 것은 세 가지이다. 첫째는 사업자로서의 역할이다. 해외 기반 모델을 붙여 상담 서비스를 제공하는 유통기업은 자체 모델을 학습하지 않았어도 이용사업자에 해당할 수 있고 사내 이용이라는 이유로 다른 법률의 책임이 사라지지도 않는다. 둘째는 고영향 여부이다. 법정 활용 영역의 업무라는 사실과 생명이나 기본권에 미치는 중대한 영향이 함께 있어야 하므로 진료 예약 상담용AI가 곧바로 고영향이 되지는 않지만 진료 판단에 실질적으로 관여하는 시스템은 별도 검토 대상이며 인사 담당자가 최종 결재한다는 사실만으로 고영향이 부정되지도 않는다. 셋째는 자동화의 수준이다. 판단 보조와 결정 확정 및 외부 실행을 구분하는 순간 감독의 강도와 실행 권한의 설계 및 복구 절차의 수준이 함께 정해진다[13][14].

어떻게 해야 하는가. 업무별로 기대 편익과 허용하지 않을 피해를 한 장에 적고 AI가 승인해도 되는 금액의 상한과 사람이 확인해야 할 조건을 수치로 정한다. 규칙만으로 처리할 수 있는 업무에 생성형AI의 재량을 줄 이유는 없으므로 규칙 처리의 대안을 먼저 검토한다. 초대규모 모델의 안전성 확보는 개발사업자의 몫이지만 고영향 판단이 끝날 때까지는 자동 결정과 외부 실행 권한을 제한해 두는 편이 안전하다. 이 단계의 결론은 업무 책임자와 보안 담당자가 함께 확인한 도입 승인 문서로 남긴다.

나. 기획과 조달: 공급자 책임의 계약화

기획과 조달에서 가장 경계할 것은 공급자의 기술을 도입하면 이용기업의 의무까지 함께 이행된다는 믿음이다. 고영향AI를 도입하는 이용사업자는 개발사업자가 이미 갖춘 위험관리와 설명 및 이용자 보호 조치를 활용할 수 있지만 사람의 감독과 문서화까지 면제되지는 않는다. 개발사업자의 협력이 노력의무에 그치는 현재 제도에서 이 공백을 메우는 수단은 계약이다[14].

고려할 것은 공급자에게서 무엇을 받아야 검증과 운영이 가능한지이다. 성능과 안전성 평가 자료, 중대한 변경의 사전 통지, 보안사고 시의 협조 범위, 계약 종료 시 데이터의 반환과 삭제 증거가 없으면 이용기업은 자신의 책무를 입증할 자료를 처음부터 갖지 못한다. 정부 매뉴얼이 공급망 위협을 별도의 영역으로 다루는 이유도 여기에 있다[2].

어떻게 해야 하는가. 계약서에 평가 자료의 제공 범위와 시점, 변경의 사전 통지 기한, 사고 협조 의무, 감사권, 데이터 반환과 삭제의 확인 방법을 구체적인 조항으로 넣는다. 개인정보가 오가는 구조라면 실제 데이터 흐름을 먼저 그린 뒤 기업의 지시에 따라 처리하는 위탁인지, 공급자가 자기 목적으로 쓰는 제3자 제공인지, 해외에서 처리되는 국외 이전인지를 기본 요건으로 확인하고 계약서의 서비스 명칭이 아니라 그 흐름에 맞는 근거를 갖춘다. 입력 자료를 공급자가 자기 모델의 학습에 쓰지 않도록 하는 조건은 보안과 개인정보 양쪽에서 함께 필요한 조항이다[17][20].

다. 데이터와 설계: 읽는 정보와 실행 권한의 분리

설계 단계의 핵심은AI가 읽는 정보와 실행할 수 있는 권한을 분리하는 데 있다. 정부 매뉴얼은AI 보안 위협을 데이터와 모델, 에이전트, 공급망, 고성능 모델의 다섯 영역으로 나누는데 기업의 실무에서 가장 자주 마주치는 것은 외부 문서에 숨겨 둔 지시문이 시스템 명령처럼 작동하는 프롬프트 인젝션과 정상 계정으로 열람 권한 밖의 자료가 흘러나오는 권한 오남용이다[2].

고려할 것은 두 가지이다. 첫째, 사내 문서를 검색해 답하는 구조는 자료에 섞인 악성 지시나 열람 권한 없는 문서의 노출까지 해결해 주지 않는다. 둘째, AI에 규칙을 지키라고 지시하는 것과 규칙을 어긴 실행이 시스템에서 차단되는 것은 다른 일이다. 모델이 아무리 지시문을 잘 따르더라도 금액 한도와 대상 계좌의 확인을 모델의 판단에 맡겨 두면 한 번의 우회로 통제가 무너진다.

어떻게 해야 하는가. 검색 단계에서 질문한 직원의 접근권한을 먼저 확인하고 권한 밖의 문서는 검색 결과에서 제외한다. 외부에서 들어온 문서의 내용에는 시스템 명령과 같은 권한을 주지 않고 도구 호출에는 업무에 필요한 최소 권한만 부여한다. 환불 업무라면 모델이 사유를 이해하는 계층과 실제 지급을 실행하는 계층을 나누고 금액 한도와 대상 계좌는 모델 밖의 업무 규칙으로 확인한다. 학습이나 검색에 쓰는 데이터는 수집 목적의 범위 안에서 쓰는지를 기본 요건으로 확인하고 계약서나 신분증 사본을 검색 자료에 넣기 전에 불필요한 식별번호를 지운다. 이 단계의 산출물은 데이터 흐름도와 권한 매트릭스이며 이것이 다음 단계의 시험 대상이 된다[2][20].

라. 검증과 배포: 실패 조건의 시험과 승인 전제의 기록

검증은 모델의 평균 정확도를 확인하는 작업에서 업무의 실패 조건을 확인하는 작업으로 넓어져야 한다. 정상 입력과 경계 상황 및 공격적 입력을 구분하고 업무 오류와 권한 위반을 따로 평가해야 하며 보안 시험은 검색 저장소와 외부 도구 및 계정 권한이 실제로 연결된 환경에서 공격자의 관점으로 수행해야 한다. 정부의 레드티밍 가이드가 시험의 범위와 절차를 제시하고 있으므로 기업은 이를 자사의 업무 시나리오에 맞추어 구체화하면 된다[2].

고려할 것은 시험의 결과보다 그 결과가 성립한 전제이다. 모델 버전과 평가 자료, 검색 범위, 도구 권한이 달라지면 같은 시험의 결과도 더 이상 유효하지 않고 실패가 없었다는 표현은 시험한 범위 안에서 관찰되지 않았다는 뜻으로 제한해야 한다. 고영향 서비스라면 사전 고지와 생성물 표시가 실제 화면에서 작동하는지, 위험관리 방안과 설명 방안 및 이용자 보호 조치가 문서로 준비되었는지를 배포 전에 확인해야 한다[13][14].

어떻게 해야 하는가. 배포 승인 문서에 시험 결과와 함께 모델 버전, 평가 자료, 검색 범위, 도구 권한을 전제 조건으로 기록한다. 업무별 허용 오류와 중단 조건을 시험 책임자와 배포 승인자가 함께 확인해 운영에서 무엇을 감시할지 정한다. 기본권 영향평가는 노력의무이지만 국가기관이 이를 거친 제품을 우선 고려하므로 공공조달을 염두에 둔 기업이라면 실질적인 거래 조건으로 보고 준비한다. 대상이 아니라는 결론에도 사용 목적과 규모라는 근거를 남겨 두면 업무가 바뀐 뒤 무엇을 다시 확인해야 하는지가 분명해진다.

마. 운영과 변경: 변경 관리와 이행 기록

인공지능 전환(AX)의 위험은 출시 시점에 고정되어 있지 않다. 모델을 바꾸지 않아도 검색 대상에 인사 자료 폴더가 추가되거나 발주 시스템이 새로 연결되는 순간 시스템이 할 수 있는 행동이 달라진다. 지시문과 검색 인덱스, 데이터 연결, 도구 권한은 모두 변경 관리의 항목이며 안내 문구를 고치는 일과 자동 승인 한도를 올리는 일을 같은 승인 절차로 처리해서는 안 된다.

고려할 것은 변경의 등급과 기록의 범위이다. 고영향AI 사업자는 위험관리 방침과 설명 방안, 이용자 보호 방안, 감독자의 성명과 연락처를 공개하고 이행 문서를 5년 동안 보관해야 한다. 다만 실제 중단 권한이 없는 감독자를 공개한 상태는 문서는 있으나 책임 구조는 작동하지 않는 상태이다. 보관 의무를 모든 입력과 원문 개인정보를 그 기간 내내 저장하라는 뜻으로 읽어서도 안 된다[14].

어떻게 해야 하는가. 변경을 안내 문구 수정 같은 경미한 변경, 검색 범위나 지시문의 변경 같은 중요한 변경, 도구 권한이나 자동 승인 한도의 변경 같은 중대한 변경으로 나누고 등급에 비례하는 재시험과 승인자를 정한다. 사용 목적이 바뀌면 도입 판단 단계로 돌아가 고영향 여부를 다시 검토한다. 어떤 버전과 규칙이 적용되었고 누가 예외를 승인했는지를 입증할 자료만 최소한으로 남기고 식별정보와는 분리해 보관한다. 신뢰성을 입증하려고 남긴 기록이 새로운 침해의 원인이 되어서는 안 되기 때문이다. 개인정보 보호책임자에게는 학습 데이터와 공급자의 재이용 현황이

보고되는 경로를 연결하고 AI 운영의 위험 수용 판단은 경영진의 결정으로 남긴다[16][17].

바. 실행 장애와 해법: 기업이 수행할 보안·안전 과업

실행을 가로막는 장애는 기술 부족만이 아니다. 처리 속도로 평가받는 고객센터와 사고 방지로 평가받는 보안 조직은 같은 오류를 다르게 해석하고 공급자가 시험 자료를 주지 않는 상황에서 도입 기한만 지키려는 조직의 검증은 형식이 되기 쉽다. 사고가 났을 때 AI의 접속을 끊는 것으로 대응이 끝났다고 여기는 것도 흔한 착각이다. 이미 지급된 환불이나 발송된 탈락 통보가 저절로 취소되지는 않기 때문이다.

이 장애를 넘기 위해 기업이 수행할 과업은 여섯 가지로 정리할 수 있다. 첫째는 단계적 권한 부여이다. 답변 초안만 제시하는 단계에서 처리 방안을 추천하는 단계로, 다시 소액 환불을 자동 승인하는 단계로 권한을 넓혀 가면 증거가 부족한 상태에서도 위험에 비례하는 실증이 가능하다. 둘째는 탐지이다. 권한 밖 자료의 접근 시도, 허용 한도를 넘는 실행 요청, 외부 문서에서 유입된 지시문의 징후, 출력 품질의 급격한 변화를 감시 지표로 두고 임계값을 넘으면 사람에게 알린다. 셋째는 격리와 차단이다. 문제가 확인되면 도구 실행 권한을 회수하고 검색 범위를 축소하며 자동 승인을 정지하되 서비스 전체를 내리기보다 위험한 기능만 끊는 범위 격리를 우선한다. 넷째는 복원이다. 검증을 통과한 모델 버전과 지시문, 검색 인덱스, 권한 설정으로 되돌리며 이를 위해서는 배포 승인 시점의 설정이 버전으로 보관되어 있어야 한다. 다섯째는 회복이다. 장애 구간에 처리된 건을 찾아 재심사하고 잘못 처리된 고객에게 연락해 바로잡는 절차와 담당자를 둔다. 여섯째는 보고와 재발 방지이다. 초대규모AI의 안전사고라면 대응 결과를 정부에 제출해야 하고 개인정보 유출이 의심되는 시점에는 정보주체 통지와 개인정보보호위원회 신고를 별도로 판단해야 하며 두 보고를 하나로 대신할 수 없다. 원인 분석의 결과는 시험 시나리오에 추가해 다음 검증에 반영한다[2][13][17].

소규모 조직에 필요한 것은 거대한 통제 조직의 복제가 아니다. 업무 책임자와 기술 담당자 및 보안·개인정보 담당자의 역할을 정하고 한 사람이 여러 역할을 맡더라도 중대한 예외를 혼자 승인하지 않도록 교차 검토를 연결하는 책임 배치이다.

사. 이용자 접점과 인간 감독: 고지와 설명 및 실질적 감독 권한

이용자에게 AI 사용 사실을 알리는 일은 가장 눈에 띄는 의무이지만 신뢰성의 전부는 아니다. 고영향AI나 생성형AI로 서비스를 제공하는 사업자는 이용자가 알 수 있도록 미리 알리고 생성물임을 표시해야 하며 실제와 구분하기 어려운 음향과 이미지 및 영상에는 명확한 고지가 필요하다. 사업자 내부 업무용에는 예외가 있으나 이는 투명성 조치에 관한 예외일 뿐 다른 안전 의무까지 없애지는 않으며 지원자의 서류 전형을 회사 안에서 처리한다는 이유로 그 결과가 내부용이 되지도 않는다[13][14].

개인정보를 처리해 사람의 개입 없이 완전히 자동화된 결정을 내리는 경우에는 기본 요건으로 정보주체의 거부와 설명 요구에 대응할 수 있어야 하고 결정의 기준과 절차를 공개해야 한다. AI 사용 사실의 고지와 개인에게 내려진 결정의 이유 설명은 서로 다른 책무이다[17].

인간 감독이 실질적 통제가 되려면 사람에게 결과를 바꿀 권한과 판단할 정보 및 충분한 시간이 있어야 한다. 화면에 승인 버튼만 두고 처리량 목표 때문에 모든 결과를 통과시키는 방식은 감독의 외형만 남긴다. 기업이 측정할 지표도 승인율보다 오류 발견과 번복 및 재처리의 적정성에 가까워야 하며 감독자의 업무량과 판단 시간을 관리하는 것 자체가 안전 과업이다. 사람을 절차에 남기는 것과 사람이 실제로 통제하는 것은 다르다[1][5].

아. 폐기: 종료 계획과 잔여 책임의 관리

폐기 단계에서는 계약의 종료일과 정보의 보유 종료일 및 이용자에 대한 책임의 종료일이 서로 다르다. 운영을 끝낸 시스템에도 고영향AI 이행 문서의 보관 기간이 남고 불필요해진 개인정보는 지체 없이 파기하되 다른 법률에 따른 보존 대상은 구분해야 한다. 세 기간을 하나로 합치지 않아야 과잉 보관과 증거 누락을 함께 막을 수 있다[14][17].

어떻게 해야 하는가. 원문 데이터와 검색 저장소, 백업, 접근 자격을 나누어 반환과 삭제 및 차단의 대상을 정하고 실제 이행의 증거를 공급자에게서 확보한다. 학습된 모델에 정보가 남을 가능성이 있다면 파일 삭제만으로 문제가 해결되었다고 약속하기는 어려우므로 재학습과 사용 제한 및 모델 교체의 가능성을 비용과 함께 검토해 둔다. 이런 선택지를 도입 시점에 확보하지 않은 기업은 계약을 끝내려는 순간 기술적 종속을 마주하게 된다. 종료 가능성을 조달의 조건으로 삼고 종료 후에 남는 이의 신청과 피해 회복의 담당자까지 정해 두어야 업무의 책임이 끊기지 않는다[20][6].

자. 생명주기별 의무와 실행 설계

이상의 내용을 생명주기별로 무엇을 갖추어야 하는지와 어떻게 실행하는지 및 어떤 증거를 남기는지로 정리한 것이 〈표2〉이다. 통합 관리의 목적은 모든 업무에 가장 무거운 절차를 적용하는 데 있지 않다. 법적 적용 여부를 먼저 가린 뒤 피해의 크기와 회복 가능성에 비례해 검증과 감독을 강화하고 각 단계의 산출물이 다음 단계의 입력이 되도록 연결하는 데 있다.

〈표 2〉 인공지능 전환(AX) 생명주기별 보안·안전관리의 핵심 구성요소와 실행 방안

생명주기	핵심 구성요소	실행 방안	산출물(증거)
도입 판단	• 사업자 역할 판단 • 고영향·생성형·초대규모 해당 여부 • 자동화 수준 설정	• 사용 목적과 영향 목록 작성 • 허용 피해와 사람 확인 조건의 수치화 • 규칙 처리 대안 검토 • 판단 전 자동 결정과 외부 실행 제한	• 도입 승인 문서
기획과 조달	• 공급자 책임의 계약화 • 데이터 흐름의 기본 요건	• 평가 자료·변경 통지·사고 협조·감사권·반환 삭제 조항 • 위탁·제3자 제공·국외 이전 구분 • 입력 자료의 학습 금지 조건	• 계약서, 데이터 흐름도
데이터와 설계	• 읽는 정보와 실행 권한의 분리 • 5대 위협 영역 대응 설계	• 검색 단계의 접근권한 확인 • 외부 문서의 권한 격리, 도구 최소 권한 • 한도와 계좌는 모델 밖 규칙으로 확인 • 불필요한 식별정보 제거	• 권한 매트릭스, 위협 대응 설계서
검증과 배포	• 실패 조건 시험과 레드티밍 • 배포 전 의무 확인	• 정상·경계·공격 입력 분리 시험 • 실제 연결 환경의 레드티밍 • 고지·표시와 고영향 조치 확인 • 허용 오류와 중단 조건 합의	• 시험 보고서, 배포 승인 문서(전제 기록)
운영과 변경	• 변경 등급 관리 • 감시 지표 • 이행 문서 보관	• 등급별 재시험과 승인 • 용도 변경 시 재분류 • 증거와 식별정보 분리 보관 • 보호책임자 보고 경로 연결, 위험 수용은 경영진 결정	• 변경 이력, 감시 기록, 이행 문서(5년)
실행 장애와 해법	• 단계적 권한 부여 • 탐지·격리·복원·회복·보고	• 감시 임계값과 통보 • 범위 격리와 승인 시점 설정 복원 • 기처리 건 재심사 • 정부 제출과 유출 신고의 별도 판단	• 대응 기록, 재심사 결과, 보고서
이용자 접점, 인간 개입 및 감독	• 고지와 표시 • 설명·거부 대응 • 감독자 권한	• 화면의 고지·표시 검증 • 결정 기준과 절차 공개 • 감독자의 변경·중단 권한 부여 • 번복·재처리 지표 관리	• 고지 화면, 감독 지표 보고
종료	• 세 종료일의 구분 • 반환·삭제·차단	• 데이터·저장소·백업·자격별 삭제 • 모델 잔존 정보 대응(재학습·교체 검토) • 종료 후 이의 처리 담당 유지	• 삭제 증거, 잔여 보관 목록

차. 기업의 현실적 한계와 소결

기업이 마주하는 가장 큰 한계는 정보의 비대칭이다. 기초 모델의 학습 연산량과 데이터 구성 및 변경 내역을 계약서에 요구 사항으로 적을 수는 있지만 공급자가 실제로 제출하도록 만들 수단은 마땅치 않다. 개발사업자의 협력이 노력의무에 그치는 현재 제도에서는 이 격차가 그대로 남는다.

판단의 경계도 유동적이다. 상담 초안에 쓰던AI를 환불 승인에 쓰는 순간 고영향 여부가 달라지고 상담 기록을 학습에 돌리는 순간 개인정보 처리 근거가 달라진다. 도입 시점에 한 번 검토하는 것으로는 부족하고 용도나 데이터가 바뀔 때마다 다시 확인해야 한다.

검증에는 또 다른 어려움이 있다. 비용은 도입 초기에 몰리는데 그 성과는 사고가 일어나지 않았다는 사실로만 확인된다. 눈에 보이는 성과가 없다 보니 일정을 앞당기라는 현업의 요구 앞에서 검증 절차가 먼저 축소되기 쉽다.

기업이 갖춰야 할 것은 세 가지이다. 사업자 역할과 사용 목적 및 자동화 수준을 따로 판단하는 습관, 도입에서 폐기까지 끊기지 않는 증거, 모델이 읽는 정보와 실행하는 권한의 분리이다. 법이 요구하는 것은 완벽한 모델이 아니라 설명할 수 있는 통제이며 그 통제는 승인 조건과 변경 이력 및 재심사 결과가 서로 연결될 때 성립한다. 즉 계도기간과 적용 연기는 준비의 시간이지 의무의 부재가 아니다.

4 결론: 민관 협력을 통한 신뢰 기반 인공지능 전환(AX) 완성 방안

가. 기업 자체 대응의 한계와 정부 지원의 필요성

기업이 스스로 감당할 수 있는 범위에는 분명한 한계가 있다. 기초 모델의 학습 데이터 구성과 연산량 및 변경 내역은 계약으로 요구할 수는 있어도 제출을 강제하기 어렵고 개발사업자의 협력이 노력의무에 머무는 한 이 비대칭은 개별 기업의 노력으로 해소되지 않는다. 공격 시나리오와 평가 자료를 기업마다 따로 만드는 일도 비용에 비해 효과가 낮다. 같은 업종의 기업들이 같은 위협을 각자 다르게 시험하면 검증 수준은 들쭉날쭉해지고 사고 경험은 공유되지 않는다. 검증 비용이 도입 초기에 집중되는 반면 그 효과는 사고가 나지 않았다는 형태로만 나타나므로 규모가 작은 기업일수록 예산과 인력을 배정하기 어렵다는 점도 제약이다.

이 한계가 정부 지원이 필요한 이유이다. 정부의 역할은 기업의 책임을 대신 보증하는 데 있지 않고 그 책임을 이행할 공통 기반을 제공하는 데 있다. 위협을 데이터와 모델, 에이전트, 공급망, 고성능 모델로 나누고 금융과 의료 등 여덟 개 분야의 시나리오를 제시한 정부 매뉴얼은 이미 그 기반의 한 축이다[2]. 다만 매뉴얼은 어떤 위협이 있고 어떻게 대응하는지를 설명할 뿐 어느 수준까지 확인해야 도입을 승인할 수 있는지는 기업의 판단에 맡겨 두고 있다. 정부가 이어서 할 일은 이 시나리오를 업종별 최소 점검 항목과 재검증 주기로 구체화하고 계약 조항 모형과 비식별 사고 사례를 축적하며 평가기관의 전문성과 독립성을 높여 기업이 내부 승인 기준으로 바로 쓸 수 있게 만드는 것이다. 캘리포니아가 독립 검증기관 등록의 틀을 만들고 EU가 조화표준으로 적합성 추정을 뒷받침하듯 국내에서도 검증의 신뢰를 제도가 받쳐 주는 구조가 필요하다.

나. 정부 인공지능 전환(AX) 지원사업과의 연계 방향

정부 지원을 신뢰성 확보와 연결하는 가장 현실적인 접점은 이미 운영 중인 인공지능 전환(AX) 지원사업이다. 초기 자원 부담이 큰 기업은 AX 원스톱 바우처를 우선 검토할 만하다. 수요기업은 AI 개발과 GPU 클라우드, 데이터 구매와 가공을 한 번에 지원받고 공급기업과 컨소시엄을 이루어 과제를 주관하며 단계평가를 거쳐 다음 단계의 지원으로 이어간다[21]. 제조 분야라면 상생형 AX 선도모델이, 데이터 확보가 관건이라면 데이터바우처가 대안이 될 수 있다. 어느 사업이든 공급기업과의 관계가 계약 관계인 동시에 컨소시엄 관계이므로 평가 자료의 제공과 변경의 사전 통지 및 데이터 반환 조건을 과제 협약 단계에서 정할 수 있고 그만큼 공급자 책임의 공백을 줄이기에 유리하다. 단계평가의 일정도 안전관리의 일정으로 함께 쓸 수 있다. 과제 기획에서 적용 법률과 위험 분류를, 중간 점검에서 데이터 처리

근거와 공급자 책임을, 완료 시점에 시험 결과와 복구 절차 및 종료 계획을 확인하는 방식이다.

한국정보통신기술협회와 전주정보문화산업진흥원이 함께 추진하는 인공지능 신뢰성(TAI) 지원사업은 제품의 검증과 인증 획득을 지원하고 한국인터넷진흥원의 ICT 중소기업 정보보호 지원사업은 정보보호 컨설팅과 보안 서비스를 제공한다. 문제는 수단의 부재가 아니라 연결의 부재이다. 구축 지원과 검증 지원이 서로 다른 기관의 별도 공고로 운영되다 보니 기업은 두 사업의 일정과 요건을 각각 맞춰야 하고 구축 과제가 끝난 뒤에야 검증 사업을 찾게 된다.

연계의 방향은 세 갈래로 제언할 수 있다. 첫째는 구축 지원사업의 인정 비용에 보안 검증과 개인정보 점검을 포함하거나 검증 지원사업과의 연계 신청 경로를 두는 것이다. 둘째는 최초 구축 이후 모델이 바뀌거나 새로운 공격이 나타났을 때의 재검증을 후속 지원으로 이어 주는 것이다. 셋째는 공급기업과 이용기업 사이의 자료 전달 양식과 공동 계약 기준을 마련하는 것으로, 반복되는 분쟁 유형을 반영하되 책임을 면제해 주는 증명서가 아니라 누구의 책임인지를 확인하는 도구로 설계해야 한다. 단계평가 항목에 위험 분류와 데이터 처리 근거 및 복구 절차의 확인을 넣으면 성과 관리와 안전관리가 한 절차로 묶인다.

다. 협력으로도 대신할 수 없는 기업의 몫

공통 기반과 지원사업이 갖춰져도 기업이 스스로 감당해야 하는 영역은 남는다. 법률 조항만으로 다루기 어려운 문제는 조직이 AI 없이도 중요한 판단을 수행할 능력을 유지하는가이다. 자동화가 오래 이어지면 상담원이 환불 규정을 스스로 판단하는 경험과 담당자가 원자료를 읽는 능력이 약해진다. 공급자가 서비스를 중단하거나 가격을 올렸을 때 업무를 이어갈 대안이 없는 기업은 협상력도 잃는다. 대체 절차의 실행 시간을 확인하고 사람의 전문성을 유지하며 서비스 종료 비용을 투자 판단에 포함하는 일은 법적 최소 의무를 넘어서는 경영의 과제이다.

기업이 설명해야 할 것은 AI가 얼마나 영리한가가 아니라 어디까지 맡겼으며 어디에서 멈추도록 했고 실패했을 때 어떻게 바로잡는가이다. 다음 투자도 더 많은 기능을 켜는 쪽이 아니라 책임의 연결을 완성하는 쪽으로 향할 필요가 있다. 경영진은 허용할 자동화의 범위와 회복에 필요한 자원을 결정하고 각 업무는 승인 조건과 변경 이력 및 이의 처리의 증거를 유지하며 정부는 이러한 실행이 기업 규모에 따라 불가능해지지 않도록 뒷받침해야 한다.

서두의 질문에 대한 답은 하나이다. AI가 내린 판단이라도 그 결과를 설명하고 책임지는 주체는 끝까지 기업이며 설명하고 멈추고 되돌릴 수 있는 범위까지가 그 기업이 감당할 수 있는 인공지능 전환(AX)의 진짜 범위이다.

참고문헌

[1] National Institute of Standards and Technology. (2023, January). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1).

[2] 과학기술정보통신부 (2026.7.8). AI 보안 위협 대응 매뉴얼.

[3] Office of Management and Budget. (2025, April 3). Accelerating federal use of AI through innovation, governance, and public trust (Memorandum M-25-21). Executive Office of the President.

[4] California Legislature. (2025, September 29). SB 53: Transparency in Frontier Artificial Intelligence Act.

[5] National Institute of Standards and Technology. (2024, July). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1).

[6] Office of Management and Budget. (2025, April 3). Driving efficient acquisition of artificial intelligence in government (Memorandum M-25-22). Executive Office of the President.

[7] The White House. (2026, June 2). Promoting advanced artificial intelligence innovation and security (Executive Order 14409).

[8] Office of Governor Gavin Newsom. (2026, September 9). Governor Newsom signs first-in-the-nation AI safeguards to protect Californians, calls on the federal government to do its part.

[9] European Parliament, & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

[10] Council of the European Union. (2026, June 29). Artificial intelligence: Council gives final green light to simplify and streamline rules.

[11] European Parliament, & Council of the European Union. (2026). Regulation (EU) 2026/1744 on the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). Official Journal of the European Union.

[12] European Parliament, & Council of the European Union. (2024). Regulation (EU) 2024/2847 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). Official Journal of the European Union.

[13] 대한민국. (2026). 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법. 국가법령정보센터.

[14] 대한민국. (2026). 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령. 국가법령정보센터.

[15] 과학기술정보통신부. (2026.1.21). 인공지능기본법 22일 시행: 생성형 AI 결과물 워터마크 표시 의무 [정책브리핑]. 대한민국 정책브리핑.

[16] 개인정보보호위원회. (2026.9.10). 개인정보 유출 사전예방 및 피해구제 강화를 위한 개인정보 보호법, 시행령, 고시, 9월 11일부터 시행.

[17] 대한민국. (2026). 개인정보 보호법. 국가법령정보센터.

[18] 대한민국. (2026). 개인정보 보호법 시행령. 국가법령정보센터.

[19] 개인정보보호위원회. (2026.8.20). 안전한 인공지능 개발을 위한 개인정보 활용 특례 신설「개인정보 보호법」 개정안 국회 본회의 통과.

[20] 개인정보보호위원회. (2025.8.6). 생성형 인공지능(AI) 개발 및 활용을 위한 개인정보 처리 안내서.

[21] 과학기술정보통신부. (2026.5.28). 2026년도 AX 원스톱 바우처 지원사업 수요기업 추가 모집 공고.

인공지능 전환(AX) 과정에서의 보안규제 기반 신뢰성 확보

발 행 처 : 정보통신산업진흥원

작 성 기 관 : 한국데이터산업협회

발 행 일 : 2026년 8월

2026년 9월 16일(보완)



인공지능 전환(AX) 과정에서의 보안규제 기반 신뢰성 확보

과학기술정보통신부
Ministry of Science and ICT

nipa 정보통신산업진흥원
National IT Industry Promotion Agency