LEGAL COMPLIANCE GUIDE

AI 기본법 5대 가이드라인 종합 Q&A 2.0

과기정통부 발간 5대 가이드라인 관련
기업 AI 담당자를 위한 핵심 체크리스트 및 실무 대응 전략

법무법인(유한)린 AI•플랫폼•테크놀로지 전문그룹

2026년 2월 10일

(최초 작성 2026. 1. 27.)

REFERENCE DOCUMENTS

핵심 가이드라인 본 가이드의 주요 작성 근거 (과학기술정보통신부 최신 발표)

고영향 인공지능 영향평가 가이드라인
과기정통부 / KISDI
2026. 1. 22.

고영향 인공지능 사업자 책무 가이드라인
과기정통부
2026. 1. 22.

최첨단 AI 안전성 확보 가이드라인
과기정통부 (초안)
2026. 1. 22.

투명성 확보 가이드라인
과기정통부
2026. 1. 22.

고영향 인공지능 판단 가이드라인
과기정통부
2026. 1. 29.

법령 및 시행령 기본 법적 근거

인공지능산업 육성 및 신뢰 기반 조성에 관한 법률 (AI …
국회 제정 법률
2026. 1. 22. 시행

동 법 시행령
2026. 1. 22. 시행

기타 참고자료 유관 기관 가이드 및 해설서

인공지능(AI) 윤리 기준 및 자율점검표
KISDI / NIA
2025. 10. 05.

TABLE OF CONTENTS

TABLE OF CONTENTS (2/2)

PART 01

법 적용 범위 및 기본 개념

AI 기본법의 적용 대상과 예외, 고영향AI 및 생성형AI의 정의,
사업자의 법적 지위 등 핵심 개념을 명확히 정리합니다.

Q1. 적용 대상과 제외 AI

적용 대상

- 대한민국 내에서 **제공, 유통 또는 운영**되는 인공지능
- **국내 이용자 또는 시장에 영향을 미치는 해외 사업자**의 인공지능 서비스
- 상업적 목적뿐만 아니라 공공기관이 도입하는 AI 시스템 전반
- 고영향AI 뿐만 아니라 생성형AI 등 법령상 의무가 부과된 모든 유형

적용 제외

시행령 제2조에 따른 특정 업무 수행 목적

- **국방·안보** 관련: 국방정보시스템, 무기체계 등
- **국가정보·보안** 관련: 대테러, 방첩, 사이버안보 등
- **범죄 수사** 관련: 정보사범 수사 등 특정 지정 업무
- 단, 지정된 업무 외의 일반 행정/민원 업무용 AI는 **적용 대상**임

실무 포인트: '전용' 여부 확인 필수

하나의 AI 모델이 국방/보안 업무와 일반 민간 서비스를 동시에 수행하는 '혼합 용도'인 경우, 법 적용이 제외되지 않을 가능성이 높습니다. 적용 제외는 엄격하게 해석되므로, 보안 업무 '전용'으로 분리 운영하는 것이 규제 리스크 관리 측면에서 유리합니다.

Q2. 고영향AI vs 생성형AI

구분	⚠ 고영향AI (High-Risk AI)	생성형AI (Generative AI)
정의	사람의 생명·신체의 안전 및 기본권에 **중대한 영향을 미칠 우려**가 있는 AI	텍스트, 이미지, 오디오, 영상 등 **콘텐츠를 생성**하는 기능을 가진 AI
판단 기준	① 특정 영역(의료, 채용 등) 해당 여부 ② 위험의 영향력, 중대성, 빈도	콘텐츠 생성 기능 여부
주요 영역/기능	생체인식, 채용, 신용평가, 의료, 교육, 법집행, 중요 인프라 관리 등	챗봇(LLM), 이미지 생성, 음성 합성, 영상 제작, 코드 생성 등
핵심 의무	• **영향평가** 실시 **(노력의무)** • **투명성 확보 (사전 고지)** • 위험관리 체계 수립 및 설명 가능성 확보	• **투명성 확보** (사전 고지) • **결과물 표시** (워터마킹/라벨링)

동시 해당 가능성 (Intersection)

생성형AI가 고영향 영역에서 활용되는 경우 **두 가지 규제 모두 적용**됩니다.

예: 생성형AI를 활용해 학생의 에세이를 자동 채점하는 경우 → **고영향AI (교육 평가) + 생성형AI (텍스트 분석/생성)**

Q3. AI 사업자 범위

법적 정의: 개발 또는 제공하는 자

인공지능을 개발하거나 인공지능을 이용하여 제품 또는 서비스를 제공하는 자를 의미합니다.

주요 사업자 유형 및 예시(서비스 구조에 따라 다를 수 있음)

모델 개발사	AI 파운데이션 모델을 직접 개발하는 기업 (예: LLM 개발사)
API 제공사	타사 모델을 API 형태로 가공하여 제공하는 중개자
서비스 사업자	AI 기능을 탑재한 최종 애플리케이션 제공사
통합/리셀러	시스템 통합(SI) 업체 또는 솔루션 재판매 파트너

의무 발생 시점: '제공' 단계

단순히 내부 연구 목적으로 개발하는 단계를 넘어, 시장에 출시하거나 타인에게 서비스를 제공하는 시점부터 법적 의무(투명성 확보, 영향평가 등)가 발생합니다.

Q4. 해외 개발 AI의 국내 적용

역외 적용 원칙: 국내 영향 기준

해외에서 개발된 AI라 하더라도 대한민국 내에서 제품 또는 서비스가 제공되거나 이용되는 경우, 본 법의 적용 대상이 됩니다. 서버가 해외에 있더라도 국내 이용자를 대상으로 영업하거나 서비스를 제공한다면 법적 의무를 준수해야 합니다.

국내 대리인 지정 필요성

국내에 주소 또는 영업소가 없는 해외 사업자의 경우, **법 제36조 제1항 및 제3항 규정에 따라** 국내 대리인(Local Representative)을 지정해야 합니다. 대리인은 규제 준수 책임을 대리하며 정부와의 소통 창구 역할을 수행합니다.

제36조(국내대리인 지정)

① 국내에 주소 또는 영업소가 없는 인공지능사업자로서 이용자 수, 매출액 등이 대통령령으로 정하는 기준에 해당하는 자는 다음 각 호의 사항을 대리하는 자(이하 "국내대리인"이라 한다)를 서면으로 지정하고, 이를 과학기술정보통신부장관에게 신고하여야 한다.

1. 제32조제2항에 따른 이행 결과의 제출
2. 제33조제1항에 따른 고영향 인공지능 해당 여부 확인의 요청
3. 제34조제1항 각 호에 따른 안전성 · 신뢰성 확보 조치의 이행에 필요한 지원(같은 항 제5호에 따른 문서의 최신성 · 정확성에 대한 점검을 포함한다)

② 국내대리인은 국내에 주소 또는 영업소가 있는 자로 한다.

③ 국내대리인이 제1항 각 호와 관련하여 이 법을 위반한 경우에는 해당 국내대리인을 지정한 인공지능사업자가 그 행위를 한 것으로 본다.

글로벌 서비스 컴플라이언스 전략

한국 법과 `EU AI Act` 등 **각국 법령은 단순한 포함 관계가 아니므로 주의가 필요합니다.** 미국(주별 규제) 등 개별 AI 규제 사항까지 고려하여 각 시장별로 정밀하게 대응해야 합니다.

실무 팁: 파트너 책임 배분

해외 본사와 한국 지사, 또는 해외 개발사와 국내 유통사 간의 계약에서 법적 책임(과태료, 시정조치 등)을 누가 부담할지 명확히 규정하는 것이 중요합니다.

Part 1 핵심 체크

01

서비스 제공 경로가 국내에 해당하는가?

- 서버 위치와 무관하게 국내 이용자 대상 여부
- 한국어 지원, 원화 결제 등 타겟팅 요소 확인
- 국내 파트너사를 통한 우회 제공 여부

02

우리의 법적 지위는 무엇인가?

- **개발사업자/제공사업자**
- API 연동하여 서비스 구현 여부 확인

03

고영향/생성형 탑다운 점검

- **영역: 의료, 채용, 신용, 생체인식, 법집행 등 구체적 확인**
- 콘텐츠 생성(텍스트/이미지) 기능 보유 여부

PART 02

고영향AI 판단 및 확인 절차

판단 2단계, 확인 요청/재확인, 절차·서류, 영역별 가이드 등
고영향AI 해당 여부를 명확히 판단하고 확인받는 프로세스를 안내합니다.

Q5. 고영향AI 판단 2단계

1

1단계: 영역 판단

법 제2조제4호 각 목의
7대 영역에 해당하는가?

- 생체인식, 채용·인사
- 의료·보건, 교육·직업훈련
- 신용평가·금융
- 법집행·공공안전, 중요인프라

NO → 비해당

YES → 2단계

2

2단계: 영향 판단

안전 및 기본권에
중대한 영향을 미치는가?

1. 영향의 중대성·심각성
2. 발생 빈도 및 지속성
3. 활용 맥락·의도된 목적

NO → 비해당

YES → 확정

CONDITION SATISFIED

고영향AI 해당

영향평가 등 책무 발생

CONDITION FAILED

고영향AI 비해당

일반 AI 규율 적용

Check Point: 확인 신청 제도 활용 권장

두 가지 요건을 모두 충족해야 고영향AI로 지정됩니다. 특정 영역에 해당하더라도 인권/안전에 미치는 영향이 경미하다면 제외될 수 있으므로, **불확실한 경우 과기정통부의 확인 절차를 적극 활용**하시기 바랍니다. (상세 기준은 **판단 가이드라인 참조**)

* 출처: 과학기술정보통신부 「고영향 인공지능 판단 가이드라인」 (2026.1.29)

Q5-보완. 판단 시 고려 요소

내부적 요소 (Intrinsic)

① 활용 맥락 및 의도된 목적

- **사용 방식** 실제 운영 환경에서의 구체적 활용 시나리오
- **자동화 수준** 완전 자동화 vs 보조적 수단 여부
- **인간 개입** 최종 의사결정에 사람의 실질적 개입 정도

② 영향의 중대성 (Severity)

생명·신체에 대한 직접적 위험 초래 여부

기본권(차별금지, 프라이버시 등) 제약의 강도와 범위

피해 발생 시 **회복 불가능성** 또는 원상복구의 어려움

외부적 요소 (Extrinsic)

③ 발생 빈도 및 지속성

- **규모** 영향을 받는 잠재적 인구(피영향자)의 크기
- **기간** 영향이 지속되는 시간 및 노출 빈도
- **반복성** 일회성 이벤트 vs 상시적/반복적 사용 여부

④ 대체 가능성 (Substitutability)

이용자가 선택할 수 있는 다른 대체 수단의 존재 여부

AI의 결정이 이용자에게 사실상 강제되거나, 거부하기 어려운 **구속력(Binding Force)**을 가지는지 여부

실무 예시: 채용 AI 케이스 분석 High Impact

단순히 이력서 형식을 검토하는 AI는 '보조적 수단'이나, **서류 전형 합격/불합격을 자동으로 판정**하거나 면접 점수를 산출하여 **채용 여부에 결정적 영향**을 미치는 경우, 대체 불가능성과 영향의 중대성이 인정되어 고영향 AI로 판단됩니다.

* 출처: 과학기술정보통신부, 「고영향 인공지능 판단 가이드라인」(2026.1.29)

Q6. 확인 신청: 언제·왜?

고영향AI 해당 여부가 불확실할 때, **확인 신청 제도**를 전략적으로 활용 가능합니다.

다만, 확인신청으로 모든 리스크가 해소될 수 있는 것은 아닙니다.

실무 TIP

WHY: 왜 신청해야 하나요?

- **불확실성 해소:** 애매한 규제 적용 여부를 정부로부터 공식 확인받아 법적 안정성 확보
- **리스크 사전관리:** 향후 규제 위반 시비나 과태료 부과 위험을 원천 차단
- **방어 수단:** 추후 문제 발생 시 '정부 확인' 결과서는 강력한 면책 근거 또는 참작 사유로 활용
- **고객 신뢰:** B2B 계약 시 고객사에게 규제 준수 여부를 증빙하는 자료로 활용

WHEN: 언제 신청해야 하나요?

- **베타 서비스 출시 전:** 본격적인 이용자 확보 전 리스크 제거
- **대형 고객 PoC/계약 전:** 계약 조건에 '규제 준수 보증' 조항이 포함될 때
- **주요 기능 업데이트 시:** AI 모델이나 데이터 처리 방식이 크게 변경될 때
- **투자 유치(Due Diligence) 전:** 투자자에게 법적 리스크 없음을 증명해야 할 때

Q7. 확인 절차 플로우

1 확인 요청

별지 서식 및 개요서 제출

- **제품 or 서비스 개요서**
- 데이터 개요
- 테스트 결과 등

사업자 → 장관

2 심사 및 검토

전문위원회 자문 및 기준 적합성 검토

- 기술적 검토
- 영향력 평가
- 필요시 보완요청

30일 이내 (최대 60일)

3 결과 회신

확인 결과 통지 (해당/비해당)

- 결과 통지서 발송
- 판단 근거 명시

장관 → 사업자

4 이의제기

결과 불복 시 재확인 요청

- 이의 신청서
- 추가 소명자료

회신 후 10일 이내

5 재회신

재검토 후 최종 결과 통보

- 전문위 재자문
- 최종 확정

요청 후 30일 이내

Tip: 심사 기간은 기본 30일이나, 자료 보완이 필요하거나 사안이 복잡한 경우 **1회에 한해 30일 범위 내에서 연장**될 수 있습니다. 확인 요청 전 자료를 충분히 준비하여 심사 지연을 방지하세요.

Q7-보완. 확인 신청 제출 서류 체크리스트

01

제품 또는 서비스 개요서

- AI 시스템의 개발 목적 및 주요 기능
- 기술적 아키텍처 및 구성도
- 적용 영역 및 활용 맥락 (Context)
- 예상 피영향자 규모 및 대상

02

학습용 데이터 개요서

- 데이터 출처 및 수집 방법 명세
- 데이터 규모, 유형 및 특성 분석
- 전처리 및 라벨링 절차 상세
- 편향성 검토 결과 및 완화 조치

03

테스트 결과 보고서

- 성능 지표 및 정확도 측정 결과
- 안전성 및 견고성 테스트 내역
- 공정성 및 편향성 평가 결과
- 위험 시나리오별 대응 테스트

* 출처: 과학기술정보통신부, 「고영향 인공지능 판단 가이드라인」 (2026.1.29) / **데이터 개요서와 테스트 결과가 핵심 심사 대상입니다.**

Q8. 결과 이의제기

기한: 결과 통지 후 10일 이내

고영향 AI 확인 결과에 이의가 있는 경우, 통지를 받은 날로부터 10일 이내에 재확인을 요청해야 합니다. 기한을 놓치면 최초 결정이 확정되므로 신속한 대응이 필요합니다.

필수 근거 자료

단순한 불복 의사 표시로는 불충분합니다. `추가 실증 데이터` , `활용 맥락 상세 설명` , `외부 전문가 소견서` 등을 첨부하여 최초 판단의 오류를 입증해야 합니다.

전략적 팁: 반증자료 사전 준비

이의제기 기간(10일)은 새로운 자료를 만들기에는 매우 짧습니다. 따라서 최초 확인 신청 단계에서부터 '고영향 아님'을 입증할 수 있는 반증 논리와 데이터를 미리 준비해 두는 것이 안전합니다.

재확인 절차 및 성격

재확인 요청이 접수되면 전문위원회의 자문을 거쳐 30일 이내에 최종 결과를 재회신 받게 됩니다. 해당 확인 결과는 참고적 성격을 가진 행정상 판단입니다(행정소송의 대상인 '행정처분'에 해당하는지 불분명)

Q9. HR/Fin/Health 판단 가이드 (산업별 사례)

구분	HR-Tech (채용/인사)	Fin-Tech (금융/신용)	Health-Tech (의료/건강)
고영향 예시 (High-Impact)	위험 영역 • 서류전형 **자동 합격/불합격** • 면접 평가 결과의 자동 점수화 • 승진·보상 결정의 자동 산정	위험 영역 • 대출 승인/거부 **자동 결정** • 신용등급 및 신용점수 산정 • 보험 인수 거부 및 요율 결정	위험 영역 • 진단·처방의 **자동 결정** • 수술 계획 수립 및 지원 • 응급환자 분류(Triage)
비해당 예시 (Low-Impact)	일반 영역 • 단순 직무 매칭 추천(지원자용) • 이력서 형식/오타 검토 • 단순 참고용 채용 통계 제공	일반 영역 • 단순 투자 참고 정보 제공 • 사기 의심(FDS) 알림만 제공 • (최종 판단은 사람이 수행)	일반 영역 • 일반적 웰니스(Wellness) 추천 • 단순 운동·식단 가이드 • 병원 예약/접수 자동화
핵심 판단 기준	인사권자의 실질적 개입 없이 고용 기회에 영향을 미치는 **자동 결정 여부가 핵심**	대출 거절 등 재산권에 직접적이고 법적 효력이 있는 **자동 결정 여부가 핵심**	의료인의 개입 없이 생명·신체에 영향을 주는 **자동 결정 여부가 핵심**

산업별 판단 Check Point

단순히 AI 기술을 사용하는 것만으로는 고영향 AI에 해당하지 않습니다. AI의 산출물이 **사람의 실질적인 검토나 개입 없이** 중대한 결과(합격/불합격, 대출 거절, 의료 진단 등)로 직결되는지 여부를 최우선으로 검토하십시오.

* 출처: 고영향 인공지능 판단 가이드라인 (과학기술정보통신부, 2026.1.29)

PART 03

고영향AI 영향평가

평가 항목, 주기, 결과 관리, 개선 프로세스 등
고영향AI 사업자에게 **권고되는 주요 사항**을 다룹니다.

Q11. 영향평가 7대 필수 항목

No.	평가 항목	상세 내용 및 검토 사항 (가이드라인 기준)
1	**영향받는 개인·집단 식별**	AI 활용으로 인해 **직·간접적인 영향**을 받게 될 이해관계자 및 **취약계층(아동, 노인, 장애인 등)**을 구체적으로 특정하고 분석
2	**기본권 유형 식별**	헌법상 보장된 기본권 중 **침해 가능성이 있는 권리**(생명권, 신체의 자유, 사생활의 비밀, 평등권, 재판받을 권리 등)를 명확히 식별
3	**사회·경제적 영향 범위**	고용, 재화/서비스 접근성, 시장 경쟁 등 사회·경제적 파급 효과의 **범위와 심각성**(피해 규모, 복구 가능성 등) 분석
4	**사용 행태 및 맥락**	AI 시스템의 의도된 목적과 실제 운영 환경, 사용자의 **오남용 가능성** 및 활용 맥락(Context)을 고려한 위험 시나리오 검토
5	**평가지표 및 산출 방식**	영향을 측정하기 위한 **정량적 또는 정성적 지표**와 구체적인 결과 산출 방법론 정의 (위험도 점수화, 전문가 평가 등)
6	**위험 예방·완화 조치**	식별된 위험을 제거하거나 수용 가능한 수준으로 낮추기 위한 **기술적(데이터 정제 등)·관리적(인간 개입 등) 통제 방안** 수립
7	**개선 이행계획**	평가 결과에 따른 사후 조치, 지속적인 모니터링 및 **구체적인 개선 이행** 일정과 책임자 지정 등 계획 수립

* 출처: 고영향 인공지능 영향평가 가이드라인 (과학기술정보통신부, 2026.1.22)

실무 Tip: 특히 **1번(피영향자 분석)**은 단순 사용자뿐만 아니라 '배제되는 집단'까지 포함해야 하며, **5번(평가지표)**은 반드시 문서화된 근거(객관적 데이터 또는 전문가 자문 결과)를 기반으로 작성해야 합니다.

Q11-보완. 영향평가 3단계 프로세스 (신설)

1

1단계: 사전 준비

주요 활동

- 평가 필요성 검토
- 평가 대상(AI 시스템) 정의
- 평가팀 구성 및 계획 수립

OUTPUT
평가 계획서, 평가 대상 정의서

2

2단계: 본 평가

주요 활동

- 피영향자 식별 및 분석
- 작동원리 및 데이터 분석
- 기본권 침해 위험 식별·평가
- 위험 완화 조치 수립

OUTPUT
영향평가 결과보고서

3

3단계: 사후 관리

주요 활동

- 후속조치 이행 점검
- 결과 문서의 체계적 보관
- 결과 요약본 공개(권장)
- 재평가 요건 발생 모니터링

OUTPUT
이행 점검표, 공개용 요약본

실무 가이드 Tip

본 평가는 **서비스 출시 전**에 완료되어야 하며, 평가 결과에 따른 위험 완화 조치가 시스템 설계에 반영되었는지 반드시 확인해야 합니다. 평가 결과는 **3년간 보관**을 권장합니다.

※ 출처: 과학기술정보통신부 「고영향 인공지능 영향평가 가이드라인」(2026. 1. 22.)

Q12. 자체평가 vs 제3자평가

구분	자체평가 (Self-Assessment)	제3자평가 (Third-Party Review)
주체	기업 내부의 전담 팀 또는 위원회 수행	외부 전문 기관, 법무법인, 컨설팅 펌 의뢰
장점 (Pros)	**속도** 신속한 의사결정 및 반영 가능 **비용** 외부 용역 비용 절감 **맥락** 서비스 특성과 내부 데이터 이해도 높음	**신뢰성** 객관적이고 중립적인 시각 확보 **전문성** 최신 법령 및 글로벌 표준 적용 용이 **대응력** 감독기관 및 고객사 요구 대응에 유리
단점 (Cons)	**객관성** 내부 편향(Confirmation Bias) 위험 **전문성** 법적·기술적 전문성 부족 가능성	**비용/시간** 높은 비용과 긴 수행 기간 **이해도** 내부 비즈니스 맥락 이해 부족 가능성
적합 대상	• 초기 단계의 AI 모델 • 위험도가 상대적으로 낮은 기능 • 잦은 변경이 발생하는 Agile 개발 과정	• **고영향AI (노력의무(권장))** • 대외 신뢰도가 중요한 핵심 서비스• 분쟁이나 규제 리스크가 높은 영역

권장 전략: 혼합형 접근 (Hybrid Approach)

핵심 서비스나 고영향AI의 경우, **평가는 내부 실무팀이 수행(자체평가)**하되, 최종 결과물에 대해 **외부 전문가의 검토(Review)를 받는 방식**이 효율성과 신뢰성을 동시에 확보할 수 있는 가장 현실적인 대안입니다.

Q13. 평가 시기 및 재시행 기준

ESSENTIAL

최초 평가 (필수)

고영향 인공지능 제품 또는 서비스의 **신규 제공 전 필수 수행**.
출시 전 잠재적 위험 요소를 사전에 식별하고 제거하는 것이 원칙.

TECHNICAL UPDATE

핵심 기능 변경 시

모델 아키텍처 변경, **학습 데이터의 대폭 추가** 등 AI 시스템의 성능이나 동작 원리에 중대한 변화 발생 시 재평가 수행 권장.

CONTEXT CHANGE

운영 환경 변경 시

서비스 이용 약관 변경, **타겟 사용자층 확대**, 운영 정책 변화 등 비기술적 요소가 이용자 권리에 영향을 줄 경우.

AD-HOC

수시 평가 (사고 시)

AI 사고 발생, **치명적 오류 발견** 또는 **규제 기관의 시정 요구**가 있을 경우 즉시 수시 평가를 통해 원인 분석 및 조치.

자율 vs 의무 판단 기준

재평가(재시행) 여부는 원칙적으로 **사업자가 자율적으로 판단**하나, 핵심 기능 변경이나 피영향자(이용자)의 기본권에 중대한 영향을 미칠 수 있는 변화가 있는 경우에는 재수행이 **강력히 권고**됩니다. (단, 단순 UI/UX 변경이나 경미한 업데이트는 제외 가능)

** 출처: 고영향 인공지능 영향평가 가이드라인 (과학기술정보통신부, 2026.1.22)*

Q14. 결과 활용 및 보관/공개

결과 및 근거 자료의 보관 (자율 권고)

법적 의무는 아니나, 신뢰성 입증을 위해 **영향평가 수행 계획서, 결과서 및 증빙자료를 체계적으로 보관·관리**할 것을 권고합니다.

보관 대상 수행 인력 현황, 회의록, 테스트 결과 로그, 개선 조치 이행 내역 등

결과 공개 범위 및 방식 (이원화)

기업 비밀 보호와 투명성 확보를 위해 공개 범위를 구분하여 운영 가능합니다.

① **전체 공개:** 결과서 전문 공개 (가장 높은 신뢰도)

② **요약 공개:** 핵심 위험 관리 방안 및 안전성 조치를 요약한 설명서(System Card 등) 공개

국가기관 제출 및 우선 고려 의무

과기정통부장관의 확인 요청 등이 있는 경우 **결과를 제출해야 하며**, 국가기관등은 AI 제품·서비스 도입 시 영향평가 결과를 **우선적으로 고려**해야 합니다(법적 근거).

실무 팁: 문서 체계 수립

단순 결과서 뿐만 아니라 '버전 관리'가 포함된 이력 관리 대장을 작성하여, 모델 업데이트 시 영향평가 재수행 여부와 그 결과를 추적 가능하도록 관리해야 합니다.

* 출처: 고영향 인공지능 영향평가 가이드라인 (과학기술정보통신부, 2026.1.22)

평가 산출물 팁

기업자율적으로 AI 영향평가 및 자율적 리스크 관리를 위해 **체계적인 문서 관리**가 권장됩니다.

실무 관리 문서

권장 산출물 (Recommended)

- **영향평가서:** 7대 평가 항목에 대한 상세 분석 결과 (자율 평가)
- **위험관리 계획:** 식별된 리스크에 대한 예방 및 완화 조치 문서
- **개선 이행점검표:** 문제점 발견 시 조치 이력 및 결과
- **내부 관리 정책:** AI 거버넌스 및 윤리 원칙 수립 문서

내부 관리용 (Internal)

- **결재 라인 증빙:** 주요 의사결정에 책임자 승인 기록 확보
- **회의록:** 위험성 검토 및 조치 결정 과정에 대한 회의 기록
- **변경 이력 관리:** 모델 업데이트, 데이터셋 변경 등 이력 추적
- **교육 훈련 기록:** 내부 임직원 대상 AI 윤리 및 법규 교육 이력

Q15. AI로 인한 기본권 침해 발생 시 대응 프로세스

1

영향 중지·완화

즉각적인 피해 확산 방지 및 임시 조치 시행

- 서비스 중단 또는 기능 제한
- 임시 안전망(Safety Guard) 적용
- 긴급 공지 및 고객 안내

2

원인 분석

문제 발생의 근본 원인 파악 및 범위 식별

- 데이터/모델/운영 로그 분석
- 오류 발생 경로 역추적
- 영향 범위(피해자 규모) 산정

3

수정·재검증

모델 및 시스템 수정 후 안전성 재확인

- 알고리즘 패치 및 재학습
- 회귀 테스트 및 안전성 검증
- 제3자 검증(필요시)

4

이해관계자 통지

관련 당사자에게 상황 공유 및 보상 안내

- 규제기관 보고(중대 사고 시)
- 피해 고객 대상 통지
- 피해 보상 절차 안내

5

재발 방지

장기적인 안전성 확보를 위한 체계 개선

- 운영 정책 및 프로세스 개선
- 모니터링 시스템 고도화
- 임직원 교육 및 훈련

Tip: 사고 대응 매뉴얼은 **사전 시뮬레이션**을 통해 실효성을 검증해야 합니다. 특히 **1단계(영향 중지)**의 골든타임을 놓치지 않도록 킬스위치(Kill-Switch) 권한을 명확히 지정해두세요. AI 영향평가는 제공 전 잠재적 위험 식별용이며, 즉각적인 대응이 필요한 사고 대응과는 구별됩니다.

Q14-보완. 타 영향평가와의 관계 (신설)

대상 제도	중복 가능 영역	중복 해소 및 통합 전략 (가이드라인)
개인정보 영향평가 (PIA)	• 데이터 수집·처리의 적절성 • 개인정보 유출 방지 조치 • 프라이버시 침해 위험 분석	• **평가 결과 상호 인정** (중복 항목 면제) • PIA 수행 시, AI 영향평가의 데이터 보호 관련 항목(제7조)을 **이행한 것으로 간주(갈음)** • 개인정보보호위원회와 공동 가이드라인 마련 예정
디지털의료제품 영향평가	• 임상적 유효성 및 안전성 검증 • 의료기기 인허가 심사 기준 • 사용자 안전 확보 조치	• **의제 처리(Deemed Compliance)** 적용 • 식약처 허가/심사를 통과한 경우, 고영향 AI로서의 안전성 확보 조치를 **이행한 것으로 인정** • 별도의 추가 평가 없이 결과서 제출로 갈음 가능
기타 평가 (정보통신망법 등)	• 이용자 보호 지침 준수 여부 • 서비스 안정성 확보 조치 • 정보보호 관리체계(ISMS) 등	• **통합 평가 창구(One-Stop) 운영** • 유사 인증/평가 결과를 증빙자료로 활용 허용 • 부처 간 협의체를 통해 중복 규제 지속 발굴 및 해소

과기정통부 가이드라인 핵심 (Integration Policy)
사업자의 부담을 최소화하기 위해 **"동일한 목적과 내용의 평가는 중복하여 요구하지 않는다"**는 원칙을 명시하고 있으며, 타 법령에 따른 평가 결과를 적극적으로 수용하여 통합 관리할 예정입니다.

* 출처: 과학기술정보통신부, 「고영향 인공지능 영향평가 가이드라인」 (2026.1.22)

PART 04

생성형AI 투명성 확보 의무

사전 고지, 결과물 표시, 예외 사항 및 책임 범위 등
생성형AI 서비스 제공자가 준수해야 할 투명성 의무를 다룹니다.

Q16. 사전 고지 3가지·예외

사전 고지 3가지 방법 (필수)

가이드라인 제3장: 이용자가 AI 사용 사실을 인지할 수 있도록 고지

계약서·약관 명시

이용약관이나 계약서에 **"생성형 AI 기술 활용"** 사실을 명확히 기재

> 예: "본 서비스의 챗봇 기능은 생성형 AI 모델에 기반하여 운영됩니다."

화면 표기

서비스 초기 화면, 대화창 상단 등 **눈에 잘 띄는 곳**에 표시

> 예: 챗봇 대화창 상단에 "AI가 생성한 답변입니다" 문구 고정 노출

인식 용이한 게시

오프라인 매장이나 키오스크 주변에 **쉽게 인식할 수 있는 방식**으로 게시

고지 의무 예외

다음의 경우 사전 고지 의무가 면제될 수 있음 (엄격 해석)

내부 업무용 사용

사업자가 **내부 업무 효율화**를 위해 임직원만 사용하는 경우

* 단, 직원에게는 사내 공지 등을 통해 알리는 것이 바람직함

활용 사실이 명백한 경우

서비스 명칭이나 기능 설명을 통해 이용자가 **이미 명확히 인지** 가능한 경우

> 예: 서비스명이 "AI 챗봇", "인공지능 비서" 등으로 명시된 경우

연구·개발 목적

일반 대중에게 공개되지 않은 **테스트, 연구 목적**의 제한된 환경

실무 팁: '고지 매트릭스' 운영 권장

제품별, 채널별(Web/App/API)로 어떤 고지 방법을 적용했는지 추적 관리하는 **'고지 매트릭스'**를 운영하세요. 특히 글로벌 서비스의 경우 국가별 규제(EU AI Act 등)에 맞춘 다국어 고지 문구 세트 관리가 필수적입니다. '명백한 경우'에 대한 판단은 보수적으로 접근하여, 가능한 **중복 고지**하는 것이 안전합니다.

출처: 과학기술정보통신부, 「투명성 확보 가이드라인」(2026.1.22)

Q17. 결과물 표시 의무 방식 (가이드라인 기준)

구분	서비스 내 제공 (In-Service) 웹/앱 UI, 플랫폼 화면	외부 반출 (Export) 다운로드, 공유, 배포
일반 생성물	• 서비스 화면 내 정보 표출 가능 (UI 표시) ※ 단, 기계 판독 방법 적용 시 별도 안내 필수	• **결과물 자체에 표시 필수** (워터마크 등) ※ 서비스 UI에 표시했더라도, 반출 시에는 파일 내 식별 정보가 포함되어야 함
딥페이크 생성물	• 일반 생성물과 동일 (실제와 구분 불가 시) ※ 오인·혼동 위험 고려	• **사람이 인식할 수 있는 방법** 필수 ※ 기계 판독만으로는 부족하며, 직관적인 식별 표지(로고, 문구)가 반드시 포함되어야 함
표시 방법 원칙	• **혼합 권장**: 사람 인식 + 기계 판독 **※ 기계 판독(메타데이터)만 사용 시 최소 1회 이상 안내 문구/음성 필수**	• **매체별 최적화** 이미지(가시적 워터마크), 음성(시작/종료 안내음), 영상(자막/로고) 등 매체 특성 반영

주의: 다운로드 기능 제공 시 의무 강화

단순히 서비스 화면에서 "AI 생성됨"을 알리는 것만으로는 부족합니다. **사용자가 결과물을 다운로드하거나 공유할 때**에도 그 사실을 알 수 있도록 **결과물 파일 자체에 워터마크나 메타데이터를 삽입**해야 법적 의무를 충족합니다.

* 출처: 과학기술정보통신부, 「투명성 확보 가이드라인」 (2026. 1. 22.)

Q18. 워터마킹 vs 메타데이터 (가이드라인 기준)

구분	디지털 워터마킹 (Watermarking)	메타데이터 (Metadata)
기술 방식	**가시적:** 이미지 내 로고 삽입 (사람 인식 가능) **비가시적:** 콘텐츠 내 식별 정보 삽입 (기계 판독)	파일 헤더/별도 영역에 생성 정보 기록 (IPTC, EXIF 등 표준 활용)
적용 시나리오	이미지/동영상 생성 시 **가시적 방법** 우선 **※ 딥페이크 생성물은 사람이 인식할 수 있는 방법만 적용 (필수)**	텍스트, 코드, 파일(PDF 등) 생성 시 기본 적용 **※ 기계 판독 방법 적용 시 1회 이상 별도 안내 문구 필수**
장단점	직관성 가시적 방법은 사용자가 즉시 인지 품질 원본 콘텐츠 품질 저해 가능성	호환성 다양한 플랫폼/뷰어 호환 (IPTC 표준) 취약성 스크린샷/포맷 변환 시 정보 소실
기술적 요구사항	• 가시적: 로고/문구 삽입 (크기/위치 고려) • 비가시적: 강건성 확보 (리사이징/압축 견딤)	• 표준 준수: C2PA, IPTC Photo Metadata • 무결성: 위변조 방지 기술 적용 권장

실무 권장: 이중 적용 전략 (Dual Layer Strategy)

가이드라인은 **'사람이 인식할 수 있는 방법(가시적)'**과 **'기계가 판독할 수 있는 방법(비가시적)'**의 병행을 권장합니다. 특히 딥페이크 위험이 있는 콘텐츠는 반드시 **가시적 표시(사람 인식 가능)**를 우선 적용해야 합니다.

** 출처: 과학기술정보통신부 「투명성 확보 가이드라인」 (2026.1.22)*

Q19. 다운로드 시 표시 의무(핵심)

구분	사람이 인식할 수 있는 방법 (가시적)	기계가 판독할 수 있는 방법 (비가시적)
텍스트	**명시적 기재** 파일 머리말(Header) 또는 문서 작성 영역 초반에 AI 생성 사실 기재	**메타데이터** 파일 속성 정보에 생성 정보 기록 * 비가시적 방법 적용 시, 다운로드 시점에 안내 필수
이미지	**워터마크** 이미지 내 식별 가능한 로고나 문구 삽입	**디지털 워터마크** 육안으로 보이지 않는 신호 삽입 **메타데이터** Exif, IPTC 등 표준 데이터 활용
동영상	**화면 표출** 영상 일부 영역에 로고 지속 표출 **자막 안내** 영상 시작 부분에 AI 생성 사실 안내	**비디오 워터마킹** 비가시적 패턴 삽입 **메타데이터** 영상 파일 헤더 정보 활용
음성	**음성 안내** 오디오 시작 부분에 멘트로 안내 (예: "AI가 생성한 음성입니다")	**오디오 워터마킹** 비가청 주파수 대역 활용
기타	파일 포맷 특성(PPT 슬라이드 등)을 고려하여 첫 화면이나 주요 영역에 명확히 표시	파일 작성자 정보 등 메타데이터 영역 활용

딥페이크(Deepfake) 관련 필수 주의사항

실제와 구분하기 어려운 가상의 결과물(딥페이크)을 제공하는 경우, 이용자의 오인 및 혼동을 방지하기 위하여 **반드시 사람이 명확히 인식할 수 있는 방법(가시적/가청적 표시)을 적용**해야 합니다. (기계 판독 방법만으로는 의무 불충족)

* 출처: 과학기술정보통신부, 「인공지능 투명성 확보 가이드라인」(2026.1.22)

Q20. B2B API 제공 시 투명성 책임 명확화

API 제공사의 책임

1. 기술적 지원

최종 사업자가 표시 의무를 이행할 수 있도록 **워터마킹**이나 **메타데이터 삽입** 기능을 기술적으로 지원해야 함

2. 고지 지원

최종 이용자 대상의 올바른 표시 방법을 안내하는 **가이드라인**을 제공하여 고객사의 규제 준수를 도움

3. 계약서 명시

계약 조항을 통해 **"최종 사업자가 이용자에게 표시할 의무가 있음"**을 명확히 규정하여 책임 소재 구분

최종 서비스 제공자의 책임

1. 표시 의무 이행

실제 이용자와 접촉하는 주체로서, 서비스 UI 및 결과물에 **AI 생성 사실**을 직접 표시해야 할 1차적 의무자

2. API 제공사와 협력

제공사가 지원하는 기술적 조치(워터마크 등)를 서비스에 **누락 없이 적용**하고 가이드라인을 준수

3. 이용자 교육

이용자가 AI 생성물의 한계와 특성을 이해하도록 **약관 및 안내문**을 통해 충분히 설명

책임의 우선순위 주의

"최종 이용자에게 직접 제품·서비스를 제공하는 사업자가 1차적인 투명성 확보 책임"을 집니다. 단순히 API만 제공하는 개발사는 기술적 지원 의무는 있으나, 이용자에 대한 직접 고지 의무는 면제될 수 있습니다.

* 출처: 과학기술정보통신부, 「인공지능 투명성 확보 가이드라인」 (2026.1.22)

PART 05

계약 및 법적 책임

계약서 핵심 조항, 공동책임, 사고 시 분담 등
법적 리스크 관리와 책임 소재를 명확히 합니다.

Q23. 최첨단 AI 안전성 의무

적용 기준 (Threshold)

- 학습 누적연산량 **10^{26} FLOPs 이상**
- 최첨단 AI 기술 적용 모델
- 생명·신체·기본권에 광범위하고 중대한 영향

02

위험관리체계 (Risk Mgmt)

- 3단계 위험 관리 프로세스 (식별 → 평가 → 완화)
- 안전사고 상시 모니터링
- 위험 대응 전담 조직 구성

03

결과 제출 의무 (Submission)

- 적용 대상 인지 후 **초기 안전성 결과 제출**
- 모델 변경/업데이트 시 추가 결과 제출 필수
- 제출처: 과학기술정보통신부

04

안전사고 보고 (Reporting)

- **24시간 이내**: 최초 보고 (사고 인지 시점 기준)
- **7일 이내**: 초동조치 보고
- **15일 이내**: 최종 결과 보고

시행령 제24조 및 안전성 확보 가이드라인 참조

* 출처: 최첨단 AI 안전성 확보 가이드라인 (과학기술정보통신부, 2026. 1. 22.)

Q24. 고영향 AI 사업자 5대 책무 NEW

구분	핵심 조치 사항 (Action Items)	필수 증빙 문서
1. 위험관리	• 수명주기 전반에 걸친 **위험 식별·평가·완화** 절차 수립 및 이행 • 위험의 식별 및 평가 결과에 따른 완화 조치 수행	위험관리 계획서 위험평가 결과보고서
2. 설명가능성	• AI가 도출한 **최종 결과 및 도출 기준** 설명 • AI 개발·활용에 사용된 **학습용 데이터 개요**(출처, 품질 등) 설명	AI 모델 설명서(Model Card) 데이터 명세서
3. 이용자 보호	• 이용자 **피드백 및 고충 처리 체계** 구축 • 서비스 모니터링 및 사고 발생 시 긴급 대응 절차 마련	이용자 보호 지침 사고 대응 매뉴얼
4. 인간 감독	• 자동화된 결정에 대한 **인간의 개입 메커니즘** 마련 • 오작동 시 즉시 중단 가능한 **킬스위치(Kill-Switch)** 또는 비상 정지 절차	인간 감독 운영 절차서 운영 로그(Log)
5. 문서화	• 안전성·신뢰성 확보 조치 결과를 증빙하는 **안전신뢰문서** 작성·보관 • 정기적인 **자가점검 수행** 및 결과 기록 유지	안전신뢰확보 문서 자가점검표

* 출처: 과학기술정보통신부 「고영향 인공지능 사업자 책무 가이드라인」 (2026.1.22)

Q25. 계약서 필수 5대 조항

순번	조항 구분	핵심 내용 및 검토 포인트
1	**AI법 준수·협력**	• 서비스 제공 시 **AI 기본법 및 관련 법령 준수 의무** 명시 • 법령 변경에 따른 서비스 수정 및 업데이트 지원 확약
2	**데이터·로그 보관**	• 규제기관 제출 요구 시 필요한 **로그 데이터 및 증빙 자료 접근권** 보장 • 데이터 보존 기간 설정 (최소 법정 의무 기간 이상)
3	**위험관리 협력**	• 고영향AI 지정 시 **영향평가 수행에 필요한 기술적 정보 제공** • 리스크 식별 및 완화 조치에 대한 상호 협력 의무
4	**책임·배상·보험**	• AI 오작동, 편향성 문제 발생 시 **손해배상 책임 범위 및 한도** 설정 • AI 관련 전문인 배상 책임 보험 가입 여부 및 증빙
5	**서브/오픈소스**	• 하청업체(Sub-processor) 변경 시 사전 통지 및 승인 절차 • 사용된 **오픈소스 라이선스 고지** 및 컴플라이언스 준수 보증

주의: 면책 조항의 한계

"AI 결과물에 대해 어떠한 책임도 지지 않는다"는 식의 **포괄적 면책 조항은 법적 효력이 제한적**일 수 있습니다. 특히 고의 또는 중대한 과실로 인한 손해, 생명·신체에 대한 침해의 경우 면책이 무효화될 가능성이 높으므로, 구체적인 책임 범위(SLA 등)를 설정하는 것이 안전합니다.

Q26. AI 사고 시 책임 소재 (케이스 분석)

구분 (원인)	주요 책임 주체	책임 판단 핵심 기준
모델 결함	**개발사 (Developer)** 알고리즘 설계 오류, 데이터 편향	• 제조물 책임법상 '설계상의 결함' 유무 • 출시 전 충분한 테스트 및 검증(V&V) 수행 여부 • **데이터 편향성 제거 노력** 입증
운영 설정 부실	**제공사 (Provider)** 파라미터 설정, 모니터링 미흡	• 서비스 운영 및 관리·감독 소홀 여부 • 이상 징후 탐지 및 **긴급 조치(Kill-Switch)** 작동 여부 • 이용자 보호 조치 이행 여부
오남용/일탈	**이용자 (User)** 약관 위반, 금지 행위, 프롬프트 해킹	• 서비스 이용약관(AUP) 및 가이드라인 위반 여부 • 고의적인 **악의적 프롬프트 주입** 시도 • 제공사가 경고/제재 조치를 취했는지 여부
복합 원인	**공동 책임** 과실 비율에 따른 분담	• 계약서상 **책임 분담 조항** (Indemnification) • 각 주체의 주의의무 위반 정도 (로그 기록 분석) • 면책 조항의 유효성 (고의/중과실 제외)

책임 배분의 결정적 요소: 증빙(Evidence)

사고 발생 시 가장 중요한 것은 **'누가 주의의무를 다했는가'**를 입증하는 것입니다.
개발사는 **모델 카드/테스트 리포트**를, 제공사는 **운영 로그/모니터링 기록**을 확보해야 면책 가능성이 높아집니다.

Q27. API 제공사 공동책임

공동책임 리스크 식별

API 제공사가 직접 고영향AI를 운영하지 않더라도, 고객사(서비스 사업자)가 해당 API를 고영향 용도(예: 채용, 대출 심사)로 사용할 경우 법적 공동책임이 발생할 수 있습니다.

리스크 완화 조치 (Mitigation)

계약과 기술적 조치를 통해 책임을 명확히 구분해야 합니다.

용도 제한 | AUP(허용 사용 정책) | 사전 통지의무 | 킬스위치(접근 차단)

증빙 및 면책 전략

API 제공사가 '선량한 관리자의 주의의무'를 다했음을 입증할 수 있는 근거를 확보해야 합니다.

- **기술 로그:** 이상 징후 모니터링 및 차단 이력
- **조치 이력:** AUP 위반 고객에 대한 경고 및 제재 기록

실무 권장사항

계약서에 "고영향AI 용도로 사용 시 사전에 서면으로 통지하고 별도 계약을 체결해야 한다"는 조항을 추가하여, 미통지 고영향 사용에 대한 책임을 고객사에게 귀속시켜야 합니다.

Q28. 오픈소스 모델 활용 책임

OPEN SOURCE RISK

오픈소스 AI 모델 활용 시 주의사항

라이선스 & 범위

상업적 이용 가능 여부 및
라이선스 조항을 철저히 확인하세요.
(Apache 2.0, MIT, RAIL 등)

책임 제한 (AS-IS)

대부분 '있는 그대로(AS-IS)' 제공되며
보증이 없으므로, 최종 책임은
활용 기업에게 귀속됩니다.

테스트/밸리데이션

자체적인 성능 검증과
안전성 테스트를 필수적으로
수행하고 문서화해야 합니다.

보안/저작권 스캔

악성코드 및 학습 데이터의
저작권 침해 요소를 사전에 스캔하여
리스크를 차단하세요.

투명성 확보

활용한 모델명과 데이터 출처를
명확히 표시하여
투명성 의무를 준수하세요.

"오픈소스니까 우리 책임 아니다"라는 주장은 통하지 않습니다. 오픈소스 모델을 서비스에 도입하는 순간, 해당 AI 시스템에 대한 법적 책임(고영향 지정, 투명성 등)은 **도입 기업**이 지게 됩니다.

PART 06

글로벌 대응 및 정합성

EU·미국·한국 동시 대응 전략

Q29. EU AI Act vs 한국 AI법

구분	EU AI Act	한국 AI 기본법
적용 범위 (Scope)	**강력한 역외 적용** EU 시장에 AI 시스템/서비스 제공 시 설립지 무관하게 전면 적용	상대적으로 **국내 중심** 국내 제공·이용 시 적용되나, 집행력의 실질적 한계 존재 가능성
제재 수위	위반 시 최대 **전 세계 매출의 7%** 또는 3,500만 유로 중 높은 금액 (금지된 AI 사용 시)	과태료 및 시정명령 중심 상대적으로 **약한 처벌 수위** (산업 육성 측면 고려)
고영향 카테고리	**명확한 리스트 기반** (Annex III) 8개 분야 구체적 열거 (생체인식, 교육, 고용, 필수 공공서비스 등)	**2단계 판단 기준** ① 특정 영역 해당 + ② 중대한 영향 (유연하지만 예측 가능성 다소 낮음)
집행 체계	**사전 적합성 평가** (CE 인증) 출시 전 엄격한 적합성 평가 의무화 제3자 인증 기관 개입	**확인 + 영향평가** (사후 관리 중심) 사업자 확인 요청 및 자율/지정 영향평가 자율규제와 책무 중심의 접근

글로벌 컴플라이언스 전략

글로벌 서비스를 지향한다면 **가장 엄격한 기준인 EU AI Act를 준수**하는 것이 안전합니다.
EU 기준을 충족하면 한국 AI 기본법의 요구사항 대부분을 자연스럽게 만족시킬 수 있습니다.

Q30. 글로벌 서비스 준수

역외 적용 범위 사전 점검

서비스 대상 국가의 AI 규제가 적용되는지 확인해야 합니다. EU AI Act와 미국 캘리포니아주 법률 등은 역외 적용(Extraterritoriality) 조항을 포함하고 있어, 해당 지역 이용자를 대상으로 할 경우 현지 법인 유무와 관계없이 규제 대상이 될 수 있습니다.

데이터 이전 및 지역화(Localization) 이슈

학습 데이터 및 이용자 데이터의 국경 간 이전 제한 여부를 검토해야 합니다. 특히 EU GDPR 적정성 결정, 중국의 데이터 보안법 등 각국의 데이터 주권 관련 법령과 AI 규제 간의 상충 가능성을 고려한 아키텍처 설계가 필요합니다.

현지 요건: 대리인, 공지, 라벨링

- EU 대리인 지정 — EU 내 사업장이 없는 경우 현지 대리인 지정 필수
- 다국어 공지 — 이용자 약관 및 AI 사용 고지의 현지 언어 지원
- 맞춤형 라벨링 — 국가별로 요구하는 워터마크 기술 표준 및 표시 방식 준수

법령 버전 관리의 중요성

각국의 AI 규제는 빠르게 신설되거나 개정되고 있습니다. 주요 진출 국가의 법령 제·개정 현황을 모니터링하고, 규제 변경 시 서비스에 즉시 반영할 수 있도록 유연한 컴플라이언스 관리 체계(Version Control)를 운영해야 합니다.

Q31. 통합 컴플라이언스 전략

01

'최고 기준' 원칙 (One High Standard)

- 가장 엄격한 규제(EU AI Act 등)를 기본 표준으로 설정
- 중복 작업을 방지하고 전사적 통일성 확보
- 하향식 적용으로 로컬 규제 자동 충족 유도

02

정책·프로세스 단일화 (Core + Adapter)

- 공통 기반(Core): 윤리 원칙, 위험 관리, 거버넌스
- 지역별 어댑터: 국가별 특화 요건(한국 고지 등)만 추가
- 유지보수 효율성 극대화 및 관리 비용 절감

03

증빙 팩토리 구축 (Evidence Factory)

- 공통 산출물(데이터셋 명세, 모델 카드) 템플릿화
- 한 번의 문서화로 다국가 규제 대응 활용
- 자동화 도구 연동으로 실시간 증빙 확보

Q32. ISO/IEC 42001 연계

국제 표준인 **ISO/IEC 42001(AI 경영시스템)**을 도입하면
국내외 AI 법규 대응을 효율적으로 관리할 수 있습니다.

국제 표준 활용 전략

전략적 가치

- **거버넌스 체계화 (AIMS)**
  AI 정책, 위험평가, 내부심사 등 체계적인 관리 프로세스 구축을 통한 전사적 리스크 통제
- **법 요건 맵핑 효율화**
  ISO/IEC 42001의 통제 항목과 AI 기본법 요구사항을 연계하여 중복 작업 최소화

실질적 혜택

- **감사 및 고객 대응 유리**
  정부 감사나 대형 고객사의 실사 요구 시 '국제 표준 인증'으로 신뢰성 입증 용이
- **글로벌 마케팅 효과**
  인증 취득 사실을 통해 AI 신뢰성과 안전성을 대외적으로 홍보하고 기업 이미지 제고

PART 07

실무 대응 로드맵

D-90 ~ D+30 실행 계획 및 거버넌스 구축, 문서화 전략 등
체계적인 실무 대응 방안을 제시합니다.

Q32. 7단계 체크리스트 (실행 로드맵)

1 2 3 4 5 6 7

Step 1
스코핑 (Scoping)
- AI 시스템 전수 조사
- 고영향/생성형 여부 1차 분류
- 적용 법령 확인

Step 2
리스크 식별 (Identification)
- 영향받는 기본권 식별
- 예상 피해 시나리오 도출
- 이해관계자 분석

Step 3
데이터/모델 점검 (Verification)
- 학습 데이터 편향성 검사
- 모델 성능/정확도 테스트
- 설명가능성 기술 검토

Step 4
거버넌스 구축 (Governance)
- AI 윤리/정책 수립
- 담당 조직/책임자 지정
- 내부 통제 프로세스 마련

Step 5
영향평가 (Impact Assessment)
- 7대 항목 정밀 평가
- 평가 보고서 작성
- 위험 완화 조치 계획 수립

Step 6
통제 이행 (Implementation)
- 투명성(고지/표시) 적용
- 안전조치/킬스위치 구현
- 이용자 보호/구제 절차 마련

Step 7
모니터링·보고 (Monitoring)
- 운영 단계 상시 모니터링
- 정기 재평가 및 감사
- 규제 당국 보고/소명

실행 포인트

각 단계는 순차적으로 진행하되, **Step 3(기술적 점검)**과 **Step 4(거버넌스)**는 병행하는 것이 효율적입니다. 특히 법 시행 전이라면 **Step 1~4**까지를 우선 완료하여 규제 대응 기반(Baseline)을 확보하는 것이 중요합니다.

Q33. AI 거버넌스 조직

실무 전담 조직 (1st & 2nd Line)

책임임원 (C-Level)
AI 윤리/신뢰성 총괄 책임, 예산/인력 배분 권한

모델리스크팀
모델 성능 모니터링, 알고리즘 편향성 검증

데이터팀
학습 데이터 품질/출처 관리, 개인정보 처리

보안팀
모델/데이터 유출 방지, 적대적 공격 방어

법무/컴플라이언스
법적 규제 대응, 계약 검토, 리스크 판단

품질(QA)팀
기능 테스트, 안전성 기준 충족 여부 확인

AI 위원회 (Decision Making)

출시 전 심의 (Launch Gate)
고영향AI 모델의 배포 전 최종 승인, 영향평가 결과 및 위험 완화 조치 적정성 검토

이슈/사고 심의 (Incident Review)
AI 사고 발생 시 대응 방안 결정, 중대 결함 발견 시 서비스 중단/리콜 의사 결정

사후 모니터링 (Post-Audit)
정기적인 운영 현황 보고 수렴, 기술/법제도 변경에 따른 정책 개정 승인

권고: 독립된 검토 라인 (Second Line of Defense)
개발 부서(1st Line)와 검증 부서(2nd Line)를 분리하여 상호 견제와 균형을 유지해야 합니다. 개발팀이 스스로 검증하고 배포하는 구조는 규제 대응과 리스크 관리 측면에서 취약하므로, 법무/리스크팀의 독립적인 검토 권한을 보장하는 것이 중요합니다.

Q34. 중소기업 지원 프로그램

AI 신뢰성 검증 및 평가 지원

정부 지정 시험·인증 기관을 통해 신뢰성 평가를 수행할 경우 비용의 일부를 바우처 형태로 지원합니다. 데이터·모델 평가에 소요되는 고비용 부담을 경감할 수 있습니다.

전문가 컨설팅 및 기술 지원

법률·기술 전문가로 구성된 컨설팅단을 통해 영향평가 수행 방법, 투명성 의무 이행, 내부 거버넌스 구축 등에 대한 맞춤형 자문을 제공받을 수 있습니다.

규제 샌드박스 (임시허가/실증특례)

혁신적인 AI 서비스가 규제로 인해 출시가 어려울 경우, 일정 기간 규제를 면제하거나 유예하는 규제 샌드박스 제도를 활용하여 시장 검증 기회를 확보할 수 있습니다.

활용 팁: 초기 단계부터 연계
제품 개발 초기 단계(PoC)부터 지원 프로그램을 신청하여 규제 준수 비용을 절감하고, 출시 전 신뢰성 검증을 완료하여 시장 경쟁력을 확보하는 것이 유리합니다.

Q35. 정부 지원 신청 절차

1
요건 확인
- 사업별 공고문 확인
- 지원 자격 충족 여부
- 제외 사유 점검

2
신청서/계획서
- 사업계획서 작성
- 증빙 서류 준비
- 온라인 시스템 제출

3
심사
- 서면/발표 평가
- 현장 실사(필요시)
- 선정 결과 통보

4
협약 체결
- 사업비 조정
- 수정계획서 제출
- 전자 협약 체결

5
집행
- 사업 수행
- 예산 집행 관리
- 중간 보고

6
정산
- 최종 보고서 제출
- 회계 감사
- 사업비 정산

혁신성 (Innovation)
기존 기술/서비스 대비 차별성 및 기술적 우수성

공공성 (Public Interest)
사회적 문제 해결 및 국민 삶의 질 향상 기여도

확산성 (Scalability)
시장 파급 효과 및 타 산업으로의 적용 가능성

문서화 필수 산출물

01

정책 문서 (Policy Docs)

- AI 거버넌스 정책 및 윤리 원칙
- 데이터 라벨링 및 품질 관리 기준
- 규제 준수 및 리스크 관리 정책

02

기술 문서 (Technical Docs)

- AI 모델 카드 (Model Cards)
- 데이터 시트 (Datasheets for Datasets)
- 성능 테스트 및 검증 리포트

03

운영 문서 (Operational Docs)

- 고영향AI 영향평가서 (AIIA)
- 모델 변경 관리 및 버전 이력
- AI 사고 대응 및 복구 플랜

PART 08

위반 시 제재 및 대응

제재 수위 이해와 즉시 대응체계

Q36. 제재 유형 및 수위

제재 유형	주요 내용 및 요건	적용 대상 및 예시
1단계 **시정명령**	• 위반 사항에 대해 **일정 기한 내 시정** 요구 (1차적 조치) • 법 제36조제1항(시정명령) 근거 • 이행 완료 시 추가 제재 면제 가능	• 정당한 사유 없이 자료 미제출 • 거짓으로 자료 제출 • 관계 공무원의 출입·검사 거부·방해·기피
2단계 **과태료**	• 시정명령 불이행 시 부과하거나, 특정 의무 위반 시 **별도 시정명령 없이 직접 부과 가능** (동시 부과 가능) **제43조(과태료)** 다음 각 호의 어느 하나에 해당하는 자에게는 **3천만원 이하의 과태료**를 부과한다. 1. 제31조제1항 위반 (고지 미이행) ***직접부과** 2. 제36조제1항 위반 (국내대리인 미지정) ***직접부과** 3. 제40조제3항 위반 (중지/시정명령 미이행)	• **즉시 과태료 대상:** 표시의무 위반, 대리인 미지정 • **시정명령 후 부과:** 자료 제출 거부 등 절차적 위반 후 불이행 시
3단계 **서비스 중지**	• **매우 예외적인 경우 시정명령으로 서비스 중지/영업정지 가능** • 사람의 생명·신체에 중대한 위험 초래 및 신속한 개선 불능시	• 안전성 기준 미달로 인명 사고 발생 우려 • 시정명령을 통해서도 위험 해소가 어려운 경우 • (시정명령의 일환으로 해석 가능)

↑ 가중 요소 (제재 강화)

- 위반 행위의 **고의성** 및 조직적 은폐 시도
- 과거 3년 내 **반복적 위반** 이력

↓ 감경 요소 (제재 완화)

- 위반 사실의 **자진 신고** 및 조사 협조
- 문제 인지 후 **즉각적인 시정 조치** 완료

Q37. 영향평가 미이행 대응

미이행 점검 포인트 (Self-Audit)

영향평가가 누락된 원인을 식별하는 것이 우선입니다.

대상 오판 고영향AI에 해당하지 않는다고 잘못 판단한 경우 **주기 미준수** 정기평가 시기를 놓치거나 대규모 변경 후 수시평가 누락

자율적 개선 권고 (Improvement)

미이행 사실 인지 시 자발적으로 개선 절차를 밟는 것이 바람직합니다.

평가 착수 영향평가 프로세스를 즉시 가동하고 문서화 시작 권장 **임시 통제** 평가 완료 전까지 고위험 기능에 대한 모니터링 강화

고객 안내 필요 시 고객사에게 상황을 투명하게 안내하고 협조 요청

재발 방지 대책 (Prevention)

지속적인 개선을 위한 내부 프로세스를 정비해야 합니다.

변경관리 연동 모델/데이터 변경 시 영향평가 검토 절차 내재화 **프로세스 개선** 배포 전 검수 단계에 영향평가 이행 여부 확인 절차 추가

법적 성격 유의

영향평가는 법적 노력의무이므로 미이행 자체가 즉각적인 법적 제재 대상이라고 단정하기는 어렵습니다. 다만, 안전한 AI 활용을 위해 자율적으로 이행하는 것이 권장됩니다.

Q38. 표시의무 위반 주의사항

VIOLATION CHECK

표시의무 위반 시 제재 및 주요 실수 유형

주체 오류

표시 의무를 **최종 이용자**에게 전가하는 것은 위법입니다. (약관으로 책임 전가 불가)

상황 오류

화면상 표시는 했으나, **'다운로드 파일'**에는 표시가 누락되는 경우가 가장 빈번한 위반입니다.

개선 방안

자동 라벨링 **파이프라인** 구축 및 배포 전 **QA 체크리스트**에 표시 검증 항목을 필수화하세요.

Solution Tip: C2PA/IPTC 표준 메타데이터를 자동 삽입하는 모듈을 백엔드에 통합하고, 워터마크가 시각적으로 보이지 않더라도 기계 판독이 가능하도록 구현해야 합니다.

Q40. 사고 긴급 대응 7단계 매뉴얼

1 탐지/신고
모니터링 알림
고객 민원 접수
이상 징후 포착
즉시 대응

2 중지/차단
서비스 일시 중단
킬스위치 가동
피해 확산 방지
긴급 조치

3 평가/분석
원인 규명
영향 범위 산정
법적 리스크 검토
원인 파악

4 소통
이용자 공지
규제기관 보고
언론 대응
대외 대응

5 수정/재검증
모델/데이터 수정
안전성 테스트
서비스 재개
복구

6 보상 검토
피해자 보상안
환불/크레딧
책임 보험 청구
사후 처리

7 사후 리뷰
재발 방지책
프로세스 개선
보고서 작성
개선

긴급 대응 Phase (1~2단계) | 분석 및 소통 Phase (3~4단계) | 복구 및 개선 Phase (5~7단계)

Golden Time: 사고 인지 후 **24시간 이내**에 초기 대응(탐지~소통)이 이루어져야 법적 책임 감경 및 평판 리스크를 최소화할 수 있습니다. 사전 시나리오 훈련이 필수입니다.

PART 09

업종별 실무 대응

각 산업 영역별 특성을 고려한 구체적인 실무 가이드와
핵심 점검 사항을 제공합니다.

생체인식 HR-Tech Fin-Tech Health-Tech Edu-Tech

A-1. 얼굴인식 출입통제 대응 실무 체크리스트

01

오인식률 및 편향 테스트

- 인종/성별/연령별 인식 정확도 격차 분석
- 조명, 각도 등 환경 변수에 따른 성능 검증
- 테스트 결과 문서화 및 내부 기준 수립

02

동의 절차 및 대안 경로 제공

- 생체정보 수집/이용에 대한 명시적 동의 획득
- 얼굴인식 거부 시 카드/비밀번호 등 대안 수단 보장
- 강제적 수집 금지 원칙 준수

03

고영향 확인 요청 (권장)

- 생체인식은 대표적인 고영향 AI 영역
- 불확실성 해소를 위해 과기정통부 확인 신청 권장
- 확인 결과에 따른 영향평가 의무 이행 준비

04

개인정보보호법 준수 연계

- 민감정보(생체정보) 안전성 확보 조치 이행
- 접근 통제, 암호화 등 기술적/관리적 보호 조치
- 개인정보 영향평가(PIA) 수행 및 결과 반영

A-2. 지문인식 용도별 판단

구분	개인기기 잠금 해제	금융/수사 신원확인
판단 결과	**일반적으로 고영향 아님** ※ 단, 민감정보 처리 시 주의 필요	**고영향 가능성 높음** ※ 재산권/신체의 자유 직접 영향
활용 목적	사용자 본인의 소유 기기 접근 권한 인증 (1:1 매칭, On-Device 처리)	대출 실행, 계좌 개설, 범죄자 식별 등 중요한 법적/경제적 행위 승인 (1:N 매칭, 중앙 서버 처리 가능성)
위험 영향	✓ 오인식 시 기기 사용 불편에 그침 ✓ 대체 수단(PIN, 패턴) 즉시 사용 가능 ✓ 타인에게 피해를 주지 않음	✓ 오인식 시 금융 거래 거절 (재산권 침해) ✓ 무고한 사람을 범죄자로 오인 (신체의 자유 침해) ✓ 사회적 신뢰도 하락 및 배상 책임 발생
실무 대응	• 생체정보 보호 가이드라인 준수 • 기기 내 안전 영역(TEE) 저장 확인 • 사용자 동의 절차 간소화 가능	• **고영향AI 확인 신청 권장** • 정기적 정확도(FAR/FRR) 테스트 필수 • 이의제기 및 대체 인증 수단 마련

판단 핵심 기준: 피해 규모와 결정의 자동성

같은 기술이라도 **"누구에게, 얼마나 중대한 영향을 미치는가"**가 핵심입니다.
단순 편의 제공 vs. **권리 제한/의무 부과** 여부를 기준으로 판단하세요.

B-1. AI 서류전형 대응 (HR-Tech)

01

데이터 편향 분석 (Bias Analysis)

- 성별, 연령, 학력별 학습 데이터 분포 균형성 검증
- 특정 집단에 대한 과대/과소 대표 여부 확인
- 민감 정보(거주지 등)의 대리 변수 포함 여부 점검

02

공정성 모니터링 (Fairness Monitoring)

- 집단 간 합격률 격차 (Disparate Impact) 상시 모니터링
- 합격자 추천 비율의 인구통계적 일관성 유지
- 편향 발견 시 모델 재학습 및 가중치 조정 프로세스

03

설명 가능성 확보 (Explainability)

- 지원자에게 구체적인 탈락/점수 사유 제시 가능 여부
- 핵심 평가 요소(키워드, 경력 등)의 영향도 시각화
- '블랙박스' 모델 지양 및 해석 가능한 모델 도입 검토

04

인간 개입 절차 (Human-in-the-loop)

- AI는 '추천'만 하고 최종 합불 결정은 채용 담당자가 수행
- 지원자의 이의제기 접수 및 재검토 채널 운영
- AI 평가 결과에 대한 사람의 정기적 샘플링 감사

B-2. 내부 인사평가 AI 실무 대응

고영향 가능성 주의 (보상/승진 영향)

외부 채용뿐만 아니라 내부 승진, 보상, 징계 등 근로자의 경제적 지위나 근로 조건에 중대한 영향을 미치는 AI 시스템도 고영향AI로 분류될 가능성이 매우 높습니다.

노조·근로자 대표 협의 및 이의절차 설계

도입 전 노동조합 또는 근로자 대표와의 사전 협의가 필수적이며, AI 평가 결과에 대해 근로자가 소명하고 재심사를 요청할 수 있는 공식적인 이의제기 절차를 마련해야 합니다.

내부용 예외 범위의 명확한 이해

"사업자 내부 업무 용도로만 사용"되는 경우 **투명성 확보 의무(사전 고지, 표시)는 면제**될 수 있으나, 고영향AI로 지정될 경우 **영향평가 및 위험관리 의무는 여전히 적용**됩니다.

실무 권장 사항
인사평가 AI는 '블랙박스'가 되어서는 안 됩니다. 평가 기준과 가중치를 투명하게 공개하고, AI 점수는 참고 자료로만 활용하며 최종 결정권은 평가자(사람)에게 있음을 명문화하세요.

Q45. 대출 심사 규제 연계 (Fin-Tech)

기존 규제 연계 (중복규제 완화)

신용정보법

제35조의2(설명의무) 준수 시

→ AI법상 **'설명가능성 확보 의무'** 이행 간주

금융소비자보호법

제10조(내부통제기준) 준수 시

→ AI법상 **'이용자 보호방안 마련'** 이행 간주

기존 컴플라이언스 체계에 AI법 요건을 통합 관리하는 것이 효율적입니다.

필수 실무 조치 (Action Items)

XAI (설명가능한 AI) 기술 도입

SHAP/LIME 등을 활용하여 '왜 거절되었는지' 구체적 요인을 산출할 수 있어야 함

결정 사유 통지 체계 고도화

단순 "종합 점수 미달"이 아닌, "연체 이력 및 소득 대비 부채 비율" 등 구체적 사유 제시

이의제기 및 재심사 절차

AI 평가에 불복 시 사람이 직접 재심사하는 프로세스(Human-in-the-loop) 보장

Fin-Tech 고영향AI 관리 팁

대출 심사 AI는 **'재산권'**에 직접 영향을 미치는 대표적인 고영향AI입니다. 과기정통부 확인 절차보다는 **'금융감독원 가이드라인'** 및 **'신용정보법'** 준수에 초점을 맞추되, AI법상의 **'위험관리 문서화'**를 보강하여 양쪽 규제를 동시에 대응하세요.

Q46. 보험 사기 탐지 AI (Fin-Tech)

구분	자동 결정형 (Automated Decision)	보조 도구형 (Assistive Tool)
역할 정의	AI가 사기 의심 건을 탐지하고, 사람의 개입 없이 **자동으로 보험금 지급을 거절**하는 경우	AI가 사기 의심 건을 탐지하여 담당자에게 **추천(Flagging)**만 하고, 최종 판단은 사람이 수행
고영향 판단	**고영향AI 가능성 매우 높음**	**고영향AI 아닐 가능성 높음**
판단 근거	개인의 **재산권(보험금 수령 권리)**에 대해 중대하고 직접적인 영향을 미침	최종 의사결정 권한이 사람에게 있으며, AI는 단순 참고 자료로 활용됨 (영향력 제한)
필수 조치	• 영향평가 의무 수행 • 거절 사유에 대한 상세 설명(XAI) 제공 • 알고리즘 투명성 확보	• 담당자의 최종 검토 절차 문서화 • AI 판단을 무시할 수 있는 권한 보장 • 오탐(False Positive) 모니터링

핵심 판단 기준: '결정권'의 소재 (Decision Authority)

단순히 "사람이 검토한다"는 형식적 절차만으로는 부족합니다. 실질적으로 담당자가 AI의 판단을 **비판적으로 검토하고 수정할 수 있는 권한과 시간**이 주어지는지가 고영향AI 판단의 핵심입니다. 이의제기 절차(Human-in-the-loop)가 실질적으로 작동해야 합니다.

D-1. 진단 보조 AI 실무 대응 (Health-Tech)

중복규제 회피 전략

디지털의료제품법 등 타 법령 준수 인정

- **QMS 적합 판정 활용**
  디지털의료제품법 제8조/12조에 따른 품질관리체계 적합 시, AI법상 **위험관리 의무 이행**으로 간주 (시행령 별표1)
- **통합 문서 관리**
  기존 의료기기 기술문서(TCF)에 AI 특화 항목(편향성, 설명가능성)만 추가하여 이중 작업 방지
- **임상 데이터 활용**
  인허가용 임상시험 결과를 AI 영향평가의 핵심 근거 자료로 갈음 가능

핵심 실무 조치

고영향 리스크 관리 및 안전성 확보

- **의사 최종 판단(Human-in-the-loop) 유지**
  AI는 '보조적 도구'임을 명시하고, 의사가 AI 의견을 기각할 수 있는 권한과 절차 보장
- **설명 가능성(XAI) 제공**
  단순 결과값뿐만 아니라 판단 근거(히트맵, 주요 변수 등)를 의료진이 이해 가능한 형태로 제공
- **실사용 데이터(RWD) 모니터링**
  출시 후 실제 임상 현장에서의 성능 변화와 오작동 여부를 지속 추적 관찰

실무 팁: 의료진 대상 교육 및 가이드라인 배포

AI의 한계점과 오남용 방지 수칙을 포함한 사용자 가이드를 배포하고, 의료진이 AI 결과를 맹신하지 않고 비판적으로 수용할 수 있도록 정기적인 교육을 제공하는 것이 중요합니다.

Q48. 건강관리 앱 경계선 케이스

일반 웰니스 vs 의료기기 구분

단순 라이프스타일, 운동, 식단 추천 기능은 생명·신체에 중대한 영향이 없어 **고영향 아님** 가능성이 높습니다. 반면, 특정 질환의 관리, 예방, 치료를 암시하거나 진단 기능을 포함할 경우 의료기기로 간주되어 **고영향 가능성** 이 높아집니다.

재분류 위험 요소

앱 내 사용하는 문구(예: '치료', '진단', '예방'), 제공하는 기능의 구체성, 타겟 대상자(일반인 vs 환자군)에 따라 일반 웰니스 앱도 의료기기로 재분류될 위험이 있습니다. 마케팅 문구 하나로도 규제 적용 여부가 달라질 수 있으므로 주의가 필요합니다.

면책 조항 명확화 (Disclaimers)

서비스 이용 약관 및 앱 실행 화면에 "본 서비스는 의료 행위가 아니며 의사의 진단을 대체할 수 없습니다"와 같은 면책 조항을 명확하고 눈에 띄게 표시해야 합니다. 이는 법적 리스크를 완화하고 이용자의 오해를 방지하는 필수적인 안전장치입니다.

실무 팁: 규제 샌드박스 및 사전 질의 활용
기능이 모호하여 의료기기 해당 여부가 불분명한 경우, 식약처의 의료기기 해당 여부 질의 제도를 적극 활용하거나 규제 샌드박스를 통해 임시 허가를 받아 안전성을 검증하는 단계를 거치는 것이 좋습니다.

E-1. 학생 평가 시스템 실무 대응

01

편향성 테스트 수행

- 지역/소득/학교 유형별 성능 차이 분석
- 학습 데이터의 특정 집단 과대/과소 대표성 점검
- 정기적인 공정성 지표 모니터링

02

교사 최종 검토 절차 마련

- AI 평가는 참고 자료로 활용 원칙 수립
- 최종 성적 산출 전 교사의 필수 검증 단계
- 학생/학부모의 공식 이의제기 채널 운영

03

평가 기준 투명화

- 평가 요소 및 가중치 공개
- AI 채점 방식에 대한 쉬운 설명 제공
- 개별 점수에 대한 상세 피드백 리포트

04

진학·진로 영향 고려

- 고입/대입 등 결정적 평가 시 위험도 재산정
- 부정적 결과가 진로에 미칠 영향 최소화
- 진로 추천 시 다양성 확보 알고리즘 적용

E-2. 강의 추천 AI

온라인 강의 추천 AI는 **고영향 AI로 분류되지 않을 가능성**이 높지만,
청소년 보호와 알고리즘 투명성 측면에서 **별도의 주의**가 필요합니다.

실무 TIP

주의 및 모니터링 포인트

- **청소년·유해 콘텐츠:** 추천 알고리즘이 미성년자에게 유해한 콘텐츠를 노출하지 않도록 필터링 강화
- **정치적·이념적 편향:** 특정 성향의 콘텐츠만 편향적으로 추천되지 않도록 다양성 지표 모니터링
- **상업적 편향:** 광고성 콘텐츠가 교육적 콘텐츠로 오인되지 않도록 명확히 구분

실무 조치 가이드

- **투명성·콘텐츠 정책 연계:** 추천 기준과 알고리즘의 작동 원리를 이용약관이나 도움말에 명시
- **방통위 가이드라인 준수:** '생성형 AI 서비스 이용자 보호 가이드라인' 등 관련 지침 준수 여부 점검
- **이용자 통제권 보장:** '추천 안 함', '관심 없음' 등 이용자가 추천 결과에 피드백할 수 있는 기능 제공

업종 공통 필수 문서 체크리스트

01

기술 명세 문서 (Technical Spec)

- **모델 카드 (Model Card):** 모델 아키텍처, 의도된 용도, 한계점, 성능 지표
- **데이터시트 (Datasheet):** 학습 데이터 출처, 수집 방법, 전처리 과정, 편향성 검토
- **테스트 리포트:** 정확도, 강건성, 안전성 테스트 결과 및 검증 이력

02

거버넌스 문서 (Governance)

- **AI 윤리 원칙 & 정책:** 조직 내 AI 개발 및 운영 가이드라인, 윤리 위원회 규정
- **위험 관리 계획서:** 식별된 리스크 목록, 완화 전략, 모니터링 계획
- **영향평가서 (AIIA):** 기본권 침해 가능성, 사회적 영향 분석 및 대응 방안

03

추적성 문서 (Traceability)

- **변경 이력 관리대장:** 알고리즘 업데이트, 데이터셋 변경 사항, 배포 버전 기록
- **라벨링 가이드 & 이력:** 데이터 라벨링 기준서, 작업자 교육 기록, 품질 검수 결과
- **의사결정 로그:** 주요 개발 단계별 의사결정 근거 및 승인 기록

04

사고 대응 문서 (Incident Response)

- **사고 대응 매뉴얼 (Playbook):** 유형별 대응 절차, 비상 연락망, 보고 체계
- **사후 분석 보고서 양식:** 원인 분석, 피해 규모 산정, 재발 방지 대책 수립 서식
- **모의 훈련 기록:** 정기적인 사고 대응 훈련 실시 결과 및 개선 사항

업종 공통 테스트 매트릭스

구분	평가 지표 (Metric)	상세 내용 및 검증 방법	릴리즈 게이트 기준 (Release Gate)
정확도 (Accuracy)	**성능 지표**	• Precision, Recall, F1-Score, AUC 등 • 업계 표준 또는 SOTA(State-of-the-Art) 모델 대비 성능 • 실제 운영 환경 데이터셋(Real-world data) 검증	**기존 모델 대비 동등 이상**
안전성 (Safety)	**위험 회피**	• 유해 콘텐츠 생성 여부 (Hate speech, Violence) • 개인정보 유출 가능성 (PII Leakage) • 적대적 공격(Adversarial Attack) 방어력	**치명적 오류 0건**
공정성 (Fairness)	**편향 분석**	• 집단별(성별, 연령, 지역 등) 성능 격차(Disparity) 분석 • False Positive/Negative Rate 균형 • 인구통계학적 동등성(Demographic Parity) 검토	**성능 격차 5% 미만**
강건성 (Robustness)	**변화 대응**	• 노이즈 데이터, 예외 입력 처리 능력 • 데이터 드리프트(Data Drift) 발생 시 성능 유지력 • 경계값(Edge Case) 테스트 결과	**성능 저하 허용 범위 내**

실무 팁: 릴리즈 게이트(Release Gate) 운영

모든 AI 모델 배포 전, 위 4가지 영역의 테스트 결과가 **사전에 정의된 임계값(Threshold)**을 통과해야만 배포를 승인하는 절차를 자동화(CI/CD 파이프라인 연동)하십시오. 결과 보고서는 영향평가 증빙자료로 활용됩니다.

업종별 종합 체크리스트

고영향 영역별 필수 확인사항

다음 7대 영역에 해당하는 경우, 아래 4가지 사항을 즉시 점검하고 조치를 취해야 합니다.

1. **고영향 해당성 확인:** 법적 기준 충족 여부 검토 및 과기정통부 확인 요청
2. **영향평가 착수:** 7대 필수 항목 기반 평가 실시 및 결과 문서화
3. **투명성·표시 체계:** 사전 고지 절차 및 결과물 표시 기술 적용

7대 고영향 관리 대상 영역

- 생체인식 (얼굴, 지문, 홍채 등)
- 채용 및 인사평가 (HR-Tech)
- 신용평가 및 대출 심사 (Fin-Tech)
- 의료 진단 및 치료 (Health-Tech)
- 교육 및 학생 평가 (Edu-Tech)
- 법집행 (범죄 예측, 증거 분석)
- 중요 인프라 (에너지, 교통, 통신)

핵심 3대 의무 요약

투명성 확보 의무

이용자가 AI 사용 사실을 인지하고
결과물의 출처를 식별할 수 있도록
보장하는 의무입니다.

- **사전 고지:** 약관/화면에 AI 사용 명시
- **결과물 표시:** 워터마크/메타데이터 삽입
- **적용:** 생성형AI 다운로드 제공 시 필수

영향평가 의무

고영향AI가 기본권과 안전에 미치는
잠재적 위험을 사전에 식별하고
관리하는 체계입니다.

- **대상:** 고영향AI (생체/채용/신용/의료 등)
- **항목:** 7대 필수 항목(기본권 영향 등)
- **주기:** 연 1회 정기 평가 + 변경 시 수시

안전성 확보 의무

고성능·고위험 AI 시스템의
기술적 견고성과 신뢰성을
유지하기 위한 조치입니다.

- **위험관리:** 위험 식별-분석-완화 체계
- **데이터 관리:** 학습 데이터 품질/편향 점검
- **거버넌스:** 책임자 지정 및 관리 조직 운영

실행 로드맵 D-90 ~ D+30

D-90 | D-60 | D-30 | D-Day | D+30

PLANNING

스코핑 & 거버넌스

적용 대상 AI 식별 및 고영향/생성형 여부 판단. 거버넌스 위원회 구성 및 R&R 확정.

ASSESSMENT

점검 & 영향평가

데이터 편향성 분석 및 모델 성능 점검. 고영향AI 대상 영향평가 착수 및 문서화.

BUILD

통제 & 체계 구축

위험 완화 조치 이행. 약관/UI 고지 문구 반영 및 표시(워터마킹) 시스템 구축.

LAUNCH

릴리즈 게이트

최종 컴플라이언스 체크리스트 확인 후 배포 승인. 법적 필수 요건 충족 여부 최종 검수.

REVIEW

모니터링 & 개선

실운영 데이터 기반 성능/편향성 모니터링. 사용자 피드백 분석 및 개선 계획 수립.

성공적인 런칭을 위한 핵심 조언

로드맵의 각 단계는 **순차적(Waterfall)**으로 보이지만, 실제로는 **반복적(Iterative)**으로 수행되어야 합니다. 특히 '영향평가' 단계에서 발견된 위험 요소는 즉시 '데이터/모델 점검' 단계로 피드백되어 수정되어야 하며, D-Day 이전까지 이 사이클을 최소 2회 이상 반복하여 리스크를 최소화하는 것이 중요합니다.

긴급 지원·프로그램

정부 지원 프로그램

AI 신뢰성 평가 지원
영향평가 비용 바우처, 데이터 품질 검증

규제 샌드박스
혁신 서비스 임시허가 및 실증특례

NIA 지능정보사회진흥원
support@nia.or.kr

정책 문의 채널

법령 유권해석
고영향AI 지정 여부, 예외 규정 해석

전문 컨설팅
중소/스타트업 대상 무료 법률 자문

과기정통부 AI정책과
1355 (AI 헬프데스크

부록 1. 주요 연락처

과학기술정보통신부
인공지능안전신뢰정책과

- **전화:** 044-202-6293
- **웹사이트:** www.msit.go.kr
- **주소:** 세종특별자치시 갈매로 477

인공지능정책센터
(NIA 한국지능정보사회진흥원)

- **전화:** 053-230-1114 (대표)
- **웹사이트:** www.nia.or.kr
- **이메일:** ai_policy@nia.or.kr

인공지능안전연구소
(AI Safety Institute)

- **전화:** 031-739-7651
- **웹사이트:** www.aisi.re.kr

법무법인(유한)린
AI•플랫폼•테크놀로지 전문그룹

- **담당자:** 구태언 AI그룹 총괄변호사
- **전화:** 02-3477-8695
- **이메일:** tekoo@law-lin.com

부록 2. 참고자료/링크

구분	자료명	주요 내용 및 출처
기본 법령	법률 AI 기본법 및 시행령	법제처 국가법령정보센터 인공지능 산업 육성 및 신뢰 기반 조성에 관한 기본법
가이드라인 ① (판단 기준)	지침 고영향 AI 판단 가이드라인	7대 영역별 상세 판단 기준 및 확인 신청 절차 과기정통부 / 고영향 AI 해당 여부 자가점검표 제공
가이드라인 ② (사업자 의무)	지침 고영향 AI 사업자 책무 가이드라인	위험관리, 설명가능성 등 5대 핵심 책무 이행 방안 과기정통부 / 신뢰성 확보 조치 및 문서화 양식
가이드라인 ③ (영향평가)	지침 고영향 AI 영향평가 가이드라인	영향평가 3단계 프로세스 및 7대 필수 평가 항목 과기정통부 / 평가 지표 및 결과보고서 작성 요령
가이드라인 ④ (투명성)	지침 투명성 확보 가이드라인	생성형 AI 결과물 표시(워터마크) 및 사전 고지 의무 과기정통부 / 텍스트·이미지·영상별 표시 방법 예시
가이드라인 ⑤ (안전성)	지침 최첨단 AI 안전성 확보 가이드라인	누적연산량 10^26 FLOPs 이상 모델의 위험 관리 과기정통부 / 안전성 테스트 및 사고 대응/보고 체계
국제 표준	표준 ISO/IEC 42001 & EU AI Act	국제 인공지능 경영시스템(AIMS) 및 EU 규제 연계 ISO 공식 웹사이트 / EU Official Journal

가이드라인 활용 팁

위 5종의 가이드라인은 AI 기본법의 하위 규정을 구체화한 핵심 문서입니다. **사업자 유형(고영향/생성형/최첨단)**에 따라 적용되는 가이드라인이 다르므로, 해당 문서를 우선적으로 숙지하시기 바랍니다.