금융분야 AI 가이드라인 개정방향

2025.12.

한국금융연구원 연구위원 백연주

KIF

목차

1

2

3

논의의 배경

AI 가이드라인: 논의의 배경

(배경) 기술양상 및 규제환경 변화 → 금융 분야 AI 가이드라인 개정 필요성

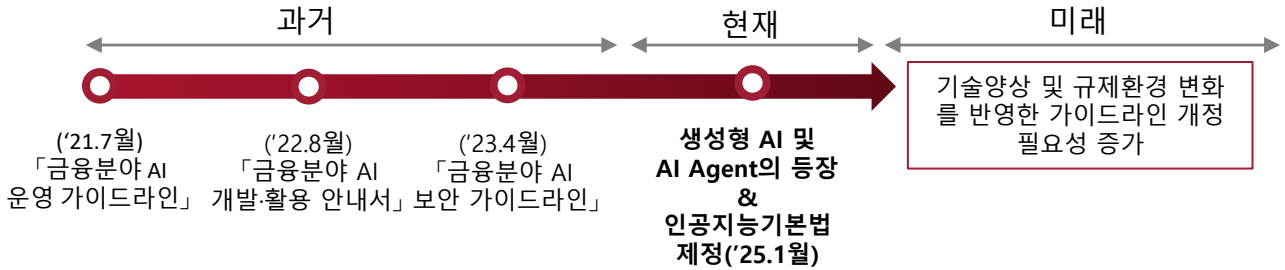
- 생성형 AI, AI Agent 등 급격한 기술 발전과 금융회사 내부의 AI 활용 범위 확대
- 2025년 1월 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제정, 9월 하위법령 초안 발표, 11월 시행령 입법예고

→ 기존 3개 가이드라인(① 운영, ② 개발, ③ 보안)*을 「금융 분야 AI 가이드라인」으로 단일화

*① 「금융분야 AI 운영 가이드라인(21.7월)」, ② 「금융분야 AI 개발활용 안내서(22.8월)」, ③ 「금융분야 AI 보안 가이드라인(23.4월)」

(적용) 금융 서비스·상품 제공 시 AI를 활용하는 금융회사 및 비 금융회사

- (업무) AI 활용 신용평가, 대출심사, 챗봇, 사기탐지시스템(FDS) 등
- (적용 대상) 금융회사(은행, 금융투자업자 등) + 비금융회사(AI 활용 결과가 금융거래 영향 O)
- 본 원칙은 **모범규준(Best Practice)**, 업권별 자율규제 형식으로 규율하고 의견을 지속 수렴하여 상시적으로 개선·보완할 계획



(참고) 금융 AI 7대 원칙(안)

- **2024년 12월, 금융위원회는 금융권 AI 개발·활용의 주요 원칙으로 금융 AI 7대 원칙을 마련**

< 금융 AI 7대 원칙(안) >

분야	7대 원칙
거버넌스	① **경영진의 역할과 책임** - **최고경영자를 포함한 경영진**은 AI 개발·활용에 대한 **관심**을 갖고 **역할과 책임**을 분담해야 함
	② **합법성** - AI 활용 전단계에서 금융·AI 등 **관련 법규**를 **준수**해야 함
	③ **보조수단성** - 현 단계에서 AI는 업무의 **보조 수단**이므로 **최종 의사결정**과 그에 따른 책임은 **임직원**이 수행함
AI 개발 단계	④ **신뢰성** - AI 개발 과정에서 **신뢰할 수 있는 데이터와 모델**을 사용해야 함
	⑤ **안정성** - AI 설계·학습 등 전과정에서 **금융 안정성** 위험을 최소화해야 함
AI 활용 단계	⑥ **신의성실성** - AI 활용 시 **금융소비자의 이익**을 최우선으로 해야 함
	⑦ **보안성** - AI 활용 시 **보안성** 기준 및 점검·개선 체계를 마련해야 함

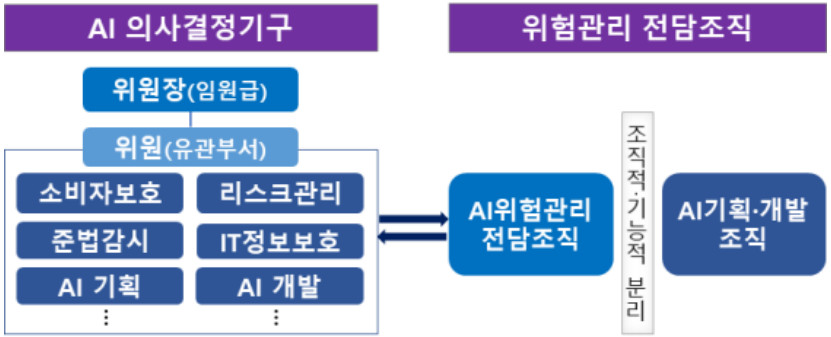
1 2 3

금융분야 AI 가이드라인: 7대 원칙

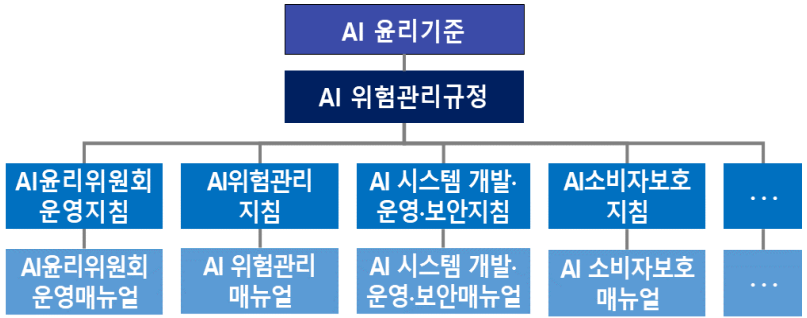
1. 거버넌스 원칙 (1/3)

- **AI 거버넌스를 통해 AI 시스템을 체계적으로 기획·개발·운영하고, 각 단계에서 발생가능한 위험을 평가·관리**
  - **AI 위험관리**를 위한 **최고 의사결정기구** 설치 (예: AI 윤리위원회) → AI 윤리기준 수립, AI 도입·활용에 **적극적 관여**
  - 독립된 **AI 위험관리 전담조직**을 설치 → **AI 업무 전반**의 **위험**을 **통제·관리**
  - AI 윤리기준을 근간으로 규정 및 지침 등 **AI 관련 내규**를 수립하고, 도입·활용 전 프로세스 관리를 위한 세부적인 **업무매뉴얼**을 마련

(예시) AI 거버넌스 체계



(예시) AI 관련 규정 체계



1. 거버넌스 원칙 (2/3)

◉ AI 위험 인식 · 측정, 위험경감, 잔여위험 평가, 위험등급 산정 등 종합 위험평가 체계 구축

- **AI 위험**의 체계적인 인식·측정·관리 등을 위해 **위험기반 접근방법**(Risk-based approach)의 **종합적인 평가체계** 구성
- **AI 7대 원칙** 중 합법성, 신뢰성, 신의성실, 보안성 4개 원칙의 **위반 가능성**을 **정량적 요소**로 **활용**하여 점수화

(예시) AI 위험평가 체계를 통한 최종 위험등급 산출

평가부문	위험 인식·측정 항목	배점	위험경감	잔여위험			위험등급 확정
합법성원칙 (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			
신뢰성원칙 (30%)	품질	6	(4)	2	∑18		
	편향성	6	(2)	4			
	공정성	6	(2)	4			
	설명가능성	6	(1)	5			
	성능	6	(3)	3			
신의성실원칙 (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			

1. 거버넌스 원칙 (3/3)

◉ 위험 수준별로 차등화된 통제·관리 수행 및 모니터링, 문서화, 교육 등 위험통제를 위한 제반 절차 마련·이행

- **모든 AI 서비스**에 **기본 통제 방안**을 적용하되, **저위험 AI 서비스**는 **완화된 통제 방안**을 적용하여 운영 효율성 도모
- **고위험 AI 서비스**는 AI 윤리위원회 **사전 승인·사후 검증**, 제3자에 의한 평가 검증, **운영단계 모니터링 강화** 등 **추가 통제**
- **AI 기본법**에 따른 **'고영향 AI'**에 대해서는 위험점수에 따른 등급 분류와 관계없이 **'고위험'**으로 분류하여 **통제 강화**

(예시) AI 서비스 위험 수준에 따른 차등화된 통제·관리 수행방안

기본 통제 방안	위험등급 분류 기준	위험 수준별 통제 방안
√ 상품·서비스 출시 전 경감조치 검증 √ 상품·서비스 운영단계 모니터링 기준 적용 보고 √ AI 세부 업무 방법에 따른 관리 (AI 위험관리 업무매뉴얼, 검증매뉴얼 등) √ 상품·서비스 위험 변경시 위험 수준 재평가		**<고위험: 추가 통제 적용>** √ AI윤리위원회 사전 승인/사후 검증 √ 제3자에 의한 평가 검증 √ 운영 단계 모니터링 강화 **<중위험: 기본 통제 및 관리 적용>** **<저위험: 통제 완화 적용>** √ 승인절차 및 작성 문서 등 축소

(참고) 금융감독원의 「금융분야 AI 위험관리 프레임워크(AI RMF)」

- **금융회사가 AI 시스템의 도입활용 전 주기에 걸쳐 유연하고 체계적으로 AI 위험을 관리할 수 있도록, ① 거버넌스, ② 위험평가, ③ 위험통제와 관련한 핵심 프로세스를 제시**

※ 거버넌스, 위험평가 및 위험통제 등의 구체적인 내용은 향후 금융감독원이 공개 예정인 「금융분야 AI 위험관리 프레임워크」(AI Risk Management Framework in Financial Sector)를 참고 가능

2. 합법성 원칙

AI 개발·이용 시 적용 법규를 사전 파악, 내부 정책 및 업무 절차에 반영 및 주기적인 점검·개선

- 공통적으로 적용되는 법령(인공지능기본법) 및 하위 법규(이하 인공지능기본법령)와 해당 업종에 따라 적용되는 관련 법령을 검토
- 법규 요구사항 → 내부 정책 및 업무 절차에 반영, 주기적 점검 → 실효성 평가·지속 개선

금융분야 관련 법령

구분		관련 법규
공통	❶ 금융소비자 보호를 위한 영업행위 규정 준수	금융소비자보호법 제10조 및 제17~22조
	❷ 정보처리시스템의 안전성 확보 및 보호대책 수립・이행	전자금융감독규정 제14조 및 제21조
	❸ 위험관리 기준 및 절차 마련	금융사지배구조법 제27조 제1항
	❹ 자동화된 결정에 대한 정보주체 권리 보장 (개인정보 처리시)	개인정보보호법 제37조의2
신용평가, 보험, 여신 등	❶ 자동화평가 등에 대한 설명 이행 (개인신용정보 처리시)	신용정보법 제36조의2 제1항
	❷ 개인신용평가에 관한 원칙 준수	신용정보법 제22조의 3
금융투자 등	❶ 전자적 투자조언장치 활용 업무의 요건	자본시장법시행령 제2조 및 금융투자업규정 제1-2조의2
	❷ 전자적 투자조언장치의 투자일임보고서작성・교부・보관등	자본시장법 제99조, 동법 시행령 제100조, 금융투자업규정 제4-78조, 4-13조

3. 보조수단성 원칙

AI 산출물에 대한 최종 책임은 임직원이 지며, 운영 전 단계에 걸쳐 임직원 개입 상황 차등화

- 인공지능을 업무의 보조수단으로 활용하고 최종 의사결정과 책임은 임직원이 수행하도록 해야 함.
- 업무 중요도, 위험수준에 상응한 의사결정 단계별로 역할과 책임에 대한 사항을 정함(예: RACI 차트).
- 단, 고영향 AI는 사람이 수행하는 의사결정에 보조적인 용도로만 사용(Human-in-the-loop)

(예시) RACI 차트를 활용한 책임 수행 체계 구성

구분	여신심사담당자	여신심사 팀장	리스크 관리부서	준법 감시인	IT운영
① 신청 접수·기본요건 확인	R	I	I	I	I
② 모델 점수 산출·권고 표시	C	I	C	I	R
③ 개입기준 판단·자료 보완	R	A	C	C	C
④ 심사의견 1차 검토·사유기재	R	I	C	I	I
⑤ 상급자 검토/복수인 확인	R	A	C	C	I
⑥ 최종결정·통지·기록	I	A/R	C	C	C

주: RACI 차트란 업무 과정에서 누가 무엇을 책임지고, 승인하고, 참조하고, 통보받을 것인지를 명확히 하는 책임분담 구조도를 의미 (Responsible: 책임수행자, Accountable: 최종책임자, Consulted: 의견제시자, Informed: 통보대상자)

(참고) 보조수단성 하 인적개입 원칙 및 활용

[관련법규] 인공지능기본법 제34조 / 동법 시행령(초안) 제26조 / 사업자책무고시(초안) 제7조

법 제34조 (고영향 인공지능과 관련한 사업자의 책무) ① 인공지능사업자는 고영향 인공지능 또는 이를 이용한 제품·서비스를 제공하는 경우 고영향 인공지능의 안전성·신뢰성을 확보하기 위하여 다음 각 호의 내용을 포함하는 조치를 대통령령으로 정하는 바에 따라 이행하여야 한다.

4. 고영향 인공지능에 대한 사람의 관리·감독

시행령(초안) 제27조 (고영향 인공지능과 관련한 사업자의 책무) ① 인공지능사업자는 법제34조제1항 각 호의 조치 중에서 다음 각 호에 해당하는 내용을 자신의 홈페이지 등에 게시하여야 한다. (이하 생략)

4. 해당 고영향 인공지능을 관리·감독하는 사람의 성명 및 연락처

사업자책무고시(초안) 제7조 (사람의 관리·감독) ① 사업자는 인공지능시스템 개발 과정에서 사람의 관리감독을 위해 다음 각 호의 조치를 이행하여야 한다.

1. 사람이 인공지능 동작에 개입할 수 있는 기준 확립
2. 사람이 즉각적으로 인공지능시스템을 정지하거나 작동을 변경할 수 있는 '긴급 정지' 기능 등의 개입 방법 마련

② 사업자는 고영향 인공지능 운영 중 사람의 관리·감독을 위해 다음 각 호의 조치를 이행하여야 한다.

1. 성능저하 및 오류 발생에 대한 정기적인 점검계획 및 방안 마련
2. 인공지능의 범위 및 수행능력에 대한 이해도를 향상시키기 위한 교육 및 훈련 제공

인공지능 의사결정에 대한 인적 개입 방법

구분	주요 내용 및 설계 방법	고위험·고영향 AI
사전 승인 절차	◦ (내용) 사람의 검토 또는 승인을 거쳐 실행 ◦ (방법) 전자 승인·이중 승인·위원회 심의 등 ◦ (기타) 사유·근거 기록 의무화	적용
AI 권고 무시/수정/반대 결정(Override)	◦ (내용) 감독자가 합당한 사유로 모델 권고를 무시·수정 또는 반대결정 권한 부여 ◦ (방법) 내부체계 확립 후 오버라이드 UI 화면 제공 ◦ (기타) 사유·증빙 등을 표준화하여 기록	필수 적용
긴급 정지 기능 및 격리	◦ (내용) 이상발생시 즉시 중단 및 문제 요소 격리, 안전모드 전환 후 롤백·재가동 ◦ (방법) 긴급정지. 안전모드 전환 등을 위한 직관적 UI 제공 * 감독자 접근권한 확보 필요	필수 적용
실시간 모니터링 대시보드	◦ (내용) 성능, 공정성, 신뢰도, 규제/제재 신호 등을 실시간 시각화 ◦ (방법) 임계치 초과 시 자동 알림·보고	권고

[참고] NIST AI RMF Playbook

4. 신뢰성 원칙

◉ 모델 성능 관리, 데이터 품질 확보, 의사결정 과정 설명, 체계적 검증 및 오류 대응 체계 등을 구축

- (모델 성능 관리) AI모델 성능 측정을 위한 명확한 지표 설정, 정기적 점검·개선
- (데이터 품질) AI 학습 및 참조에 사용하는 데이터 및 AI 시스템에 입력되는 데이터의 품질 검증·확인
- (설명가능성) AI 의사결정 과정과 결과 ← 이해관계자의 합리적 이해가 가능하도록 설명 가능한 형태로 제공(신뢰성 ↑)

AI 설명 제공 예시

- 결과: 거절
- 이용된 기초정보(해당 고객의 주요 원본 정보)

항목명	설명	값
제1금융권 대출 건수	제1금융권 대출 건수	18
리볼빙 잔액 비율	신용한도 대비 리볼빙 잔액 비율	63
거래 연체 비율	연체되지 않은 거래 비율	94
상환 비율	과거 정상 상환된 신용 거래 비율	96
	... (이하 생략) ...	

- 주요 판단 사유: 제1금융권 대출 건수(부정적), 리볼빙 잔액 비율(부정적), 거래 연체 비율(긍정적), 상환 비율(긍정적)

자동 설명 생성

고객 대응 업무 담당자가 해석하여 설명 생성

고객님은 연체되지 않은 거래 비율 94% 및 과거 정상 상환된 신용 거래 비율 96%가 긍정적인 요인이었으나, 제1금융권 대출 건수 18건 및 신용한도 대비 리볼빙 잔액 비율 63%가 부정적으로 작용하여 **대출이 거절**되었습니다.

(참고) 데이터 품질 관리·편향성 예시

(예시) 데이터 전 처리 시에 처리해야 할 요소

항목	설명
노이즈 (Noise)	· 측정 과정에서 무작위로 발생하는 측정값의 오류
이상치 (Outlier)	· 나머지 데이터와 현저히 다른 특성을 보이는 값 · 데이터 입력·측정 오류/실험 오류로 발생할 수 있지만, 일부 예외 특성을 갖는 값일 수 있음
결측치 (Missing Value)	· 전산오류 및 미입력 등의 이유로 누락된 측정값
불일치 값 (Mismatch Value)	· 동일 개체에 있어, 측정 데이터가 다르게 나타나는 경우
중복 (Duplicate)	· 모든 속성 및 값이 동일한 경우
바이어스 (Bias)	· 측정 장비에서 측정하는 값과 실제 값과의 차이점
아티펙트 (Artifact)	· 외부 요인으로 인해 반복적으로 발생하는 왜곡이나 에러 ※ (예시) 카메라를 이용한 영상 데이터 획득에 있어, 렌즈의 얼룩에 의해 지속적인 왜곡 발생 등
오염 (Poisoning)	· 악의적인 목적으로 변조한 데이터

(예시) AI 편향성의 원인

□ **치우친 표본(Skewed Sample)**: 우연히 초기 편향이 발생하는 경우 시간이 지남에 따라 편향 증폭
※ 예) 초기 범죄율이 높은 곳으로 더 많은 경찰관을 파견하는 경향이 있고, 그러한 지역에서 범죄율에 대한 기록이 더 높아질 확률이 높음

□ **오염된 사례(Tainted Example)**: 축적된 데이터에 존재하는 사람의 편견을 알고리즘에서 특별히 교정하지 않고 유지하는 경우 동일한 편향이 복제됨
※ 예) 구글 뉴스 기사에서 남성-프로그래머의 관계는 여성-주부와의 관계와 매우 유사한 것으로 밝혀짐(Bolukbasi et al., 2016)

□ **제한된 속성(Limited Feature)**: 데이터의 특정 속성에 대해 소수 그룹에 대해서는 제한되거나 낮은 신뢰도의 정보만 수집

□ **표본 크기의 불일치(Sample Size Disparity)**: 소수 그룹에서 제공되는 학습 데이터가 대다수 그룹에서 제공되는 학습데이터보다 훨씬 적은 경우 소수 그룹을 정확히 모델링 할 가능성이 낮음

□ **대리 변수의 존재(Proxy)**: 학습 시 공정성 측면에서 민감한 데이터 속성(인종, 성별 등)을 사용하지 않더라도 이를 대리하는 다른 속성(이웃 등)이 항상 존재할 수 있어, 이러한 속성이 포함되어 있으면 편향이 계속 발생

자료: 기계학습 공정성 관련 연구 동향(소프트웨어정책연구소, `20.2월)

5. 금융안정성 원칙

AI의 설계, 학습 등 전 과정에서 금융안정성 위험을 최소화

- (금융안정위험 평가·관리) AI 시스템이 금융시장 전반, 금융안정에 미치는 영향 등을 평가하고 관리하는 방안 마련
- (안전장치 마련) AI 시스템 오작동 시 백업모형 활용, 사후 개입을 위한 비상정지 장치, 회생계획 설계 등 안전장치 마련
- (제3자 IT리스크 관리) AI 모델 외주개발 또는 오픈소스 기반 AI 활용 → 별도 제3자 IT리스크 관리방안 마련
- (감독당국 정보공유 및 보고) **시스템위험**을 초래할 수 있는 AI 사고 발생 시 감독당국에 보고하고, AI 활용구조를 감독당국이 사전에 파악할 수 있도록 **정보 공유** 필요

(예시) AI 관련 제3자 계약 체결 시 포함 사항

① 제3자가 제공하는 서비스 등에 대해 기능, 위험 등에 대한 명확한 설명
② 사이버 보안, 데이터 및 프라이버시 보호에 관한 내용
③ 서비스를 제공하는 지역, 국가 장소
④ 제3자의 파산 및 사업 운영 중단 또는 계약 종료 시 해당 서비스, 데이터 등에 대한 금융회사 등의 접근 및 복구, 반환의 보장에 관한 내용
⑤ 사고 발생 시 정보 공유 및 사고처리, 배상책임과 관련한 내용
⑥ AI 관련 피해 발생 시 책임 소재
⑦ AI 관련 법규 명령 및 정책 등의 준수 의무
⑧ 안전하고 신뢰할 수 있는 AI 개발·활용과 관련 금융회사 등의 행동강령, 정책 등 준수

6-7. 신의성실성 및 보안성 원칙

(신의성실성) AI활용 對고객 서비스 제공 → 소비자 이익 최우선 고려

- (이해상충 방지) 대 고객 서비스 AI 활용 시, 이해상충 문제 발생 방지를 위한 관리·감독장치 마련
- (소비자 보호대책 마련) AI활용과정에서 충실한 소비자 보호 → AI활용사실 사전고지, 피해대응을 위한 절차마련

(보안성) AI시스템 보안성 확보 → AI 시스템 고유 보안위협 식별 & 대응방안 마련

- (AI특화 보안 위협 식별·관리) AI시스템에 특화된 **보안 위협** 체계적 **식별, 대응전략 마련**
- (AI 특화 공격 탐지·대응) 식별된 AI특화 **공격 탐지, 차단, 대응체계 구축**
- (AI자산 보호 및 관리) **핵심자산**(데이터, 파라미터 등)의 무단 접근·유출·변조되지 않도록 **보호대책** 적용
- (외부 모델 및 데이터 검증) **외부 도입 모델·데이터 → 보안** 및 신뢰성 검증 수행 → 공급망 위험 최소화
- (기존 보안관리 AI확장) 기존 IT보안체계 기반의 **AI시스템 보안 확장** 적용
- (AI시스템 보안성 검증, 운영관리) **개발** 단계부터 보안성의 **체계적 검증, 지속적 관리**

1

2

3

결론

결론

- **본 가이드라인은 금융업권의 모범규준 (Best Practice)으로서 업계의 자율적인 노력을 권장**
  - 개별 금융회사의 상황에 따라 가이드라인의 내용을 적절히 반영할 필요가 있음
  - AI 관련 위험을 관리하기 위해 금융회사 간 경험을 공유하고 업권별 특성을 반영하여 자율규제를 하는 것이 원칙

- **AI 관련 기술 및 정책 논의가 발전함에 따라 본 가이드라인도 지속적으로 개정할 예정**
  - AI 기술이 전격적인 도입이 시작된 지 얼마 되지 않아 금융분야 관련 정책 방식에 대해 현재도 활발히 논의되고 있음
  - 향후 각국 당국과 지식공유를 통해 금융분야 위험 사례들을 선제적으로 파악하고 사고 예방을 위해 장치를 마련할 필요
  - 기술 변화에 따른 보안 기준 변경 등 세부 내용은 지속적으로 개정할 예정

- **인공지능기본법 하위법령의 세부내용 및 금융업권의 의견을 지속적으로 반영하여 개정 예정**
  - 25년 11월 인공지능기본법 시행령의 입법예고가 있었으나 아직 기타 고시, 가이드라인 등의 구체적인 내용은 확정되지 않은 상황
  - 향후 정책 방향이 구체화 되면 이를 반영하여 본 금융분야 AI 가이드라인도 개정할 예정
  - 또한 업권의 의견을 지속 수렴하여 본 가이드라인을 개정하는 것이 큰 방향임.

감사합니다

2025.12.

한국금융연구원 연구위원 백연주

KIF