디지털금융협의회
데이터 분과회의

금융분야 AI 가이드라인 및 주요 검토 필요사항

2021. 7. 8.

금 융 혁 신 기 획 단
금 융 데 이 터 정 책 과

차

Ⅰ. 금융분야 AI 규율 형식

□ 해외 주요국은 통상 **비규제적 지침·가이드라인 형태**로 AI 위험을 규율하나, 최근 **법률적 규제** 도입을 위한 움직임도 대두

○ 규율형식 차이에도 불구, **AI 활용원칙-거버넌스 구축-감사·평가 절차** 등 내용적인 측면에서는 대부분*이 유사

* 예외 : 최근 EU 「AI규제(초안)」의 경우, 인간의 기본권 등에 미치는 영향에 따라 AI위험수준을 3단계로 나누고 최고위험단계(Unacceptable risk)는 **서비스 금지** 등 **강력한 규제**를 적용

규제강도	구분	예시	
강 ↕ 약	법적 규제	AI 규제 초안	EU 집행위
		금융투자업의 AI 및 머신러닝 활용규칙 초안	국제증권관리위원회
		AI 및 기계학습 시스템 보고 의무	인도 증권위
	지침, 가이드라인	美 기업의 AI 및 알고리즘 이용에 관한 지침	美 FTC
		AI 윤리가이드라인 & 신뢰할 수 있는 AI 자체평가	EU집행위
		AI 및 데이터 보호 가이드라인	英 ICO
		AI 기반 의사결정에 대한 규제지침	英 ICO
		자동화된 의사결정에 대한 지침	캐나다 정부
	전략, 보고서	싱가포르 금융 AI윤리원칙	싱가포르
		AI윤리 프레임워크	호주
		빅데이터 및 고급분석 보고서	EU 은행청
		AI 규제 권장사항	獨 데이터 윤리위원회

□ 소비자 및 금융산업·시장 건전성 보호를 위해 AI 운영의 **방향성**을 **제시**하되, AI 활성화를 위해 **'약한 규제'** 형식의 가이드라인 필요

* 「신용정보법」상 **자동화신용평가 대응권**, 「자본시장법」상 **로보어드바이저 운영** 등 개별 법률에 **기초적인 AI 활용 규제 수단**이 있는 점을 고려

○ '강한 규제'에 해당하는 법규마련, 금감원 행정지도가 아닌 **모범규준**(Best Practice), **업권별 자율규제** 형식으로 **2단계 규율**

❶ 금융업권·서비스별 AI 활용사례 및 잠재위험 등이 상이함을 고려하여 **모범규준**으로 **총론 규제**를 설정 (→ **가이드라인**)

❷ 실제 운영상 **불확실성**을 제거하기 위해 **업권별 자율규제**로 **서비스**별 **실무지침** 등 운영(→ **가이드라인을 바탕으로 마련**)

□ 시장 초기에는 **필요최소한**의 **규제**를 **적용**하고, AI 활용이 활성화된 이후 **기능별·업권별 체계적·통일적 법률 형식**으로 **전환 검토**

Ⅱ. AI 가이드라인 주요내용

가이드라인 추진경과

□ ('20.7.~) 「**금융분야 AI Working Group**」 논의를 통해 금융업권 AI 실제 활용 현황 및 애로사항, 규제시 고려사항 등 논의

□ ('20.10.~'21.3.) **연구용역**을 통해 AI 가이드라인 **초안 마련**

* 「금융분야 AI 활성화를 위한 가이드라인 등 마련」 고학수(서울대) 외 6인

□ ('21.4.13) **디지털금융협의회**를 통한 **가이드라인 초안** 논의

□ ('21.5.~6.) **6개 금융업권협회 의견수렴** 등을 통한 **가이드라인** 보완

가. 목적과 범위

□ (**목표**) 금융분야 AI 시스템의 개발, 사업화 및 활용의 전 과정에서 **신뢰성**을 **제고**하여 **AI 활성화**에 **기여**

□ (**적용**) 금융권 **AI 시스템 전반**을 **대상**으로 하되 **금융기관**이 AI서비스의 특성을 감안하여 **가이드라인 적용 여부 자율결정**

√ 가이드라인 적용 대상

- (**대상업권**) AI를 금융거래 및 금융서비스 등에 적용한 **全금융업권**

  → 실무지침의 경우 **AI 활성화 정도** 등을 감안하여 **은행, 보험, 금융투자업권** 등 **대형 금융업권** 위주로 마련하되,

  · **금융연관분야**라 하더라도 **AI활용**에 따르는 **금융서비스** 등에 **미치는 위험**이 **크다고 판단**되는 경우(예:CB사, 전자금융업 등) **확대 적용**

- (**활용분야**) AI를 접목한 **全금융서비스**를 대상으로 하되, **금융소비자**에 미치는 영향이 없는 **Back Office 관련 분야**(예:RPA 등)의 경우 **제외**

나. 거버넌스의 구축

□ (**조직 내 AI윤리 마련**) 조직 내 가치, AI 활용 상황 등을 고려하여 **인공지능 이용**에 관한 **내부윤리 원칙**과 **기준**을 수립

□ (**책임·권한의 정의**) AI시스템의 위험 평가·관리를 위한 **구성원**의 **역할·책임·권한**을 구체적으로 **정의**(필요시, AI윤리위 별도 설치)

√ AI 조직의 역할 예시

① AI 활동에 대한 책임과 감독
② AI 정책 및 절차 관리
③ AI 활용에 대한 영향평가
④ AI 서비스에 대한 자체평가
⑤ 소비자 권리보장을 위한 체계 운영(이의제기 처리, AI서비스 공시 등)
⑥ 기업 內 AI 책임 문화 확산 촉진
⑦ AI 모델 및 학습데이터 거버넌스 관리
⑧ AI 관련 이슈 발생시 감독당국과의 설명 등 소통

□ (**위험관리정책 마련**) AI 시스템의 위험에 대한 **평가·관리 정책** 마련

○ **개인권리**에 **중대한 위험**을 **초래할 수 있는 AI 활용사례**에는 **내부통제** 및 **승인절차**를 **마련**하고, **승인책임자 지정**을 권고

√ 개인권리에 중대한 위험을 초래할 수 있는 AI 활용사례란?

- AI 활용에 따라 개인의 **금융거래계약의 체결·유지** 등에 **직접적인 차별**이 **발생하는 경우** (예 : 신용평가, 대출심사, 보험심사 등)
- 다만, 실제 사례별·기능별 차이를 감안하여 **금융회사**가 **개인권리**에 **중대한 위험**을 초래할 수 있는지 **자체평가**하여 결정

< 참고 : EU 집행위의 위험단계 구분 및 금융분야 예시(Annex 기준) >

Risk Level	정의	금융분야 예시
용인불가 위험	인간의 기본권에 명확한 위협으로 작용하는 시스템	-
고위험	인간의 생명·건강·기본권 등을 위험하게 할 수 있는 AI시스템	신용평가 등
제한된 위험	AI 사용에 심각한 위험이 없는 AI	챗봇
최소한의 위험	AI 사용에 거의 위험이 없어 자유롭게 활용 가능	-

√ 승인책임자의 지위 등? 회사의 AI 활용사례별 특성에 맞추어 책임있는 업무수행이 가능한 **지위**(임원 등)로 하되, **유사 직책***과 **겸직 허용**

* 예 : 최고위험관리책임자(CRO), 신용정보보호관리인, 최고정보보호책임자(CISO) 등

√ AI 위험관리정책 사항 (예시)

AI 도입 및 적용 절차	- AI 도입 시 적합성 검토 및 영향평가(위험식별 등) 실시 - AI 적용 시 학습데이터 관리(데이터 품질 평가, 비식별 조치 등) - AI 모델 개발 절차(보안성 준수, 설계 및 테스트 절차 등) - AI 조달 및 계약 시 고려사항 등
AI 자체 평가	- AI 편향성 평가 절차(학습데이터 평가, 편향성 완화조치 평가 등) - AI 성능 평가(AI 성능 측정지표, AI 전략 달성여부 등) - 보안성 평가(AI 서비스 보안조치, 학습데이터 보안조치 등) - AI 내부 감사 절차 등
금융소비자 보호	- AI 적용기술 및 데이터 유형 등 공개 - AI에 대한 금융소비자 권리(설명요구권 등) 및 사고시 통지절차 등 안내 - 금융소비자 이의제기시 처리절차 및 기준 - AI 결과에 대한 설명요구 수준 및 절차·기준 등
AI 운영·관리	- AI 서비스 백업 및 장애시 처리절차 - AI 서비스에 대한 인적개입 수준 및 개입 절차 - AI 성능·결과·보안성 모니터링 기준 - AI 기술에 대한 사내 교육 훈련 기준 - AI 모델 개선(업데이트) 및 재학습 절차 등

다. 기획·설계 단계

□ (**기획**) AI 활용목적이 **윤리원칙**에 **부합**하는지 여부, AI시스템 활용의 **사회적·경제적·문화적 영향, 잠재적 피해 가능성** 등 평가

□ (**설계**) AI가 인간의 의사결정과정을 대체하는 경우, AI시스템을 **감독·통제**하고 **책임성**을 유지할 수 있도록 **시스템**을 설계

√ <참고> 보험상품 설명시 음성봇 활용 요건 (보험업법 시행령 입법예고안)

① **설명속도, 음량** 등을 **조절할 수 있는 기능**을 제공할 것
② 소비자가 **AI 음성봇 사용 중단을 요청**하면 **즉각 중단**
③ 소비자가 설명내용에 대해 **질문이나 추가설명을 요청**하면, **설계사가 실시간으로 직접 응대**할 수 있는 시스템을 갖출 것
④ **음성봇**이 상품설명을 한다는 사실 및 ①·②·③에 대해서 **소비자에게 미리 안내하고 동의**를 받을 것

라. 개발 단계

□ **(데이터)** AI 학습 데이터의 **품질 검증·개선** 및 **최신성 유지** 등

* 예시 : 데이터 출처의 정확한 파악, 데이터 규격 문서화, 데이터 품질 검증(누락, 중복, 불일치 등), 편향되지 않은 충분한 학습데이터 확보, 무결성 검증 및 주기적 갱신 등 관리 활동 수행

□ **(개인정보)** 사생활 정보, 민감정보 등을 활용하는 경우 **비식별 조치** 등 **충분한 안전조치** 후 **개인정보 활용 필요성** 등을 평가

√ 국제표준화기구(ISO), AI 관련 데이터 생명주기별 프라이버시 위협 고려사항

구 분	고려 사항
데이터 확보	- AI 성능을 위한 데이터 확보와 개인정보보호를 위한 데이터 최소화 원칙 간 관계 고려 - 공격자의 데이터 스토리지 손상 위협에 대한 대비
데이터 전처리 및 모델링	- AI를 활용하여 일반 정보에서 민감정보를 추론하거나 비식별화된 데이터의 재식별 가능성 등이 있으므로 이에 대비
모델 쿼리(Query)	- 공격자가 모델 탈취 공격으로 민감정보를 유출하거나 본래 목적과 다르게 AI 모델을 악용할 수 있으므로 이에 대비

□ **(설명가능성)** 신뢰성, 통제가능성 확보 등을 위하여 **설명가능한 AI 기술 도입** 노력

마. 평가·검증 단계

□ **(성능)** AI 시스템 오류 유형간 통계적 상충관계* 등을 고려하여 **적합한 성능 목표 및 성능평가지표를 선정**하고 **충족여부** 확인

* 예: 은행이 대출부적격자를 걸러내기 위해 여신심사를 강화하는 경우 대출적격자(True)에 대한 여신이 거절(False)되는 경우도 증가

√ 서비스 특성별 성능 판단기준(예시)

○ **소비자**에 **금융거래 기회**를 **제공**하는 기능을 수행하는 경우(예 : 신용평가, 대출심사, 보험심사 등) : **False Negative 오류**(True를 False로 구분) **최소화**

○ **사기탐지**, **규제 미준수** 등 **위법·부당사례 탐지** 기능을 수행하는 경우(예 : FDS, Reg-tech) : **False Positive 오류**(False를 True로 구분) **최소화**

□ (공정성) AI 시스템 **공정성 평가지표**를 선정·측정, 불균형이 **발견된 경우 공정성 개선을 위한** 기술적·관리적 노력 제고

√ 서비스 특성별 공정성 판단기준(예시)

ㅇ **특정 계층·집단**에 대한 **결과적 평등**을 고려할 필요가 있는 경우(예 : 금융 소외계층 금융접근성 제고) : **인구통계적 동등성 기준 적용**
(예 : 집단간 대출승인율 동등)

ㅇ **소비자**에 **금융거래 기회**를 **제공**하는 기능을 수행하는 경우(예 : 신용평가, 대출심사, 보험심사 등) : **기회의 균등 기준**을 적용(예 : 집단간 재현율* 동등)

* 재현율(True Positive Rate) : 전체 대출적격자 중 AI가 대출적격자로 정확하게 예측한 비율

□ (설명가능성) **상황에 맞는 설명**이 도출되는지 여부를 확인하고, 설명가능성을 합리적 수준으로 개선하기 위해 노력

바. 도입·운영·모니터링

□ 대고객 AI시스템 운영시 고객에게 AI 이용을 **사전고지**하고, 자동화평가(Profiling) 대응권* 등 **소비자의 권리** 및 **이의신청·민원제기 방식** 등 **권리구제 방안**을 고지

* AI 등 자동화평가 결과에 대해 설명요구, 이의제기, 기초데이터의 제출·정정, 결과의 재산정을 요구할 수 있는 권리 (「신용정보법」 §36-2)

□ AI시스템 성능을 주기적으로 **모니터링**하고, **재학습 필요성 여부**를 검토하는 등 **성능 개선 가능성 확인** 및 **개선 추진**

□ AI시스템 이용자에 의한 **오용·악용 가능성**을 **방지**, AI 개발 환경의 **보안취약성 상시통지 시스템 마련** 등

√ 국제표준화기구(ISO), AI위험 완화 조치 전략 (☞참고2)

- AI 시스템 운영에 따르는 위험을 완화하기 위한 10가지 전략*을 제시하고, 서비스 생명주기 전반에 대해 **위험완화 조치의 주기적 수행 권고**

* 투명성, 설명가능성, 통제가능성, 편향성 완화, 프라이버시, 신뢰성·복원력·견고성, 하드웨어 결함 완화, 기능적 안전성, 테스트·평가, 활용·적용성

사. AI 업무위탁에 관한 특례

□ (**지침 마련**) AI 시스템 개발·운영 등을 **외부기관**에 **위탁**할 경우 **수탁기관**이 준수하여야할 **위험관리지침**을 **마련·공유**

○ **AI 윤리 및 위험관리정책의 주요내용을 포함**하여 마련함으로써 수탁기관도 위탁기관의 AI 운영원칙에 따라 시스템 개발·운영

* 소규모 금융회사로 자체적인 AI 시스템 운영 없이 외부기관 위탁만으로 AI시스템을 개발·운영하는 경우 위험관리지침으로 위험관리정책을 갈음

□ (**주기적 점검**) 위탁한 AI시스템 개발·운영이 **위험관리지침**에 따라 이루어졌는지 **주기적인 보고·점검** 체계 구축·운영

○ 개인권리에 중대한 위험을 초래할 수 있는 AI 활용사례에 대해서는 **내부통제** 및 **승인절차**에 준한 **사전 점검**을 거칠 수 있도록 함

* 예시 : AI 개발·운영 계획(사용 데이터, 잠재적 위험에 대한 발생가능성에 대한 평가 및 조치내역 등) 등에 대한 위탁기관의 사전확인, AI 시스템 관련 소비자 피해 발생시 조치 및 보고 절차 마련 등

√ EU「신뢰할 수 있는 AI 자체평가」세부내용 中

- 제3자(공급업체, 이용자, 근로자 등)가 AI시스템의 잠재적인 취약성, 위험, 편향을 보고할 수 있는 절차를 마련했는지?
- 이러한 보고절차가 위험관리 절차의 개선에 반영되는지?

□ (**소비자 피해**) 소비자 피해 발생에 따른 **책임주체**를 **명확히** 하고, 위탁계약 체결시 **손해배상 지연** 등을 방지하기 위한 **명확한 책임조항** 및 **손해배상 처리 절차** 등을 기재

참고1 금융분야 AI 운영 가이드라인 (안)

1. 목적과 적용 범위

가. 가이드라인은 금융분야에서의 인공지능(이하 'AI'라 한다.) 시스템의 개발, 사업화 및 활용과 관련한 기획·설계, 평가·검증, 도입·운영 및 모니터링의 전 과정에서 신뢰성을 제고하여 AI 활성화를 제고하고 금융서비스에 대한 고객신뢰를 확보하는데 기여하는 것을 목적으로 한다.

나. 가이드라인은 금융서비스 및 금융상품의 제공을 위한 업무에 AI 시스템을 직·간접적으로 활용(금융회사 내부 직원관리, 단순 업무 효율화 등 AI 시스템 활용으로 고객에 미치는 영향이 없는 경우를 제외한다.)하거나 활용하고자 하는 금융회사, 상품추천·신용평가 등 금융연관 서비스 제공을 위한 업무에 AI시스템을 직·간접적으로 활용하거나 활용하고자 하는 비금융회사(이하 '금융회사 등'이라 한다.) 등에 적용한다.

다. AI 시스템이란 특정 목표가 주어진 상태에서, 데이터를 획득하여 환경을 인식하고, 획득된 데이터를 해석하며, 지식을 추론하거나 정보를 처리하고, 해당 목표를 달성하기 위한 최선의 행동을 결정함으로써 물리적 또는 디지털 차원에서 작동하는 인간이 설계한 소프트웨어 또는 하드웨어 시스템을 의미한다.

라. 금융회사 등은 AI 시스템이 활용된 서비스의 특성 및 고객 특성, AI 시스템이 활용된 서비스의 고객 수 등을 종합적으로 고려하여 AI 활성화 및 금융서비스에 대한 고객신뢰 확보라는 가이드라인의 취지를 훼손하지 않는 범위 내에서 가이드라인의 적용 범위 등을 조정할 수 있다.

2. 거버넌스의 구축

가. 금융회사 등은 조직이 추구하는 가치와 주된 AI 활용 맥락 등을 고려하여 AI 활용에 관한 윤리원칙과 기준을 수립한다.

나. 금융회사 등은 AI 시스템의 잠재적 위험을 평가하고 이를 관리하기 위하여 구성원의 역할·책임·권한 등을 AI 시스템의 전 과정에 걸쳐 구체적으로 정의한다. 금융회사 등은 AI 윤리원칙과 기준에 맞는 조직 관리를 위하여 AI 윤리위원회를 별도로 설치할 수 있다.

다. 금융회사 등은 AI 시스템의 전 과정에 걸쳐 AI 활용에 따라 나타날 수 있는 잠재적 위험을 인식·평가하고, 이를 관리·최소화하는 방안을 검토하는 등 AI 활용으로 인한 잠재적 위험을 관리하는데 필요한 위험관리정책을 마련한다. 위험관리정책은 소비자 권리보장을 위한 시스템 운영, AI 모델 및 학습데이터의 관리, AI 시스템 관련 문제 발생 시 감독당국과의 소통, 회사 내 AI 책임 문화 확산의 촉진 등의 업무처리에 관한 내용을 포함한다.

라. 금융회사 등이 개인에 대한 부당한 차별 등 개인의 권익와 안전, 자유에 대한 중대한 위험을 초래할 수 있는 서비스(이하 '고위험 서비스'라 한다.)에 대해 AI 시스템을 활용하는 경우, 적절한 내부통제 활동 및 승인절차를 마련하고, 승인 책임자를 지정한다, 승인책임자는 책임있는 업무 수행이 가능한 지위로 하되 최고위험관리책임자, 신용정보보호·관리인, 최고정보보호책임자 등 유사업무와 겸직할 수 있다.

3. AI 시스템의 기획 및 설계 단계

가. 금융회사 등은 AI 시스템의 활용 목적이 윤리원칙에 부합하는지 검토하고, 활용 맥락을 고려하여 AI 활용으로 나타날 수 있는 사회적, 경제적, 문화적 영향 및 잠재적 피해 가능성을 평가하여야 한다.

나. 금융회사 등은 AI 시스템의 목적 및 특성, 고객의 특성 등을 고려하여 탄력적으로 AI 시스템을 활용할 수 있다. 다만, AI 시스템이 인간의 의사결정을 전면적으로 대체하거나, 중요 의사결정을 대체하는 경우 금융회사 등은 AI 시스템을 효과적으로 감독·통제하고 책임성을 유지할 수 있도록 AI 시스템을 설계한다.

4. AI 시스템의 개발 단계

가. 금융회사 등은 올바른 AI 학습을 위하여 데이터의 출처, 품질, 편향성 등을 조사·검증하고 주기적인 데이터 갱신 등 데이터 품질 개선을 위한 방법을 검토한다.

나. 금융회사 등은 AI 시스템이 「개인정보 보호법」 제23조제1항 및 시행령 제18조에 따른 민감정보, 또는 이와 유사한 사생활 관련 정보 등을 활용하는 경우 사전 동의 획득 또는 비식별조치 등 안전한 정보 활용을 위한 충분한 조치를 거쳐야 하며, 해당 정보 활용의 필요성을 평가하고, 데이터 처리 과정에서 해당 정보의 재식별, 유출, 악용가능성이 없도록 한다.

다. 금융회사 등은 관련 법령 등에 따라 고객에 대한 설명의무가 있는 금융서비스 등에 AI 시스템을 활용하는 경우 또는 고위험 서비스에 AI 시스템을 활용하는 경우에는 개발 단계에서부터 설명 가능성을 고려하고, 가용한 설명 가능한 인공지능 기술 등을 확인하여 이를 도입하기 위한 노력을 기울인다.

5. AI 시스템의 평가 및 검증 단계

가. 금융회사 등은 AI 윤리원칙, AI 시스템의 목적, 오류 사례에 따른 고객 영향 및 잠재적 피해의 정도, AI 성능 측정 지표의 상충관계 등을 종합적으로 고려하여 AI 시스템의 적절한 성능 목표 수준 및 성능 측정 지표를 선정·관리한다.

나. 금융회사 등은 AI 윤리원칙, AI 시스템의 목적, 공정성 평가 지표별 고객 영향 및 잠재적 피해의 정도, AI 공정성 평가 지표의 상충관계 등을 종합적으로 고려하여 AI 시스템의 적절한 공정성 목표 수준 및 공정성 판단 지표를 선정·관리한다. 선정된 공정성 판단 지표에 따라 불균형이 발견된 경우, 공정성을 개선시킬 수 있는 기술적·관리적 노력을 기울인다.

다. 금융회사 등은 관련 법령 등에 따라 고객에 대한 설명의무가 있는 금융서비스 등에 AI 시스템을 활용하는 경우 또는 고위험 서비스에 AI 시스템을 활용하는 경우 설명가능 인공지능 기술 등 적절한 인공지능 기술을 투명하게 적용하여 맥락에 맞는 설명이 도출되는지 여부를 확인하고, AI 시스템의 안정성·신뢰성 등을 훼손하지 않는 범위 내에서 설명가능성을 합리적인 수준으로 개선하기 위해 노력해야 한다.

6. AI시스템의 도입, 운영 및 모니터링 단계

가. 금융회사 등은 대고객 AI 시스템 운영시 고객에 AI 이용 여부, 설명·이의제기권 등 관련 법령에 따른 소비자의 권리, 이의신청·민원제기 방식 등 AI 시스템의 성격에 맞추어 적절한 권리구제 방안을 고지해야 한다.

나. 금융회사 등은 도입된 AI 시스템의 성능을 주기적으로 모니터링하고, 데이터 재학습 필요성 검토 등 성능 개선 가능성을 확인한다.

다. 금융회사 등은 AI 시스템에 고객 또는 제3자에 의한 데이터 오염 공격, 적대적 공격 등 오용·악용 가능성이 있는지 여부를 확인하고, 가용한 기술 범위 내에서 오용·악용 사례를 최소화할 수 있는 방안을 도입하기 위해 노력한다. 금융회사 등은 오픈소스 기반 AI 개발 프레임워크 등 AI 개발 환경의 보안 취약성에 관해 상시적으로 통지를 받을 수 있는 절차를 반영하고, 최선의 보안 시스템을 구축하기 위하여 노력한다.

7. AI 시스템 업무위탁에 관한 특례

가. 금융회사 등은 AI 시스템의 개발·운영 등을 외부기관에 위탁하고자 할 경우, 수탁기관이 동 가이드라인 및 가이드라인에 기초하여 마련된 AI 윤리원칙 및 위험관리정책을 준수하여 AI 시스템을 개발·운영할 수 있도록 하기 위한 위험관리지침을 마련하고 금융회사가 직접 AI 시스템을 개발·운영하는 경우에 비해 AI 시스템 운영에 따르는 위험이 확대되지 않도록 한다.

나. 금융회사 등은 외부기관에 의한 AI 시스템 개발·운영이 위험관리지침에 따라 이루어졌는지 주기적인 보고·점검 체계를 구축·운영하고, 고위험서비스에 대해서는 AI 개발·운영계획 등에 대한 금융회사 등의 사전확인, 소비자 피해 발생시 조치 및 보고 절차 마련 등 엄격한 사전 점검이 이루어질 수 있도록 한다.

다. 금융회사 등과 외부기관은 AI 시스템 개발·운영 등에 따라 소비자 피해가 발생한 경우 손해배상 지연 등을 방지하기 위한 명확한 책임 조항 및 손해배상 처리 절차 등을 마련한다.

참고2 AI 위험 완화 전략 세부 내용(ISO)

구분	주요 내용
투명성	- 동일 조건에 반복 가능한 결과를 도출하는 AI 시스템 구축 - 데이터, 특징, 모델, 알고리즘, 학습방법 등 품질을 보장하기 위한 방법과 절차를 외부에 공개 - AI 시스템에 대한 외부의 제3자 평가를 수용 - 공정성 및 사회적 윤리 가치를 고려하여 AI에 반영 - 수집 데이터, 데이터 출처, 수집 근거 등을 제공 - AI 처리 과정 및 결과에 대한 상세한 설명 제공 및 사람의 개입 요청 시 대응 - AI 시스템에 등급, 기호, 아이콘, 마크 등을 사용 - 일반 이용자가 이해할 수 있도록 AI의 동작방식을 설명 - AI 알고리즘이 활용 목적에 부합
설명 가능성	- AI 시스템을 활용하기 전에 일반적인 특성 및 기능 관련 정보를 이용자에게 사전 설명 - AI가 결과 도출 후 결정 관련 특성 및 특징에 대한 사후 설명 - 사후설명은 전반적 AI 처리 과정을 설명하거나 개별 입력값 및 출력값에 대한 AI 모델의 처리 절차를 세부적으로 설명* * ① 인과관계 설명 ② 기능적 설명 ③ 결과의 타당성 설명 - 결과의 타당성 설명은 AI 시스템에 대한 기술적 설명을 넘어 관련 법·규제 및 기업의 처리 절차 등을 설명 - 이해관계자의 요구, 데이터, AI 시스템 유형 및 제반 환경* 등을 고려하여 설명의 수준 고려 * 민감정보 활용, AI 실패가 개인에게 중대한 영향, AI가 이용자의 자율성을 제한, 성평등과 같은 사회적 이슈 발생 가능성 등 - 일관된 설명을 제공하는지에 대해 평가 수행
통제 가능성	- 운영자가 AI의 제어권을 넘겨받을 수 있는 메커니즘 구현 - 최종 의사결정 및 AI 시스템 개선시 사람의 개입 절차 고려
편향성 완화	- 적절한 임계값 설정, 법규제 등을 고려한 시스템 요건 정의 - 데이터 출처 및 데이터 완전성 분석, 데이터 수집 절차 검토 - AI 학습시 편향을 탐지하고 완화하기 위한 관련 기법 적용 - 편향성 탐지 테스트 및 기술 적용 - 정기적인 운영 검토를 통해 편향 관련 문제 식별

구분	주요 내용
프라이버시	- K-익명성, 차분 프라이버시 등 개인정보 비식별 처리 기법 적용 - 비식별 정보의 재식별 위험성 관리
신뢰성, 복원력 및 견고성	- AI 시스템이 필요한 기능을 수행할 수 있도록 신뢰성 확보 - AI 시스템의 내결함성(Fault Tolerance) 확보 - AI 시스템의 복원력 및 견고성* 확보 * 잡음(Noise)이 많은 대규모 데이터를 이용한 학습 등을 통해 달성
하드웨어 결함 완화	- 인프라 이중화(Redundancy) 및 결함에 안전한(Fail-safe) 하드웨어 구현* * 전자장치 안전관련 시스템의 기능 안전성 표준(IEC 61508) 참고
기능적 안전성	- 시스템의 기능 안전성을 보장하기 위해 시스템 통제* 관련 표준 등을 참고하여 구현 * AI 결과를 모니터링하여 허용범위 내에 있지 않은 경우 제어하는 등
테스트 및 평가	- 소프트웨어 검증을 위해 인공신경망에 대한 평가* * AI 결과 도출의 불확실성, 학습데이터의 안정성 등 평가 - AI 시스템에 대한 실증 테스트* * 벤치마킹, AI 결과에 대한 전문가 검토, AI 입출력 테스트 등 - AI 결과와 사람이 도출한 결과를 비교 평가 - 가상환경 및 실제 환경에서의 AI 시뮬레이션 테스트 - AI 시스템 목적에 따라 측정지표를 설정하여 견고성 테스트* * AI 시스템이 예측 가능하고 허용 범위 내 결과를 출력하는지 점검 - 프라이버시 측정기준 수립 및 평가 - AI 시스템의 예측가능성 평가* * AI의 행동을 추론하여 실제 결과와 맞는지 비교 평가
활용 및 적용성	- 각종 법·규제 및 표준 준수 - AI 시스템에 라벨(기본적 안내) 표시* * AI 목적, AI 활용 위험, 재교육 빈도, 평가 수행여부, 데이터 출처 등 - 인지과학 등 사람과 AI 간 상호작용 고려

Ⅲ. 향후 추진계획

□ (7.8.) 「금융분야 AI 가이드라인」 **발표** 및 **금융업권 전파**

○ 다만, 회사별 AI 윤리 마련, 위험관리정책 수립 등이 내실있게 운영되도록 가이드라인 이행을 위한 **충분한 준비기간**을 **부여**

□ (3분기 중) 최종 가이드라인을 바탕으로 업권별 특성 등을 반영한 **'금융업권/서비스별 세부실무지침'** 마련 추진

○ (금융업권) AI 서비스 활용도가 높은 **은행·금투·카드업권** 등을 중심으로 **금융업법상 규제와의 정합성을 고려**한 지침 마련

- 기존 금융업법상 규제를 AI 서비스에 적용할 경우, 나타날 수 있는 **규제 불확실성을 해소하는 보완장치**를 **마련**함에 초점

* 예 : 법령상 규제 준수로 볼 수 있는 적정한 AI 설계·운영 절차 등

○ (서비스별) **신용평가, 챗봇 등 고객응대, 설명의무 수행** 등 **주요 기능·서비스**에 대해서는 **세부준칙** 마련

□ (연내) 「금융분야 AI 가이드라인」 **전면 시행**

○ AI 기술의 발전, 금융분야 규제 변화, 금융분야의 AI 수용도 등을 고려하여 **상시적**으로 「금융분야 AI 가이드라인」 **개선·보완**

< 향후 추진일정 >

<table>
<tr><th></th><th>7월</th><th>3분기</th><th>연내</th></tr>
<tr><td>AI 가이드라인</td><td>가이드라인 최종안 마련 및 발표</td><td colspan="2">AI 가이드라인 적용 준비 → 전면 시행</td></tr>
<tr><td>실무지침</td><td>실무지침 제정반 구성·운영</td><td>실무지침 마련</td><td>실무지침 발표·전파</td></tr>
</table>