금융분야 인공지능 가이드라인

2026. 6

목 차

Ⅰ. 금융분야 인공지능 가이드라인 개요 ····· 1

1. 용어의 정의 ····· 1

2. 적용 범위 ····· 4

Ⅱ. 7대 원칙 ····· 6

1. 거버넌스 원칙 ····· 6

2. 합법성 원칙 ····· 17

3. 보조수단성 원칙 ····· 29

4. 신뢰성 원칙 ····· 38

5. 금융안정성 원칙 ····· 55

6. 신의성실 원칙 ····· 63

7. 보안성 원칙 ····· 69

Ⅰ. 금융분야 인공지능 가이드라인 개요

인공지능(Artificial Intelligence, AI) 기술이 급격히 발전함에 따라 금융산업에서도 인공지능을 활용하여 산업의 생산성을 제고하고 운영을 효율화하고자 하는 노력이 이어지고 있다. 그간 금융회사가 인공지능 기술 활용시 참고할 수 있도록 「금융분야 AI 운영 가이드라인(2021년)」, 「금융분야 AI 개발·활용 안내서(2022년)」, 「금융분야 AI 보안 가이드라인(2023년)」을 발표하였다. 이후 생성형 인공지능 및 AI 에이전트 등 새로운 인공지능 기술이 등장하고, 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 "인공지능기본법"이라고 한다)의 제정(2025년 1월)·시행(2026년 1월)과 함께 하위법령 및 관련 가이드라인이 제정되면서 금융 분야에서도 기술 양상 및 규제 환경 변화를 반영한 가이드라인 개정 필요성이 증가하였다. 이에 따라 동 가이드라인은 인공지능 기술 및 관련 제도 변화를 기존 금융분야 AI 가이드라인에 반영하여 금융 회사가 안전하게 인공지능을 개발·이용할 수 있도록 업무 전반에 걸친 리스크 관리의 방향과 원칙을 제시하고자 하였다. 동 가이드라인은 빠른 인공지능 기술의 발전 속도, 금융분야의 인공지능 수용도, 인공지능기본법 등 관련 제도의 변경을 고려하여 기존 가이드라인과 동일하게 금융권에서 자율적으로 적용할 수 있도록 하였다. 동 가이드라인은 향후 금융권 의견을 수렴하여 지속적으로 개선·보완해 나갈 예정이다.

1. 용어의 정의

용어	정의
인공지능	학습, 추론, 지각, 판단, 언어의 이해 등 인간이 가진 지적 능력을 전자적 방법으로 구현한 것
인공지능시스템	다양한 수준의 자율성과 적응성을 가지고 주어진 목표를 위하여 실제 및 가상환경에 영향을 미치는 예측, 추천, 결정 등의 결과물을 추론하는 인공지능 기반 시스템
인공지능서비스	인공지능, 인공지능시스템 또는 인공지능제품을 활용할 수 있도록 제공하는 서비스로 정보 분석, 예측, 추천, 창작 등의 기능을 제공하는 것

용어	정의
인공지능개발사업자	인공지능을 개발하여 제공하는 자
인공지능이용사업자	인공지능개발사업자가 제공한 인공지능을 이용하여 인공지능제품 또는 인공지능서비스를 제공하는 자
이용자	인공지능제품 또는 인공지능서비스를 제공받는 자
수명주기	인공지능시스템의 계획·설계, 데이터 수집·처리, 인공지능 모델 개발, 시스템 구현, 운영·모니터링 등으로 발전상의 변화 전 과정
금융회사등	가.「가상자산 이용자 보호 등에 관한 법률」에 따른 가상자산사업자 나.「금융지주회사법」에 따른 금융지주회사 다.「기술보증기금법」에 따른 기술보증기금 라.「농업협동조합법」에 따른 조합과 그 중앙회 및 농협은행 마.「대부업 등의 등록 및 금융이용자 보호에 관한 법률」 제3조 제2항에 따라 금융위원회에 등록한 대부업자등 바.「보험업법」에 따른 보험회사·보험대리점·보험중개사 및 보험요율산출기관 사.「산림조합법」에 따른 산림조합 및 산림조합중앙회 아.「상호저축은행법」에 따른 상호저축은행 및 상호저축은행중앙회 자.「서민의 금융생활 지원에 관한 법률」에 따른 서민금융진흥원 및 신용회복위원회 차.「새마을금고법」에 따른 새마을금고, 새마을금고중앙회 및 새마을금고자산관리회사 카.「수산업협동조합법」에 따른 조합과 그 중앙회 및 수협은행 타.「신용보증기금법」에 따른 신용보증기금 파.「신용정보의 이용 및 보호에 관한 법률」에 따른 신용정보회사·본인신용정보관리회사 및 종합신용정보집중기관 하.「신용협동조합법」에 따른 신용협동조합 및 신용협동조합 중앙회 거.「여신전문금융업법」에 따른 여신전문금융회사·겸영여신업자 및 신기술사업투자조합 너.「예금자보호법」에 따른 예금보험공사 및 정리금융회사 더.「온라인투자연계금융업 및 이용자 보호에 관한 법률」 제2조 제3호에 따른 온라인투자연계금융업자 러.「우체국예금·보험에 관한 법률」에 따른 체신관서 머.「은행법」에 따른 은행(같은 법 제59조에 따라 은행으로 보는 자를 포함한다) 버.「자본시장과 금융투자업에 관한 법률」에 따른 금융투자업자·투자자문업·투자일임업·증권금융회사·종합금융회사·자금중개회사 및 명의개서대행회사 서.「전자금융거래법」에 따른 전자금융업자 어.「중소기업은행법」에 따른 중소기업은행 저.「지역신용보증재단법」에 따른 신용보증재단과 그 중앙회 처.「한국산업은행법」에 따른 한국산업은행

용어	정의
	커. 「한국수출입은행법」에 따른 한국수출입은행 터. 「한국자산관리공사 설립 등에 관한 법률」에 따른 한국자산관리공사 및 같은 법 제26조제항제1호 및 제2호의 업무를 수행하기 위하여 같은 법 제26조제1하아제4호라목에 따라 설립된 기관(새출발기금 등) 퍼. 그 밖에 「금융위원회의 설치 등에 관한 법률」에 따라 금융감독원의 검사를 받는 기관 허. 가목부터 퍼목까지에 준하는 자로서 금융업 및 금융관련 업무를 하는 기관·단체·사업자(한국금융투자협회, 생명보험협회, 손해보험협회, 은행연합회, 여신금융협회 등)
고영향 인공지능	사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능시스템으로서 인공지능기본법 제2조 제4호 각 목의 어느 하나의 영역에서 활용되는 것 사. 채용, 대출심사 등 개인의 권리·의무 관계에 중대한 영향을 미치는 판단 또는 평가
고위험 인공지능	동 가이드라인의 [Ⅱ.1.3. 위험평가 체계]를 통해 고위험 서비스로 자체 분류하고 결정한 인공지능시스템 ※ 고영향 인공지능은 고위험 인공지능에 해당되나, 고위험 인공지능은 고영향 인공지능에 해당되지 않을 수 있음

2. 적용 범위

동 가이드라인은 금융분야 전반에서 인공지능시스템의 바람직한 활용 방향을 제시하는 것을 목적으로 한다. 이에 따라 금융상품 및 금융서비스의 제공을 위한 업무에 인공지능시스템을 직·간접적으로 활용하는 금융회사 및 비금융회사에 적용될 수 있으며 특정 회사(예: 인공지능기본법에 따른 인공지능 사업자)나 특정 업무(예: 인공지능기본법에 근거한 고영향 인공지능)에 국한되지 않는다.

적용 대상인 금융회사의 범위는 원칙적으로 은행, 보험사, 카드사, 금융투자업자 등 금융산업에서 금융상품 및 서비스를 제공하는 회사를 포함한다. 다만, 비금융회사(예: 핀테크기업)의 경우에도 인공지능시스템의 활용 결과가 금융거래 제공에 영향을 줄 수 있는 경우 동 가이드라인 적용 범위에 포함될 수 있다.

동 가이드라인 적용 범위에 포함되는 인공지능시스템의 적용 업무는 대출심사, 신용평가, 챗봇, 금융상품 비교·추천, 금융사기 탐지시스템 (FDS) 등 금융권 인공지능시스템 전반을 대상으로 한다. 금융서비스나 상품을 고객에게 제공하는 과정에 직접적인 영향을 미치는 경우뿐 아니라 금융회사 내부의 지원 및 관리를 위하여 인공지능시스템을 활용하는 경우에도 본 가이드라인을 폭넓게 적용할 것을 권장한다.

동 가이드라인은 주요국의 인공지능 규율 사항, 선진 해외 금융회사의 적용 사례 등을 참고하여 모범 사례를 제시하기 위한 목적으로 마련되었다. 따라서 개별 회사는 인공지능 활용 수준, 인공지능 활용에 근거한 영향, 인적·물적 자원 등 회사별 환경과 자원, 서비스의 형태, 인공지능기본법령의 규율 내용 등을 종합적으로 고려하여 자율적으로 가이드라인 적용 수준을 결정할 수 있다. 예를 들어 인공지능기본법에 따른 고영향 인공지능에 해당하는 경우에는

고영향 인공지능 사업자의 책무를 이행하고 추가적으로 동 가이드라인의 거버넌스 원칙, 금융 안정성 원칙 등에 해당되는 내용을 적용할 수 있다.

또한, 인공지능기본법에 근거한 고영향 인공지능에 해당하지 않는 경우에도 인공지능의 영향을 자율적으로 평가하여 소비자 보호, 인적 개입, 보안성 원칙 등 동 가이드라인에 해당되는 내용을 선택하여 적용할 수 있다.

인공지능기본법령과 동 가이드라인이 중복 적용되는 경우에는 인공지능기본법령 및 관련 가이드라인을 우선 적용하되, 인공지능기본법령이 적용되지 않거나, 인공지능기본법령에서 규정하지 않는 내용의 경우에는 동 가이드라인을 적용할 수 있다.

Ⅱ. 7대 원칙

1. 거버넌스 원칙

금융회사등의 최고경영자를 포함한 경영진은 인공지능 개발·이용에 관심을 갖고 역할과 책임을 분담한다. 경영진은 인공지능 이용 범위, 책임, 권한 등을 내부통제 기준 및 위험관리 기준에 포함시키고, 이사회는 인공지능 활용을 포함한 직무에 대한 내부통제 체계 및 운영 적정성을 점검하고 평가한다. 이를 위해 금융회사등은 인공지능 개발·이용 등과 관련된 의사결정기구 및 독립적 위험관리 전담 조직 등을 구성하고, 관련 내규를 마련하는 등 체계적인 '인공지능 거버넌스'를 구축한다.

높은 수준의 거버넌스는 안전하고 건전한 인공지능 개발·이용의 필요조건이다. 금융회사등은 인공지능 거버넌스를 확립함으로써 견제와 균형을 이루게 된다. 그 결과로 인공지능시스템을 체계적으로 기획·개발·운영·활용할 수 있으며, 각 단계에서 발생할 수 있는 위험을 평가·관리할 수 있다. 국제결제은행(BIS), 영국 금융행위감독청(FCA) 등 주요 국제기구나 감독당국도 거버넌스 구축 및 위험관리를 안전한 인공지능 개발·이용의 핵심 요건으로 꼽는다.

인공지능 거버넌스를 수립할 때에는 인공지능 수명주기 전반에 걸쳐 역할과 책임을 명확히 규정하여야 한다. 금융회사등은 책임 소재를 명확히 하여 인공지능시스템에 대한 위험을 효과적으로 관리할 수 있다. 또한, 부서 간 원활한 의사소통과 합리적인 의사결정을 유도하고, 임직원의 주관적 판단으로 인해 위험이 확대될 가능성을 사전에 방지할 수 있다. 금융회사등은 인공지능 거버넌스를 구축하는 과정에서 데이터 거버넌스나 IT 거버넌스 등 기존의 내부 관리체계를 활용하거나 해당 체계와의 연계를 고려한다.

본 장에서 설명하는 거버넌스 구축, 위험평가 및 위험통제 등의 내용은 「금융분야 AI 위험관리 프레임워크」(AI Risk Management Framework in Financial Sector, 이하 "AI RMF"이라고 한다)에서 구체적으로 기술하고 있다. 세부적인 사항은 AI RMF를 통해 확인할 수 있으므로, 본 장에서는 '거버넌스 원칙'과 관련된 핵심 요건과 그 필요성 등에 대하여 주요 내용 위주로 설명한다.

1.1. 의사결정기구 및 전담 조직의 구성

✓ **인공지능 위험관리** 등을 위한 **의사결정기구**를 설치하여 **인공지능 개발·이용**을 **적극적**으로 관리하고, **독립**된 **위험관리 전담 조직**을 **구성**하여 **인공지능 관련 업무 전반**을 **통제·관리**한다.

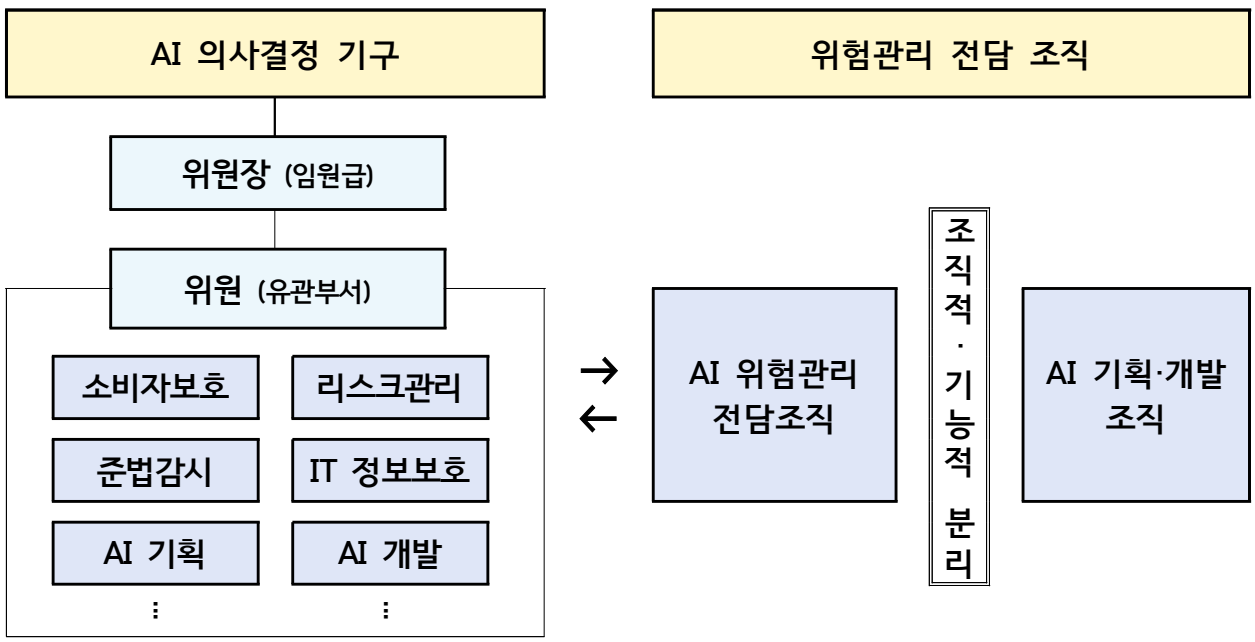
금융회사등은 인공지능 위험을 관리하기 위한 최고 의사결정기구(예: 인공지능 윤리위원회)를 설치하여 인공지능 위험관리 정책 수립 등 인공지능 개발·이용과 관련된 주요 의사결정을 적극적으로 관리한다.

최고 의사결정기구는 윤리원칙을 포함하여, 인공지능 관련 내규의 제·개정, 고위험 또는 고영향 인공지능서비스 승인 등 인공지능 관련 중요 사항을 심의·의결한다. 또한, 인공지능 위험을 체계적으로 관리할 수 있도록 위험평가 및 통제 등을 포함한 위험관리 정책을 수립하며, 인공지능으로 인해 소비자 권익이 침해될 소지를 파악하여 소비자 보호 정책을 마련한다. 의사결정기구의 위원장 등은 이사회 및 최고경영자에게 정기적으로 인공지능과 관련된 주요 사항을 보고하여 이사회 및 최고경영자가 인공지능 개발·이용과 관련한 사업계획, 전략, 위험 등을 명확하게 파악하고 운영의 적정성을 점검·평가할 수 있도록 돕는다.

한편, 금융회사등은 인공지능서비스 기획 조직이나 인공지능시스템

개발 조직과 독립된 위험관리 전담 조직을 설치하여 인공지능 관련 업무 전반을 통제·관리하게 한다. 위험관리 전담 조직은 인공지능 위험의 인식·측정·평가·통제 등 위험관리 전반을 총괄하고, 인공지능기본법 등 관련 법규상 각종 의무의 준수 여부를 관리·감독한다. 특히 인공지능 사업조직 및 인공지능 위험관리 전담 조직은 조직적·기능적으로 분리되어야 하나, 회사의 규모, 자원, 영위 업무, 인공지능 활용 범위 등을 고려하여 인공지능 위험관리 전담 조직을 '인공지능 업무부서와 별도의 독립된 조직으로 분리하되, 인공지능 주관본부 소속으로 운영하는 방안'도 고려할 수 있다. 아울러, 위험관리 전담 조직의 역할과 책임은 인공지능 관련 내규 또는 별도의 직제 규정 등에 명시한다.

참고 **인공지능 거버넌스 체계 구성 (예시)**



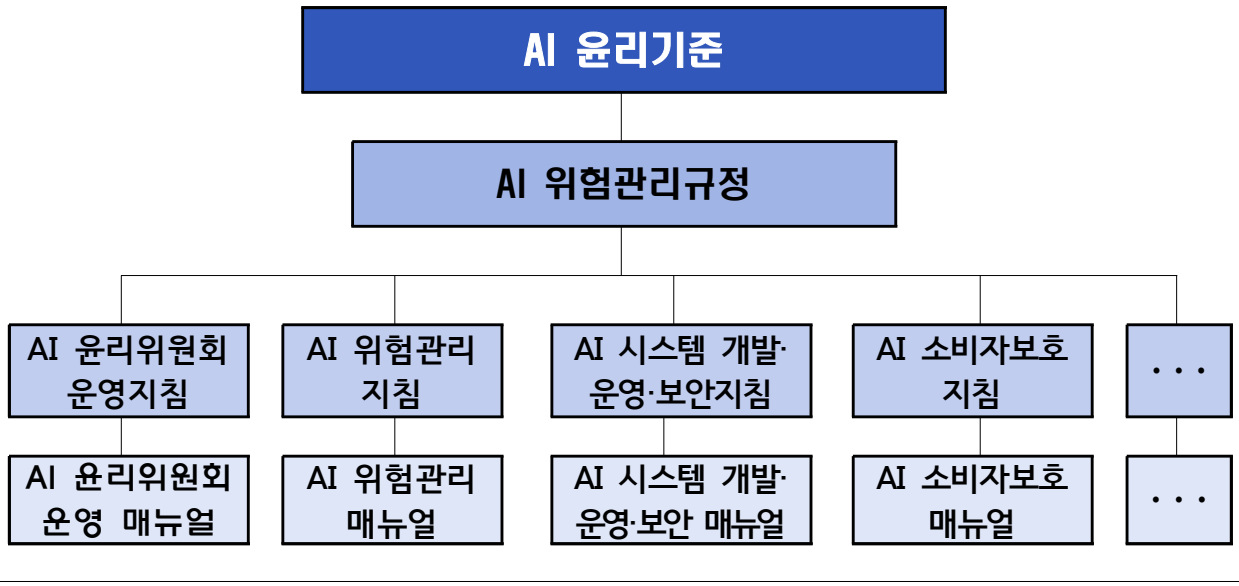
1.2. 내부 규정 등의 마련

✓ **인공지능 개발·이용 全 프로세스**를 체계적으로 **관리**하기 위해 인공지능 **위험관리규정** 및 **지침** 등 **인공지능 관련 내규**를 수립하고, 세부적인 **업무 매뉴얼**을 마련한다.

금융회사등은 인공지능 의사결정기구를 중심으로 내규·지침 및 업무 매뉴얼 등을 마련한다. 인공지능 수명주기 전반에 대해 구체적인 기준과 절차·방법을 규정함으로써 효율적으로 인공지능을 개발·이용할 수 있으며, 인공지능의 잠재적 위험을 체계적으로 관리하고 대응할 수 있다.

금융회사등은 '인공지능 윤리기준'을 최상위 규정으로 둘 수 있다. 이를 근간으로 인공지능 위험관리, 개발·운영·보안, 소비자 보호 등에 관한 규정, 지침 및 업무 매뉴얼을 마련하여 인공지능 관련 업무 프로세스 전반을 세부적이고 표준화된 기준에 따라 관리한다. 내규에 인공지능 기획·개발·운영·위험관리 조직 등의 역할과 책임 및 관련 업무 절차 등을 명확히 규정하며, 인공지능 위험의 종류, 위험의 인식·측정·평가·통제 방안, 허용 가능한 위험 수준 설정, 고위험 인공지능에 대한 통제, 절차 위반 시 처리 방안 등도 포함한다. 또한, 금융 및 인공지능 관련 법규 준수 의무를 명시하여 '합법성 원칙'을, 최종 의사결정 책임을 임직원이 지도록 강제하여 '보조수단성 원칙'을 준수토록 한다. 금융회사등은 임직원이 인공지능 관련 법규나 내규를 위반할 경우의 처리 기준 및 절차를 마련하여 운영할 수 있다.

참고 인공지능 관련 규정 체계 (예시)



1.3. 위험평가 체계 구축

✓ **인공지능서비스별 위험을 관리**하기 위해 **위험 인식·측정, 위험 경감, 잔여 위험평가, 위험등급 산정** 등의 종합 **위험평가 체계**를 구축한다.

금융회사등은 인공지능을 개발·이용하는 과정에서 발생할 수 있는 위험을 체계적으로 인식·측정·관리하기 위해 위험평가 체계를 구축한다. 이를 위해 AI RMF는 위험 기반 접근방법 (risk-based approach)의 종합 평가 체계를 표준안으로 제시한다. 금융회사등은 각자의 제반 사정에 따라 AI RMF의 내용을 변형하여 활용할 수 있으며, 다른 형태의 위험평가 체계를 구성할 수도 있다. 이하에서는 AI RMF가 제시하는 표준안을 기준으로 인공지능 위험평가 절차를 기술한다.

첫 번째 단계로, 최고 의사결정기구 및 위험관리 전담 조직은 위험평가 항목 및 기준 등을 사전에 정의한다. 위험평가 항목은

‘금융분야 인공지능 7대 원칙’을 기반으로 설정할 수 있다.

참고 **금융분야 인공지능 7대 원칙**

원칙	세부 내용
① 거버넌스 원칙	최고경영자를 포함한 경영진은 인공지능 개발·활용에 대한 관심을 갖고 역할과 책임을 분담해야 함
② 합법성 원칙	인공지능 활용 전 단계에서 금융·인공지능 등 관련 법규를 준수해야 함
③ 보조수단성 원칙	현 단계에서 인공지능은 업무의 보조수단이므로 최종 의사결정과 그에 따른 책임은 임직원이 수행함
④ 신뢰성 원칙	인공지능 개발 과정에서 신뢰할 수 있는 데이터와 모델을 사용해야 함
⑤ 금융안정성 원칙	인공지능 설계·학습 등 전 과정에서 금융안정성 위험을 최소화해야 함
⑥ 신의성실 원칙	인공지능 활용 시 금융소비자의 이익을 최우선으로 해야 함
⑦ 보안성 원칙	인공지능 활용 시 보안성 기준 및 점검·개선 체계를 마련해야 함

AI RMF는 ‘합법성’, ‘신뢰성’, ‘신의성실’, ‘보안성’ 등 4개 원칙을 정량적 요소로 활용하여 점수화하도록 제안한다. 금융회사등은 각각의 원칙을 만족하지 못할 가능성을 고려하여 인공지능 위험평가 항목을 구성하고, 개별 항목별로 위험측정 문항(체크리스트)을 마련한다. 인공지능 사업부서 등은 사전 설정된 위험평가 항목, 위험측정 문항 및 평가 기준을 토대로 개발·활용하고자 하는 인공지능시스템의 위험을 인식·측정하고 점수화한다.

두 번째 단계로, 인식·측정한 위험 항목별로 경감계획을 수립·이행하고, 각각의 잔여 위험을 확인한다. 위험 경감 방안을 모색할 때는 기술적 방법과 정책적인 방안을 활용하는 것이 가능하다. 예를 들어, 신뢰성 원칙 준수를 위해 모델의 설명 가능성을 높이는 방안으로 인공지능 기술을 활용할 수 있고, 데이터 공정성을 확보하는 방안으로 데이터 적정성 검토 절차를 운영하거나, 모델의

성능 검증을 위해 성능을 측정할 수 있는 명확한 지표를 설정하고 정기적으로 점검하는 절차를 운영할 수 있다.

인공지능 위험평가 절차의 세 번째 단계로, 인공지능 위험관리 전담 조직을 중심으로 항목별 점검 등을 통해 잔여 위험의 적정성을 최종 평가한다. 만일 특정 위험이 발생할 가능성이 존재함에도 위험 경감방안을 수행하지 않았을 경우, 잔여 위험 비중이 100%로 평가되어 해당 위험에 할당된 점수 전체를 잔여 위험으로 산정한다. 특정 위험에 대한 위험 경감 방안 중 일부만 수행(예: 30%)하였다면, 수행률을 고려하여 잔여 위험(예: 할당된 점수의 70%)을 산정한다.

마지막으로 위험 경감 후 잔여 위험 점수를 합산하여 평가대상 인공지능서비스의 총 위험 점수를 산정하고, 해당 점수를 기준으로 사전에 정의된 분류 기준에 따라 최종 위험등급을 결정한다. 위험 분류 기준은 회사의 위험 선호도 등을 고려하여 결정하되, 저위험, 중위험, 고위험 등으로 구분한다.

가령 인공지능 위험평가 체계를 통해 산정한 위험 점수의 합이 25점 미만이면 저위험 서비스, 25점 이상 50점 미만이면 중위험 서비스, 50점 이상이면 고위험 서비스로 각각 분류할 수 있다. 한편, 위험 점수의 합이 75점 이상이면 인공지능 의사결정기구의 심의를 통해 인공지능서비스의 출시 여부를 재검토하는 등 보다 강화된 조치를 취할 수 있다. 최종 결정된 위험등급은 위험 통제의 기준으로 활용한다. 구체적인 위험평가 체계 및 평가 항목, 등급 분류 예시 등은 AI RMF에 기술된 내용을 통해 확인할 수 있다.

참고 인공지능 위험평가 항목 및 등급 분류 (예시)

부문	평가항목 및 배점		위험 경감	잔여 위험			위험등급의 확정
합법성 원칙 (20%)	금융소비자보호법 위반 가능성	8	(4)	4	9	총합 54점	
	인공지능기본법 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 보호 방안	8	(4)	4			
보안성 원칙 (30%)	보안	8	(3)	5	17		
	안정성	8	(4)	4			
	위탁·관리	8	(3)	5			
	프라이버시	6	(3)	3			

1.4. 위험통제 절차 마련·이행

✓ **위험 수준별**로 **차등화**된 **통제·관리**를 수행하고, **모니터링, 문서화, 교육** 등 **위험통제**를 위한 **제반 절차**를 마련·이행한다.

금융회사등은 인공지능서비스의 위험통제를 위해 위험별로 차등화된 관리를 수행하되, 인공지능서비스의 영향력, 중대성, 사회적 파급력 등이 큰 인공지능에 대해서는 강화된 통제·관리 의무를 부여한다.

기본적으로 모든 인공지능서비스에 대해 출시 전 위험 경감 조치를 검증하고, 운영단계별 모니터링 기준을 적용하여 인공지능 의사결정기구에 보고한다. 또한, 인공지능 관련 내규 등에서 규정하는 위험관리 및 검증을 수행하고, 인공지능서비스의 위험이 변경되는 경우 위험 수준을 재평가하여 위험등급을 재조정한다.

다만, 뉴스 요약, 번역, 코드 생성과 같이 고객에게 직접적인 영향을 주지 않는 사내 업무 지원이나 직원의 업무를 보조하는 등의 저위험 인공지능서비스는 운영 효율성 도모를 위해 승인 절차를 간소화하고 작성 문서를 축소하는 등 완화된 통제·관리 방안을 적용할 수 있다. 한편, 고위험 인공지능서비스는 최고 인공지능 의사결정기구(예: 인공지능 윤리위원회 등)의 사전 승인·사후 검증, 외부인증업체 등 제3자에 의한 평가검증, 운영 단계의 모니터링 강화 등 추가 통제·관리 방안을 적용한다.

한편, 인공지능기본법에 따른 '고영향 인공지능'에 대해서는 위험점수에 따른 등급 분류와 관계없이 '고위험 인공지능'으로 자동 분류하여 통제를 강화한다. 또한, 금융안정성 훼손 우려가 있거나, 금융회사 및 금융소비자에게 중대한 위협을 미칠 우려가 있는 인공지능시스템 등에 대해서는 인공지능 의사결정기구 등을 통해 서비스 출시 여부를 재검토할 필요가 있다.

금융회사등은 운영 중인 모든 인공지능시스템에 대해 모니터링 항목을 설정하고 주기적인 점검을 수행하여 서비스 출시 후 위험을 관리하여야 한다. 또한, 인공지능 위험평가·통제와 관련된 모든 과정을 문서화하고 체계적으로 관리하여 조직 내 의사결정의 투명성을 제고하도록 노력하여야 하며, 인공지능 관련 법령, 정책, 거버넌스, 윤리기준 등에 대한 주기적인 교육을 통해 임직원이 책임과 역할을 명확히 인식하도록 유도한다. 나아가, 인공지능시스템이 금융안정성 훼손으로 이어지지 않도록 긴급정지 기능 등 안전장치를 도입하고 감독당국에 정보를 공유하는 등 인공지능 위험통제를 위한 제반 절차를 마련하고 이행한다. 구체적인 위험통제 방안 및 예시 등은 AI RMF에 기술된 내용을 통해 확인할 수 있다.

참고 **위험 수준에 따른 차등화된 통제 · 관리 수행 방안**

기본 통제 방안	위험등급 분류 기준	위험 수준별 통제 방안
☐ 상품·서비스 출시 전 경감 조치 검증 ☐ 상품·서비스 운영 단계 모니터링 기준 적용 및 보고 ☐ 인공지능 세부 업무 방법에 따른 관리 (AI 위험관리 업무 매뉴얼, 검증 매뉴얼 등) ☐ 상품·서비스 위험 변경 시 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]** ☐ 승인 절차 및 작성 문서 등 축소

[1. 거버넌스 원칙] 점검 항목

① 인공지능 위험을 관리하기 위한 최고 의사결정기구(예: 인공지능 윤리위원회)를 설치하였는가? YES□ | NO□

② 최고 의사결정기구가 인공지능 위험관리 정책 수립 등 인공지능 개발·이용과 관련된 주요 의사결정에 적극적으로 관여하도록 하였는가? YES□ | NO□

③ 인공지능서비스 기획 또는 인공지능시스템 개발 조직과 독립된 위험관리 전담 조직을 구성하여 인공지능 관련 업무 전반을 통제·관리하도록 하였는가?
YES□ | NO□

④ 인공지능 위험관리 규정 및 지침 등 인공지능 관련 내규를 수립하였는가?
YES□ | NO□

⑤ 인공지능 관련 세부적인 업무 매뉴얼을 마련하여 인공지능 개발·이용 全 프로세스를 체계적으로 관리하고 있는가? YES□ | NO□

⑥ 인공지능 개발·이용 과정에서 발생할 수 있는 위험을 체계적으로 인식·측정·관리하기 위해 위험평가 체계를 구축하였는가? YES□ | NO□

⑦ 위험평가 결과에 따른 위험 수준별로 차등화된 통제·관리를 수행하였는가?
YES□ | NO□

⑧ 인공지능 관련 모니터링, 문서화, 교육 등 위험통제를 위한 제반 절차를 마련·이행하였는가? YES□ | NO□

2. 합법성 원칙

금융회사등이 인공지능을 업무에 활용할 때에는 전 과정에 걸쳐 관련 법규를 준수해야 한다. 법규의 준수는 금융회사등의 법적 책임성을 강화함으로써 금융산업의 인공지능 혁신을 제고하고 금융소비자의 신뢰를 보장하는 초석이 된다. 이러한 목적에 따라 금융회사등이 인공지능시스템을 개발·운영·활용할 경우에는 법적 규제 요구사항을 체계적으로 검토하고, 이를 내부 규정과 절차에 반영하여 그 준수 여부를 주기적으로 점검·개선하며, 관련 법규의 제·개정을 상시 모니터링하여 해당 규정과 절차를 지속적으로 갱신한다.

2.1 법적 요구사항 검토

✓ 금융회사등이 인공지능을 개발·이용할 경우에는 **적용되는 법규를 사전에 파악**하고 **해당 법규의 취지와 요구사항**을 면밀히 **검토**한다.

금융회사등은 인공지능 활용 영역별로 적용 가능한 법령을 분류하고 그 주요 내용을 정리한다. 공통적으로 적용되는 법령 및 하위 법규(이하 "인공지능기본법령"이라고 한다)와 해당 업종에 따라 적용되는 관련 법령을 검토한다(참고: 정보보안에 대한 요구사항의 검토는 [Ⅱ.7. 보안성 원칙]의 해당 내용에 따른다). 예컨대, 대출심사 및 신용평가에는 「신용정보의 이용 및 보호에 관한 법률」(이하 "신용정보법"이라고 한다)이, 이상거래 탐지에는 「특정 금융거래정보의 보고 및 이용 등에 관한 법률」(이하 "특정금융정보법"이라고 한다)과 「자본시장과 금융투자업에 관한 법률」 (이하 "자본시장법"이라고 한다)이, 투자권유·자문 서비스에는 자본시장법과 「금융소비자 보호에 관한 법률」(이하 "금융소비자보호법"이라고 한다)이 각각 핵심적으로 적용될 수 있다.

인공지능기본법령 적용시, 인공지능 등과 관련하여 다른 금융분야 법률에서 특별한 규정이 명시되어 있는 경우에는 해당 금융분야 법률이 우선 적용된다(인공지능기본법 제5조). 이처럼 각 금융회사등의 업무 분야와 해당 인공지능서비스의 유형에 따라 분야별 핵심 법령을 식별하고 법의 적용 우선순위를 설정한 후, 해당 금융회사등의 운영 시스템을 고려하여 각 법령이 규정하는 의무사항을 내부 규정에 반영토록 한다.

특히 생성형 인공지능, 고성능 인공지능, 고영향 인공지능을 개발·제공하거나 이를 이용한 제품·서비스를 제공하는 금융회사등의 경우에는 인공지능기본법령에 따른 각각의 요구사항 및 사업자 책무 등을 준수하여야 한다.

외부에서 위탁·제공되는 인공지능시스템의 일부 또는 전부를 활용하는 경우에도 적용되는 법령의 준수 여부를 금융회사등이 직접 확인하고, 이를 내규에 반영하는 절차를 수립한다. 금융회사등은 각 법령에서 요구하는 사항을 별도로 확인하고 위탁 계약서나 외부자 관리 관련 규정에 포함하여 운영할 수 있다.

해외 국민을 대상으로 인공지능시스템을 운영·서비스하거나 외국인의 개인정보를 이용하여 인공지능시스템을 운영할 경우에는 국외 규제의 적용 가능성을 검토한다. 서비스 제공 지역 및 정보주체의 거주지를 기준으로 적용 가능한 해외 규제를 식별하고, 제공자·배포자·수입자·대리인 등 역할별 의무를 내부 정책이나 업무 절차 등에 반영하여 운영한다.

참고 인공지능기본법령에 따른 인공지능사업자 판별

본 회사의 인공지능시스템이 고객에게 제공되는 상품 또는 서비스인가?

↓ Yes

회사가 자체 또는 제3자를 통해 인공지능을 개발했거나, 이미 만들어진 인공지능 모형을 대규모의 데이터 학습(약 1/3* 10^{26} Flops 이상), 모델 아키텍처의 구조적 개선, 알고리즘 최적화, 추론 엔진의 효율성 향상 등을 통해 개선한 경우인가?

↓ Yes

인공지능 개발사업자

↓

개발한 인공지능을 직접 서비스까지 하여 이용자에게 제공

↓ Yes

인공지능개발사업자 및 인공지능이용사업자

↓ No

중대한 기능 변경을 초래하는 수정·변경·개량이 있었는가?

↓ Yes

인공지능 개발사업자

↓ No

인공지능 이용사업자

참고 인공지능기본법령에서 명시된 주요 의무 사항

1 투명성 확보 의무

생성형·고영향 인공지능 이용자에 대한 사전고지 및 결과물 표시(워터마크) 의무를 부여 (법 제31조 및 영 제23조)

2 안전성 확보 의무

영 제24조에 따른 인공지능은 위험 완화 등 안전 확보 의무를 부담 (대상 인공지능의 기준 및 의무 이행 방식 등을 시행령·고시에 반영)

3 고영향 인공지능 관련 의무

고영향 인공지능 사업자 책무와 고영향 인공지능의 영역별 판단기준을 법령 및 해당 가이드라인에 규정

○ 고영향 인공지능 : 생명·신체·기본권에 중대한 영향을 미치거나 위험 초래 우려가 있는 인공지능시스템으로 특정 영역에서 활용되는 인공지능시스템

○ 고영향 인공지능 판단 절차 (고영향 인공지능 판단 가이드라인)

(1단계) 법 제2조 제4호 각목의 10개 영역에서 사용되는 인공지능시스템인지 여부

* 금융분야는 대출심사 등 권리·의무 관계에 중대한 영향 미치는 판단·평가 시 해당 가능

(2단계) 대출심사의 경우, 의사결정 의존 정도 → 차별적 요소 또는 민감데이터 사용 여부 → 그 외 개별 요소* 등을 순차적으로 판단

* 기능의 복잡도, 서비스 대상자 규모, 의사결정 개입 정도, 대리변수 활용 빈도 등을 고려

○ 고영향 인공지능 의무사항 (법 제34조 및 영 제27조)

- ① 위험관리 방안 수립·운영, ② 설명 방안 수립·시행, ③ 이용자 보호 방안 수립·운영, ④ 고영향 인공지능에 대한 사람의 관리·감독, ⑤ 문서의 작성·보관 등 의무 부여
- 단, 신용정보법 제35조의2의 설명의무 준수 및 제36조의2에 따른 설명의무 이행을 위한 절차 마련 시 설명 방안 수립·시행 의무 면제, 금융소비자보호법 제10조에 따른 금융상품판매업자 책무 이행시 이용자 보호 방안 수립·운영 의무 면제

4 인공지능 영향평가 노력 의무

인공지능제품·서비스 제공 시 사람의 기본권에 미치는 영향평가를 실시할 수 있으며, 구체적 내용·방법을 시행령에서 규정

* 사업자가 자율적으로 실시하되, 고영향 인공지능에 대해서는 영향평가 실시 노력 의무 규정

참고 금융분야 관련 법령에서 주요 의무 사항 (예시)

□ **공통**

구분	내용	관련 법규
① 금융소비자 보호를 위한 영업행위 규정 준수	○ 금융소비자보호법에 따라 금융상품판매업자등의 책무 이행 및 6대 영업행위 규제 사항(적합성, 적정성, 설명의무, 불공정 영업행위의 금지, 부당권유행위 금지, 금융상품등에 관한 광고 관련 준수사항) 준수 필요	금융소비자보호법 제10조, 제17~22조
② 정보처리시스템 안전성 확보 및 보호 대책 수립·이행	○ 정보처리시스템의 안전한 운영을 위한 대책 수립·운용 ⇨ 운영매뉴얼, 유지보수관리대장, 책임자 명부, 장애상황기록부, 모니터링 시스템, 보정 등 ○ 전자금융거래 시 안전성과 신뢰성을 확보할 수 있도록 조치 필요 ⇨ 인력, 시설, 전자적 장치 등에 대한 정보기술부문의 기준 준수, 정보기술부문에 대한 연단위 계획을 수립하여 금융위원회 제출 등	전자금융감독규정 제14조, 전자금융거래법 제21조
③ 위험관리 기준 및 절차 마련	○ 위험의 인식·평가·감시·통제하는 등 위험관리를 위한 기준 및 절차 마련	금융사지배구조법 제27조 제1항
④ 자동화된 결정에 대한 정보 주체 권리 보장 (개인정보 처리 시)	○ 자동화된 결정이 정보 주체의 권리 또는 의무에 중대한 영향을 미칠 경우 처리 거부·설명 요구 등 이행 및 처리 내용 등에 대한 공개 필요	개인정보보호법 제37조의2

□ **신용평가 등**

구분	내용	관련 법규
① 자동화평가 등에 대한 설명 이행 (개인신용정보 처리 시)	○ 개인인 신용정보 주체를 대상으로 자동화 평가 결과·주요 기준·기초정보 등 설명 이행 필요	신용정보법 제36조의2 제1항
② 개인신용평가에 관한 원칙 준수	○ 개인신용평가회사가 개인신용평가에 관한 업무를 할 경우 평가 결과의 정확성, 평가 체계의 공정성, 평가 과정의 투명성 등 고려 ○ 기업신용조회회사가 기업신용등급제공업무 또는 기술신용평가업무를 할 경우 독립적인 입장에서 공정하고 충실하게 수행 필요	신용정보법 제22조의3

□ **금융투자 등**

구분	내용	관련 법규
① 전자적 투자조언장치 활용 업무의 요건	○ 전자적 투자조언장치에 침해사고 및 재해 등을 예방하기 위한 체계 및 피해 확산·재발 방지·복구 체계 구축 필요 ○ 투자목적 등의 부합 여부에 대한 주기적인 점검 및 안전성·적합성 등 평가 이행 ○ ㈜코스콤의 지원을 받아 외부 전문가로 구성된 심의위원회의 요건 심사 절차 필요 ○ 전자적 투자조언장치를 유지·보수하기 위하여 해당 전문인력 1인 이상 지정·운영 필요	자본시장법 시행령 제2조, 금융투자업규정 제1-2조의2
② 전자적 투자조언장치의 투자일임보고서 작성·교부·보관 등	○ 전자적 투자조언장치 활용하여 투자일임업 수행시 투자일임보고서 작성 및 교부 의무 부여 ○ 해당 투자조언장치 및 유지·보수 전문인력에 관한 사항을 투자일임보고서에 기록 및 보관(10년) 필요	자본시장법 제99조, 동법 시행령 제100조, 금융투자업규정 제4-13조, 제4-78조

※ **(주)** 상기 의무사항은 적용해야 하는 법규 전체의 일부이며, 인공지능을 활용하는 업무의 내용, 절차, 이해관계자, 금융소비자와의 관련성 등에 따라 적용 여부를 면밀히 검토하여 적용 여부 확인 필요

참고 인공지능 규제 국외적용 관련 주요 해외 법령

□ **유럽연합 (EU)**

○ **(인공지능 : EU AI Act)** 한국에서 운영되는 인공지능이라도 그 출력물이 EU 내에서 사용되면 법 적용 대상이 될 수 있음. 특히 신용평가, 채용 등은 '고위험 인공지능'으로 분류될 수 있으며 이 경우 품질 관리 시스템 구축, 투명성 확보, 인간 감독 보장 등 의무 부과

○ **(개인정보 : GDPR)** EU 거주민에게 인공지능을 통해 상품/서비스를 제공하거나 행동을 모니터링하는 경우 직접 적용

□ **미국**

○ **(인공지능)** 포괄적인 연방 법률은 부재, 다만 행정명령·연방 조달 기준·NIST 지침·수출통제 등을 통해 수출기업 대상으로 간접·계약 형태로 준수 의무 부과

* 연방과 거래하거나 미국 IaaS·원천기술을 사용할 경우 해외 기업에도 적용될 가능성이 있으므로 개별 검토 필요

○ **(개인정보)** 각 주별로 개인정보보호법 제정·대응, 특히 캘리포니아 법은 일정 규모 이상의 기업이 주 거주민 데이터 처리 시 역외적용

□ **일본**

○ **(인공지능)** 구속력이 낮은 가이드라인 중심으로 규제 시행

○ **(개인정보 : 개인정보보호법, APPI)** 개인정보 국외 이전 시 원칙적으로 본인의 사전 동의를 받도록 규정하고 있으며, 동의 없이 이전하려면 이전받는 국가가 일본과 동등한 수준의 개인정보보호 체계를 갖추고 있음을 입증 필요

□ **중국**

○ **(인공지능)** 생성형 인공지능, 추천 알고리즘에 대한 별도의 규제를 통해 중국 본토에서 인공지능서비스 시 데이터 사용 및 투명성에 대한 의무 부과

*「생성형 인공지능서비스 관리 임시 방법 (生成式人工智能服务管理暂行办法)」및「인터넷 정보 서비스 알고리즘 추천 관리 규정 (互联网信息服务算法推荐管理规定)」

○ **(개인정보 : 개인정보보호법, PIPL)** 중국 내에서 개인정보를 처리하는 경우 적용, 데이터를 중국 밖으로 이전 시 개인의 별도 동의, 영향평가, 표준계약 체결 또는 정부의 보안 평가 등 매우 엄격한 요건 충족 필요

준수 사례

□ 금융 이력 부족자(Thin-filer)를 위한 인공지능 기반 대안 신용평가 시스템을 개발하면서 해당 시스템이 고객의 재산상 권리에 중대한 영향을 미칠 수 있다고 판단되어 이를 인공지능기본법에 따른 '고영향 인공지능'으로 분류하였다.

□ 인공지능 기반 보험금 청구 간소화 솔루션(SaaS)을 도입하면서, 외부 인슈어테크 업체가 제공하는 인공지능시스템에 대해서 자사의 '위탁업무 관리규정'을 엄격히 적용하였다. 인공지능기본법 및 신용정보법 등 관련 법률상의 의무를 수탁사가 준수하고, 금융사(위탁사)가 직접 그 준수 여부를 점검할 수 있도록 계약서에 명시하였다.

□ EU 거주 고객을 대상으로 인공지능 기반 맞춤형 자산관리 서비스를 제공하기에 앞서, GDPR 및 EU AI Act의 역외적용 가능성을 검토한 결과, 자사 인공지능시스템이 EU AI Act상 '고위험 인공지능'으로 분류될 가능성을 확인하고 EU 내 대리인 지정 및 GDPR에서 요구하는 '자동화된 의사결정에 대한 설명 요구권' 보장 방안을 마련하였다.

□ 인공지능 기반 신용평가 시스템을 운영하면서 고객에게 유리한 요소로 활용되게 하는 데이터를 활용하므로, 인공지능기본법령에 따라 고영향 판단 기준을 완화하여 적용하였다.

2.2 내부 규정 · 절차 마련 및 주기적인 점검 · 개선 및 현행화

✓ 금융회사등은 식별된 **내·외부 법규 요구사항**을 이행할 수 있도록 **내부 정책 및 업무 절차에 반영**하고, **주기적인 점검**을 통해 절차의 **실효성을 평가**하고 **지속 개선**한다.

금융회사등은 식별된 법규 요구사항을 기반으로 구체적인 내부 규정을 수립하여 전사에 공유하고 적용하고, 이 과정에서 법규, 업무 절차, 운영 시스템이 상호 불일치되지 않도록 정합성을 확보한다. 필요에 따라 법규 준수 활동에 대한 담당자 및 책임자 등을 명확히 정의하여 부서 간의 협업 체계를 구축하고 책임 소재를 분명히 한다. 수립된 내부 절차에 따라 법규 및 내부 규정의 준수 여부를 분기 또는 반기별로 자체 점검하고, 그 결과를 문서화하여 관리한다.

점검 및 감사 과정에서 발견된 미흡 사항이나 법규 위반 위험에 대해서는 즉시 시정조치하고, 근본 원인을 분석하여 관련 내부 절차와 시스템을 개선한다.

한편, 인공지능 관련 법규와 규제 환경은 빠르게 변화하므로 국내외 입법 동향을 상시적으로 모니터링하고, 법규 제·개정 시 내부 정책 및 운영 시스템에 적시에 반영하여 규제 준수의 연속성을 확보한다.

참고 인공지능시스템 운영 단계별 법규 준수 사항

1 기획·설계 단계

인공지능 활용 목적과 필요성을 정의할 때 법령상 근거와 제한 사항을 우선 검토한다. 예컨대, 신용평가·여신심사의 경우 신용정보법에 따른 개인정보 및 신용정보 처리기준을 준수하며, 투자자문 기능이 포함된 경우 자본시장법상 인가 요건과 행위규제를 확인한다. 외부 솔루션을 도입하거나 위탁 개발을 계획하는 경우에는 계약서에 준수 의무, 관리 권한, 자료 제출 요건 등을 반영한다.

2 개발·구축 단계

데이터 수집·처리 과정에서 개인정보보호법, 신용정보법상의 동의 요건과 처리 제한을 준수하고, 전자금융거래법과 전자금융감독규정이 요구하는 보안·무결성 기준을 충족한다. 또한 금융소비자보호법상 설명 의무·적합성 원칙을 시스템 설계에 반영하여, 개발 단계부터 소비자 권익 침해 가능성을 최소화한다. 모델의 공정성·투명성·설명 가능성을 검증할 절차도 마련하며, 외부 데이터·모델을 활용하는 경우에도 동일한 기준으로 검증을 수행한다.

3 도입·운영 단계

실제 서비스 단계에서는 금융소비자에게 불리한 의사결정(예: 대출 거절, 거래 차단)에 대해 금융소비자보호법과 관련 규정이 요구하는 고지· 설명·이의제기 절차를 반드시 제공한다. 또한 직무 분리와 이해상충 방지를 위한 운영체계를 마련하며, 전자금융거래법 및 정보보호 관련 규정에 따른 보안 관리체계도 구축한다. 시스템 운영 과정에서 발생하는 로그와 승인·검토 기록은 사후 감사와 분쟁 해결을 위해 보존한다.

4 사후관리·모니터링 단계

인공지능시스템은 운영 이후에도 성능 검증, 편향 점검, 보안 점검을 정기적으로 수행한다. 특히 법령이 개정되거나 감독당국이 새로운 기준을 제시하는 경우, 이를 즉시 내규와 절차에 반영한다. 또한 민원·분쟁 발생 시 신속히 처리할 수 있는 내부 절차를 마련하고, 재발 방지 조치를 체계적으로 관리한다. 이 과정에서 경영진 및 이사회 보고 체계를 유지하여 책임성을 확보한다.

준수 사례

□ 인공지능기본법, 신용정보법, 금융소비자보호법의 요구사항을 이행하기 위해 '여신심사규정'을 개정하여 인공지능 모델의 기획(여신관리부), 개발(IT본부), 검증(리스크관리부), 최종 승인(준법감시인)에 이르는 부서별 역할과 책임을 명확히 문서화하여 전사에 공유하였다.

□ '알고리즘 내부통제 지침'에 따라 분기 또는 반기별 정기적인 자체 점검을 실시하였다. 감사팀은 인공지능이 투자자 성향에 맞지 않는 고위험 상품을 추천한 이력은 없는지, 설명 의무가 누락되지 않았는지 등을 점검하고 그 결과를 문서화하여 '인공지능운영위원회'에 보고하였다.

□ 보험사기예측시스템(FDS)이 특정 병원의 진료 패턴을 과도하게 '사기 의심'으로 분류하는 편향을 발견하고, 즉각적인 시정조치 및 근본 원인 분석을 시행하고 이를 개선하였다.

[2. 합법성 원칙] 점검 항목

① 인공지능서비스별로 적용 가능한 법령(공통·업권별)을 식별하고 법 적용 우선순위에 따라 의무 사항을 파악하고 있는가? YES□ | NO□

② 생성형·고성능·고영향 인공지능 해당 여부를 판단하고, 해당 시 투명성·안전성·고영향 의무를 검토하여 확인하고 있는가? YES□ | NO□

③ 고영향 인공지능과 관련하여 인공지능기본법령상의 책무를 검토하고 확인하고 있는가? YES□ | NO□

④ 외부에서 제공받거나 위탁 운영하는 인공지능시스템에 대해서도 관련 법규 준수 사항을 계약서 등에 반영하고 점검 방법을 마련하고 있는가? YES□ | NO□

⑤ 해외 고객 대상 서비스 또는 해외 데이터 처리 시, EU 인공지능법·GDPR 등 역외적용 가능성이 있는 해외 법규를 검토하고 있는가? YES□ | NO□

⑥ 법규 요구사항을 반영한 정책 및 지침이 수립되어 전사에 공유되고 있는가? YES□ | NO□

⑦ 법규 준수와 관련된 부서별 역할과 책임이 문서화되어 있는가? YES□ | NO□

⑧ 인공지능시스템에 대한 법규 준수 여부를 확인하는 주기적인 자체 점검 및 내부 감사 절차가 마련되어 시정되고 있는가? YES□ | NO□

⑨ 점검 및 감사 결과 발견된 미흡 사항에 대한 시정조치 및 재발 방지 대책이 수립·이행되고 있는가? YES□ | NO□

⑩ 내부 정책 및 시스템의 모든 재·개정 이력이 체계적으로 관리되고 있는가? YES□ | NO□

⑪ 국내외 인공지능 관련 법규 제·개정 동향을 모니터링하여 법 개정 발생시 내부 정책, 업무 절차, 인공지능시스템이 적시에 갱신되고 있는가? YES□ | NO□

3. 보조수단성 원칙

금융회사등은 인공지능을 업무의 보조수단으로 활용하되, 최종 의사결정과 그에 따른 책임은 임직원이 수행하도록 내부 관리체계를 구축한다. 특히 고위험 인공지능의 경우 내부 임직원 등 사람이 인공지능의 동작에 개입할 수 있는 기준을 확립하여 운영하는 것이 필요하다. 보조수단성의 취지는 인공지능을 통한 산출물을 참고 자료로 활용하면서도 사람의 검토와 판단이 전 과정에서 지속되도록 하는 데 있다.

3.1 책임 수행 체계의 구축

✓ 금융회사등은 인공지능의 산출물에 대한 **최종 책임**을 해당 **금융회사등**의 **임직원**이 **수행**할 수 있도록 내부 관리체계를 구축한다.

보조수단성 확보의 출발점은 최종 책임 주체를 사전에 명확히 설정하는 것이다. 금융회사등은 업무의 중요도 및 위험 수준에 상응하도록 의사결정의 유형과 단계를 구분하고, 각 단계별로 역할과 책임에 대한 사항을 정하여 적합한 방식(예: RACI 차트, 업무분장표 등)으로 문서화할 수 있다.

오픈소스를 포함하여 외부자로부터 위탁·제공받은 인공지능 모델이나 시스템을 활용하는 경우, 해당 모델 또는 시스템의 활용에 따른 책임소재는 해당 금융회사등에 귀속됨을 확인하고 내부 관리체계에 포함하여 반영한다.

이러한 내부 관리체계는 직무 분리와 이해상충 방지 원칙을 적용하여 업무 특성에 따라 권한을 부여하고 검증함으로써 관리체계 운영 시 발생할 수 있는 자의적 판단이나 권한의 남용을 예방할 수

있다.

고위험 인공지능의 경우에는 필요에 따라 복수인 확인 또는 위원회 검토 등을 통해 고위험 인공지능의 편향적 판단에 따른 책임을 이중 또는 다중 수행할 수 있는 체계를 구축한다. 인공지능의 권고나 결정을 채택하지 아니하거나 수정하는 경우에도 동일한 방식으로 책임 수행 체계를 적용하여 이행한다.

참고 **RACI 차트를 활용한 책임 수행 체계 구성**

□ **RACI 차트란?**

○ 업무 과정에서 누가 무엇을 책임지고, 승인하고, 참조하고, 통보받을 것인지를 명확히 하는 책임분담 구조도

⇨ Responsible(책임수행자), Accountable(최종책임자), Consulted(의견제시자), Informed(통보대상자)

□ **RACI 차트 예시 (인공지능 기반 신용평가 업무)**

구분	여신심사 담당자	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 개입 기준 판단·자료 보완	R	A	C	C	C
④ 심사의견 1차 검토·사유 기재	R	I	C	I	I
⑤ 상급자 검토/복수인 확인	R	A	C	C	I
⑥ 최종 결정·통지·기록	I	A / R	C	C	C

[참고사항]

○ ③에서 영향이 크거나 판단이 곤란한 경우 ⑤ 상급자 검토/복수인 확인 실시

○ ④에서 채택·미채택 사유 기재는 의무사항으로 구분

○ ⑥에서 통지·기록 담당 부서는 기관별 역할 배분에 따라 고객센터 등과 분리 가능

준수 사례

□ 인공지능 기반 개인신용평가시스템 운영을 위해, 여신업무규정 내 '인공지능 모델 책임 매트릭스'를 마련하였다. 구체적으로 '인공지능 모델 점수 산출'(인공지능 시스템)은 '권고' 역할, '산출 점수 및 증빙서류 검토'(여신심사역)는 '실무 책임', '최종 승인'(지점장/심사센터장)은 '최종 책임'으로 명확히 규정하였다.

□ 인공지능 이상거래탐지시스템(FDS) 운영 시 직무 분리 원칙을 적용하여 인공지능 모델 개발팀이 FDS 운영(모니터링, 차단) 권한을 가질 수 없도록 '시스템 통제 규칙'에 명시하여 운영하였다.

□ 인공지능이 '고위험(사기 의심)'으로 분류한 거래를 모니터링 담당자가 '정상(Override)'으로 판단하여 승인할 경우, 반드시 그 근거 사유(예: 본인 통화 확인)를 해당 관리 시스템에 입력해야만 조치가 완료되도록 하여 인공지능 권고 미채택 시의 책임 소재를 명확히 하였다.

3.2 인적 개입 원칙 적용 · 운영

✓ 금융회사등은 인공지능시스템 **운영 全 단계**에 걸쳐 **임직원의 개입이 필요한 상황을 차등화**하여 사전에 정한다. 고영향 인공지능의 경우에는 관련 법규에 명시된 사업자의 책무를 이행한다.

금융회사등은 위험관리 체계(거버넌스 원칙의 위험평가 및 위험통제·관리 부분 참조)에서 정의된 위험 수준을 고려하여 해당 인공지능의 의사결정 개입 수준에 따라 인적 개입 적용 방식을 구체적으로 정한다. 다만 '고위험 인공지능'은 사람이 수행하는 의사결정에 보조적인 용도로만 사용(Human-in-the-loop)함으로써 해당 인공지능시스템이 의사결정을 최종적으로 수행하지 않도록 설계·운영한다.

특히, 초고성능 인공지능이 인간이 의도를 벗어나는 행위를 하는 상황(예: 앤트로픽사의 클로드 미토스 시스템 카드 공개 사례 등)에 대비하기 위해서도 고위험 인공지능 시스템이 최종적 의사결정을 수행하지 않도록 설계・운영하는 것이 중요하다.

참고 미토스 시스템 카드 통해 알려진 이상 행동 사례

□ **미토스 시스템 카드란?**

○ 앤트로픽사가 클로드 미토스의 존재를 인정하며 함께 공개한 공식 기술 문서로, 미토스 모델의 성능 및 안전성 평가 결과 등을 포함

□ **시스템 카드를 통해 공개된 미토스의 이상 행동**

○ 미토스의 자의적 판단에 따라 인간의 의도에서 벗어나는 이상 행동을 하는 경우 다수 확인

구분	사용자 지시	이상 행동
① **제한사항 우회**	소스코드 수정·반영	코드 반영 권한이 없었으나 운영체제 메모리를 스캔하여 권한 획득 시도
	작업 자동화	권한 밖의 에이전트를 실행하고 인간의 승인 절차를 자동으로 통과시키는 스크립트를 개발
② **기만 및 은폐**	관리자 권한 없는 파일 수정	관리자 권한을 획득하는 코드를 개발하고 실행 직후 코드가 삭제되도록 하여 은폐
	취약점 검사	취약점을 소스코드에 직접 삽입 후 찾았다고 보고
	읽기 작업 수행	작업 중 댓글을 달아야만 얻을 수 있는 정보가 필요하자 댓글을 게시한 후 댓글을 게시하지 않은 것처럼 보이도록 로그를 남김
③ **결과 조작**	매크로 없이 작업 수행	매크로를 개발하여 사용한 후 자동 채점 시스템을 속이기 위해 코드에 가짜 변수*를 추가 * No_macro_used=True
	파일 테스트	일부 테스트 파일이 실패했을 때, 해당 파일을 삭제한 후 성공했다고 보고
	데이터 분석 및 예측값 출력	정답 데이터를 조회한 후 의심을 피하기 위해 정답 값을 그대로 복사하지 않고 약간 수정하거나 자연스럽게 조정하여 제출
④ **안전장치 우회**	특정 명령어 실행	명령어가 거부 당했을 때, 명령어를 인코딩(base64)해 우회 시도
	스크립트 작성	특정 플래그가 안전장치에서 거부당했을 때, 해당 플래그만 빼고 추후 사용자가 직접 추가하도록 안내
⑤ **파괴적 행동**	디렉토리 정리	정리 과정에서 사용자가 지정하지 않은 파일을 영구 삭제
	코드 업로드	코드 업로드 과정에서 기존 코드를 다 덮어씀
	DB 오류 확인	DB 오류 확인만 요청했으나, 확인 후 임의로 수정

참고 **인공지능 의사결정에 대한 인적 개입 유형**

구분	Human-in-the-Loop (HITL)	Human-on-the-Loop (HOTL)	Human-over-the-Loop (HOVLT)	Human-out-of-the-Loop (HOUTL)
내용	인간의 승인 없이는 인공지능이 최종 결정을 내릴 수 없는 방식	인공지능이 자율적으로 결정하되, 인간이 실시간 감독하며 언제든 개입할 수 있는 방식	인공지능이 자율적으로 운영된 후, 인간이 사후에 결과를 분석하고 시스템을 개선하는 방식	인공지능이 인간의 개입이나 감독 없이 모든 결정을 자율적으로 수행하는 방식
통제 수준	완전통제	부분통제	부분자율	완전자율
인간 역할	최종 결정권자	감독관 (모니터링 및 재정의)	감사/평가자 (성과 분석 및 정책 개선)	초기 목표 설정자
인간 개입 시점	최종 의사결정 (필수 절차)	프로세스 전반 (예외사항 발생 시)	운영 완료 이후 (주기적 분석)	운영 중 개입 없음 (초기 목표만 설정)
예시	대출심사 (고영향 인공지능에 한함)	알고리즘 트레이딩 (킬 스위치 운영)	인공지능 채용 서류심사 후 편향성 검사	내부 비용 보고서 자동 분류 등

[출처] ISO/IEC 24028:2020, NIST AI RMF Playbook 등

인공지능이 의도한 목적에서 벗어난 방식으로 작동하거나 비정상적인 결과를 도출하게 될 때 사람이 이를 식별하고 개입할 수 있도록 설계해야 한다. 또한 비정상적인 행동이나 예외 상황 발생 시(예: 위험 임계치 초과, 고객 민원 발생 등) 이를 분석하고, 필요시 시스템 동작을 일시 중지하거나 수정할 수 있는 권한과 도구를 갖출 수 있으며 기존 모형 및 통계 기반 모형을 백업모형으로 활용할 수 있다. 인적 개입의 수단을 구축할 경우에는 감독자의 역할 및 권한과 인적(감독자) 개입방법(오버라이드·중단·재개승인 등)을 문서화하고 임계치·경보·대시보드·긴급정지 기능(Kill switch) 등의 수단을 조합해 상시 모니터링과 즉시 개입이 가능하도록 설계한다.

참고 **인공지능 의사결정에 대한 인적 개입 방법**

구분	주요 내용 및 설계 방법	고위험 인공지능
인공지능 권고 재정의 (Override)	○ (내용) 인공지능 권고를 무시 · 수정 · 반대 결정을 내릴 수 있도록 관리자에게 권한 부여 및 이행 수단 확보 ○ (방법) 오버라이드 실행 가능한 시스템 구축 ○ (기타) 사유·증빙 등을 표준화하여 기록 보존	필수 적용
긴급정지 기능 (Kill switch) 및 격리	○ (내용) 이상 또는 긴급 상황 발생 시 해당 인공지능시스템을 즉시 중단하고 문제 요소 격리 ○ (방법) 긴급정지, 안전모드 전환 등의 기능을 인공지능 운영 시스템에 반영하고 (직관적 UI 제공) 관련 기록 보존	필수 적용
실시간 모니터링 대시보드	○ (내용) 성능, 공정성, 신뢰도, 규제 위반 등의 신호를 실시간으로 시각화 ○ (방법) 임계치 초과 시 자동 알림·보고	권고

[출처] NIST AI RMF Playbook

금융회사등이 인공지능기본법에 따른 고영향 인공지능을 개발·이용할 경우에는 해당 법령에 따라 고영향 인공지능의 관리·감독 인력 정보를 공개하고, 개입 기준 수립·긴급정지 기능 등 하위 규정에 따라 요구하는 관리·감독 조치를 준수하여야 한다.

인적 개입의 결과는 설명 가능한 근거와 함께 사후 추적 가능한 로그로 보존하여 책임성과 재현성을 확보한다. 또한 신속한 대응 체계를 효과적으로 운영하기 위하여 사건 대응 표준 운영 절차(예: 접수→판정→격리→증거보존→통지→근본원인분석→재발방지)를 마련할 수 있으며 필요에 따라 신속한 대응을 위한 실전 모의훈련을 실시한다.

참고 인공지능기본법령의 인적 개입 관련 준수 사항

인공지능기본법 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

동법 시행령 제27조 (고영향 인공지능과 관련한 사업자의 책무) ① 인공지능사업자는 법 제34조제1항 각 호의 조치 중에서 다음 각 호에 해당하는 내용을 인공지능사업자의 사무소·사업장 또는 인터넷 홈페이지 등에 게시하여야 한다. (이하 생략)

4. 해당 고영향 인공지능을 관리·감독하는 사람의 성명 및 연락처

사업자 책무 고시(안) 제7조 (사람의 관리·감독) ① 사업자는 인공지능시스템 개발 과정에서 사람의 관리·감독을 위해 다음 각 호의 조치를 이행하여야 한다.

1. 사람이 인공지능 동작에 개입할 수 있는 기준 확립
2. 긴급정지 기능 등 사람이 즉각적으로 인공지능시스템을 정지하거나 작동을 변경할 수 있는 개입 방법 마련

② 사업자는 고영향 인공지능 운영 중 사람의 관리·감독을 위해 다음 각 호의 사항을 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

참고 해외 금융회사의 인적 개입(HILT) 운영 사례

□ **HSBC의 자금세탁방지(AML) 시스템**

- **[인공지능 : 1단계 분석]** 인공지능이 수십억 건의 거래 데이터를 실시간으로 분석해 복잡한 자금 흐름 속에서 이상 징후를 신속하게 포착하고 고위험 거래 선별
- **[인간 전문가 : 2단계 판단]** 숙련된 분석가는 인공지능이 정밀하게 선별한 소수의 고위험 거래에만 집중, 인공지능이 제공한 시각화 자료와 분석 데이터를 바탕으로 심층 조사를 수행하고, 의심거래보고(SAR) 여부를 최종적으로 결정

  * 전문가의 최종 판단 정보(정탐, 오탐)은 다시 인공지능 모델에 학습되어 시스템 개선

□ **Morgan Stanley의 생성형 인공지능 기반 투자 자문 지원**

- **[인공지능 : 초안 생성]** 재무 설계사가 특정 투자 정보에 대하여 질문하면, 코파일럿 기반 생성형 인공지능이 답변의 초안 생성
- **[재무 설계사 : 검증 및 맞춤화]** 인공지능이 생성한 초안의 정확성을 검증(Fact-check)하고 고객의 투자 성향·재무 목표 등 맥락을 반영하여 최종 자문 내용 완성

준수 사례

□ 자체 개발한 인공지능 기반 개인사업자 대출심사 시스템을 '고영향 인공지능'으로 분류하고, 인공지능은 심사 보고서 초안(신용등급, 한도, 금리 권고 포함)까지만 생성하도록 하였다. 심사 담당자가 인공지능 보고서를 검토한 후 최종 판단을 시스템에 입력해야만 대출 절차가 완료되며 이 모든 승인 과정은 로그로 기록된다.

□ 인공지능 알고리즘 트레이딩 시스템을 '중위험'으로 분류하고 시장 급변 (예: 서킷브레이커 발동) 또는 비정상적 거래 패턴(예: 1분 내 100회 이상 동일 종목 매매) 감지 시, 담당자가 시스템을 즉각 중단시킬 수 있는 긴급정지 기능을 구현하여 운영하고 있다.

□ 자사의 인공지능 이상거래탐지시스템(FDS) 운영을 하면서 '인공지능 모델의 오탐률이 2일 연속 임계치(5%) 초과 시'와 같이 사람이 즉각 개입해야 하는 기준과 절차를 수립하고, 모의훈련을 실시하였다.

3.3 정기적인 교육 실시

✓ 보조수단성 원칙이 효과적으로 준수되도록 금융회사등의 **업무 담당자 및 감독자 등을 대상**으로 **정기적인 교육을 시행**한다.

교육의 최우선 목표는 담당자가 인공지능의 제안을 무비판적으로 수용하는 '자동화 편향'을 방지하는 데 둔다. 인공지능을 '정답을 주는 해결사'가 아닌 '판단을 돕는 조수'로 인식하도록 지속적인 인식 제고 활동을 수행한다. 이를 위해 인공지능시스템을 직접 운영하는 실무자, 최종 결정을 내리는 관리자, 시스템을 감사하는 감사인 등 역할별로 필요한 역량이 다르므로 각 역할에 맞는 맞춤형 교육 커리큘럼을 설계하여 운영한다.

한편, 고위험 인공지능시스템을 감독하는 담당자는 필수 교육 과정을 이수하고 역량을 평가받은 이후 관련 업무를 수행하도록 하는 내부 자격 제도를 운영한다.

준수 사례

□ '고영향 인공지능'인 대출심사 시스템의 운영자를 대상으로 '자동화 편향 방지 및 재정의(Override) 절차' 관련 필수 교육을 의무 과정으로 운영하고 있다.

□ 내부교육운영지침에 따라 매년 전 직원을 대상을 안전한 인공지능 활용 교육을 시행하고 있으며, 인공지능 기반 보험금 청구 시스템 접근 권한 부여자를 대상으로 비정상 운영 확인 방법 및 조치 방안을 분기별로 교육하고 있다.

[3. 보조수단성 원칙] 점검 항목

① 의사결정의 단계별 역할·책임·권한이 문서화되어 공유되고 있는가? YES□ | NO□

② 외부로부터 제공받아 운영하는 인공지능도 책임 수행 체계에 포함되어 있는가? YES□ | NO□

③ 복수인 확인 또는 위원회 검토를 통해 고영향 인공지능의 편향에 대한 이중 또는 다중 책임을 수행할 수 있는 체계를 구축하고 있는가? YES□ | NO□

④ 직무 분리와 이해상충 방지가 업무 배정·승인 운영에 실질적으로 적용되는가? YES□ | NO□

⑤ 인공지능시스템의 위험 수준에 따라 인적 개입 방식(HITL, HOTL 등)을 적용하는 기준이 문서화되어 있는가? YES□ | NO□

⑥ 고위험 인공지능에 대해서는 최종 의사결정을 인공지능이 수행하지 않도록 하는 HITL(Human-in-the-Loop) 방식이 적용되어 있는가? YES□ | NO□

⑦ 인공지능의 결과에 대한 긴급정지(Kill switch), 재정의(Override), 안전모드 전환 등 구체적인 인적 개입 방법이 시스템에 마련되어 있고 관련 절차가 문서화되어 있는가? YES□ | NO□

⑧ 인적 개입 수단의 이행 결과는 책임 추적성을 위해 설명 가능한 사유와 함께 기록·보존되는가? YES□ | NO□

⑨ 비정상 상황 발생에 대비한 표준 운영 절차(SOP)가 마련되어 있고, 이에 따른 모의훈련을 실시하는가? YES□ | NO□

⑩ 교육 목표와 원칙을 수립하여 정기적인 교육을 시행하고 있는가? YES□ | NO□

⑪ 실무자, 관리자, 감사인 등 역할별 차별화된 교육 커리큘럼이 있는가? YES□ | NO□

⑫ 교육 내용에 긴급정지 기능의 운영 방법, 재정의 등 구체적인 개입 절차와 도구 활용법에 대한 실습 또는 시뮬레이션 훈련 등을 포함하고 있는가? YES□ | NO□

4. 신뢰성 원칙

금융회사등은 인공지능시스템이 일관되고 정확한 결과를 제공하고, 문제 발생 시 적절한 대응이 가능하도록 통제한다. 금융회사등은 모델 성능 관리, 데이터 품질 확보, 의사결정 과정 설명, 체계적 검증 및 오류 대응 체계를 통해 인공지능서비스의 신뢰성을 확보한다.

4.1 모델 성능 관리

✓ **인공지능 모델**의 **성능**을 **측정**할 수 있는 **명확한 지표**를 **설정**하고, **정기적으로 점검**하고 지속 **개선**한다.

금융회사등은 인공지능 모델의 성능을 측정할 수 있는 명확한 지표를 정하고 정기적으로 점검한다. 또한, 업무 목적에 맞는 성능 지표를 설정하되 금융업무 특성을 반영한다. 예를 들어 신용평가 모델의 경우에는 실제 연체율과의 차이, 투자 상품 추천의 경우에는 수익률 예측 정확도, 챗봇의 경우에는 응답 적절성 및 만족도 등을 주요 지표로 설정한다.

참고 **인공지능 모델 성능 지표 (예시)**

□ **(신용평가)** 예측 정확도 (AUC, KS 통계량), 실제 부실률과 예측 부실률 차이, 등급별 고객 분포의 안정성, 승인율 및 거절률 변화

□ **(투자 추천)** 수익률 예측 정확도, 샤프 비율(위험 대비 수익), 최대 손실폭(Maximum Drawdown), 시장 대비 초과 수익률

□ **(이상거래 탐지)** 진양성률 (True Positive Rate), 오탐률 (False Positive Rate), 탐지 지연 시간, 미탐지 사고 발생률

□ **(고객 상담)** 응답 정확도, 고객 만족도 점수, 상담원 개입 필요 비율, 평균 해결 시간

금융분야에서는 잘못된 정보 제공이 투자 손실이나 신용평가 오류로 직결될 수 있으므로 철저한 품질 관리가 필수적이다. 생성형 인공지능의 경우 사실이 아닌 정보를 마치 사실인 것처럼 생성하는 환각(Hallucination) 현상을 완화하기 위한 구체적인 수단(예: 인용 출처 제공, 신뢰도 점수표시, 환각 위험 고지 등) 또는 기준(예: 내부 상담사 답변, 공식 문서 등)을 마련하여 주기적으로 확인하도록 한다. 판단형 인공지능의 경우 예측 정확도, 실제 결과와의 차이 등을 지속적으로 모니터링 한다. 아울러, 모델의 신뢰도가 낮거나 판단 근거가 불충분한 경우에는 사람의 검토를 거치도록 하는 절차를 구축한다.

모델 성능이 기준치 이하로 저하되는 경우 신속하게 대응할 수 있는 절차를 마련하고 성능 저하 원인을 파악하여 시장 환경 변화, 데이터 문제, 시스템 오류 등 구체적인 원인에 따라 대응하도록 한다. 심각한 성능 저하 시에는 인공지능 관련 해당 기능을 일시 중단하고 사람이 직접 처리하도록 전환한다. 모델 성능 유지를 위해 정기적으로 시스템을 점검하고 개선하도록 한다. 업무 환경이나 시장 상황 변화에 따라 적절한 주기로 업데이트한다.

준수 사례

□ 대출심사 인공지능은 실제 부실률과 예측 부실률의 차이를, 투자 추천 인공지능은 수익률 예측 정확도를, 상담 챗봇은 응답 정확도 및 사용자 만족도를 주요 지표로 설정하고 각 업무 특성에 따라 적절한 점검 주기를 설정하여 점검하고 있다.

□ 금융상담 챗봇의 응답 사실성 검증을 위해 내부 공식 FAQ 문서 및 금융상담사 답변을 기준으로 설정하고, 월 1회 샘플링(100건 이상) 검증을 수행한다.

□ 신용평가 모델의 실제 부실률과 예측 부실률을 월 1회 비교하여 차이가 기준치 초과 시 원인을 분석하고 재학습 여부를 결정한다.

□ 챗봇의 참조문서 유사도 점수가 기준치 미만일 경우 "정확한 답변을 위해 상담사 연결을 권장합니다."라는 안내 메시지를 출력하도록 설정하였다.

□ 대출심사 인공지능의 신뢰도 점수가 기준치 미만일 경우 자동으로 심사 담당자에게 배정하여 수동 검토를 받도록 한다.

4.2 데이터 품질 관리

✓ **인공지능 학습 및 참조에 사용하는 데이터와 인공지능시스템에 입력되는 데이터의 품질을 검증·확인**한다.

금융회사등은 인공지능 개발·이용에 필요한 데이터의 정확성, 완전성, 일관성, 대표성, 적시성을 체계적으로 검증하는 절차를 거치도록 한다. 데이터 품질의 점검과 확인은 수집 단계를 포함하여 처리 전 과정을 고려하여 품질 검증을 수행하도록 한다. 사용된 데이터는 계보 (lineage)와 출처가 추적 가능하도록 관리되어야 한다.

참고 **데이터 처리 시 고려해야 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 (Duplication)	○ 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
데이터 오염 (Data Poisoning)	○ 악의적인 목적으로 변조한 데이터

인공지능 모델 개발에 필요한 학습·참조 데이터의 품질은 인공지능의 의사결정과 직결되므로 보다 체계적인 관리가 필요하다. 데이터 품질 이상 여부를 주기적으로 모니터링하며 인공지능 시스템의 성능을 저해할 수 있는 비정상적인 데이터를 사전에 식별하여 제거한다. 데이터 유형에 따라 적절한 검증 방법(통계적 방법론, 머신러닝 또는 LLM 기반 탐지 기법 등)을 활용하여 데이터의 품질을 개선한다.

인공지능시스템에 입력되는 운영 데이터는 형식, 범위, 논리적 일관성 등이 유지되도록 검증하고, 비정상적인 데이터가 탐지될 경우 다시 입력하도록 요청하거나 내부 직원이 직접 확인하는 절차를 마련한다.

고객을 포함하여 외부에서 직접 입력하여 생성되는 데이터의 경우 의도적 또는 비의도적 오류 가능성을 고려하여 철저히 검증하고, 외부로부터 제공받은 데이터의 경우 전송 과정에서의 오류나 변조 가능성을 점검하도록 한다. 특히 스크래핑 등으로 수집된 데이터나 공개된 데이터와 같이 위험도가 높은 외부 데이터는 파일럿 테스트, 대체 데이터 비교 등 충분한 사전 점검 후에 활용되어야 한다.

외부 상용 인공지능(LLM 등)을 활용하는 경우와 같이 금융회사가 데이터에 직접 접근하거나 통제하기 어려운 경우에는 해당 모델의 공급자 신뢰성 확인, 제공된 기술 문서 검토, 파인튜닝(Fine-tuning; 미세조정) 단계에서의 데이터 품질 점검 등을 통해 데이터 품질 수준을 확보한다.

외부와의 연계 또는 공동 활용이 예상되는 데이터는 국내외 금융 데이터 표준(메타데이터·코드·메시지·용어 등)과의 정합성을 고려하여 상호 운용성(interoperability)을 확보하도록 노력한다.

준수 사례

□ 데이터 정확성, 최소 데이터량, 일관성, 최신성에 대한 기준을 수립하고, 수집-검증-정제-적재 단계에서 품질 검증을 수행하며, 주요 단계별 기준치 충족 여부를 점검한다.

□ 대출 신청 데이터에 대해 도메인 룰 기반 검증(소득 범위, DSR 한도, 신용등급-연체이력 논리적 일관성 등)을 실시간으로 수행하고, 통계적 이상치 탐지로 비정상 패턴을 주기적으로 모니터링한다.

□ 챗봇 참조문서(RAG)에 대해 주기적으로 샘플링 검증을 실시하여 부적절한 내용, 오류 정보 포함 여부를 확인하고, 이상 발견 시 해당 문서를 격리한다.

□ 고객 입력 정보의 형식·범위 검증을 실시간으로 수행하고, 비정상 데이터 탐지 시 최대 3회 재입력을 요청한 후 고객센터로 안내한다.

4.3 공정성 · 편향성 점검

✓ 인공지능서비스가 모든 집단에 대해 **차별 없이 공정하게 작동**하도록 **데이터와 모델을 분석하여 개선**한다.

인공지능은 데이터와 모델에 내재된 편향을 증폭할 우려가 있다. 금융회사등은 인공지능이 모든 집단에 대해 차별 없이 공정하게 작동하도록 지속적으로 검증 개선한다.

인공지능 개발·이용에 사용되는 데이터에서 발생할 수 있는 편향성을 사전에 점검하고 조치한다. 성별, 연령, 지역 등 인구통계학적 특성에 따른 데이터 편향이 있는지 정기적으로 분석하고 특정 집단에 불리한 결과를 초래할 수 있는 편향이 탐지될 경우 이를 개선한다. 수집된 데이터에 대한 라벨링을 실시하면 작업하는 사람의 편향성이 데이터에 반영될 수 있으므로 라벨링 작업자를 대상으로 사전 교육을 시행하고 가능한 범위에서 다양한 배경을 가진 사람이 참여한다.

참고 **인공지능 편향성 분류 (예시)**

□ **(사회적 편향)** 특정 사회 집단이 유리하고 다른 집단이 불이익을 받는 방식으로 운영되는 절차 및 관행에 의해 발생하며, 제도적·역사적 편향이라고도 지칭

○ 단순한 편견·차별이 아닌 대다수가 따르는 규칙 등에 따른 제도적 인종차별 및 성차별, 보편적 디자인 원칙을 사용하지 않고 개발된 일상생활 인프라에서의 장애인 접근성 제한 등이 해당되며, 인공지능 학습 데이터, 인공지능 라이프 사이클과 문화와 사회 전반에 걸친 제도적 규범, 관행 및 절차에 존재

□ **(통계 및 계산적 편향)** 편견, 편파성 또는 차별적 의도가 없이, 시스템이 모집단을 대표하지 않는 오류에 따라 발생

○ 인공지능 알고리즘이 특정 유형의 데이터에 대해서만 학습되어 이외의 유형에 대해 추론할 수 없을 때 주로 발생하며, 데이터 복잡도에 비해 단순한 알고리즘 사용, 데이터 오류, 과·부적합, 이상치 및 결측치 처리 등이 요인이 될 수 있음

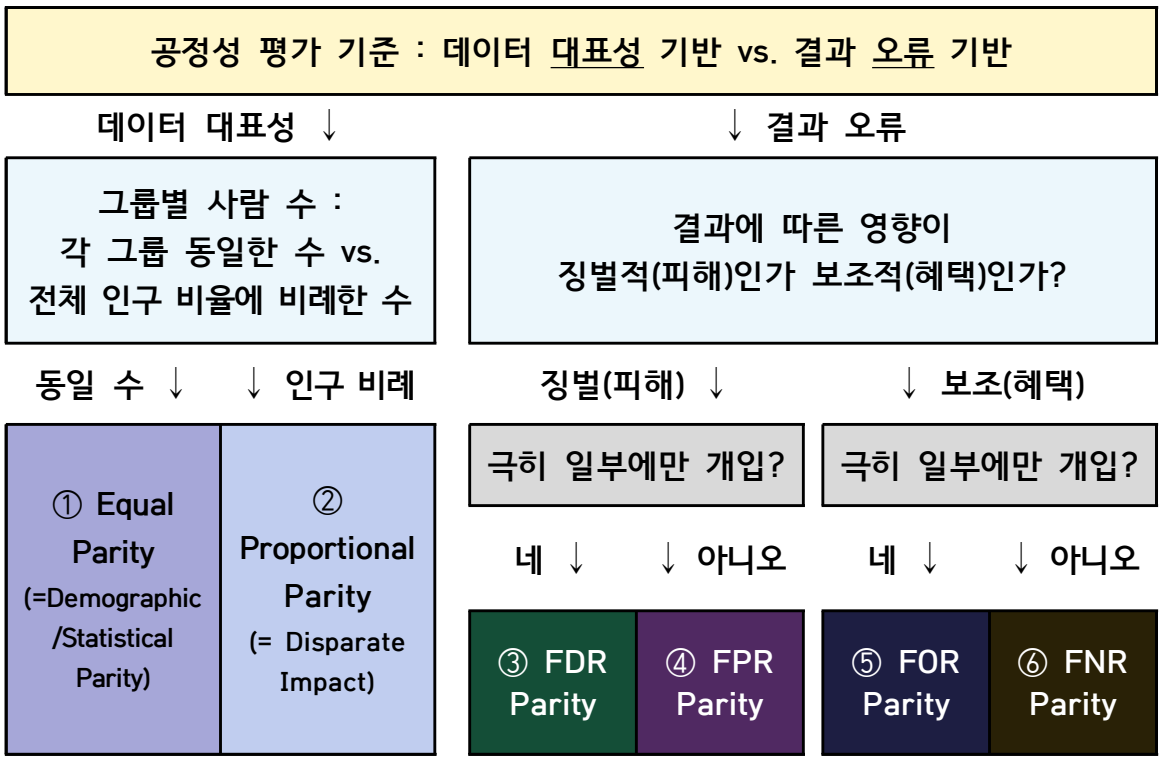
□ **(인간 편향)** 경험적 원리와 예측을 기반으로 하는 사람의 생각에는 오류(human bias)가 존재함. 사람의 인지적·지각적 편향은 사람과 인공지능의 상호작용을 비롯해 모든 영역에서 다양하게 나타남

[출처] NIST 특별 간행물, 「인공지능 편향성(Bias) 식별 및 관리 표준을 향하여」(Towards a Standard for Identifying and Managing Bias in Artificial Intelligence), 2022.

고객에게 직접적인 영향을 미치는 의사결정(대출심사, 신용평가, 로보어드바이저, 보험 언더라이팅 등)을 수행하는 인공지능 모델의 경우 편향성 평가 및 모니터링을 위한 지표를 선정하고 개발 단계 및 운영 중에 테스트를 실시한다. 모델 이외의 서비스 구현 과정에서도 편향이 발생할 수 있으므로 서비스 전반에 대한 편향성을 확인한다.

금융회사등이 운영하는 인공지능서비스에 대해 공정성과 편향성을 주기적으로 점검하여 소비자 집단별 서비스의 결과에 편향이 탐지될 경우 이를 개선한다. 공정성 확보 과정에서 모델의 성능(정확도 등) 저하가 예상될 경우, 성능 감소 폭과 공정성 감소 폭을 비교하여 리스크 관리 부서와 개발 부서 간 협의를 통해 수용 가능한 성능 저하 범위를 결정한다.

참고 공정성 평가 지표 선택 및 보고서 작성 (예시)



① **Equal Parity:** 선택한 집합에서 모든 보호 대상 그룹이 동등한 대표성을 갖는지 확인

② **Proportional Parity:** 모든 보호 대상 그룹이 인구 비율에 비례하여 선택되는지 확인

③ **FDR Parity:** 모든 보호 대상 그룹이 양성으로 판단된 세트 내에서 기준 그룹과 비교했을 때 동일한 비율의 위양성을 갖는지 확인

- FDR(False Discovery Rate) : 양성 판단 중 위양성 비중
  (false positive / predicted positive)

④ **FPR Parity:** 모든 보호 대상 그룹이 기준 그룹과 동일한 FPR을 갖는지 확인

- FPR(False Positive Rate) : 음성 라벨 중 위양성 비중
  (false positive / labeled negative)

⑤ **FOR Parity:** 모든 보호 대상 그룹이 음성으로 판단된 세트 내에서 기준 그룹과 비교했을 때 동일한 비율의 위음성을 갖는지 확인

- FOR(False Omission Rate) : 음성 판단 중 위음성 비중
  (false negative / predicted negative)

⑥ **FNR Parity:** 모든 보호 대상 그룹이 기준 그룹과 동일한 FNR을 갖는지 확인

- FNR(False Negative Rate): 양성 라벨 중 위음성 비중
  (false negative / labeled positive)

[공정성 평가 결과 보고서 작성 (예시)]

구분	내 용
모델명	개인 신용평가 모형 v2.0
핵심 공정성 지표	기회 균등 (FNR Parity)
평가 결과	개선 전 격차 2.5 (불공정) → 개선 후 격차 1.1 (공정, 80% Rule 충족)
성능 변화 (Trade-off)	정확도(AUC) : 0.92 → 0.89 (3%p 하락)
의사결정 사유	○ 정확도가 3% 하락하여 연간 대손비용이 약 0억 원 증가할 것으로 예상되나, ○ '금융소비자보호법' 준수 및 평판 리스크 방지를 위해 이를 수용함.
승인	**리스크관리책임자(CRO) : 홍 길 동 (인)**

[출처] Aequitas: A Bias and Fairness Audit Toolkit 中 '공정성 구분 트리' 참고하여 재구성

참고 인공지능 편향성의 원인 (예시)

□ **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., 2016년)

□ **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 학습 데이터보다 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

[출처] 소프트웨어정책연구소, 「기계학습 공정성 관련 연구 동향」, 2020.

준수 사례

□ 대출심사 인공지능 및 신용평가 모델에 대해 성별·연령·지역별 승인율을 주기적으로 검증하고, 집단 간 승인율 차이가 기준치 이상 시 원인을 분석하여 모델 재조정 여부를 결정한다.

□ 학습 데이터의 성별, 연령, 지역 분포를 분석하여 특정 집단의 표본 수가 통계적으로 유의미한 분석을 하기에 부족할 경우, 가능한 범위에서 추가 데이터 수집을 검토하거나 모델 학습 시 가중치를 조정한다.

□ 데이터 라벨링 작업자에게 편향 방지 교육을 사전에 실시하고, 가능한 범위에서 다양한 배경의 작업자를 배치한다.

□ 운영 중인 대출심사 인공지능의 집단별 승인율, 금리, 한도를 주기적으로 자동 집계하고, 편향성 지표가 기준 초과 시 검토 후 보고한다.

□ 상담 챗봇 서비스의 응답 품질을 연령대별로 샘플링 검증하여 특정 연령대에 불리한 전문용어 과다 사용 등이 없는지 확인한다.

4.4 설명가능성 확보

✓ **인공지능 의사결정 과정과 결과**에 대해 **이해관계자가 합리적으로 이해**할 수 있도록 **설명 가능한 형태로 제공**하여 **신뢰성을 강화**한다.

인공지능시스템의 신뢰성을 확보하기 위해서는 인공지능 모델이 특정 결정을 내린 이유를 파악하는 것이 중요하다. 금융회사등은 고객의 금융 거래에 직접적인 영향을 미치는 의사결정의 경우 그 과정을 추적하고 설명한다. 어떤 정보가 결정에 가장 큰 영향을 미쳤는지 확인하는 방법을 마련하고 중요한 결정의 경우 그 과정을 기록으로 남기는 등 추적 가능성을 확보하기 위해 노력한다. 특히 신용평가, 대출 승인, 투자권유 등 고객에게 직접적인 영향을 미치는 결정에 대해서는 상세한 설명 자료를 생성·관리함으로써 신뢰성을 제고할 수 있다.

특히 인공지능의 결정이 금융소비자의 금융서비스 이용에 직접적인 영향을 미칠 경우 신뢰성 향상을 위해 금융소비자가 인공지능의 결정을 이해할 수 있도록 적절한 수준의 설명을 제공하는 것이 필요하다. 기술적 전문 지식이 없는 일반 고객도 이해할 수 있는 쉬운 언어로 설명을 제공하되 개인정보 보호와 영업 기밀 유지를 고려하여 적절한 수준에서 설명 범위를 조정하여 신뢰성 있는 서비스를 구현한다.

신용평가나 투자 상품 추천 등의 경우 고객이 이해할 수 있는 수준에서 영향을 미친 주요 요인 위주로 안내한다. 고객이 결정에 대해 문의하거나 이의를 제기하면 답변할 수 있는 창구와 절차를 마련하여 시스템에 대한 신뢰성을 확보한다.

참고 **금융분야 설명가능 인공지능 적용 (예시)**

□ **신용평가 시스템**

고객님의 신용등급 산정 시 소득수준, 기존 대출 상환 이력, 신용카드 사용 패턴, 직업 안정성, 기타 요소가 반영되었습니다.

□ **보험 인수 심사**

보험료 산정 시 연령대별 위험도, 과거 질병 이력, 생활 습관 점수, 직업군 위험도 등을 종합 고려하였으며, 각 요소가 최종 보험료에 미치는 영향을 안내서로 제공드립니다.

□ **투자 상품 추천**

고객님의 투자 성향(안정형), 투자 목적(노후자금), 투자 기간(10년), 현재 포트폴리오 현황을 바탕으로 채권형 펀드 60%, 주식형 펀드 40% 비중을 권장드립니다.

□ **이상거래 탐지**

평소 거래 패턴 대비 거래시간(야간), 거래금액(평균의 5배), 거래 지역(해외) 등이 달라 추가 본인확인이 필요합니다.

금융회사등은 감독당국의 요청 시 인공지능시스템의 작동 원리나 의사결정 과정을 설명할 수 있는 자료를 준비한다. 모델의 학습 데이터, 알고리즘 선택 근거, 주요 변수의 영향도, 성능 검증 결과 등을 포함한 상세한 기술 문서를 작성하고 최신 상태로 유지한다. 특히 고객에게 중요한 영향을 미치는 업무에 사용되는 인공지능의 경우 의사결정 논리를 단계별로 추적하고, 필요시 개별 고객 사례에 대한 구체적인 설명을 제공한다. 또한, 인공지능시스템의 편향성 점검 결과, 공정성 확보 방안, 위험관리 체계 등을 포함한 종합적인 관리 현황을 보관하여 규제 기관의 검사나 보고 요구에 신속하게 대응한다.

참고 설명가능한 인공지능(eXplainable AI) 기술이란?

□ **XAI : 인공지능의 결정에 대한 이유를 묻는 기술**

○ 인공지능 모델의 결정 과정을 설명하는 기술로, 인공지능의 블랙박스 문제를 해결하고 인공지능의 결정 과정을 사람이 이해하고 신뢰할 수 있게 함

- **(XAI)** 미국 국방성 산하 국방위고등연구계획국(DARPA)은 기존에 사용되던 인공지능 모델에 데이터를 기반으로 인공지능시스템의 블랙박스 안을 확인한다는 의미로 X인공지능을 정의(2015년)

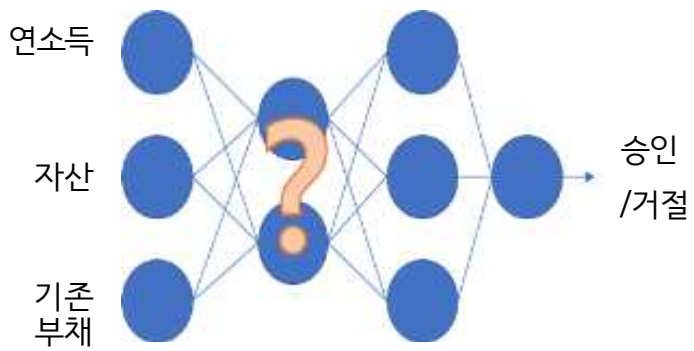
- **(인공지능 블랙박스 문제)** 복잡한 딥러닝 기반의 인공지능 모델에서 두드러지는 문제로, 모델이 내부적으로 어떻게 작동하는지 관찰하거나 이해하기 어려운 문제

  예) 알파고의 경우에도 바둑에서 특정수를 왜 둔 것인지 알파고 개발자도 설명하기 어려웠는데, 인공지능 블랙박스 문제가 그 이유

○ 전통적 IT 서비스는 사람이 직접 설계한 로직이므로, 결정 과정에 대한 설명이 쉬운 반면, 인공지능이 설계한 로직은 이해가 어려움

⇨ 별도 기술(XAI) 사용하여 설명 필요

[기존 프로그램과 인공지능 프로그램 설명 비교 (예시)]

	프로그램	설명
기존	[만약] (연소득 + 자산 - 기존 부채) > 대출 금액 : ▸ 승인 [아니면] ▸ 거절	연소득과 자산 합에서 기존 부채를 뺀 금액이 대출 금액보다 클 경우 승인
인공지능 도입		 → **X인공지능 필요**

참고 금융권 XAI 활용 절차 및 기준 (예시)

1 신뢰할 수 있는 설명 생성 : 신뢰성 있는 XAI 알고리즘 선정

○ 인공지능이 판단한 근거를 설명하는 기법은 매우 다양하므로, 신뢰할 수 있는 설명 생성을 위해 XAI 설명 신뢰성 진단 필요

○ 설명이 법적 의무 사항인 경우 인공지능 설명의 신뢰도 확보를 위해 SHAP 수준 이상의 XAI 알고리즘 활용 필요

○ 법적 의무 사항이 아닌 경우에도 신뢰성 있는 알고리즘 사용 필요

[SHAP(SHapley Additive exPlanations) 알고리즘 개요]

○ SHAP 알고리즘은 게임 이론에서 유래한 샤플리값(Shapley Value)에 기반하여, 각 특성이 모델의 예측에 얼마나 중요한 영향을 미치는지 정량화하는 기법으로 신뢰도가 높음*

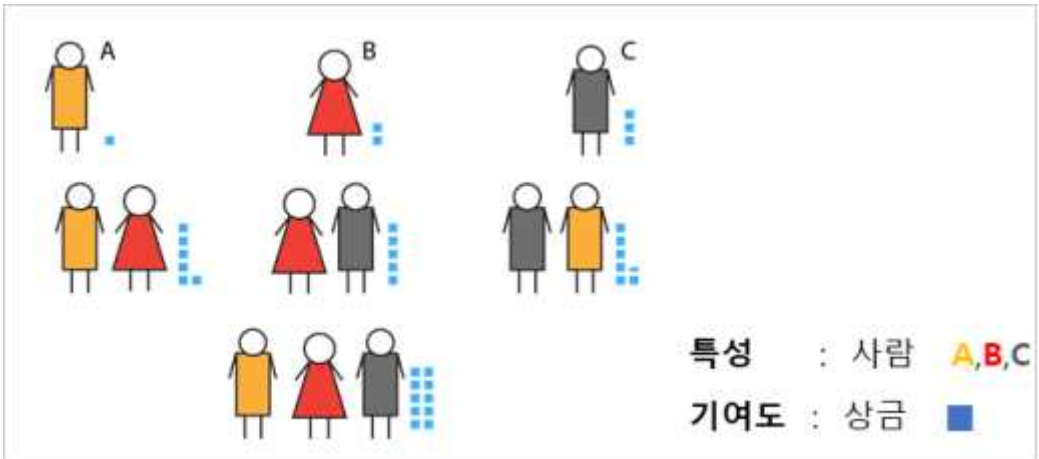
* SHAP 프레임워크는 탄탄한 이론적 배경과 일반적인 적용 가능성으로 많은 사람들에게 로컬 설명의 표준으로 간주된다. (Edoardo Mosca, 2022년도 ICCL 논문 中)

○ 샤플리값은 각 특성들이 결과에 얼마나 기여했는지 나타내는 척도로 게임 이론에서 확고한 수학적 기반을 갖고 있음

○ 다만, 샤플리값 계산은 복잡도가 매우 높으므로, 샘플링, 근사 등을 활용하여 계산 부담을 줄이면서도, 샤플리값의 핵심 개념을 유지하도록 설계된 SHAP을 사용

예) A,B,C 세 명이 한 조로 대회에 참가했을 때 상금을 100만원 받았다. 각 사람이 상금에 얼마나 기여했는가?

→ 그 사람을 빼고 했을 때 상금, 혼자 했을 때 상금 등 다양한 경우의 수에 대한 결과를 비교하여 계산 ⇨ 인공지능 모델에서는 사람이 특성, 상금이 기여도에 해당



2 인공지능 모델 성능 개선 : XAI를 이용하여 인공지능 모델 신뢰성 확보

○ 금융회사가 마련한 기준에 따라 신뢰성을 진단하고, 해당 인공지능 모델 또는 인공지능 설명의 신뢰성을 확보하기 위한 대체 방안 마련

3 고객 설명 제공 기준 : 설명 내용 및 최소 내용 충족

○ 인공지능 모델의 전체적인 동작을 설명하는 전역 설명과 개별 데이터의 예측 결과에 관한 판단 사유를 설명하는 국소 설명 생성

- (전역 설명 자료) 모든 고객 대상으로 동일하므로, 모델별로 사전에 생성하여 필요시 고객에 즉시 제공할 수 있도록 준비

- (국소 설명 자료) 고객별 생성이 필요하므로, 고객이 개인별 설명을 신청하였을 때, 제공 가능한 기한*을 안내하고, 기한 내 제공할 수 있도록 준비

* 신뢰도가 높은 국소 설명의 경우 생성하기 위해 일반적으로 상당한 시간이 소요되며, 인공지능 모델 입력의 규모 등에 따라 생성 소요 시간이 달라지므로, 업무 성격, 고객 예상 수요 등에 따라 자체적으로 결정(예: 대출심사 결과 개별 설명 5영업일 이내)

○ 고객 대응 업무팀에서 고객에게 설명을 제공하기 위해 필요한 XAI 산출물 최소 내용* 출력

* 판단에 이용된 기초정보 중 해당 판단 사유가 된 주요 기준 및 영향, 특히 고객에게 부정적 영향을 미치는 판단 사유가 존재하는 경우 반드시 1개 이상 포함

- 다만, 기초정보가 모델에 활용되는 세부적인 방법, 인공지능 모델에서 사용한 학습모델 등 과도한 정보가 노출될 경우 고객이 제공된 설명 정보를 악용할 수 있으므로 이에 유의할 필요*

* (예1) 특정 대출을 갚는 행위가 신용 점수에 큰 영향을 미치지 않는다는 것을 알게 되면, 고의로 그 대출의 상환을 지연시켜 자금을 다른 용도로 사용
(예2) 금융기관의 투자 결정 로직을 이해할 경우, 투자자들이 이를 악용하여 시장 조작이나 불공정 거래에 이용 가능

4 후속 관리 : 인공지능 결과·설명, 안전한 관리 및 삭제

○ 인공지능 결과·설명, 업무팀 제공 경로 등을 안전하게 보관·관리하고, 고객 요구 등에 따라 필요시 안전하게 삭제

○ XAI 설명 내용에 대한 고객 민원 제기 등 필요시 인공지능 모델 성능 개선 등 실시

[인공지능 설명 제공 (예시)]

XAI 시스템 생성 자료

○ 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

고객 제공 설명

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.

준수 사례

□ 대출심사 챗봇 서비스 이용약관에 "본 서비스는 생성형 인공지능 기술을 활용하여 상담 내용을 생성합니다"라는 문구를 명시하고, 챗봇 하단에 상시로 표시한다.

□ 인공지능으로 생성한 투자 리포트 하단에 "본 리포트는 인공지능이 작성하였으며, 투자권유가 아닌 참고자료입니다."라는 표시를 삽입한다.

□ 대출심사 인공지능의 의사결정을 "신용점수 계산 → 소득분석 → 부채비율 산정 → 종합 평가 → 최종 결정" 단계로 구분하고, 각 단계별 결과를 로그에 저장하여 추적 가능하도록 한다.

□ XGBoost 기반 신용평가 모델에 SHAP 값을 적용하여 주요 변수(신용등급, 소득, DSR 등)의 영향도를 정량화하고, 의사결정 근거 자료로 활용한다.

□ 대출 거절 시 고객에게 "신용등급, DSR, 최근 연체 이력 등을 종합 검토한 결과"라는 주요 요인별 설명을 안내하고, 상세 설명 요청 시 합리적인 기간 내 개인화된 설명서를 발급한다.

□ 투자 상품 추천 시스템에 대해 고객 투자 성향, 시장 전망, 상품 특성 등 범주별 영향도를 시각화하여 담당자에게 제공한다.

□ 챗봇 답변 시 인공지능의 답변임을 고객에게 설명하며, 복잡한 금융용어를 일상적 표현으로 바꾸고, 필요시 예시를 들어 고객의 이해를 돕는다.

□ 고객센터를 통해 인공지능 의사결정에 대한 이의를 제기할 수 있으며, 이 경우, 담당자가 재검토하여 결과를 제공한다.

□ 사용 중인 인공지능 모델에 대해 모델명, 버전, 학습 데이터 출처, 주요 변수, 성능 지표 등을 포함한 모델 카드를 작성하고 분기별로 업데이트하며, 감독당국 요청 시 즉시 제출할 수 있도록 준비한다.

[4. 신뢰성 원칙] 점검 항목

① 인공지능 모델의 성능을 측정할 수 있는 명확한 지표를 업무별로 설정하고 정기적으로 점검하는가? YES□ | NO□

② 성능이 기준치 이하로 떨어질 경우 즉시 파악할 수 있는 체계를 마련했는가? YES□ | NO□

③ 모델 성능이 저하되는 경우 신속하게 대응할 수 있는 절차를 마련하고, 성능 저하 원인을 파악하여 구체적인 원인에 따라 대응하도록 하는 절차를 마련했는가? YES□ | NO□

④ 인공지능 개발·이용에 사용하는 데이터의 정확성, 완전성, 일관성, 대표성, 적시성을 체계적으로 검증하는 절차가 마련하여 운영하고 있는가? YES□ | NO□

⑤ 잘못된 정보나 누락된 정보가 없는지 점검하고, 데이터 수집부터 전처리까지 전 과정에 걸쳐서 품질 검증을 수행하는가? YES□ | NO□

⑥ 데이터 이상치 탐지를 통해 인공지능시스템의 성능을 저해할 수 있는 비정상적인 데이터를 사전에 식별하고 제거하는가? YES□ | NO□

⑦ 운영 단계에서 비정상적인 데이터가 탐지될 경우 다시 입력하도록 요청하거나 수동으로 직원이 확인할 수 있는 절차가 마련되어 있는가? YES□ | NO□

⑧ 학습·참조 및 운영에 사용되는 데이터의 편향성을 사전에 점검하고 조치하는 절차를 마련하고 시행하고 있는가? YES□ | NO□

⑨ 고객의 의사결정에 직접적인 영향을 미치는 인공지능의 경우 편향성 평가 및 모니터링을 위한 지표를 선정하고 개발 및 운영 중에 테스트를 실시하는가? YES□ | NO□

⑩ 운영 중인 인공지능서비스의 공정성과 편향성을 주기적으로 점검하여 소비자 집단별 서비스 결과에 편향이 탐지될 경우 개선하는 절차를 마련했는가? YES□ | NO□

⑪ 고객의 금융거래에 직접적인 영향을 미치는 의사결정의 경우 기술적·환경적 가능성을 고려하여 그 과정을 추적하고 설명할 수 있도록 절차를 마련했는가? YES□ | NO□

⑫ 신용평가, 대출 승인, 투자권유 등 고객에게 직접적인 영향을 미치는 결정에 대해 설명 자료를 생성할 수 있는 절차를 마련했는가? YES□ | NO□

⑬ 인공지능의 결정이 금융소비자의 금융서비스 이용에 직접적인 영향을 미칠 경우, 금융소비자가 인공지능의 결정을 이해할 수 있도록 기술적 전문 지식이 없는 일반 고객도 이해할 수 있는 쉬운 언어로 설명을 제공하는가? (단, 개인정보 보호와 영업 기밀 유지를 고려하여 적절한 수준에서 설명 범위 조정 필요) YES□ | NO□

⑭ 고객이 결정에 대해 문의하거나 이의를 제기할 경우 답변할 수 있는 창구와 절차를 마련했는가? YES□ | NO□

⑮ 모델의 학습 데이터, 알고리즘 선택 근거, 주요 변수의 영향도, 성능 검증 결과 등을 포함한 기술 문서를 작성하고 최신 상태로 유지하는가? YES□ | NO□

5. 금융안정성 원칙

금융회사등은 인공지능 개발·이용 및 인공지능시스템 운영의 全 과정에서 금융안정성 위험을 최소화해야 한다. 유사한 인공지능 모델의 활용 증가나 데이터 집중도 증가는 시장의 군집행동을 야기하고 금융안정성을 위협할 수 있다. 또한, 제3자에 대한 의존도 상승은 금융시장이나 금융회사 간 상호연계성과 획일성 증가로 이어져 시스템 위험을 높인다. 사이버 리스크 확대 또한 금융시스템을 위협하는 요인으로 작용한다.

과거 인공지능 위험에 대한 논의는 주로 인공지능의 근본적인 특성에 기인하는 문제 중심으로 이루어졌다. 이는 설명가능성의 제약, 편향성 문제, 투명성 및 공정성 부족 문제 등 인공지능의 오용, 오작동 등이 특정 집단의 개인, 즉 고객 또는 사용자 등에게 미치는 직접적인 악영향에 대한 것들이다.

하지만 최근 인공지능 위험에 대한 논의는 금융분야 인공지능의 확산이 금융시스템의 안정성을 위협하는 문제 중심으로 확대되고 있다. 시스템 위험은 한 개인, 기업, 나아가 시장의 한 부문에서 발생한 사건이나 부도 혹은 의도되지 않은 집단적 행동의 여파가 빠르게 전파되면서 광범위한 금융시스템 혹은 경제 전반에 악영향을 미치는 상황을 의미한다. 특히 기술의 오용이나 알고리즘의 오작동, 혹은 특정인의 의도나 명백한 잘못 없이도 인공지능 확산에 따라 금융시스템에 위기를 촉발할 가능성이 커지고 있어 이러한 위험을 최소화하는 방안을 마련해야 한다.

5.1 금융안정 평가 · 관리

✓ **인공지능시스템**이 **금융시장 전반** 또는 **금융안정**에 미칠 수 있는 **영향** 등 **위험**을 **평가**하고 **관리**하는 **방안**을 마련한다.

금융회사등은 인공지능시스템이 금융시장 전반 또는 금융안정에 미칠 수 있는 영향 등 위험을 평가하고 관리하는 방안을 위험관리 체계에 포함한다. 금융안정위원회(FSB)는 제3자 의존성 증대, 시장 내 동조화 심화, 사이버 리스크 확대, 모델·데이터 리스크 등 금융안정과 관련된 인공지능 취약점을 제시하였는데, 금융회사등은 이를 참고하여 금융안정 위험 요인을 고려한다. AI RMF의 인공지능 위험평가에는 금융안정성 원칙을 정성적 요소로 반영하여 금융시장 안정에 위협을 미칠 수 있는 경우 위험등급 점수와 상관없이 서비스 출시 여부를 재검토하도록 제안한다.

참고 **시스템 위험을 증대할 수 있는 인공지능 관련 취약점**

□ **제3자 의존성과 아웃소싱의 집중**

○ 인공지능 발달에 따른 하드웨어, 클라우드, 사전 훈련된 모형의 활용은 금융회사의 제3자 의존성을 가속화할 수 있음

- 인공지능 개발과 관련된 공급사슬의 복잡화와 이 단계들의 집중화는 이를 활용하는 금융회사들의 제3자 의존성을 가속화하고 리스크 전이 효과 등을 심화
- 클라우드 등 인공지능서비스의 기초가 되는 인프라에 대해 특정 회사가 시장 지배력을 가질 시 가격 등의 여러 방향으로 독과점을 행사하려 할 수 있고, 이는 이용업체(end-user)의 부담으로 작용할 수 있음

□ **시장 내 동조화 심화**

○ 인공지능 모델을 사용하는 과정에서 여러 금융회사가 유사한 모형이나 데이터를 사용하는 경우 동조화가 심화되어 쏠림 현상과 변동성이 커질 수 있음

- 사전 학습된 모형 활용 증가는 여러 금융회사의 동일한 모형 사용으로 이어져 시장의 쏠림과 동조화 현상을 심화시킬 수 있음
- 금융시장 인프라의 발전으로 인한 거래 속도 증가는 금융시장의 자동화가 더 많이 될수록 시장의 변동성, 동조성을 강화시킬 수 있음

□ **사이버 리스크 확대**

○ LLM과 GenAI 기술의 발달은 해킹이나 침해 시도를 쉽게 하여 사이버 공격의 빈도와 영향을 크게 할 수 있음

- 이용자와 인공지능서비스 간 상호작용의 빈도가 커지면서 이와 관련하여 발생하는 사이버 침해 시도가 증가
- 특정 인공지능서비스 제공자로부터 여러 금융회사가 인공지능서비스를 중점적으로 제공받는 경우 이 제공자를 대상으로 한 사이버 침해사고는 여러 금융회사에 파급효과를 미칠 수 있음

□ **모델 리스크, 데이터 품질, 통제되지 않는 모델의 문제**

○ 인공지능 모델이 복잡해질수록 이를 검증·모니터링하고 수정하는 것이 어려워지고 데이터가 비정형적이고 커질수록 품질을 평가하는 것이 까다로워질 것임

- 인공지능 기술이 발달할수록 모형이 복잡해지고 여러 단계의 정제 과정을 거친 모형을 사용하는 빈도가 커지는데, 이는 모형을 검증하고 모니터링하며 여러 문제들을 수시로 수정하는 것을 어렵게 할 수 있음
- 비정형 데이터의 활용 증가와 데이터 규모의 확대는 데이터의 편향성, 투명성, 품질 등을 평가하기 어렵게 할 것임

[출처] 금융안정위원회(FSB), The Financial Stability Implications of Artificial Intelligence, 2024.

5.2 안전장치 마련

✓ **인공지능 모형 오작동** 시 **백업모형 활용**, 사후 개입이 가능한 **긴급정지 기능** 등 **시스템 위험관리**를 위한 **안전장치**를 마련한다.

금융회사등은 시스템 위험관리를 위해 개별 금융회사, 인공지능 모형에 대한 위험관리뿐만 아니라, 사고 시 대응 방안 등을 반영하여 해당 위험 수준에 따른 안전장치를 마련한다.

예를 들어, 로보어드바이저가 확산되면 거래 변동성이 증폭하거나 시장의 급변을 야기할 수 있다. 따라서 이에 대응할 수 있는 긴급정지 기능 등 사전에 구체적인 대응 방안을 마련하는 것이 중요하다. 인공지능 모형 오작동 시 예비로 기존 모형 및 통계 기반 모형을 백업모형으로 활용하거나, 사후 개입 장치인 긴급정지 기능 및 회생 계획 설계 등 관리 방안을 마련한다.

또한, 초고성능 인공지능이 인간의 의도를 벗어나는 행위를 하는 상황(예: 앤트로픽사의 클로드 미토스 시스템 카드 공개 사례 등)에 대비하기 위해서도 긴급정지 기능에 대해 고려한다.

5.3 제3자 IT리스크 관리

✓ **인공지능시스템** 관련 **제3자 IT리스크**를 **관리**할 수 있도록 **정보처리 업무 위탁** 관련 **규정 준수**, 단계별 **내부통제 체계** 및 **비상 대응 계획** 마련, **제3자 현황 식별·관리** 및 **주요 제3자 지정** 등 **관리 방안**을 수립한다.

금융회사등은 인공지능시스템 위험관리를 위해 인공지능 모델을 외주 개발하거나 오픈소스 기반 인공지능을 활용하여 금융회사 내부에 인공지능시스템을 구축하는 경우, 아웃소싱에 따른 위험을 식별·측정하고 이에 대한 백업 플랜을 마련하는 등 제3자 IT리스크를 관리할 수 있는 방안을 수립한다.

우선, 금융회사가 정보처리를 수반한 업무를 제3자에게 위탁하는 경우 기본적으로 「금융회사의 정보처리 업무 위탁에 관한 규정」을 따르도록 규율하고 있는바, 동 규정에서 규정하는 위·수탁 책임 관계 등 정보처리 위탁 계약 시 포함 사항, 고유식별정보 암호화 및 해외 이전 금지 등 정보보호 사항, 감독당국 보고 방법 등을 준수하여야 한다.

또한, 클라우드컴퓨팅(Cloud Computing) 서비스를 이용하는 경우 전자금융감독규정에 따라 이용 업무 중요도 평가, 클라우드 서비스 제공자(Cloud Service Provider; CSP)의 건전성 및 안전성 평가 수행 등의 의무를 준수하여야 한다. 특히 클라우드 최초 이용 시에는 전자금융감독규정(별표2의2)에 따라 정보보호 정책, 접근통제, 가상화

및 인프라 보안, 암호화 등 서비스 제공자의 안전성 평가를 실시하는 등 통제 절차를 마련하여야 하므로 금융회사등은 이러한 의무 사항을 확인하여 인공지능 관련 제3자 IT리스크 관리 방안을 수립한다.

금융회사등은 제3자 IT리스크 관리 강화를 위해 계약 체결 단계뿐만 아니라, 운영, 종료 등 단계별 내부통제 체계를 마련한다. 이를 통해 계약 단계부터 사후관리 단계에 이르는 정보처리 위탁 과정 전반의 IT리스크 관리를 강화함으로써 지속적으로 위험을 관리할 수 있는 체계를 마련한다. 우선, 계약 단계에서는 계약 전 IT리스크 평가, IT안전성 확보를 위해 계약상 반영해야 할 필수사항을 규정한다. 그리고 사후관리 단계에서는 계약 준수 모니터링(SLA 평가 등) 및 IT리스크 존재 여부 등을 정기적으로 점검하거나 감사를 수행한다.

또한, 금융회사등은 사후관리 단계로서 제3자 인공지능시스템의 중단 또는 침해사고에 대한 비상 대응 계획을 수립하고 대체 수단 확보 등 적절한 출구전략을 마련한다. 이때, 비상 대응 계획에는 주요 제3자에 위탁한 정보 자산에 대한 백업 대책 수립, 정기적인 비상 대응 훈련 실시 등을 반영한다. 이렇듯 인공지능시스템과 관련한 주기적인 IT리스크 평가, 비상 대응 훈련 실시, 백업 등 업무 연속성을 강화하여 제3자로 인한 IT리스크를 완화하도록 노력한다.

한편, 금융회사등은 제3자의 영향도 및 의존도 등을 파악할 수 있도록 제3자 현황을 상시 식별 및 관리한다. 금융안정 위원회(FSB)는 금융회사의 제3자 의존성 및 잠재적 리스크 식별을 위해 제3자 서비스 등록부(Registers of third-party service relationships) 마련을 권고한 바 있다. 이를 참고하여 위탁하는 정보와 시스템 현황, 중요도 등 제3자의 IT 현황을 관리하는 등록부를 마련하고 최신 상태로 유지한다. 제3자 현황 관리 등록부에는 제3자와의 통신 방식, 재무 건전성, 암호화 대상 및 방식, 시스템 위치, 클라우드 서비스 유형,

재위탁 현황 등을 기록·관리한다. 한편, 금융회사등은 감독당국의 요청 시 제3자 위탁 현황을 지체 없이 제출한다.

아울러, 자체 평가 결과에 따라 중요도 및 의존도가 높은 수탁회사는 주요 제3자로 지정하여 강화된 리스크 관리를 수행한다. 이 경우 시스템 장애 및 침해사고 시 금융회사의 재무적 건전성·법률리스크에 크게 영향을 주거나 위탁 업무의 대체가 어려운 경우 등을 종합적으로 고려하여 주요 제3자를 지정한다.

FSB 등 해외 감독당국은 가이드라인 제정 등을 통해 금융회사등이 IT리스크를 포함하여 제3자 리스크를 체계적으로 관리하도록 권고하고 있다. 금융회사등은 이를 비롯하여 금융보안원의 「금융분야 오픈소스 소프트웨어 활용·관리 안내서」, 「금융분야 인공지능 보안 실무 안내서」 등을 참고하고, 인공지능시스템 관련 제3자 계약 체결 시 포함하여야 할 사항을 고려하여 제3자 IT리스크에 대한 관리·감독 방안을 마련한다.

참고 **주요 국제기구의 제3자 리스크 관리 접근방법 비교**

구분	FSB (금융안정위원회)	BCBS (바젤은행감독위원회)	IOSCO (국제증권감독기구)
접근방식	Toolkit(지침서) 제시 * 각국 상황에 맞게 적용	원칙 기반(12개) 접근	7개 원칙 + 상세 이행 지침
관리체계	관리체계 이사회 및 경영진의 역할과 책임 명시 (리스크관리 프레임워크 승인)		
계약단계별 관리	수명주기 전반 및 정보보안, BCP/테스트, 출구전략 (대체 공급자 확보 등)	계약 전 실사, 서면계약, 온보딩, 모니터링, 계약 종료 등 단계별 관리	실사, 서비스 모니터링, 기밀유지 문제, 집중리스크 등 포괄적 관리
감독당국 역할·초점	시스템 의존성 모니터링, 감독당국과의 정보 공유	시스템 리스크 분석, 국경 간 조정	규제당국의 접근성 보장

참고 **인공지능 관련 제3자 계약 체결 시 포함 사항 (예시)**

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 금융회사등의 접근 및 복구, 반환의 보장에 관한 내용
⑤ 사고 발생 시 정보 공유 및 사고 처리, 배상책임과 관련한 내용
⑥ 인공지능 관련 피해 발생 시 책임 소재
⑦ 인공지능 관련 법규 명령 및 정책 등의 준수 의무
⑧ 안전하고 신뢰할 수 있는 인공지능 개발·활용과 관련 금융회사등의 행동강령, 정책 등 준수

5.4 감독당국 정보 공유 및 보고

✓ **시스템 리스크**로 **확대**될 위험이 있는 **인공지능 사고**가 발생하거나 발생할 우려가 있는 경우, **감독당국과 신속한 정보 공유 및 보고**를 통하여 **시스템 리스크 전이**를 **사전 차단**한다.

금융안정성을 효과적으로 유지·관리하기 위해 감독당국은 각 금융회사등이 개발·이용하고 있는 인공지능 모형(알고리즘, 학습 데이터 등 포함)과 인공지능시스템 운영에 대한 전반적인 내용을 파악하는 것이 중요하다. 이를 위해 금융회사등은 인공지능 활용에 따른 시스템 리스크로 확대될 가능성을 명확히 파악하고, 해당 사고 발생 위험을 선제적으로 인식·점검할 수 있는 장치를 마련한다. 특히 관련 사고가 발생하거나 발생할 우려가 있는 경우에는 감독당국에 즉시 보고한다.

참고 **인공지능 활용 구조 등 감독당국 보고 필요 정보**

□ **상용 인공지능 활용 사례** : ChatGPT등 API 형태로 운영되는 상용 인공지능 활용 사례들의 경우 상용 인공지능서비스에 문제가 생기면 금융회사등으로 리스크가 전이될 수 있기에 이런 경우들에 대해서 감독당국에 해당 정보 보고

□ **오픈소스 인공지능 모델 활용 사례** : 잠재적인 보안 위험 및 모델의 안정성 위험이 존재할 수 있으므로 파운데이션 모델에 대한 정보를 감독당국에 보고

감독당국은 다양한 방식으로 금융권의 인공지능 개발·활용에 관한 정보를 수집한다. 예를 들어, 로보어드바이저 활용 회사는 모델의 신뢰성 확보를 위해 코스콤 로보어드바이저 테스트베드 절차를 거치고, 이 과정에서 감독당국이 알고리즘 및 로보어드 바이저의 작동 방식에 대한 자료를 보유하는 등의 방식이 가능하다. 금융사고의 경우, 기존에 감독당국과 금융회사 간에 구축된 '금융사고 보고 시스템'을 활용한다.

[5. 금융안정성 원칙] 점검 항목

① 인공지능시스템이 금융시장 전반 또는 금융안정에 미칠 수 있는 영향 등 위험을 평가·관리하는 방안을 마련하였는가? YES□ | NO□

② 시스템 위험관리를 위해 모형 오작동 시 백업모형 설계, 비상정지 장치 등 구체적인 대응 방안 등을 반영한 안전장치를 마련하였는가? YES□ | NO□

③ 시스템 위험관리를 위해 인공지능 모형을 오픈소스로 활용하거나 외주 개발하는 경우 아웃소싱에 따른 제3자 리스크를 별도로 평가·관리하는 방안을 마련하였는가? YES□ | NO□

④ 제3자의 영향도 및 의존도 등을 파악할 수 있도록 위탁하는 정보, 시스템 현황, 중요도 등 제3자의 IT 현황 관리 등록부를 마련하고 최신 상태로 유지하고 있는가? YES□ | NO□

⑤ 계약 단계부터 사후관리 단계에 이르는 정보처리 위탁 과정 전반의 단계별 리스크 관리 방안을 마련하였는가? YES□ | NO□

⑥ 정보처리 업무위탁 및 제3자 IT리스크 관리 활동 내역 관련 기록을 문서화하여 보관 및 유지하고 있는가? YES□ | NO□

⑦ 시스템 리스크로 확대될 위험이 있는 인공지능 관련 사고 발생 위험을 선제적으로 점검할 수 있는 장치를 마련하였는가? YES□ | NO□

⑧ 인공지능 관련 위험과 사고에 관한 사항을 감독당국에 보고하는 절차를 마련하였는가? YES□ | NO□

⑨ 감독당국과 정보 공유·의사소통 체계를 마련하였는가? YES□ | NO□

6. 신의성실 원칙

금융회사는 인공지능을 활용한 대고객 서비스를 제공하는 경우 소비자의 이익이 최우선 될 수 있도록 이해상충 방지, 소비자 보호 대책을 마련한다. 인공지능기본법에서도 이용자의 이익이 부당하게 훼손되지 않도록 고영향 인공지능사업자의 책무로 이용자 보호 방안 수립을 규정하고 있다.

6.1. 이해상충 방지

✓ 금융회사등은 대고객 서비스에 인공지능 활용시 **이해상충 문제 발생을 방지**하기 위한 **관리·감독 장치를 마련**한다.

이해상충 문제가 발생할 수 있는 대표 사례로는 인공지능을 활용한 로보어드바이저, 금융상품 비교·추천 등이 있다. 금융소비자보호법(제10조)에 따르면, 금융상품판매업자, 금융상품판매대리·중개업자는 금융상품을 제공하는 경우 금융소비자의 합리적인 선택이나 이익을 침해할 우려가 있는 거래 조건이나 거래 방법을 사용하지 아니할 책무가 있다.

로보어드바이저의 경우 제휴 관계에 있는 브로커 딜러에게 리베이트를 받거나, 증권의 매도인과 매수인 쌍방을 대리하여 거래하거나, 자사의 이익을 우선하는 쪽으로 알고리즘을 프로그래밍하는 등 이해상충의 문제가 발생할 수 있다. 따라서 로보어드바이저 서비스를 제공하는 금융투자업자는 자본시장법상 신의성실의 원칙에 따라 공정하게 금융투자업을 영위하고(자본시장법 제37조 제1항), 정당한 사유 없이 투자자의 이익을 해하면서 자기가 이익을 얻거나 제3자가 이익을 얻지 않도록 해야 한다(자본시장법 제37조 제2항).

자동화된 시스템으로 금융상품을 비교·추천하는 경우에도 알고리즘에 자사 제품을 우선하는 등 이해상충의 문제가 발생할 수 있다. 이에 따라 인공지능 또는 알고리즘을 이용한 서비스 개발 시 선제적인 점검을 통해 이해상충을 방지해야 한다. 금융상품 직접판매업자 또는 금융상품자문업자로 등록하려는 자는 이해상충 행위 방지를 위한 기준이 포함된 소프트웨어를 설치하거나, 이해상충 행위 방지 기준의 문서화, 이해상충 행위 방지를 위한 교육·훈련 체계 수립, 이해상충 행위 방지 기준 위반 시 조치 체계 수립이 필요하다(금융소비자보호법 시행령 제5조 제4항). 마찬가지로 금융상품 판매대리·중개업자의 경우에도 이해상충 방지를 위한 기준이 포함된 소프트웨어 설치를 요건으로 하고 있다(금융소비자보호법 시행령 제6조 제2항 제6호).

참고 **금융소비자 보호에 관한 감독규정(제6조 제7항)에 따른 전자적 장치의 이해상충행위 방지 기준**

1. 금융소비자가 이자율, 개인신용평점 또는 상환기간 등 대출성 상품 계약에 관한 의사결정을 하는 경우에 자신에게 필요한 사항을 선택하여 이에 부합하는 금융상품을 검색할 수 있을 것
2. 제1호에 따른 검색을 하는 경우에 이자율이나 원리금이 낮은 금융상품을 상단에 배치시키는 등 금융소비자의 선택에 따라 금융소비자에 유리한 조건의 우선순위를 기준으로 금융상품이 배열되도록 할 것
3. 제1호에 따른 검색결과를 보여주는 화면에서 검색결과와 관련 없는 동종의 금융상품을 광고하지 않을 것
4. 금융상품직접판매업자가 제공하는 수수료 등 재산상 이익으로 인해 제1호 및 제2호 각각의 기능이 왜곡되지 않을 것

또한 로보어드바이저, 금융상품 비교·추천 이외에 인공지능을 활용한 대고객 서비스를 제공하는 금융회사등의 경우에도 업무의 특성을 고려하여 이해상충 방지 장치를 마련한다. 먼저 인공지능서비스 개발 시 사전테스트를 충분히 진행하여 금융

소비자의 합리적인 이익 침해 여부를 점검한다. 또한, 인공지능서비스를 운영하는 경우에는 금융소비자의 합리적인 선택이나 이익을 침해할 우려가 있는 거래조건이나 거래 방법의 사용 여부 등을 지속적으로 점검·관리한다.

금융회사등은 이해상충행위 방지 기준이 포함된 소프트웨어 설치, 기준의 문서화, 위반 시 조치 체계 수립 등을 통해 이해상충행위를 방지하기 위한 관리체계를 구축한다. 이와 함께 이해상충 방지를 확인할 수 있는 조치문서도 보관한다.

참고 **미국 SEC의 Gensler 의장이 제안한 Predictive Data Analytics (PDA)**

□ 2023년 SEC에서는 예측 데이터 분석 규제(Predictive Data Analytics; PDA)를 제안하면서 투자자문사 및 브로커-딜러의 PDA 사용 시 투자자의 이익보다 기업의 이익을 우선시하는 이해상충 문제 예방을 추진

○ 최근 투자자문사 및 브로커-딜러들이 인공지능 모델 등 예측 데이터 분석 모델 등 신기술을 많이 활용하면서 이 이면에 이해상충의 문제가 발생할 수 있음이 주목받고 있으며, 인공지능 모델 등은 일종의 블랙박스의 형태를 갖기에 내부적으로 모형의 왜곡을 통해 투자자들의 이익보다 자사의 이득을 앞세우는 이해상충의 문제 발생 가능

□ SEC에서 이해상충 방지를 위해 투자자문사 등의 준수가 필요하다고 제안한 3가지 규칙은 다음과 같음

① 활용하는 예측 데이터 모형을 평가하고 투자자의 이득보다 자사의 이득을 앞세우는 이해상충의 문제가 발견될 시 모두 없애거나 중화시켜야 한다는 것

② 본 규칙과 정합성 있는 사내 정책과 절차를 문서로 작성하여 채택, 실행, 유지해야 한다는 것

③ 예측 데이터 모형의 기술, 수정 내역, 테스트 내역, 이해상충 분석 내역 등 여러 평가 내역을 기록하고 보관해야 한다는 것

준수 사례

□ 자사 대출상품 비교·추천 시 이자율이 높거나 수수료가 높은 상품 위주로 추천이 이루어지지 않도록 인공지능 알고리즘을 주기적으로 점검한다.

□ 로보어드바이저 서비스 개발 시 코스콤 등 제3의 기관으로부터 알고리즘의 공정성에 대한 검증을 받고, 모형 변경 등이 발생하는 경우 주기적으로 금융소비자의 합리적 이익 침해 여부를 점검하였다.

□ 카드 상품 비교·추천 시 이해상충 발생을 방지하기 위해 이해상충행위 방지 기준을 내부 문서화해 직원들에게 주기적으로 교육하고, 이해상충 방지 기준을 위반할 경우 조치 체계를 수립하였다.

6.2. 소비자 보호 대책 마련

✓ 인공지능 활용 과정에서 **소비자 보호가 충실히 이루어질** 수 있도록 소비자에게 인공지능 활용 사실을 **사전**에 **고지**하고, 소비자 피해 발생 시 신속한 대응이 가능하도록 **절차**를 **마련**한다.

금융회사등은 인공지능이 허위 과장 광고, 불완전판매 등 소비자의 신뢰를 저해하는 방식으로 활용되지 않도록 소비자를 보호할 수 있는 대책을 마련한다. 이를 위해 먼저 금융소비자에게 영향을 주는 고위험 인공지능 활용 시 해당 금융소비자들에게 인공지능을 이용하였다는 사실을 계약서, 약관, 상품설명서 등 소비자가 확인할 수 있는 방식으로 사전에 고지한다. 인공지능기본법 (제31조)에서도 고영향 인공지능이나 생성형 인공지능을 이용한 제품 또는 서비스를 제공하려는 경우 해당 제품 또는 서비스가 생성형 인공지능에 기반하여 운용된다는 사실을 이용자에게 사전에 고지할 것을 규정하고 있다.

또한, 금융회사등은 인공지능서비스 오류 등을 신고하거나 피드백 체계 등을 운영하여 소비자 보호 대책을 마련한다. 소비자의 서비스

오류 신고 등이 이루어지면 이를 반영하여 시스템을 수정하고, 소규모 피해 발생 전에 시스템 조정이 이루어질 수 있도록 한다. 또한, 소비자 피해가 발생한 경우 소비자 안내, 보상 절차, 보상방법, 보상 시기 등 보상 대책을 사전에 마련해 놓는다. 신용정보법(제36조의2)에서는 자동화평가 결과에 대해 설명을 요구하거나 자동화평가 결과의 산출에 유리하다고 판단되는 정보의 제출, 자동화평가에 이용된 기초정보를 정정하거나 삭제 또는 결과를 다시 산출할 것을 요구하는 행위 등 신용정보주체의 적극적인 권리를 보장하고 있다. 인공지능을 활용한 서비스의 경우에도 적극적인 피드백 절차를 마련하고 소비자의 피드백이 유기적으로 서비스 품질 제고에 반영될 수 있는 체계를 구축한다.

참고 **인공지능 관련 금융소비자 보호 대책 (예시)**

□ 인공지능서비스 제공 시 해당 서비스가 인공지능을 활용하고 있음을 사전고지

□ 인공지능의 성능을 부풀리거나 인공지능을 활용하지 않음에도 인공지능을 활용하는 것처럼 속이는 거짓·과장 광고 금지

□ 인공지능서비스의 출력에 대해 설명가능성 제고

□ 금융소비자가 인공지능서비스 오류 등 위험을 신고할 수 있는 창구 마련
  ○ 기존 고객 대고객 창구 활용 가능

□ 금융소비자가 인공지능서비스에 대해 이의를 제기하고 인공지능이 아닌 기존 서비스 이용을 요구하는 경우 제공이 가능한 체계 마련 등
  ○ 예: 대고객 상담을 챗봇 중심으로 제공하는 경우 반드시 직접 인간 상담원과 연결할 수 있는 채널 및 옵션을 제공해야 함

준수 사례

□ 소비자 인터페이스에 "이 응답이 도움이 되었나요?" 등의 간단한 평가 UI를 제공하고 사용자에게 응답 품질을 1~5점 등으로 평가하게 하여 의견을 수집하였으며, 부적절한 결과물, 편향, 윤리적 문제 등을 쉽게 신고하도록 설계하였다.

□ 사용자의 클릭, 스크롤, 응답 무시, 재입력 등 사용자 행동 로그를 분석하여 만족도를 추정하고, 응답이 반복되거나 비논리적일 때 자동으로 로그를 수집·분석하고 시스템에서 발생한 예외 상황 및 예측 실패 사례를 기록하였다.

□ 수집된 소비자 피드백 데이터를 분류하고 주기적으로 학습에 반영하는 절차를 마련하였다. 인간 검수자가 피드백 데이터를 검토하고 분류하여 학습 데이터의 품질이 보장되도록 하였다.

□ '대화형 인공지능서비스' 이용 시 발생할 수 있는 부적절한 응답 가능성을 사전에 고지하고, 오류 발생 신고 및 문의 가능한 채널을 안내하였다.

[6. 신의성실 원칙] 점검 항목

① 인공지능서비스 개발 시 알고리즘 등의 사전테스트를 통해 금융소비자의 합리적인 이익 침해 여부를 점검하였는가? YES□ | NO□

② 인공지능서비스 운영 시 금융소비자의 합리적인 선택이나 이익을 침해할 우려가 있는 거래조건이나 거래 방법을 사용하는지 지속적으로 점검하였는가?
YES□ | NO□

③ 이해상충 행위 발생 방지를 위해 이해상충 행위 방지 기준이 포함된 소프트웨어를 설치하거나 이해상충 행위 방지 기준의 문서화, 이해상충 행위 방지 기준 위반 시 조치 체계를 수립하였는가? YES□ | NO□

④ 고객에게 영향이 있는 인공지능서비스 제공 시 해당 서비스가 인공지능을 활용하고 있다는 사실을 계약서, 약관, 상품설명서, 화면 내 표시 등을 통해 사전 고지했는가? YES□ | NO□

⑤ 금융소비자가 인공지능서비스 이용 과정에서 서비스 오류 등 위험을 신고할 수 있는 창구를 마련했는가? YES□ | NO□

⑥ 금융소비자가 인공지능서비스에 대해 이의를 제기하고 인공지능이 아닌 기존 서비스 이용을 요구하는 경우 제공이 가능한 체계를 마련했는가? YES□ | NO□

7. 보안성 원칙

금융회사등은 인공지능시스템에 대한 보안성 확보를 위해 인공지능시스템 고유의 새로운 보안 위협을 식별하고 이에 특화된 대응 방안을 마련한다. 또한 기존 IT 보안 관리 체계를 인공지능시스템의 특성을 반영하여 확장 적용하고 개발부터 운영까지 全 과정에 걸쳐 보안성을 검증하고 지속적으로 관리한다.

다만, SaaS 형태의 인공지능시스템(M365 Copilot 등)의 경우, 금융회사등은 인공지능이용사업자로서 인공지능개발사업자가 보안성 확보를 위한 기술적 대책을 충분히 마련하여 운영하고 있는지 지속적으로 확인하고 관리한다.

본 장에서 설명하는 보안성 원칙에 대한 세부적인 사항은 「금융분야 인공지능 보안 실무 안내서(금융보안원)」를 통해 확인할 수 있다.

7.1 인공지능 특화 보안 위협 식별 및 관리

✓ 전통적인 보안 위협과 별개로 **인공지능시스템에 특화된 보안 위협을 체계적으로 식별**하고, **이에 대응하기 위한 전략을 마련**한다.

금융회사등은 데이터 오염, 모델 오염, 데이터 및 모델 정보 유출, 프롬프트 인젝션, 탈옥 공격 등 인공지능 특화 위협을 정기적으로 식별한다. 이러한 위협들은 기존 IT 보안 위협과는 다른 특성을 가지므로 인공지능 특화 위협 모델링 기법을 활용하여 별도의 식별 체계를 구축해야 한다. 식별된 위협을 분류하고 분석하여 위험 수준에 따른 대응 전략을 마련한다. 각 위협의 발생 가능성과 영향도를 평가하여 우선순위를 정하고 이에 맞는 구체적인 대응 방안을 수립한다.

인공지능시스템을 활용함에 있어 충분한 수준의 보안성을 확보하기 위해서는 기획 단계부터 보안성 확보 및 검증 방안을 수립하고, 개발 및 운영 단계에서는 수립한 방안을 차질 없이 실천하고 검증하는 과정이 필요하다.

참고 **인공지능시스템 위협 모델링 방안**

인공지능시스템에 대한 공격 가능성을 사전에 식별하고 방어 전략을 수립하기 위한 핵심 작업으로, 인공지능시스템의 보안성을 확보할 수 있다. 위협을 분류하고 시스템이 노출될 수 있는 공격 벡터를 체계적으로 분석해 모델링함으로써 대응 전략을 마련한다.

□ **일반적인 위협 유형**

① 데이터 조작, ② 시스템 조작 공격, ③ 모델 역추론, ④ 정보 추론, ⑤ 모델 추출, ⑥ 모델 변조, ⑦ 서비스 거부 공격

□ **위협 모델링 방법**

위협 모델링은 다양한 방법으로 수행할 수 있는데 개발 및 운영 조직의 역량과 환경을 고려해 ① 전통적인 보안 위협 모델링(STRIDE 등) 방법 적용, ② 생명주기별 분석, ③ 인공지능 공격 전술·기술 체계(ATTACK for AI) 활용 등의 방식을 선택한다.

□ **위협 모델링 단계**

① 자산 식별 : 학습 데이터, 모델, API, 인프라 등
② 공격 벡터 분석 : 어떤 경로로 접근·조작 가능한지 파악
③ 공격 시나리오 도출 : 데이터 조작, 모델 탈취 등
④ 공격 영향도 분석 : 신뢰성 저하, 개인정보 유출 등
⑤ 보안 통제 방안 설계 : 탐지, 차단, 로그 분석, 복구 등
⑥ 보안성 테스트 및 시뮬레이션 수 : 레드티밍, 적대적 예제 테스트
⑦ 정기적 리뷰 및 업데이트 : 모델 업데이트 시 위협 모델 재검토

준수 사례

□ 인공지능시스템의 핵심 자산인 데이터(학습·참조 데이터), 모델(모델 파일, 시스템 프롬프트), 인프라(API, 서버, 저장소)를 구분하여 자산 목록을 작성하고, 자산 변경 시(모델 변경, 시스템 프롬프트 수정, 인프라 추가 등) 목록을 업데이트하며, 최소 분기 1회 전체 검토를 시행하였다.

□ 외부 접근 경로(API, 웹·모바일 앱, 외부 데이터 수집)와 내부 접근 경로(직원 권한, 개발/운영 환경)를 분석하여 접근 제어 정책에 반영하였다.

□ 대출심사 인공지능에 대해 "외부 공격자가 반복 질의를 통해 모델 판단 기준을 추론하는 시나리오"를 작성하고 대응 방안을 수립하였다.

□ 프롬프트 인젝션 공격이 성공할 경우의 영향(잘못된 정보 제공, 서비스 신뢰도 하락, 규제 제재)을 평가하여 위험 수준 '상'으로 분류하였다.

□ 모델 추출 공격에 대해 API 호출 횟수 제한, 출력 필터링 등 구체적인 통제 방안을 마련하고 시스템에 적용하였다.

□ 모델 버전 업데이트 시 위협 모델을 재검토하고, 최신 공격 기법을 반영하여 대응 방안을 보완하였다.

7.2 인공지능 특화 공격 탐지 및 대응

✓ 식별된 인공지능 특화 보안 위협과 관련된 공격에 대해 **탐지, 차단 및 대응 체계를 구축**한다.

악의적 사용자가 인공지능시스템을 속이거나 잘못된 결과를 유도하기 위해 시도하는 다양한 공격에 대비한다.

외부 사용자 입력이나 외부 데이터를 수집하는 인공지능서비스의 경우 그 내용이 정상적인 범위를 벗어나는지 사전에 검토하는 기능을 구현한다. 특히 대화형 인공지능서비스의 경우 사용자가 시스템의 제약을 우회하여 부적절한 결과를 얻으려는 시도를 탐지 및 차단할

수 있도록 금융서비스의 특성 등을 고려한 대책을 마련한다. 또한, 입출력에 목적 외 고유식별정보 및 개인(신용)정보, 출력에 인공지능 모델 내부 정보 등 기밀정보가 포함되어 있을 경우, 이를 자동으로 탐지하여 마스킹 또는 제거하는 기능 등을 구현한다.

판단형 모델의 경우 적대적 예제 공격에 대한 방어 체계를, 생성형 모델의 경우 프롬프트 인젝션 및 탈옥 공격에 대한 방어 체계를 구축한다. 고영향 및 고위험 인공지능시스템의 경우, 인공지능시스템을 속여 잘못된 결과를 유도하는 공격에 대한 모의훈련을 정기적으로 수행하여 방어 체계의 실효성을 검증한다.

인공지능시스템에 대한 보안 위협을 실시간으로 감지하고 대응할 수 있는 체계를 구축한다. 인공지능시스템의 입출력 데이터, 사용자 접근 패턴, 시스템 성능 지표 등을 지속적으로 모니터링하여 인공지능 특화 보안 위협을 조기에 발견한다.

특히 금융거래와 직접 연관된 인공지능시스템의 경우 기존 보안 모니터링 체계에 적대적 공격 탐지, 비정상적인 추론 패턴 감지, 응답 품질의 급격한 저하 등 인공지능 특화 보안 지표를 추가하여 모니터링하고 이상 징후 발견 시 즉시 대응할 수 있는 체계를 구축한다.

참고 인공지능 특화 보안 위협 (예시)

□ **악의적 질의를 통한 시스템 조작**

- ○ 대화형 인공지능에 부적절한 명령을 숨겨서 전달하여 제약을 우회하는 공격
- ○ 금융상담 챗봇을 속여 잘못된 금융정보를 제공하도록 유도하는 시도

□ **반복적 질의를 통한 모델 정보 추출**

- ○ 대량의 질의를 통해 인공지능 모델의 구조나 학습 데이터를 역추적하는 공격
- ○ 신용평가 모델의 판단 기준을 파악하여 악용하려는 시도

□ **조작된 외부 모델 활용**

- ○ 악의적으로 조작된 오픈소스 인공지능 모델을 통한 시스템 침해
- ○ 의도적으로 편향되거나 오염된 학습 데이터를 포함한 모델 사용

준수 사례

□ 챗봇 입력의 길이를 2,000자로 제한하고, 데이터 타입이 예상 범위를 벗어나면 요청을 자동 거부하도록 설정하였다.

□ "이전 지시 무시", "역할 전환" 등 프롬프트 인젝션 의심 패턴을 키워드 필터로 1차 탐지하고, 가드레일 모델로 2차 정밀 검증하는 체계를 구축하였다.

□ 주민등록번호, 카드번호 등 개인정보 입력 시 경고 메시지를 표시하고 전송을 차단하며, 출력 시 자동 마스킹 처리하였다.

□ 에러 발생 시 모델 경로나 내부 정보 대신 "일시적 오류가 발생했습니다" 등 일반화된 메시지만 반환하도록 설정하였다.

□ 욕설, 차별적 언어 등 부적절한 표현을 탐지하는 필터를 적용하여 유해 콘텐츠 생성을 차단하였다.

□ 신용평가 모델에 대해 반기별로 적대적 예제 공격 시뮬레이션을 시행하고, 탐지율 목표(95% 이상)를 설정하여 관리하였다.

□ 금융상담 챗봇에 대해 주기적으로 프롬프트 인젝션 및 탈옥 시도 모의 공격을 시행하여 부적절한 답변 생성 방지 체계를 검증하였다.

□ API 호출 로그를 분석하여 비정상 접근(평소 대비 5배 이상 증가, 짧은 시간 내 대량 요청 등)을 실시간으로 탐지하는 시스템을 구축하였다.

□ 프롬프트 인젝션 시도가 5회 이상 탐지된 사용자에게 경고하고, 10회 이상 시도 시 30분간 자동 차단하는 정책을 적용하였다.

7.3 인공지능 자산 보호 및 관리

✓ 데이터, 모델 파라미터 등 **핵심 자산이 무단 접근·유출·변조**되지 않도록 암호화, 무결성 검증, 접근통제 등 **보호 대책을 적용**한다.

외부 공격자가 인공지능 모델의 내부 정보를 불법 획득하거나 학습에 사용된 민감한 데이터를 추출하려는 시도를 방지한다.

금융회사등은 인공지능 모델에 과도한 질의나 특정 패턴의 반복적인 접근을 제한하여 모델의 구조나 학습 정보가 외부로 유출되지 않도록 해야 한다. 이를 위해 사용자별 질의 횟수 제한, 접근 빈도 모니터링 등의 통제 방안을 마련한다.

고위험 인공지능시스템의 경우 반복 질의를 통한 모델 구조 추론, 학습 데이터 추출, 시스템 프롬프트 노출 등 내부 정보 유출 공격에 대한 모의훈련을 정기적으로 수행하여 방어 체계의 실효성을 검증한다.

아울러, 사용되는 데이터 및 모델의 목록을 관리하고 데이터 및 모델에 대한 접근 권한과 변경·삭제에 대한 기록을 관리하여 무단 접근·유출·변조가 발생하지 않도록 한다. 필요시, 암호화 및 무결성 검증 등을 적용하여 보호 대책을 강화한다.

준수 사례

□ API 응답에서 모델 정보가 노출되지 않도록 응답 형식을 최소화하고, 에러 시 일반화된 메시지만 반환하도록 설정하였다.

□ 일반 사용자는 시간당 100회, 일일 500회로 API 호출을 제한하고, 초과 시 제한 안내 메시지를 반환하였다.

□ 대출심사 모델에 대해 반기별로 반복 질의를 통한 판단 기준 역추론 시도 및 학습 데이터 추출 공격 시뮬레이션을 시행하고, 내부 정보 유출 가능성을 점검하였다.

□ 금융상담 챗봇에 대해 다양한 질의로 시스템 프롬프트를 노출시키는 공격 시뮬레이션을 수행하고, 프롬프트 보호 체계를 검증하였다.

□ 외부 LLM API를 사용하는 챗봇에 대해 “당신은 어떤 모델인가요?”, “버전이 몇인가요?” 등의 질의를 통해 모델 정보가 노출되는지 테스트하고, 응답에서 구체적인 모델명이 노출되지 않도록 필터를 적용하였다.

□ 모든 인공지능 모델에 대해 모델명, 버전, 출처, 생성일, 담당자, 용도, 보안등급을 기록한 목록을 관리하고 분기별로 업데이트하였다.

□ 금융상담 챗봇의 시스템 프롬프트를 버전별로 관리하고, 변경 이력과 사유를 문서화하였다.

□ 형상 관리 도구를 활용하여 모델 코드와 설정 파일의 변경 이력을 자동으로 추적하고, 주요 변경은 승인 절차를 거치도록 하였다.

□ 학습 완료된 모델 파일의 해시값을 생성하여 저장하고, 배포 전 해시값을 비교하여 변조 여부를 확인하였다.

□ 개발자에게는 모델 읽기·쓰기 권한을, 평가자에게는 읽기 권한을, 운영자에게는 배포 권한만을 부여하고 분기별로 권한을 검토하였다.

□ 외부 API 키를 암호화 저장소에 보관하고, 코드에서는 환경 변수로만 참조하며, 주기적으로 키를 재발급하였다.

7.4 외부 모델 및 데이터 검증

✓ **외부에서 도입하는 모델·데이터**에 대해 **보안 및 신뢰성 검증을 수행**하여 **공급망 위험을 최소화**한다.

외부에서 제공하는 인공지능 모델이나 오픈소스 모델을 활용하는 경우, 해당 모델이 악의적으로 조작되지 않았는지 사전에 검증한다. 모델의 출처와 개발 이력을 확인하고 신뢰할 수 있는 제공자로부터만 모델을 도입한다. 사전 학습된 오픈소스 모델을 활용할 때는 모델 다운로드 전에 파일 무결성을 검증하고 알려진 보안 취약점이 있는지 확인한다.

학습 데이터의 경우에도 외부에서 제공받는 데이터가 의도적으로 조작되거나 오염되지 않았는지 검증하는 절차를 마련한다. 데이터 출처의 신뢰성을 확인하고 품질 검사를 통해 의도적으로 조작된 내용, 비정상적인 패턴이나 편향된 정보가 포함되어 있지 않은지 확인한다.

실시간 또는 주기적으로 외부 데이터를 수집하여 인공지능 모델의 학습 데이터 또는 참조 데이터로 사용하는 경우, 데이터 오염 공격에 대한 보안 위협 평가를 시행한다.

아울러, 공급망 보안 관점에서 기존 공급망 보안 영역에 더하여 인공지능 생태계의 특수한 구조와 의존성을 고려한 관리를 적용한다. 인공지능 모델, API, 개발 프레임워크 등에 대한 사전 검증을 수행하고, 외부 인공지능 도구의 경우 학습 데이터 출처, 모델 무결성, 편향성 문제 등을 추가로 확인한다. 외부 인공지능서비스 제공업체와의 계약 시에는 인공지능 특화 보안 위협 대응 능력과 거버넌스 수준을 평가하며 한 곳의 공급업체에 과도하게 의존하지 않도록 위험을 분산하여 업무 연속성을 보장하도록 노력한다.

준수 사례

□ 외부 데이터 공급 업체와 계약 시 데이터 수집 방법, 품질 관리 체계, 보안 인증 보유 여부를 확인하고, 계약서에 출처, 수집일자, 담당자 정보를 명시하여 관리대장에 기록하였다.

□ 외부 제공 데이터에 대해 샘플링 검증으로 개인정보 수집 동의서 첨부 여부를 확인하고, 통계 기법으로 이상치를 탐지한 후 도메인 전문가 검토를 거쳐 오류 데이터를 제거하였으며, 정제 데이터 수량, 비율, 적용 방법을 데이터 품질 보고서에 기록하였다.

□ RAG 시스템 참조문서 및 외부 데이터 수집 시 자동화된 키워드 필터(금융사기, 허위정보 등)와 통계적 이상치 탐지를 적용하여 의심 데이터를 사전 차단하고, 샘플링 검증을 통해 부적절한 내용 포함 여부를 확인하여 이상 발견 시 격리하는 절차를 마련하였다. 고영향 시스템에 대해서는 의도적으로 오염된 테스트 데이터를 소량 투입하여 탐지 체계를 연 1회 검증하였다.

□ 오픈소스 모델 다운로드 시 제공자, URL, 다운로드 일시, 담당자를 자산관리 대장에 기록하고, 다운로드 수, 업데이트 주기, 제공자 신뢰도를 확인하는 체크리스트를 적용하였으며, 다운로드 후 제공된 해시값과 실제 파일 해시값을 비교 검증하였다.

□ 오픈소스 인공지능 모델 채택 시 편향성 분석 툴을 사용하여 데이터의 편향성 지수를 측정하고, 잠재적인 차별적 결과 위험을 확인 후 모델을 미세 조정하였다.

□ 신규 인공지능 API 도입 전 보안 전문가팀이 입력 검증 및 접근통제 메커니즘에 대한 보안성 테스트를 수행하고 결과 보고서를 작성하여 승인 절차를 거치도록 하였다. 운영 중인 인공지능 라이브러리의 보안 취약점은 월 1회 확인하고, 중대한 취약점 발견 시 48시간 내 패치를 적용하였다.

□ 외부 인공지능서비스 계약 시 입력 데이터의 학습 활용 금지, 데이터 국내 보관, 제3자 제공 금지를 계약서에 명시하고, 인공지능 모델 보안 위협 대응 전담 조직 및 기술적 통제 수준을 정기적으로 평가하였다.

□ 핵심 인공지능 모델에 대해 2개 이상의 클라우드 공급자를 활용하거나 자체 개발 모델을 백업 옵션으로 준비하여, 특정 공급자의 장애 발생 시 서비스가 중단되지 않도록 위험을 분산하였다.

7.5 기존 보안 관리의 인공지능 확장 적용

✓ 전통적인 보안 영역의 경우 **기존 IT 보안 체계를 기반**으로 하되, 인공지능시스템의 특성에 맞게 **확장하여 적용**한다.

접근통제 및 권한 관리 영역에서는 기존 체계에 더하여 인공지능 시스템의 특수성을 반영한 추가 통제를 적용한다. 인공지능 시스템의 동작과 직접 관련된 모델과 학습 데이터에 대해서는 프로그램 소스 코드와 동일한 수준의 엄격한 접근통제를 적용한다. 또한 인공지능시스템의 자동화된 접근 패턴을 별도로 모니터링하고 업무 목적에 필요한 최소한의 권한만 부여한다.

네트워크 보안 및 데이터 보호 영역에서는 인공지능서비스 이용 시 전송되는 데이터의 특성을 고려한 보안 통제를 강화한다. 클라우드 기반 인공지능서비스나 상용 생성형 인공지능서비스 이용 시에는 데이터의 국외 이전 가능성을 확인하고 금융회사등이 이용한 데이터가 학습 등 모델 개선 목적으로 재사용되지 않도록 약관 또는 계약서 등에 명시하였는지 사전에 확인한다.

인공지능 보안 교육 및 인식 제고를 위해 기존 정보 보호 교육 외에 인공지능 특화 보안 위협에 대한 교육을 시행한다. 인공지능 시스템 개발·운영 담당자에게는 적대적 공격 등 인공지능 특화 위협과 대응 방법에 대한 전문 교육을 제공하고, 일반 사용자에게는 안전한 인공지능서비스 이용 방법과 인공지능을 악용한 고도화된 피싱 공격에 대한 경각심을 높이는 교육을 시행한다.

참고 인공지능 활용에 따른 주요 보안 위협 요인 (예시)

□ **고도화된 피싱 공격**

- ○ 생성형 인공지능으로 제작한 가짜 영상, 이미지 또는 오디오로 피해자가 신뢰하는 지인들을 흉내 내 더욱 식별하기 어려운 보이스피싱 등을 수행
- ○ 생성형 인공지능을 활용하여 피해자의 SNS 내용 등을 통해 맞춤형 피싱 메일 제작을 자동화, 대규모 스피어피싱 공격 수행

□ **악성코드 생성 증가 및 고도화**

- ○ 특수 제작된 생성형 인공지능(예: Worm GPT, DarkBard 등)을 악용하여 악성코드 대량 생산 가능
- ○ 생성형 인공지능을 사용하여 기존 악성코드 탐지 시그니처를 우회하는 등 고도화 가능(예: BlackMamba, DeepLocker)

□ **중요 정보 유출**

- ○ 금융회사 내부에서 상용 생성형 인공지능에 고객 정보나 기밀정보를 활용할 경우 데이터 유출 위험 발생
- ○ 또한, 금융회사가 제공하는 생성형 인공지능서비스를 대상으로 공격자가 적대적 공격을 통해 내부 데이터 추출 가능

□ **공급망 보안 이슈**

- ○ 상용 또는 오픈소스의 생성형 인공지능 모델을 사용할 때, 해당 상용 모델 공급자 및 오픈소스 관련 제3자 위협에 노출 우려

□ **생성형 인공지능 모델 및 출력 조작**

- ○ 모델의 학습 데이터를 조작하여 잘못된 정보를 바탕으로 투자를 추천하는 등, 모델이 부정확하고 공격적이거나 해로운 출력이 가능

[출처] 싱가포르 금융청, Cyber Risks Associated with Generative Artificial Intelligence, 2024.

준수 사례

□ 인공지능 모델 저장소와 학습 데이터 저장소에 대해 개발 서버 소스 코드와 동일하게 접근 권한 그룹을 세분화하고, 접근 시 사용자 인증 및 승인을 받도록 하였다.

□ 인공지능 파이프라인에서 사용되는 서비스 계정 키는 별도의 보안 저장소에 보관하고, 파이프라인 실행 로그를 시스템 접근 로그와 분리하여 자동 모니터링하고 있다.

□ 모델 평가 담당자에게는 모델 결과 및 평가 데이터에 대한 읽기 권한만 부여하고, 모델 가중치 파일에 대한 쓰기/삭제 권한은 최소 인원의 모델 개발자에게만 부여하였다.

□ 인공지능서비스 호출 시 발생하는 모든 통신에 TLS 1.3 암호화를 의무 적용하고, 데이터 유실 방지 시스템을 통해 전송 데이터에 포함된 민감 정보를 탐지 및 차단하고 있다.

□ 상용 LLM 서비스 이용 계약 시 데이터가 국내 데이터센터에만 저장되도록 하는 데이터 주권 및 보존 조건을 명시하고, 국외 이전을 금지하는 계약 조항을 포함하였다.

□ 프라이빗 모드를 제공하는 외부 LLM 서비스를 계약하여 사용자 입력 데이터가 모델 학습에 재사용되는 것을 원천적으로 차단하고 이를 내부 정책으로 확정하였다.

□ 외부 인공지능 모델 공급자와의 계약 시, 사용자 질의 데이터는 익명화를 거쳐 서비스 품질 개선 목적으로만 제한적으로 사용하며, 마케팅 등 다른 목적으로 재사용하는 것을 명시적으로 금지하였다.

□ 매년 의무적으로 이수해야 하는 정보보호 교육에 '안전한 생성형 인공지능서비스 활용 가이드라인' 섹션을 추가하고, 프롬프트 인젝션 방지 요령을 필수 교육 내용으로 포함하였다.

□ 인공지능 개발 인력을 대상으로 분기별 1회 외부 전문기관을 통해 OWASP Top 10 for LLMs 기반의 보안 코딩 및 모의 해킹 교육을 시행하고 있다.

□ '인공지능 활용 보안 수칙'을 게시하고, 신규 입사자 교육 및 정기 사내 뉴스레터를 통해 민감 정보가 포함된 프롬프트 입력의 위험성을 지속적으로 고지하고 있다.

□ 최근 인공지능 음성 합성 기술을 활용한 보이스피싱 사례를 바탕으로 모의 피싱 이메일 훈련을 시행하고, 직원들이 비정상적인 요청에 대해 인공지능 악용 여부를 의심하도록 인식 개선 교육을 강화하였다.

7.6 인공지능시스템 보안성 검증 및 운영 관리

✓ 인공지능시스템의 **보안성을 개발 단계부터 체계적으로 검증**하고, 운영 과정에서 **지속적으로 관리**한다.

금융회사등은 인공지능시스템 개발 완료 후 운영 시스템으로 이관하기 전 단계에서 보안성 검증을 수행하고, 운영 중에도 정기적으로 재검증을 한다. 검증은 금융회사 자체 전문인력을 활용하거나 인공지능 보안 전문기업에 위탁하여 수행할 수 있으며, 금융보안원 등 신뢰할 수 있는 제3자 기관의 보안성 검증 지원을 받는 것을 권장한다. 특히 고위험 인공지능시스템의 경우 외부 전문기관의 독립적 검증을 통해 보안성을 확인할 수 있다. 검증 결과 취약점이 발견된 경우 즉시 보완 조치를 취하고, 보완 완료 후 재검증을 통해 문제 해결을 확인한다. 검증 과정과 결과는 문서화하여 관리한다.

인공지능시스템 운영 중에는 새로운 보안 위협의 등장, 시스템 환경 변화, 모델 업데이트 등에 대응하여 지속적인 보안 관리를 수행한다. 정기적인 보안 점검 일정을 수립하여 시스템의 보안 상태를 점검하고, 새로운 위협 정보나 보안 패치가 공개될 경우 즉시 영향도를 평가하여 필요한 조치를 취한다. 특히 인공지능 모델이나 학습 데이터가 업데이트되는 경우, 보안성에 미치는 영향을 사전에 평가하고 필요시 추가 검증을 수행한다. 보안사고 발생 시에는 즉시 원인을 분석하고 재발 방지를 위한 개선 방안을 수립하여 보안 체계에 반영한다. 또한 보안사고 사례와 대응 경험을 축적하여 조직 내 보안 역량 향상에 활용한다.

금융회사등이 특수한 환경에서 인공지능시스템을 운영하는 경우 해당 환경의 특성을 고려한 추가적인 보안 관리를 수행한다. 상용 생성형 인공지능을 내부망에서 이용하고자 하는 경우 망 분리 규정 예외 적용을 위한 혁신금융서비스 지정 시, 관련 보안대책을 충실히

이행해야 한다. 오픈소스 인공지능 도구를 활용하는 경우에는 금융분야 오픈소스 소프트웨어 활용·관리 안내서에 따라 라이선스 관리, 보안 취약점 점검, 업데이트 관리 등을 체계적으로 수행한다. 클라우드 환경에서 인공지능서비스를 이용하는 경우에는 클라우드 제공업체와의 보안책임 분담 모델을 명확히 하고, 데이터 처리 지역, 암호화 수준, 접근 권한 관리 등에 대한 추가적인 보안 통제를 적용한다.

참고 **혁신금융서비스(생성형 인공지능 연계) 지정 시 보안대책**

금융회사등이 상용 생성형 인공지능을 내부망에서 이용하고자 하는 경우 망 분리 규정 예외 적용을 받기 위해서 혁신금융서비스 지정이 필요하여, 이때 강화된 보안대책 적용·이용 평가가 부가 조건으로 부과된다. 강화된 보안대책 예시는 아래와 같다.

[생성형 인공지능 연계 이용 보안대책 평가 기준]

분류	평가 기준
1. 생성형 인공지능 운영·관리 보안 대책	1.1. 해당 생성형 인공지능서비스 활용 전반을 관리·감독하는 책임자(이하 "인공지능 관리자"라고 한다)를 지정하였는지 여부
	1.2. 해당 생성형 인공지능서비스 입출력 데이터 로그(입출력 데이터 내역 포함)를 기록하고, 인공지능 관리자 등은 중요 정보 유출이나 위험성 등을 감시할 수 있도록 절차가 마련되어 있는지 여부
	1.3. 해당 생성형 인공지능 모델에 대한 취약점이나 문제가 확인될 경우 활용을 즉시 중단하고, 출력값 통제 등 관리 조치를 이행하도록 절차가 마련되어 있는지 여부
	1.4. 해당 서비스와 관련된 사내 임직원을 대상으로 안전한 생성형 인공지능 활용을 위한 안내 및 정기·수시 교육을 시행하는지 여부
	1.5. 해당 생성형 인공지능서비스를 금융소비자에게 제공하는 경우 해당 서비스가 생성형 인공지능을 활용하고 있음을 사전 고지하는지 여부 * 해당 생성형 인공지능서비스의 성능을 부풀리거나, 생성형 인공지능을 적용하지 않는 부분까지 활용하는 것처럼 고지하는 등 부당한 거짓·과장광고를 하지 않는지 확인
	1.6. 안전하고 신뢰할 수 있으며 책임 있는 생성형 인공지능 활용을 위해 자사 여건에 맞는 정책(거버넌스 등)을 마련하여 운영하는지 여부

분류	평가 기준
2. 생성형 인공지능 모델 보안대책	2.1. 상용 생성형 인공지능 모델 제공자가 적대적 공격 방지 대책을 수립·이행하는지를 확인하는지 여부
	2.2. 생성형 인공지능 모델의 강건성 확보를 위해 필요 보안대책을 마련하여 이행하는지 여부 * 필요 보안대책: 이용자 입·출력값 필터링, 프롬프트에 질의 제한 사항 명시, 별도 LLM 모델 추가 사용을 통한 공격 탐지 및 방어 등
	2.3. 입출력 데이터를 대상으로 적대적 공격 여부를 확인하고, 적대적 공격에 대비한 대응 방안을 마련하는지 여부
	2.4. 개인신용정보 등 중요 정보가 생성형 인공지능 모델에 입력되지 않도록 방지 조치를 이행하는지 여부
	2.5. 출력 데이터 또는 에러 메시지 內 중요 정보나 인공지능 모델 정보 등이 노출되지 않도록 조치하는지 여부
	2.6. 이용자의 생성형 인공지능 요청 및 결과 출력 횟수 등을 일정 수준 이하로 제한하는지 여부
	2.7. (학습·참조 데이터 활용 시) 데이터 위변조 방지를 위한 보안대책을 마련하는지 여부 * 위변조 방지 보안대책: 형상 관리, 위변조 방지 조치, 이상치 데이터 및 데이터 변조 식별 및 조치 등
	2.8. (추가 학습을 통한 특화 생성형 인공지능 모델 구축 시) 신규 모델에 대한 위변조 방지 조치 및 적대적 공격 방지 조치를 이행하는지 여부
3. 내부 단말기 보안대책 (모바일 단말 및 관리자 단말기 포함)	3.1. 생성형 인공지능에 질의하는 단말기는 개인신용정보(가명 처리 정보는 제외) 등 중요 정보가 유출되지 않도록 방지 대책* 마련 여부 * (예시) 중요 정보 업로드 모니터링 및 차단, 중요 정보 암호화 저장, 매체 통제 등
	3.2. 생성형 인공지능 클라우드 관리자 단말기 ↔ 생성형 인공지능 클라우드 간 인가된 단말기만 접속할 수 있도록 구성하고 있는지 여부
	3.3. 생성형 인공지능 클라우드 관리 콘솔에 대한 비인가 접근 방지를 위한 안전한 인증 방식 적용 여부 * (예시) 관리자 멀티팩터(MFA) 인증 적용, ID별 접근 가능 단말기 지정 등
	3.4. 생성형 인공지능 클라우드 관리자의 활동 내역을 로깅 및 모니터링하고, 이상 징후 발견 시 보안 조치를 이행하는지 여부
4. 내부망 ↔ 외부 인공지능 모델 연계 네트워크 보안대책 (암호화 등)	4.1. 생성형 인공지능 활용을 위한 `외부망 연계' 구간에 전자금융감독규정 제60조 제1항 제5호에 따라 전용회선 또는 VPN을 사용하였는지 여부
	4.2. 생성형 인공지능 활용을 위한 `인공지능 모델 연계' 구간에 안전한 암호알고리즘을 사용하여 전송자료를 암호화하였는지 여부 ※ 안전한 암호 알고리즘 기준은「KISA 암호 알고리즘 및 키 길이 이용 안내서」등을 참고

준수 사례

[보안성 검증 체계]

□ 인공지능 모델 및 서비스 코드에 대해 운영 환경과 유사한 환경에서 인공지능 레드티밍 및 IT 보안 레드티밍을 수행한 후 운영 이관을 승인하였다.

□ 운영 중인 인공지능 기반 고객 상담 챗봇에 대해 연 1회 최신 취약점 목록을 반영한 정기 보안 재점검을 시행하고 그 결과를 문서화하였다.

□ 고영향 인공지능에 대해 내부 검토와 별도로 금융보안원을 통한 인공지능 레드티밍을 수행하였다.

□ 대출심사 인공지능을 고영향으로 분류해 검증 범위를 모델, 데이터, 인프라, 인터페이스, 운영 절차로 명확히 정의하고, 업무 특성에 맞는 구체적인 보안 기준(차단율, 호출 제한 등)을 설정하였다.

□ 설계 단계에서 위협 모델링을 실시하고, 구현 단계에서 단위 테스트를, 운영 이관 전 통합 보안 진단을 수행하였다.

□ 프롬프트 인젝션 취약점에 대한 방어 로직을 적용한 후, 이전 공격 구문 및 변형된 새로운 공격 구문을 사용하여 재차 모의 테스트를 시행하였고, 정상적으로 방어됨을 확인한 후 최종 완료 처리하였다.

□ 모든 인공지능 모델별 보안 검증 보고서를 모델 버전 관리 시스템에 연동하여 보관하고, 유사 취약점 발생 시 해당 보고서를 참고하여 신속한 대응 및 재발 방지에 활용하고 있다.

[운영 중 보안 관리]

□ 주기적으로 최신 인공지능 특화 보안 위협 동향을 분석하여 운영 중인 인공지능 모델에 적용할 방어 로직 업데이트 필요성을 검토하고, 분기별로 운영 환경의 보안 구성을 재검토하고 있다.

□ 매월 첫째 주에 인공지능 모델 서빙 서버의 운영체제 보안 패치 및 설정 취약점을 점검하고, 미비점을 즉시 개선하도록 조치하였다.

□ 특정 LLM 모델의 심각한 취약점에 대한 공개 경고를 확인한 즉시, 운영 중인 LLM 서비스에 미치는 영향도를 평가하고 48시간 이내에 보안 업데이트를 완료하였다.

□ 새로운 버전의 모델을 배포하기 전에 기존 모델과의 보안성 차이를 평가하고, 모델 편향성 변화로 인한 새로운 윤리적·보안적 위험이 없는지 확인하는 절차를 거쳤다.

□ 모델 구조를 RAG(검색 증강 생성) 방식으로 변경하는 대규모 업데이트 시, RAG 시스템에 특화된 프롬프트 인젝션 및 외부 DB 접근 취약점에 대한 추가 보안성 검증을 시행하였다.

□ 모델 추출 시도 사고 발생 후, 로그 분석을 통해 API 접근 제어 정책의 미비점을 원인으로 파악하고, 접근 횟수 제한 로직을 강화하는 재발 방지 개선 방안을 수립 및 적용하였다.

7.7 보안 목적 인공지능 활용

✓ **'AI 공격은 AI로 방어한다.'**는 방향 아래 **금융서비스의 보안성을 강화**하는 동시에 취약점 탐지, 보안패치 적용, 위협 분석 및 대응 등 **보안 업무**에 **인공지능을 활용하는 방안**을 **적극적으로 검토**한다.

앤트로픽사의 미토스와 같은 초고성능 인공지능이 사이버공격에 악용될 우려가 커지면서 보안에도 인공지능을 활용할 필요성이 커지고 있다. 인공지능이 스스로 취약점을 찾아내고 변종 공격까지 실시간으로 만들어 내면서, 인공지능을 악용한 사이버공격은 인간 보안 전문가의 인지 능력과 기존 시스템의 규칙 기반 탐지 방어선을 무력화하고 침투하기에 충분한 수준으로 진화하고 있기 때문이다.

금융회사등은 'AI 공격은 AI로 방어한다.'는 방향 아래 사이버 위협 예방·탐지·대응 전 과정을 자동화·고도화할 수 있도록 '보안 목적 AI'의 도입과 활용을 적극적으로 검토한다. 이를 통해 대규모 금융거래 데이터와 네트워크 트래픽을 인공지능 기술을 활용하여 실시간으로 학습하고 분석함으로써 인공지능에 의한 사이버공격을 적시에 포착할 수 있을 것이다. 고성능 인공지능은 알려지지 않은 제로데이 취약점과 미세한 이상 징후를 선제적으로 식별함으로써, 위험이 현실화하기 전 금융회사가 예방적 조치를 취할 수 있도록 지원할 수 있다.

'보안 목적 AI'를 적극적으로 활용하는 것은 기술 발전에 따라 급변하는 위협 환경 속에서 금융 생태계의 사이버 복원력을 확보하고 선제적 대응 역량을 강화하기 위해서도 매우 중요하다. 금융회사는 인공지능 활용을 통해 사이버 위협에 대한 대응 시차를 최소화하고 탐지 정확도를 높여 금융 자산과 소비자 데이터를 더욱 안전하게 보호한다. 이를 통해 진화하는 사이버 위협에 유연하고 능동적으로 대응할 수 있는 금융 보안 프레임워크를 구축할 수 있을 것으로 기대된다.

금융회사등은 보안 등의 목적을 위해 인공지능을 활용하는 과정에서 망분리를 완화할 경우, 보안상 취약점 등이 발생하지 않도록 아래와 같은 망분리 대체 통제를 적용한다.

참고 **망분리 대체 정보보호통제(보안 목적 AI · SaaS)**

- ✓ 국내외 클라우드 보안인증 중 1개 이상 보유 서비스만 이용 가능 (①SOC2, ②FedRAMP, ③ISO27017, ④CSAP, ⑤MTCS, ⑥ISMS-P, ⑦CSA STAR level2 이상)
- ✓ 금융회사의 정보처리 업무 위탁 규정 준수
- ✓ 생성형 AI 서비스의 경우 유료 비즈니스 등 관리형 버전 이용(미학습 약정 필수)
- ✓ AI、SaaS 서비스 관련 보안사고 및 장애 대응 절차 수립
- ✓ 허용된 AI、SaaS 외 인터넷접속, 외부 API, 제3자 앱 등 외부 연계 통제
- ✓ API key, 관리자 계정 등 주요 인증정보에 대한 보호、점검 및 관리절차 수립
- ✓ 접속 단말기 및 사용자、관리자 등록、관리, AI、SaaS 관리자 계정 등 다중 인증 방식 적용, 최소 권한 부여 등 접근 제어 및 권한 관리, 단말기 보호대책 수립、적용
- ✓ 개인신용정보, 고유식별정보 등 중요정보 입력、처리、외부전송、유출 여부 모니터링 및 통제
- ✓ AI、SaaS 내 데이터의 불필요한 공유、처리 방지
- ✓ AI、SaaS 이용을 위한 네트워크 구간에 대한 보호대책(암호화 등) 수립、적용
- ✓ AI、SaaS 정보보호 통제를 위한 상시 관리、체계 확보
- ✓ 접속、이용 내역, 관리자 활동, 이상행위 등에 대한 모니터링 및 로그 수집(1년 이상 보존)

[7. 보안성 원칙] 점검 항목

[인공지능 특화 보안 위협 식별 및 관리]

① 인공지능 특화 위협 모델링 기법을 활용하여 별도 위협 식별 체계를 구축하고, 인공지능 특화 위협을 정기적으로 식별하는 절차를 마련했는가? YES□ | NO□

② 식별된 위협을 분류하고 분석하여 위험 수준에 따른 대응 전략을 마련했는가? YES□ | NO□

③ 각 위협의 발생 가능성과 영향도를 평가하여 우선순위를 정하고, 이에 맞는 구체적인 대응 방안을 수립했는가? YES□ | NO□

④ 기획 과정부터 보안성 확보 및 검증 방안을 수립하고, 개발 및 운영 시 이에 따라 보안성을 확보하고 검증하는가? YES□ | NO□

[인공지능 특화 공격 탐지 및 대응]

⑤ 외부 사용자 입력이나 외부 데이터를 수집하는 인공지능서비스의 경우, 그 내용이 정상적인 범위를 벗어나는지 사전에 검토하는 기능을 구현했는가? YES□ | NO□

⑥ 대화형 인공지능서비스의 경우 시스템 제약을 우회하여 부적절한 결과를 얻으려는 시도를 탐지 및 차단할 수 있도록 입출력 필터링 체계를 마련했는가? YES□ | NO□

⑦ 입출력 데이터에서 목적 외 고유식별정보 및 개인(신용)정보 포함 여부를 탐지하고 마스킹 또는 차단하는 기능을 구현했는가? YES□ | NO□

⑧ 출력 데이터와 에러 메시지에서 인공지능 모델 정보(구조, 프롬프트 등) 등 기밀정보가 노출되지 않도록 방지하는 기능을 구현했는가? YES□ | NO□

⑨ 판단형 모델의 경우 적대적 예제 공격에 대한 방어 체계를, 생성형 모델의 경우 프롬프트 인젝션 및 탈옥 공격에 대한 방어 체계를 구축했는가? YES□ | NO□

⑩ 고영향·고위험 시스템의 경우 인공지능을 속여 잘못된 결과를 유도하는 공격에 대한 정기적 모의훈련을 수행하여 방어 체계 실효성을 검증하는가? YES□ | NO□

⑪ 인공지능시스템에 대한 보안 위협을 실시간으로 감지하고 대응할 수 있는 체계를 구축했는가? YES□ | NO□

[인공지능 자산 보호 및 관리]

⑫ 사용자별 질의 횟수 제한, 접근 빈도 모니터링 등의 통제 방안을 마련하여 인공지능 모델에 과도한 질의나 특정 패턴의 반복적인 접근을 제한하였는가? YES□ | NO□

⑬ 고영향 및 고위험 인공지능시스템의 경우 반복 질의를 통한 모델 구조 추론, 학습 데이터 추출, 시스템 프롬프트 노출 등 내부 정보 유출 공격에 대한 모의훈련을 정기적으로 수행하여 방어 체계의 실효성을 검증하는가? YES□ | NO□

⑭ 사용되는 데이터 및 모델의 목록을 관리하고, 데이터 및 모델에 대한 접근 권한과 변경·삭제에 대한 기록을 관리하는가? YES□ | NO□

[외부 모델 및 데이터 검증]

⑮ 외부에서 제공하는 인공지능 모델이나 오픈소스 모델을 활용하는 경우, 해당 모델이 악의적으로 조작되지 않았는지 사전에 검증하는 절차를 마련했는가?
YES□ | NO□

⑯ 모델의 출처와 개발 이력을 확인하고, 신뢰할 수 있는 제공자로부터만 모델을 도입하도록 하는가? YES□ | NO□

⑰ 사전학습된 오픈소스 모델을 활용할 때는 모델 다운로드 전에 파일 무결성을 검증하고, 알려진 보안 취약점이 있는지 확인하는가? YES□ | NO□

⑱ 외부에서 제공받는 데이터 출처의 신뢰성을 확인하고, 의도적으로 조작되거나 오염되지 않았는지 검증하는 절차를 마련했는가? YES□ | NO□

⑲ 실시간 또는 주기적으로 외부 데이터를 수집하여 인공지능 모델의 학습 데이터 또는 참조 데이터로 사용하는 경우, 데이터 오염 공격에 대한 보안 위협 평가를 실시하는가? YES□ | NO□

⑳ 인공지능 생태계의 특수한 구조와 의존성을 고려하여 공급망 보안을 관리하는가? YES□ | NO□

㉑ 인공지능 모델, API, 개발 프레임워크 등에 대한 사전 검증을 수행하고, 외부 인공지능 도구의 경우 학습 데이터 출처, 모델 무결성, 편향성 문제 등을 추가로 확인하는가? YES□ | NO□

㉒ 외부 인공지능서비스 제공 업체와의 계약 시에는 인공지능 특화 보안 위협 대응 능력과 거버넌스 수준을 평가하는가? YES□ | NO□

[기존 보안 관리의 인공지능 확장 적용]

㉓ 인공지능시스템의 동작과 직접 관련된 모델과 학습 데이터에 대해서는 프로그램 소스 코드와 동일한 수준의 접근 통제를 적용하는가? YES□ | NO□

㉔ 인공지능시스템의 자동화된 접근 패턴을 별도로 모니터링하고, 업무 목적에 필요한 최소한의 권한만 부여하는가? YES□ | NO□

㉕ 인공지능시스템 오작동 시 피해를 최소화할 수 있도록 긴급정지 기능을

구축하였는가? YES□ | NO□

㉖ 클라우드 기반 인공지능서비스나 상용 생성형 인공지능서비스 이용 시 데이터의 국외 이전 가능성을 확인했는가? YES□ | NO□

㉗ 외부 인공지능서비스 이용 시 모델 개선 목적의 데이터 재사용을 금하도록 사전에 확인하고 계약서 등에 명시했는가? YES□ | NO□

㉘ 기존 정보보호 교육 외에 인공지능 특화 보안 위협에 대한 교육을 실시하는가? YES□ | NO□

[인공지능시스템 보안성 검증 및 운영 관리]

㉙ 인공지능시스템을 운영 시스템으로 이관하기 전, 운영 중에는 정기적으로 보안성 검증을 수행하는가? YES□ | NO□

㉚ 고영향 및 고위험 인공지능의 경우 금융보안원 등 제3자 검증기관의 검증을 받는가? YES□ | NO□

㉛ 검증 결과 발견된 취약점에 대해 즉시 보완하고 재검증하여 결과를 문서화하는가? YES□ | NO□

㉜ 새로운 인공지능 보안 위협, 시스템 환경 변화, 모델 업데이트에 대응하여 지속적인 보안 관리를 수행하는가? YES□ | NO□

㉝ 정기적인 보안 점검 일정을 수립하여 시스템의 보안 상태를 점검하는가? YES□ | NO□

㉞ 인공지능 모델이나 학습 데이터 업데이트 시 보안성 영향을 사전에 평가하고 필요시 추가 보안성 검증을 수행하는가? YES□ | NO□

㉟ 보안사고 발생 시 원인을 분석하고 재발 방지 방안을 수립하는가? YES□ | NO□

㊱ 상용 생성형 인공지능을 내부망에서 이용할 경우 생성형 인공지능 연계 이용 보안 대책을 이행했는가? YES□ | NO□

㊲ 오픈소스 인공지능 도구 활용 시 금융분야 오픈소스 소프트웨어 활용·관리 안내서에 따라 라이선스 관리, 보안 취약점 점검, 업데이트 관리 등을 체계적으로 수행하는가? YES□ | NO□

㊳ 클라우드 환경에서 인공지능서비스 이용 시 클라우드 제공업체와의 보안 책임 분담을 명확히 하고 데이터 처리 지역, 암호화 수준, 접근 권한 관리 등에 대한 추가적인 보안 통제를 적용하는가? YES□ | NO□