“금융분야 인공지능 활성화를 위한 가이드라인 등 마련”

보고서 요약본

서울대학교 산학협력단(연구책임자 : 고학수 교수)

제1장 서설

본 보고서는 향후 AI가 가장 활성화·고도화될 분야 중 하나가 될 것으로 예상되는 금융분야에서 AI의 활성화를 촉진하기 위한 법제도의 정비, 기반(특히 공적 인프라)의 설계, 가이드라인의 제정 방안을 논의한다. 금융분야의 특성상 각 개별 서비스별 특성을 고려한 상향식(bottom up) 접근이 필요하므로, 이 보고서는 먼저 금융서비스 산업에서 AI의 활용이 특히 중요해질 분야로 이미 제도가 어느 정도 정립된 전자적 투자조언장치(robo-adviser)와 알고리즘거래(algo trading)·고빈도거래(HFT) 이외에도 ① 챗봇(chatbot) 등 AI 기반 대화형 에이전트(conversational agent)를 활용한 고객 응대 및 접점 확대(금융상품 추천·판매 포함), ② AI 기반 자동화 신용평가, 대출심사, 보험심사, ③ 사기탐지(Fraud Detection System; FDS)에 주목한다.

본 보고서는 이들 서비스에서의 AI 활용에 대한 장애요소를 규명함으로써, 해당 AI의 연구 개발과 사업화를 촉진하기 위한 규제 혁신 방안을 모색한다. 특히, 금융소비자의 권익 보호(설명과 알 권리, 차별 금지, 프라이버시, 이해충돌의 방지 등)와 이를 통한 금융시장에 대한 신뢰성 유지, 공정성, 포용 금융 등의 사회적 가치 또한 놓치지 않도록 법제도와 기반을 어떻게 설계·구축·운용할지 검토한다. 관련 기술과 서비스의 발전 동향, 국내외의 사례·입법례·가이드라인, 법적 쟁점에 대한 최신의 논의들, 현장의 수요 등을 종합적으로 고려하여 법개정 및 기반 설계를 제안한다. 그 구체적인 내용은 다음과 같다.

제2장 금융분야 AI 활용 현황

제2장에서는 금융분야 AI 활용 사례 및 예시를 살펴본다. 국내외의 금융업권 및 기타 금융관련 AI 활용사례 및 예시를 살펴봄으로써 (전자적 투자조언장치와 알고리즘거래·고빈도거래 이외에도) ① 챗봇 등 AI 기반 대화형 에이전트를 활용한 고객 응대 및 접점 확대, ② AI 기반 신용평가, 대출심사, 보험심사, ③ 사기탐지(FDS)에 논의를 집중해야 할 필요성을 강조한다.

금융 AI의 개념

AI가 무엇인지에 대해 다양한 논의가 전개되어 왔지만 단일하게 합의된 정의는 아직 존재하지 않는다. 오늘날에도 발전도상에 있는 기술이며 사용되는 맥락에 따라 다양하기 때문이다. 그럼에도 이러한 논의들을 종합하면, AI는 기계가 사람처럼 인지활동을 할 수 있는 능력이라고 정의할 수 있다. AI가 금융영역에서 어떻게 활용되는지를 체계적으로 정리하면 **지각(Sense), 추론(Reason), 학습(Learn), 조치(Take Action)로 구분**하기도 하는데, 이 관점에서 볼 때 금융 AI 또한 결국 소비자의 특성, 금융서비스 이용 양태나 시장 상황 등과 관련된 데이터를 **지각**하고, 이로부터 소비자의 신용 위험, 인지나 행동, 시장 성과 등을 **학습**하고 **추론**하며, 이를 통해 금융거래 체결, 고객 응대 등 **행동**을 내리는 일련의 과정으로 파악할 수 있다. 현재 AI를 **대표하는 기법은 머신러닝**이다. 완전히 새로운 수학적 모델이라기보다는, 전통적인 통계학의 회귀모형(이산값의 경우 로지스틱회귀모형)이 고도화(다양한 비선형모델의 사용, 분산 통제 기술의 고도화, 정규화 등)된 것으로 이해하는 것이 타당하다.

금융 분야에서 AI가 초래한 질적 변화

금융분야에서는 그 어떤 산업보다 일찍이 데이터 기반의 계량적, 통계적 모델들이 광범위하게 사용되었다. 금융 시스템의 주요한 기능은 저축을 투자로 전환하는 것이며, 금융기관(financial intermediaries)은 여기에서 발생하는 도덕적 해이(moral hazard)와 역선택(adverse selection)과 같은 **정보(information)의 문제**와 예금과 투자를 적절하게 연결(matchmaking)해 주는 **채널(channel)의 문제**를 극복하고자 고민해 왔다. 그러나 본 보고서는 현재 진행 중인 변화가 기존의 통계적 기법으로 해결되던 영역과는 질적으로 다르다는 견해를 취한다. 지금까지 금융기관이 기술발전을 선도하며 새로운 경쟁자의 도전에 슬기롭게 극복해 왔던 것과는 달리, 빅테크와 플랫폼으로 대변되는 신규 진입자는 산업의 지형을 전면적으로 바꾸어 놓고 있기 때문이다. 정보의 맥락에서 비금융·비정형데이

터가 활용되고 있으며, 채널의 측면에서는 포털과 스마트폰이 금융기관 지점들이 수행하던 역할을 대체하고 있기 때문이다. 이들은 지속되어온 **양적인 변화의 연장이 아니라 질적으로 새로운 변화**로 이해될 필요가 있다.

금융 AI의 핵심 활용 분야

금융 AI의 활용 분야는 **고객 응대 및 고객 경험 개선**(enhancing customer interaction and experience), **은행 업무처리의 효율성 개선**(enhancing the efficiency of banking process), **보안 및 위험관리 개선**(enhancing security and risk control) 등으로 분류할 수 있다. 국내 법제상 제도화가 어느 정도 이루어진 영역의 활용 사례로서 전자적 투자조언장치(robo-advisor), 알고리즘거래(algo trading)와 고빈도거래(HFT)를 들 수 있다. 반면 국내 법제상 제도화가 완전히 이루어지지는 아니한 영역의 활용 사례로서 고객 응대 및 접점 확대와 금융기관 내부 프로세스에서의 AI 활용(신용평가, 대출심사, 보험심사) 및 금융 보안·위험 관리에서의 AI 활용, 특히 사기거래탐지(FDS)를 고려해볼 수 있다. 요컨대 **향후 법제도와 인프라 설계에 있어 특히 중점적인 검토가 필요한 분야로 챗봇 등 AI 기반 대화형 에이전트를 활용한 고객 응대 및 접점 확대, AI 기반 신용평가, 대출심사, 보험심사, 사기탐지(FDS)**를 들 수 있다. 논의의 구체화를 위해 이하에서는 이들 분야에 집중하여 AI 활성화를 위한 법제도 개선 및 기반 구축 방안을 논의한다. AI라는 신기술은 혁신을 뜻하고 금융의 효율성을 증진시킬 것이라는 기대가 있으나 이러한 신기술 악용은 금융시장의 공정성과 안정성을 해할 수 있으므로 기술적 특성을 고려한 금융규제 개선을 통해 혁신과 시장 건전성, 그리고 규제의 간명성간 조화를 이룰 필요가 있다.

제3장 금융분야 AI 활성화를 위한 법·제도 개선방안

제3장에서는 이들 분야에서의 법제도 개선방안을 마련한다. 금융업권 법령 및 기타 금융관련 법령상의 AI 활용을 저해하는 규제 및 개선방안을 발굴하며, 해외의 금융분야 AI 활성화를 위한 법제도 또는 가이드라인 사례를 조사한다.

금융규제의 목적과 금융 AI

금융규제는 전통적으로 새로운 시도를 통한 혁신 촉진과 소비자보호 및 건전한 시장질서 유지라는 일정부분 서로 상충되는 정책 목표를 조화시키는 것으로 이해된다. 이들의 상충관계가 **트릴레마**(Trilemma)를 이룬다는 프레임워크가 제시되었는데, 그 축은 **금융 혁신(Financial Innovation), 시장 건전성(Market Integrity), 규제의 간명성(Rules Simplicity)**이다. AI는 사전에 작동결과를 알기 어려운 점(소위 black-box model의 문제), AI이 한 행위의 책임소재가 불명확한 점, 데이터 축적에 따라 모델이 지속적으로 업데이트되는 점 등의 기술적 특성이 있다. 이러한 기술적 특성으로 인해 금융규제 트릴레마가 증폭될 수 있다는 우려가 있다.

금융 AI 도입의 장애요소 개관

금융 AI 도입의 장애요소로서 AI 도입을 위한 자원(전문인력, 양질의 데이터)의 부족, 규제 불확실성(개인정보 보호 규제, AI 윤리 규제의 불투명성), 책임 소재의 불명확성, 설명가능성의 미비 등이 금융 AI 도입의 주로 제시된다. AI 도입을 위한 **자원 부족**의 문제는 전문인력의 부족 문제와 품질 높은 데이터의 부족 문제로 구분해 볼 수 있다. 금융기관의 AI 도입을 저해하는 또다른 요소는 **규제의 불확실성**으로 개인정보 보호 규제가 대표적이며, AI 윤리 원칙이 금융 분야에 있어 어떻게 적용되는 것인지 현재의 법 제도 맥락에서 어떻게 해석되고 적용될 것인지에 관한 불투명성이 크다. 나아가 인간의 개입 없이 AI에 의해 투자판단이 이루어지는 경우나 금융상품판매에 AI가 활용되는 경우 금융상품판매규제의 적용과 그 위반에 따른 **책임 문제**가 대두될 수 있다. 마지막으로 AI가 명확한 설명을 제공하지 못한 채 의사결정 과정에서 광범위하게 도입될 경우 **설명불가능성(inexplicability) 또는 불투명성(opaqueness)의 위험**도 발생할 수 있다.

본 보고서는 이러한 금융 AI 도입의 장애요소를 극복하기 위하여 다음과 같은 방안을 중점적으로 검토한다.

첫째, AI 기반 고객 응대 및 접점 확대와 관련하여, 대화형 에이전트를 통해 금융소비자에게 실질적인 설명의무를 다하여 불완전판매를 해소할 수 있을지 여부는 선험적으로 판단될 사항이 아니라 금융소비자의 다양한 금융 독해력(literacy)과 인간과 AI 간 상호작용(human-AI interaction)에 따른 리스크의 실제적 인지 정도에 따라 달리 나타날 것이다. 따라서 금융기관들이 대화형 에이전트의 출시 전 이를 실제 시연하여 자발적으로 참여한 금융소비자의 리스크 인지도를 측정할 수 있도록 **테스트베드(test bed)**을 조성하여 제공하고, 감독당국이 이러한 실증적 측정 결과에 따라 금융소비자에 대한 설명의무 이행 여부

를 판단할 필요가 있다.[1] 이러한 테스트베드가 마련된 이후에는, 금융소비자보호법 제19조의 개정 등을 통해 감독당국으로부터 권한을 위임받은 기관이 테스트베드에서의 실증적 측정 결과에 따라 금융소비자에 대한 설명의무 이행의 수단으로 사용될 수 있을지 여부를 인증(향후 불완전판매 소송 등에서 증거로도 활용 가능)할 수 있는 법적 근거를 두는 것이 바람직하다.

둘째, AI 기반 신용평가, 대출심사, 보험심사 등의 경우 (2020년 개정 신용정보법이 다루고 있는 사회관계망 서비스 등에 공개된 데이터 뿐 아니라) 정부 보유 또는 비금융 데이터와의 결합 가능성을 추가적으로 제고하여 정부의 금융이력부족자(thin-filer)들에 대한 기존 **포용금융 정책**을 더욱 심화, 발전시킬 필요가 있다. 전자정부법 개정안 제43조의2의 본인행정정보의 전자적 제공 제도(일명 "공공 마이데이터")가 입법될 경우 특히 금융이력부족자들의 금융 접근성 확보에 유용하게 활용될 수 있을 것으로 생각되므로, 그 실효적 운용을 위한 기반을 구축해야 한다. 이와 더불어 신용정보법 제17조의2를 개정함으로써 특히 금융이력부족자의 경우 비신용정보집합물 결합을 허용하고 신속하고 간명하게 결합을 허용하는 패스트트랙 제도를 도입하는 방안을 고려할 수 있다. 한편, 피심사자에 대한 공정성(비차별성)의 확보를 위한 다양한 상황에 대한 독립성·분리성·충분성 등 평가지표(metrics)와 이를 계량적으로 측정할 기반 및 설명가능성 확보를 위한 기반의 구축이 필요할 수 있다.[2] 이 때, 이를 뒷받침하기 위하여 금융소비자보호법 제15조를 개정하여 금융상품판매업자등의 금융소비자 부당 차별 여부를 측정할 수 있는 기반의 구축 및 제공 근거를 마련하고, 동법을 개정하여 신용정보법 제36조의2 자동화평가 규정을 기타 영역(대출심사, 보험심사) 자동화평가에 확대하되, 자동화된 평가나 심사의 설명가능성 검증을 위한 기반 구축 및 제공 근거를 마련하는 방안을 고려할 수 있다.

셋째, 사기탐지의 경우 가명정보 처리, 결합과 이동성의 적용 범위의 모호성으로 인해 데이터 활용 활성화에 어려움이 있으므로, 신용정보법 제32조 제6호를 개정하여 **사기탐지**를 동의 없는 개인신용정보 처리 근거로 명시하는 등 동의 없는 데이터 처리 범위의 과감한 확대를 고려할 필요가 있다. 단, 이를 위해 일정한 기술적·물리적·관리적 보호조치가 전제되어야 할 것이다.

넷째, **전자적 투자조언장치**의 경우 기존 법제가 정비되어 있으나 알고리즘 등에 의한 시세조종 등 불공정거래 행위의 유발 가능성을 심사 요건으로 추가하는 방안을 고려할 수 있다. 알고리즘거래·고빈도거래의 경우 한국거래소 시장감시위원회의 위험관리 가이

---

1 구체적인 기술적 기반 조성 방안은 본 보고서 제4장에서 검토한다.

2 이에 대하여는 본 보고서 제5장 및 제6장에서 보다 상세히 논의한다.

드라인을 법제화하고, 이 과정에서 주문착오 사고 발생 시 손실 분담 기준을 민법상 착오의 특칙으로 법제화하는 것이 도움될 수 있다.

제4장 금융분야 AI 활성화를 위한 인프라 마련 방안

제4장에서는 금융 AI의 활성화, 금융 AI의 신뢰성 확보, 금융소비자 행동의 고려 등을 위한 데이터/테스트 인프라의 구축 방안을 논의한다. 본 연구에서는 특히 데이터 인프라를 중심으로 검토한다.

공공데이터 확충

금융 AI 학습에 사용되는 데이터는 주로 금융기관이 자체적으로 보유(in-house)하고 있는 데이터이다. 그러나, 개별 금융기관이 자체적으로 보유한 데이터만으로는 금융 AI 활성화에 있어 한계가 존재할 수 있다. 그 이유로는 학습데이터의 불균형 문제, 선별적 라벨 문제, 오픈소스 데이터 및 공통 과제의 부족 등이 있다. 금융 AI 활성화 방안에 있어 데이터 활용 확대를 위한 법제도적 인프라로는 비금융정보의 활용 확대, 마이데이터 사업 등 금융소비자의 전송권에 기반한 데이터 활용 확대, 비식별 조치를 통한 데이터 활용 확대 등으로 이미 법개정을 통해 상당부분 개선된 영역이다. 법제도 이외의 데이터 인프라 마련에 있어 정부의 역할이 필요한 지점은 공공재적 성격을 갖는 금융 AI용 데이터셋을 구축하여 공중에 공개하거나 수요 금융기관·스타트업·연구자 등에 제공하는 것이다.

데이터 인프라 구축의 핵심적 전제 사항은 특히 신용정보주체의 프라이버시에 대한 충분한 보호가 이루어져야 한다는 점이다. 이는 금융 AI의 활용 확대에 대한 공중의 신뢰를 제고하는 측면에서도 중요한 정책적 과제가 된다. 프라이버시 보호를 달성하면서도 금융정보를 활용하기 위한 방안으로는 일반적으로 비식별조치를 통한 데이터 활용이 주로 고려되어 왔다. 최근의 국내 시도로는 (1) 2019년 익명처리된 금융 데이터셋(CreDB)의 구축 및 공개, (2) 2020년 가명·익명처리 기술 활용에 대한 가이드라인 마련 및 가명정보의 결합에 관한 지원이 있으며, 추가적으로 (3) 프라이버시 보호와 데이터 유용성 확보를 위한 기술 수단인 합성 데이터(synthetic data) 활용을 검토할 수 있다.

본 보고서는 산업적으로 특히 공공데이터셋 구축이 절실한 분야로 금융 대화형 에이전트를 위한 **금융 말뭉치(corpus)**와 **사기탐지(FDS) 데이터셋**에 주목한다. 금융 챗봇이 금융소비자에 대하여 적절하게 설명의무를 이행하고 적정한 판매행위를 할 수 있도록 하기 위해서는 금융소비자의 수요와 금융 독해력(financial literacy)에 따라 정확한 대화 내용을 출력하는 것이 가능해야 하고, 이를 위해서는 기반이 되는 양질의 훈련 데이터가 필수적이다. 한글 구어 데이터, 특히 금융 상품 판매에 특유한 전문적인 대화 내용이 축적된 말뭉치 데이터가 금융기관들의 협력 하에 빠르게 구축될 수 있도록 인프라를 구축할 필요가 있다. 또한 사기탐지 학습·검증 데이터 허브를 구축하여 금융기관들이 사기탐지에 필요한 데이터를 협력적으로 축적할 수 있도록 지원할 필요가 있다. 앞 장에서 살펴보았듯이 신용정보법 제32조 제6호를 개정하는 등의 방법으로 사기탐지를 동의가 필요 없는 개인신용정보 처리 근거로 명시할 경우, 상기 허브를 통한 금융기관 간의 사기탐지 학습·검증 데이터의 결합과 교환이 가속화될 것으로 전망된다.

AI 신뢰성 표준 제정

한편 금융 AI 활성화를 위해서는 사회적 신뢰성 확보가 필요하다. 이를 위한 접근법은 AI 신뢰성을 제품이나 서비스의 '품질'과 유사한 특성으로 보는 방안과 조직 내에서 신뢰성 확보를 위해 이루어지는 지속적인 '프로세스' 차원에서 접근하는 방안이 있다. 이상의 두 가지 접근법을 고려할 때, 금융 AI 신뢰성을 확보하기 위한 지원 방안은 (i) AI 신뢰성의 기술적 구현을 위한 지원과 (ii) AI 신뢰성에 대한 내부 통제 및 검증 프로세스 구축을 위한 지원으로 구분해 볼 수 있을 것이다.

국제표준화기구(International Organization for Standardization, "ISO")의 "AI 신뢰성 개요" 기술보고서는 AI 특유의 보안 위협으로 (1) AI의 오동작을 초래하는 데이터 오염 공격(data poisoning), (2) AI 시스템을 악용하는 적대적 공격(adversarial attack), (3) AI 모델 도용 공격(model stealing)을 제시한다. 또한 국제전기전자기술자협회(Institute of Electrical and Electronics Engineers, "IEEE")도 인공지능 신뢰성 기술 표준을 마련하고 있다. 정부나 전문기관은 이러한 국제적 기술 표준의 제정 절차에 참여하거나, 그 발전 상황을 예의주시하면서, 금융기관에 있어 중요한 기술 표준을 소개하고 그 적용을 도울 필요가 크다.

금융독해력 판단 인프라

챗봇 등 AI에 기반한 고객 응대 및 접점 확대 활용 사례에 있어 금융 소비자의 행동을 분석하고 그 이해도를 측정하는 것이 중요하다. 특히 고객 접점에 있어서의 AI 활용에서 주요한 장애요소가 되는 것은 비대면 금융상품 판매시 설명의무와 관련된 법적 리스크(불완전판매 등)인데, 이를 해결하기 위해 금융 소비자의 금융 독해력이나 금융상품에 대한 이해 수준을 평가하는 방법이 더욱 고도화될 필요가 있다. 금융 독해력 측정이 가능하다면 개인에 맞는 개별화된(personalized) 설명이 가능해질 수 있다. 이때 실증적 기법을 활용함으로써 고객이 얼마나 금융상품의 리스크를 잘 이해하고 있는지 확인할 수 있다.

이러한 금융 소비자의 이해도 평가 작업은 금융 서비스 산업 전체에 있어 공공재와 같은 역할을 할 수 있으므로, 원칙적으로 공적 인프라 마련 차원에서의 접근이 바람직할 것으로 판단된다. 만약 정부 또는 공공기관이 위와 같은 실험을 수행할 경우 금융상품 이해도 평가 과정과 그 결과에 대한 투명성을 보장할 필요도 있다.

감독당국이나 그로부터 위임받은 기관이 AI 에이전트에 의한 설명의무 준수 가능 여부를 실증적으로 측정하고, 해당 실험에 의한 설명 효과가 충분한 것으로 검증된 경우에는 이를 **금융소비자에 대한 설명의무를 이행할 수 있는 수단으로 인증하는 제도**도 고려해 볼 수 있다. 나아가, 금융상품 판매 프로세스에 대해 위와 같은 실증적 측정 및 검증을 수행하는 제도가 도입되면, 사후적으로 불완전판매 분쟁이 발생할 경우에 실증적 측정 및 검증의 결과를 과학적 증거로 활용하는 것도 가능할 수 있다.

제5장 금융분야 AI 정확성·공정성 심사 방안

제5장에서는 금융 AI의 신뢰성을 확보하기 조건으로 통계적 정확성과 공정성을 검토한다. 다양한 통계적 정확성 지표 중 어떠한 지표를 기준으로 AI의 성능을 평가할 것인지, 여러 지표들 간의 상충 관계를 어떻게 해소할 것인지는 쉽지 않은 문제이다.

통계적 정확성과 그 상충관계

통계적 정확성 판단을 위해 기본적으로 사용되는 지표는 네 가지로서, True Positive Rate (재현율, 민감도), True Negative Rate (특이도), Positive Predictive Value (정밀도, 양성예측도),

Negative Predictive Value(음성예측도)이다. 한편 이들 기준 간에는 상충관계가 발생하는데 어떤 AI 모델의 정확성을 평가하는데 있어 실무상으로 널리 활용되는 방법은 수신자 조작 특성(Receiver Operating Characteristic, ROC) 곡선을 활용하는 것이다. ROC 곡선은 한도(threshold) 값을 약간씩 변화시켜 가면서 그 정확도가 어떻게 달라지는지를 표시하며 가로 축은 False Positive Rate(FPR), 세로 축은 True Positive Rate (TPR = 재현율(recall))이다.

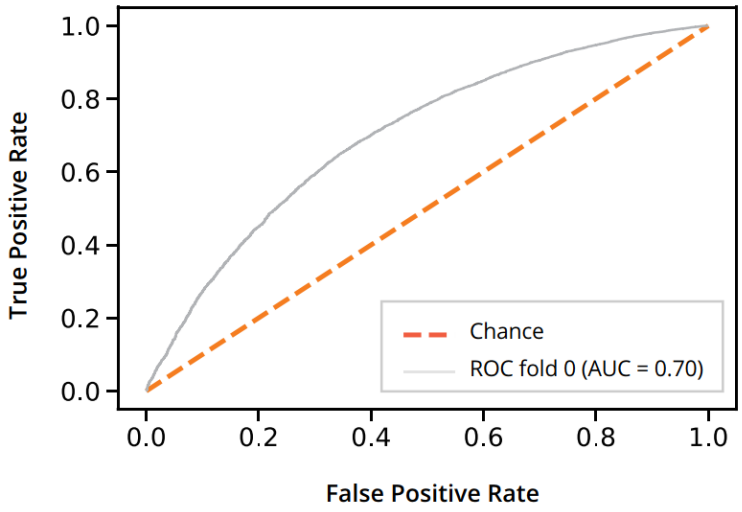

Figure 3.8: ROC curve of model performance on the test set.

[MAS Veritas 프로젝트 예시 사례의 ROC 곡선]

ROC 곡선은 False Positive 오류와 False Negative 오류 간의 상충 관계를 보여준다. ROC 곡선에서 우상단으로 갈수록 False Negative 오류는 줄어들고(그 결과 True Positive Rate가 1에 접근), 그 대신 False Positive 오류가 증가한다(그 결과 False Positive Rate도 1에 접근). 반대로 ROC 곡선에서 좌하단으로 갈수록 False Negative 오류가 증가하고(그 결과 True Positive Rate가 0에 접근), False Positive 오류는 감소한다(그 결과 False Positive Rate도 0에 접근). 따라서 ROC 곡선은 위로 볼록할 수록 모델의 성능이 우수한 것으로 볼 수 있다.

AI 시스템의 통계적 정확성을 확인할 수 있는 여러 지표가 존재하며 그 지표 간에는 불가피한 상충관계가 존재하므로 **금융기관은 AI 시스템을 도입하고자 할 경우 지표간 상충관계에 대한 의사결정을 내릴 필요**가 있다. 따라서 금융기관은 이러한 상충관계 판단에 대한 내부적 의사결정 프로세스를 구축하고, 그 프로세스에 따른 결정을 정당화할 수 있는 근거를 마련할 필요가 있다. 따라서 AI 신뢰성 구축을 위한 인프라 제시를 위해서는 ① 금융기관이 위와 같은 상충관계에 대한 의사결정 프로세스를 구축하는 것에 관한 가

이드라인을 제시하는 한편, ② 해당 평가를 수행할 수 있는 적절한 방법론을 제시하는 것이 바람직하다.

금융기관이 통계적 정확성 지표간 상충관계에 관한 의사결정을 내릴 때 오류에 따라 어떤 위험(이용자 해악)이 발생할 수 있는지를 기준으로 사안에 따라 적절한 판단을 내릴 수 있어야 한다. 가령 **금융 AI가 신용평가, 대출심사, 보험심사 등과 같이 금융소비자에 대한 금융 기회를 제공하는 기능을 수행하는 사례**에서는 상대적으로 **False Negative 오류(적격자에 대한 기회 부인)를 최소화**하는 것이 중요할 수 있다. 이 경우에는 통계적 정확성 지표 중 재현율(recall)을 높이는 것이 중요하다. 반면 **금융 AI가 사기 탐지(FDS) 또는 금융 규제(Reg-Tech) 분야와 같이 위법·부당 사례를 탐지하는 기능을 수행하는 사례**에서는 상대적으로 **False Positive 오류(잘못된 사고 탐지)의 최소화**가 중요할 수 있다. 이 경우 통계적 정확성 지표 중 정밀도(precision)를 높이는 것이 중요하다.

공정성 평가 및 통제방법

금융 AI가 인구통계 집단(성별, 나이, 지역 등)에 대해 차별적으로 작동하거나, AI의 활용 결과 개인의 기본권이 침해될 수 있다는 우려가 제기되고 있다. AI가 집단간 불균형을 발생시킨 사례에 대해서는 다양한 맥락에서 반복적으로 지적되어 왔는데 그 원인으로는 크게 4가지 단계가 지적되고 있다. (1) 우선 AI 모델에 대한 문제 설정(framing) 단계에서 AI에 어떤 변수를 입력 값으로 활용하는 것이 개인의 자유를 부당하게 제약하거나, 해당 정보를 활용하는 자체가 공정성 관념에 부합하지 않는다는 비판이 제기되는 경우이다. (2) 데이터 수집 단계에서 대표성이 결여된 데이터가 수집되거나, 학습 데이터 자체에 편향성이 내포된 경우 차별 문제가 발생할 수 있다. (3) 또 다른 가능성은 AI의 학습 단계에서 차별적 작동을 학습하게 되는 경우이다. 이는 입력 변수에서 차별적 변수가 제거되더라도 AI가 통계적 정확도를 높이기 위해 차별적으로 작동하게 되는 문제이다. (4) 마지막으로 AI 모델이 설계된 목적이 이용자의 이익과 일치되지 못하는 경우에 발생하는 문제이다.

AI에 의한 집단간 차별대우를 시정하려는 노력 중 가장 기초적인 형태는 AI 시스템이 해당 차별적 속성을 입력 값으로 고려하지 않도록 하는 것이다(Unawareness 기준). 예를 들어, 학력을 차별적 변수로 보아 신용평가모형의 입력 변수로 활용하는 것을 허용하지 않는 것이다. 이는 학력 속성에 대해 Unawareness 기준을 적용한 사례로 이해될 수 있다. 하지만, 데이터셋에서 차별적 속성을 제거하더라도 AI 시스템은 학습 과정에서 다른 특

성에 중복적으로 반영된 차별적 속성에 관한 정보를 활용할 수 있다. 이처럼 차별적 속성을 대체하는 다른 속성을 대체변수(proxy)라고 한다. 일반적으로 대체변수의 영향으로 Unawareness 기준만으로는 충분하지 못하다는 평가가 많다.

이에 대해 AI의 출력 결과값을 기준으로 공정성 심사를 진행할 필요가 있다는 논의가 활발하게 진행중이다. AI의 출력 결과에 대한 공정성 심사 기준 중에서 주로 언급되는 기준으로는 3가지가 있다. (1) 인구통계적 동등성(Demographic Parity)과 독립성 조건은 단순히 집단간 취급율이 동일할 것을 요구하는 것이다. (2) 기회의 균등(Equal Opportunity)과 분리성 조건은 집단간 실제 적격자중 적격자 취급을 받는 비율인 True Positive Rate(재현율)가 동등할 것을 요구하는 것이다. (3) 마지막으로 예측도 동등성(Predictive Parity)과 충준성 조건은 집단별로 AI 모델의 예측 정확도가 동등할 것을 요구하는 것이다.

한편 **이들 공정성 요건을 모두 만족하는 모델을 구축하는 것은 수학적으로 불가능**하며 따라서 AI 모델의 목적과 예상되는 이용자 해악 등의 맥락을 고려하여 적절한 공정성 평가방법을 활용해야 한다. 이러한 상충 관계에 관한 의사결정은 지속적으로 이루어져야 하고 그 판단 과정을 문서화하여 관리함으로써 AI 활용에 대한 정당화 근거를 마련할 필요도 있다.

제6장 금융분야 AI 설명가능성 구현 방안

제6장에서는 AI 또는 머신러닝 모델에 대한 기술적 설명 방법을 살펴보고 해외사례를 검토한 다음 국내 금융 분야에서 설명가능 AI를 구현하기 위한 지원 방안을 살펴본다.

기술적 설명 방법

AI 또는 머신러닝 모델을 설명하는 방법은 (1) 설명가능한 모델을 사용하는 방법과 (2) '블랙박스' 모델을 설명하는 방법으로 구분된다. 설명가능한 모델의 대표적인 예는 현재 사용되는 신용평가 모델로 각 변수에 고정된 가중치가 부여된다.

블랙박스 모델에 전체에 대한 설명방법으로는 부분 의존성 플롯(Partial Dependence Plot, "PDP"; 관심을 갖는 변수의 값을 변화시키면서 머신러닝 모델의 기대값을 그래프화한 것)과 개별 조건부 기대 플롯(Individual Conditional Expectation, "ICE"; 개별 관측치에 있어 관

심있는 특성의 입력 값을 변경시켰을 때 출력 값이 어떻게 변화하는지를 함께 표시한 것)가 사용된다. 한편 예측 결과에 특정 변수가 얼마나 중요한 역할을 하는지를 파악하기 위해 민감도 분석(Sensitivity Analysis)이 활용되기도 한다.

한편 매우 복잡한 결정경계선을 갖는 모델도 국소적으로는 선형 모델로 근사할 수 있다는 장점에 기초해 국소적으로 대상 범위를 축소시켜 모델을 설명하려는 접근법도 존재한다. 국소적 대리 모델을 찾는 방법으로 최근 널리 활용되는 것은 LIME(Local Interpretable Model-agnostic Explanation)이다. 또한 협조적 게임 맥락에서 공정한 보상 분배 방안을 연구하는 과정에서 제안된 경제학적 개념인 섀플리 밸류(Shapley value)를 이용한 변수의 AI모델이 출력한 결과값에 대한 영향력 분석도 널리 사용되고 있다.

마지막으로 예제 기반 설명 방식에는 (1) 설명 대상 사례를 설명할 수 있는 가급적 단순하면서도 넓은 범위를 포괄하는 규칙을 생성하는 앵커(Anchor) 방식, (2) 모델 전체에 대한 설명과 개별 사례에 대한 설명을 동시에 할 수 있다는 장점을 가진 MMD-critic 기법, (3) 학습 데이터에 변화가 생기는 경우 모델이 어떻게 달라질지를 평가하는 Influence Instances 기법, (4) 주어진 사례에 대해 결과값이 달라질 수 있는 최소한의 변화가 무엇인지를 검토하는 counterfactual explanation 방법 등이 있다.

해외 사례

금융 AI 기술에 설명가능성 연구를 적용하려는 해외 사례로는 (1) 영국의 중앙은행 Bank of England (이하 "**BoE**")가 2019년 8월 발행한 "금융 분야에 있어서의 머신러닝 설명가능성: 채무불이행 위험 분석" 사례와 (2) 미국 FICO사가 2018년 진행한 설명가능 머신러닝 경진대회(Explainable Machine Learning Challenge) 사례를 들 수 있다.

BoE는 채무불이행 위험 모델을 설명하기 위하여 다음 네 가지 유형의 설명 방식을 제시했다. (1) 개별 예측에 대한 설명, (2) 모델 전체에 대한 설명, (3) 선형 모델과의 차이점을 통한 설명, (4) 모델의 구체적 작동 방식에 대한 설명이다. BoE는 특히 클러스터링을 활용하여 모델의 작동 방식을 설명하는 방법을 소개하였다.

FICO 사의 경진대회는 **모범적인 설명 사례(best practice)**를 제시하고, 참가자들은 그러한 설명을 도출할 수 있는 기술적 방안을 구현하는 것을 목적으로 하였다. 이러한 설명은 여러 기술적 방법을 통해 구현될 수 있는데, 예컨대 Recognition Award를 수상한 듀크대학 팀은 블랙박스 모델을 직접 설명하는 방식이 아니라, 블랙박스 모델을 해석 가능한

모델로 대체하는 방식을 제안하였다. 듀크 대학 팀은 선형 모델링을 기반으로 하여 해석 가능한 비선형적 요소를 삽입한 모델을 구성함으로써, 블랙박스 모델만큼의 정확성을 달성하면서도 해석 가능한 모델을 도출하였다.

설명가능 AI 지원방안

설명가능 AI 기술의 확산을 위해서는 우선 국내 금융기관의 AI 연구자나 개발자들에게 설명가능 AI 기술의 현황을 소개하는 작업이 필요하다. 이러한 작업을 참조하여 금융 AI를 개발하는 실무 담당자들이 참조할 수 있는 기술 핸드북 또는 매뉴얼 등을 작성하는 방안을 고려할 수 있다. 또한 FICO 경진대회 사례와 같이 구체적인 모범적인 설명 사례(best practice)를 제시하는 가이드라인을 포함시키는 것도 고려할 수 있다. 특히, 현재 설명가능 기법에 대한 사례는 주로 신용평가에 집중되고 있으나, 그 이외에도 금융상품 추천, 사기 탐지, 금융 독해력 예측 등의 여러 적용 사례에서 모범적인 설명 수준이 무엇인지 제시하는 것이 바람직하다.

FICO사의 2018년 설명가능 AI 경진대회 사례와 같이 국내 연구자들의 관심을 북돋우기 위한 경진대회를 고려할 수 있다. 이러한 경진대회 확산을 위해서는 참여자에게 학습 데이터를 제공할 필요가 있는데, 이러한 데이터는 두 가지 유형으로 구분하여 제공하는 방안을 고려해 볼 수 있다. 즉, (1) 익명처리 또는 합성 데이터 기법을 이용하여 재식별이 불가능한 형태로 처리한 다음 일반 공중에 공개하는 **교육 목적 데이터**와 (2) 가명처리를 거쳐 연구 목적(경진대회 참여)으로 제한된 경진대회 참여자에 대해서만 제공하는 **연구 목적 데이터**로 구분하여 제공하는 것이다. 교육 목적 데이터는 캐글 등을 통해 온라인으로 공개하여, 교육 목적으로 국내 AI 관련 대학 또는 실무자 교육, 금융 공학 분야 연구자들의 위한 교육 자료로 활용하도록 할 수 있다.

제7장 금융분야 AI 윤리 가이드라인

제7장에서는 해외 금융당국 등이 발간한 AI 윤리 가이드라인 사례를 조사하며, 국내 금융 AI 윤리 가이드라인의 제정 방향 및 주요 내용을 제시한다.

가이드라인 제정 방향

AI 윤리 가이드라인 제정과 관련해 참고할 수 있는 AI 윤리원칙 이행을 위한 체크리스트의 대표적 사례는 EU 집행위원회 산하 인공지능 고위전문가 그룹(High-Level Expert Group on Artificial Intelligence, "AI HLEG")이 2020년 7월 발표한 "신뢰할 수 있는 인공지능을 위한 평가 리스트(The Assessment List for Trustworthy Artificial Intelligence)" (약칭 'ALTAI')이다. 외국 금융당국에 의한 AI 윤리 가이드라인으로는 싱가포르 통화청(Monetary Authority of Singapore, "MAS")의 사례가 주목할 만하다. MAS는 금융분야 인공지능 공정성, 윤리, 책임성, 투명성 원칙 준수를 위하여 2021년 1월 금융분야 인공지능 공정성 평가에 관한 케이스 스터디 사례를 발표하면서 14 항목의 구체적인 체크리스트 문항을 제시하였다.

이상을 포함한 해외의 사례는 다음과 같은 시사점을 준다. (1) **거버넌스 수립의 중요성** – AI 윤리원칙을 구현하기 위해서는 AI 위험 평가 · 관리 업무를 수행할 조직을 수립하고 책임자를 정하며, 내부 절차를 마련하는 등의 거버넌스 확립을 중시하고 있다. (2) **위험 기반 접근** – AI 시스템이 어떠한 위험을 제기하는지를 우선 평가하고, 이에 비례하는 조치가 이루어질 것을 요구하는 것으로, 빠르게 변화하는 기술에 맞추어 새로운 위험 평가를 수행할 수 있고, AI 윤리의 측면에서도 새로운 문제가 부각될 경우 그에 맞추어 유연하게 판단하고 대응할 수 있다는 장점이 있다. (3) **맥락에 따른 비교 형량의 필요성** – AI 윤리원칙을 구현하는데 있어서는 여러 가치들 간의 '비교 형량'(즉, trade-off 결정)이 불가피하고, 비교 형량에 따른 의사결정을 내리기 위해서는 AI가 활용되는 '맥락(context)'이 중요하게 고려되어야 한다. (4) **금융 AI 활용에 제약이 되지 않도록 유의할 필요성** – 본 가이드라인이 복잡한 절차와 승인, 과다한 문서화를 요구함으로써 금융기관의 AI 활용을 제약하는 방향으로 작동하지 않도록 유의할 필요가 있다.

가이드라인 주요 내용

금융분야 인공지능 활성화를 위한 가이드라인(안) [첨부 1]은 금융분야에서의 인공지능(AI) 시스템의 개발, 사업화 및 활용과 관련한 기획·설계, 개발, 평가·검증, 도입·운영 및 모니터링 전 과정에서 신뢰성을 제고하여 AI의 활성화에 기여하는 것을 목표로 한다. 주요 내용으로는 AI 시스템의 잠재적 위험을 평가하고 이를 관리할 거버넌스 구축(제2장), AI 시스템의 기획 및 설계 단계에서 AI 활용 사례에 대한 사회적 영향 평가 및 인간 통제 가능성 확보(제3장), AI 시스템 개발 단계에서 학습 데이터 품질 검토, 개인정보 활용

정당성 평가 및 AI 시스템의 설명가능성 고려 모델 선정(제4장), AI 시스템의 평가 및 검증 단계에서 성능, 공정성, 비차별성 및 설명가능성 평가(제5장), AI 시스템의 도입, 운영 및 모니터링 단계에서 성능에 대한 감독, 안전성 및 보안에 대한 평가(제6장)가 있다.

*　　*　　*　　*　　*