정 보 보 호 학 회 지
제 32 권 제3호, 2022. 6

금융분야 AI의 윤리적 문제 현황과 해결방안

이 수 련*, 이 현 정**, 이 아 람***, 최 은 정****

요 약

우리 사회에서 AI 활용이 더욱 보편화 되어가고 있는 가운데 AI 신뢰에 대한 사회적 요구도 증가했다. 특히 최근 대화형 인공지능'이루다'사건으로 AI 윤리에 대한 논의가 뜨거워졌다. 금융 분야에서도 로보어드바이저, 보험 심사 등 AI가 다양하게 활용되고 있지만, AI 윤리 문제가 AI 활성화에 큰 걸림돌이 되고 있다. 본 논문에서는 인공지능으로 발생할 수 있는 윤리적 문제를 활용 도메인과 데이터 분석 파이프라인에 따라 나눈다. 금융 AI 기술 분야에 따른 윤리 문제를 분류했으며 각 분야별 윤리사례를 제시했고 윤리 문제 분류에 따른 대응 방안과 해외에서의 대응방식과 우리나라의 대응방식을 소개하며 해결방안을 제시했다. 본 연구를 통해 금융 AI 기술 발전에 더불어 윤리 문제에 대한 경각심을 고취시킬 수 있을 것으로 기대한다. 금융 AI 기술 발전이 AI 윤리와 조화를 이루며 성장하길 바라며, 금융 AI 정책 수립 시에도 AI 윤리적 문제를 염두해두어 차별, 개인정보유출 등과 같은 AI 윤리 규범 미준수로 파생되는 문제점을 줄이며 금융분야 AI 활용이 더욱 활성화되길 기대한다.

Ⅰ. 서 론

우리 사회에서 AI 활용이 더욱 보편화되어가고 있다. 빠르게 발전하는 인공지능 기술이 인간 삶에 영향을 주고 있으며 향후 인공지능의 도움을 더 많이 받게 될 것으로 전망한다. 가트너가 발표한 2020 하이프 사이클을 살펴보면 AI는 The Peak of Inflated Expectations단계를 지나 우리 일상의 구성요소로 자리잡아가고 있다[1]. 또한 2021년 가트너 전략기술 9에 따르면 팬데믹 이후 사회의 회복적 탄력을 위한 기술로 AI 엔지니어링을 선정했다[2]. 빠르게 발전하는 인공지능은 인간의 삶에 영향을 주고있으며 팬데믹을 극복하는 과정에서도 큰 도움을 주고 있다. 방역 과정, 치료제 개발에도 AI 기술이 활용되고 있으며, '언택트' 속에서 경제와 사회를 유지시키는 데에도 AI가 적지 않은 도움을 주고 있다.

한편 AI 활용이 많아짐과 동시에 AI 신뢰에 대한 사회적 요구도 증가했다. AI의 발전은 기업에게는 생산비용 절감, 개인에게는 개개인 맞춤 서비스 등과 같은 이익을 가져다주었다. 하지만 AI는 오용, 악용될 때 재앙으로 돌아올 수 있는 양날의 검이 되기도 한다. 개인정보유출, 불평등, 인권에 대한 영향 등 다양한 분야와 이해관계가 얽힌 문제가 생길 수 있다. 특히 최근 대화형 인공지능 '이루다' 사건으로 AI 신뢰에 대한 사회적 논의가 뜨거워졌다. '이루다'는 장애인, 여성, 노인 등에 대한 혐오 발언과 개인정보 유출 문제로 사회적 물의를 일으켰으며 2021년 1월 정식 오픈 후 3주만에 서비스가 중단되었다. 한 언론에서는 만 18세 이상 남녀 1000명을 대상으로 인공지능(AI)과 윤리성에 대한 조사를 진행했다. 조사 결과 10명 중 9명은 AI가 도덕성을 갖추어야 한다고 응답했고, 절반은 AI 윤리 문제로 인한 피해 방지가 AI 기술 발전보다 더 중요하다고 응답했다[3]. 이러한 AI윤리 문제 해결의 사회적인 요구에 따라 정부에서는 2021년 5월 "사람이 중심이 되는 인공지능을 위한 신뢰할 수 있는 인공지능 실현전략"을 발표했다. 또한 기업에서도 AI 개발의 윤리 원칙을 세우는 등, 기술 수준의 고도화에만 초점을 맞추던 흐름에서 AI 신뢰성에 대해 주목하고 있다.

금융분야에서도 로보어드바이저, 대출 심사, 보험심사, 사기 탐지등 AI가 다양하게 활용되고 있다. AI 로 생산 비용도 절감하고 고객에게 맞춤형 금융 상품을 제공하는 등 여러 이익을 받고 있지만 금융권에서도

* 서울여자대학교 정보보호학과 (학생, dolpong@swu.ac.kr)
** 코스콤 네트워크서비스부 (차장, hjlee@koscom.co.kr)
*** 쿠팡 Security & Privacy Policy (과장, alee5@coupang.com)
**** 서울여자대학교 정보보호학과 (교수, chej@swu.ac.kr)

AI 도입 이후 AI 신뢰와 관련된 문제가 발생하고 있다. Apple카드에서는 신용 한도를 책정하는 과정에서 남편과 자산 및 계좌를 공유하는 여성들이 남편보다 현저하게 낮은 신용공여한도를 부여해 논란을 빚었다. 신용한도 심사 외에도 대출 심사, 보험 심사 알고리즘에서도 성별, 소득 수준 등에 따른 차별적인 결과로 문제가 된 사례가 많다. 이러한 AI 신뢰 문제들은 금융분야 AI도입 활성화에 큰 걸림돌이 되고 있으며, 반드시 해결해야하는 문제이다.

로보어드바이저, 보험심사, 대출심사와 같은 금융산업은 소비자 이해관계가 첨예한 분야로, AI로 인한 윤리문제 발생시 더욱 큰 파장을 일으킬 수 있다. 본 논문에서는 금융분야 AI 활성화를 위해서 반드시 해결해야 하는 문제인 AI 신뢰, 특히 윤리 문제를 극복하기 위한 방안을 제시한다. 먼저 금융AI의 발전 현황과 현재 직면한 윤리적 문제를 살펴보고 파생될 수 있는 윤리적 문제들을 특정 기준에 따라 분류한다. 마지막으로 해외에서의 대응방식과 우리나라의 대응방식을 소개하며 해결방안을 제시한다. 본 연구에서는 금융 AI 기술 분야에 따른 윤리 문제를 분류하였다. 또한 각 분야별 기술에서 고려할 윤리사례를 제시했으며, 윤리문제에 따른 대응방안을 제시했다. 본 연구를 통해 금융 AI 기술 발전에 더불어 윤리문제에 대한 경각심을 고취시킬 수 있을 것으로 기대된다.

Ⅱ. 금융AI 발전 현황

인공지능 구현 수준이 크게 발전하면서 은행·보험·카드사 등 금융권에서도 AI 기술을 활용한 서비스를 제공하고 있다. 금융회사들이 제공하는 서비스를 이용하는 고객들은 높은 수준의 정보와 서비스를 편하게 이용할 수 있게 되었다. 동시에 인공지능의 투명성, 보안 등에 대한 사회적 요구가 늘어남에 따라 금융 분야의 인공지능 활성화를 위해서는 인공지능 기술의 신뢰성 보장이 필수가 되었다. 하지만 로보어드바이저, 보험·대출 심사, 알고리즘 트레이딩등 금융권에서 제공하는 AI 서비스에는 신뢰성 보장을 막는 장애 요소가 존재한다.

로보어드바이저는 AI가 고객의 성향을 참고하여 빅데이터와 알고리즘으로 투자 포트폴리오를 제공하는 서비스이다. 로보어드바이저의 신뢰도를 위협하는 요소로 해킹, 데이터 편향, 블랙박스 알고리즘이 있다. 해킹으로 시세 조종을 할 수 있으며, 데이터 편향으로 고객에게 제공되는 정보에 질적인 차이가 발생할 수 있다. 또 블랙박스 알고리즘 특성상 고객에게 제공되는 결과에 대해 설명하지 못하는 문제도 생긴다. 보험·대출 심사에서도 블랙박스 알고리즘으로 인한 설명 불가능의 문제가 발생한다. 이 외에도 보험·대출 심사에서는 비금융 데이터의 활용 비율이 증가하면서 나이, 성별, 학력, 주거지역 등의 데이터로 인한 데이터 편향성이 고객에 대한 차별적인 판단으로 이어질 수 있다. 마지막으로 알고리즘 트레이딩에서 인공지능은 1초에 수백, 수천 번의 주문을 넣어 대량의 주식거래를 만들기 때문에 주가 조작의 위험이 있다. 따라서 가짜정보, 위장 거래로 시세 조종이 발생할 수 있다. 또한 이러한 시세 조종이 발생한 경우 법적 책임의 주체가 모호해지는 문제도 발생한다.

Ⅲ. 금융 AI 윤리 문제

3.1. 기존 연구

오요한 등[4]은 빅데이터에 기반을 둔 인공지능 알고리즘이 사법, 치안, 국가 세 분야에서 차별적인 결과를 가져온 사례를 관련된 학술 연구를 검토하며 분석했다. 컴파스 알고리즘을 예로 들며, 맥락과 분야에 따라 유연하게 적용해야 한다고 서술했는데, 알고리즘의 공정성은 기준과 척도의 설정에 따라 다양한 결과로 이어지기 때문에, 절대적인 완벽한 공정성은 없다는 주장이다. 임용 등[5]은 데이터의 존재와 이용 용이성에 따라 4가지 유형으로 분류하여 유형별 문제를 제시했다. 이와 같은 학습 데이터가 인공지능의 공정한 판단에 영향을 주는 현상 즉 알고리즘적 차별이 발생한다고 보고, 인공지능 정책 수립을 위해서는 데이터 규율과의 연관성을 고려해야 한다고 주장했다. 양종모[6]는 인공지능 알고리즘으로 인한 차별적인 법적 의사결정과 그로 인한 문제에 대해 논의하고 사전적 방법을 중심으로 규제 방안을 제시했다. 인공지능 의사결정 알고리즘 설계 단계에서 윤리적 기준을 세워 그에 따라 설계를 해야 하며 정적 평가와 동적 평가를 통해 알고리즘 분석 평가를 해야 한다고 주장했다. 신영진[7]은 인공지능 서비스에서의 개인정보보호를 위한 개선 상황을 법제적 기준과 처리 과정 기준에서 제시했다. 법제적 기준에서는 법률 재정비, 위험관리체계 구

축 및 운영을 방안으로 제시했고 처리 과정에서는 데이터셋 참조모델 표준화, 데이터셋 품질관리를 방안으로 들었다. 또한, 비식별 조치의 명확한 기준을 세워 재식별이 불가능하도록 관리해야 함도 덧붙였다. 김승래[8]는 현행 법 제도는 국가 경쟁력 확보에 걸림돌이 된다며 우리나라 금융 데이터산업 활성화를 위한 현행 규제를 고찰하고 정책 보완의 방향을 제시했다. 먼저 데이터 금융혁신과 관련하여 규제가 많다는 점을 꼬집었고, 다음으로 보안 정책과 데이터 활용 정책의 적절한 균형점을 찾아야 한다고 주장했다.

3.2. 윤리적 문제 국내외 사례

마이크로소프트는 사람과 대화를 나누는 인공지능 채팅봇 테이를 선보였다. 테이는 어떤 데이터를 입력받느냐에 따라서 행동 양식이 바뀌는 채팅봇이다. 이를 악용하여 일부 극우 성향 사용자들이 욕설, 차별적 발언, 정치적 발언 등을 하도록 유도했다. 테이의 이런 발언이 물의를 일으키자 MS는 문제가 된 테이의 일부 트윗과 공개 메시지 등을 16시간 만에 삭제하고 운영을 일시중지했다[9]. 두 번째 예로는 얼굴인식 프로그램의 인종적 편향성이다. 미국 NIST는 지난 2019년 보고서에서 얼굴인식 소프트웨어가 여성과 흑인 얼굴보다 백인 남성 얼굴에 훨씬 더 잘 작동한다고 밝혔다. 이에 따라 미국기관들과 글로벌 기업들이 인공지능의 알고리즘 편향에 적극적으로 대처하여 얼굴인식기술의 편견문제가 개선되고 있으며 오류율도 2년마다 절반씩 줄어드는 것으로 나타났다[10]. 세 번째 예로는 트위터의 자동 이미지 자르기 기능이 있다. 자동 이미지 자르기는 사용자가 찍은 사진을 트위터에 올리는 과정에서, AI 알고리즘이 핵심 인물을 자동으로 선택해 잘라내서 미리보기용 사진을 만드는 기능이다. 조사 결과, 트위터 AI 알고리즘은 흑인보다는 백인을, 남성보다는 여성을 우선적으로 핵심 인물로 선택했다. 트위터는 모바일 앱을 업데이트하고, 트위터 홈페이지에서 AI 알고리즘을 적용한 자동 이미지 자르기 기능을 중단할 계획이라고 밝혔다[11]. 네 번째 예로는 COMPAS가 있다. COMPAS는 피고인의 정보를 종합해 재범가능성을 판단하여 구속 여부를 제안하는 인공지능 알고리즘이다. 미국의 언론사 프로퍼블리카는 미국 내 여러 주 법원에서 사용하는 AI 시스템 COMPAS가 흑인에 대해 불리한 판단을 내린다는 사실을 보도했다. COMPAS는 명시적으로 인종을 판단 기준으로 삼지 않는다고 밝혔지만, 프로퍼블리카는 COMPAS가 플로리다에서 체포된 범죄자 1만 명을 대상으로 재범 가능성을 판단한 결과 흑인의 재범 가능성이 백인보다 2배 이상 높게 예측했다고 주장했다[12].

3.3. 제도적 대응 조치 현황

그동안 기술 수준의 향상을 우선으로 초점을 맞추어 왔지만, 최근은 인공지능으로 인한 편향, 차별, 프라이버시 문제 등 윤리적인 문제가 이슈가 되고 있다. 인공지능 윤리의 중요성을 인식하고 우리나라에서는 가이드라인과 준수 원칙, 포럼 등을 통해 인공지능 윤리 문제에 대응해왔다. 2018년 4월에는 지능정보사회 윤리가이드라인을 발표하며 공공성, 책무성, 투명성, 통제성의 공통원칙(PACT)을 바탕으로 개발, 사용 등의 제부 지침을 제시했다[13]. 2019년 11월에는 방송통신위원회에서 이용자중심의 지능정보사회를 위한 원칙을 발표했다. 주요 IT기업과 전문가의 의견수렴 과정을 거쳐 만들어진 이 원칙은 AI기술과 같은 신기술 도입시 초래할 수 있는 위험을 대비하기 위해 지능정보사회의 모든 구성원들이 고려할 투명성과 설명과 능성, 책임성, 차별금지 등 공동의 기본원칙을 제시했다[14].

지능정보사회윤리가이드라인과 원칙은 AI에 특화된 조치가 아닌 클라우드, 빅데이터, IoT, 블록체인 등과 같은 지능정보기술을 대상으로 했기 때문에 AI와 특화된 원칙 혹은 가이드라인이 필요하다. 2020년 12월에는 "사람이 중심이 되는 인공지능 윤리기준"을 발표했는데 인간 존엄성, 사회의 공공선, 기술의 합목적성의 3대 기본원칙을 바탕으로 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성의 세부 요건을 제시했다[15]. 인공지능에 특화된 윤리기준이지만 법이나 지침이 아닌 자율 규범이기 때문에 위 기준은 권고가 될 뿐 의무적으로 지켜야 할 사항은 아니다.

3.4. 활용 서비스별 윤리적 문제

3.4.1. 로보어드바이저

로보어드바이저는 로봇(robot)과 자문(advisor) AI로 고객 성향에 따라 자산을 배분해주는 서비스다. 전통적 자문 서비스였던 프라이빗 뱅크(PB)서비스와 비교하면 거래 수수료가 낮고, 초기 투자금액도 몇백만 원 단위로 낮다. 때문에 고액 자산가를 중심으로 운영되었던 PB 서비스와 달리 로보어드바이저의 주 고객은 일반 개인 투자자이다. 2020년 3월 기준, 로보어드바이저 서비스 가입자 수는 18만 6천 명에 달하며, 2018년부터 이용자 수가 빠르게 증가했다. 2020년 1분기에만 5만 명이 새로 가입한 것으로 집계되었다. 규모 역시 매년 가파르게 몸집을 키워나가고 있으며, 2025년에는 30조 원의 시장 규모를 가질 것으로 전문가들은 예상하고 있다[16].

로보어드바이저는 자산 운용 주체에 따라 자문형 로보어드바이저와 일임형 로보어드바이저로 나뉜다. 자문형 로보어드바이저는 AI 알고리즘을 통해 고객에게 알맞은 투자 포트폴리오를 추천하는 서비스이다. 고객은 로보어드바이저의 추천을 바탕으로 직접 자산을 운용한다. 일임형은 자산 운용까지 모두 로보어드바이저가 전담하는 서비스이다. 이 경우 로보어드바이저가 직접 거래를 하는데, 이 과정에서 문제가 생긴다. 로보어드바이저에는 자체적인 감시 기능이 없다. 또한, MS 단위의 매우 짧은 시간 안에 자동으로 고속 거래를 하는 로보어드바이저 특성상 시장 안정성, 공평성에 악영향을 줄 수 있다. 또한, 건전한 가격 형성도 훼손시킬 수 있다.

자문형과 일임형 두 로보어드바이저에서 공통으로 사용되는 기술은 크게 인공지능 모델 기반 서비스와 전문가들이 직접 규칙을 설정하고 상황을 설정한 알고리즘 기반 서비스로 나뉜다. 본 논문에서 다루는 것은 인공지능 모델 기반의 로보어드바이저이다. 인공지능(AI) 모델 기반 서비스는 먼저 데이터를 사전에 머신러닝과 딥러닝기법으로 학습하여 모델을 생성한 뒤 실시간으로 들어오는 고객과 시장 데이터를 변수로 넣어 결과를 도출한다. 여기서 치명적인 문제점이 생긴다. 머신러닝과 딥러닝으로 생성된 모델은 그 자체가 블랙박스라는 점이다. 모델 내부는 무수히 많은 네트워크로 이루어져 있어 단순히 기준 금리 변동, 해외 증시 주가 변동 등과 같은 외부 파라미터만으로는 로보어드바이저가 내린 결론까지 이르는 과정을 해석하기 힘들다. 즉 결과에 대한 원인을 찾기가 어렵다. 이러한 설명가능성의 결여는 자산 투자에 대한 책임 소재 불분명의 문제까지 이어지게 한다.

3.4.2. 보험심사

보험 심사 영역도 AI 기술이 활용되고 있다. 보험사들은 AI 기술을 보험 가입 심사, 보험금 지급 심사에 활용하면서 비용을 크게 절감시키고 있다. 특히 국내 한 보험회사는 AI로 보험금 심사의 비중을 높여 약 80억 원의 비용 절감 효과를 얻었다[17]. 보험사들은 AI를 통해 운영 비용 절감뿐 아니라, 서비스 질의 경쟁력 확보에도 이익을 얻고 있다. 소액보험 청구 같은 저위험 심사의 경우는 AI 시스템을 통해 처리 효율성을 높이고, 심사자는 고위험 심사에 집중할 수 있게 된 것이다.

그러나 보험 심사 역시 AI 기술과 관련한 윤리 문제를 안고 있다. 먼저 AI가 과연 공정한 판단을 내렸는가에 대해서는 많은 의문이 있다. AI의 판단 근거가 되는 데이터의 비중 중 특정 집단에 유리한 데이터가 높으면 AI가 편향된 판단을 하게 된다. 실제 미국에서는 자동차보험료를 산출하는 과정에서 같은 조건일 때 흑인 운전자가 백인 운전자보다 더 많은 보험료를 산정한 사례가 있었다[18]. 다음으로는 AI의 블랙박스 특성으로 인한 설명 불가의 문제가 있다. AI의 보험 심사결과로 보험금이 지급이 되지 않았을 경우 사유에 관해서 설명할 수 없다는 것이다. 보험금 미지급 통보 시 소비자에게 그 사유에 관해서 설명하지 않으면 신뢰의 문제가 생기기 때문에 AI를 이용한 보험금 지급을 더욱 활성화하기 위해서는 위 문제를 반드시 해결해야 한다.

〔표 1〕 시뮬레이션 피라메터

분류	문제
로보어드바이저	시장 안전성, 공평성 훼손 결과에 대한 설명 불가능 투자 책임 소재 불분명
보험심사	편향적 판단 심사결과 설명 불가능

3.5. 데이터 파이프라인에 따른 윤리적 문제 발생 가능성

데이터 파이프라인은 크게 수집, 전처리, 모델링, 시각화의 네 가지로 분류된다. 자료수집단계에서는 인공지능의 알고리즘을 만들고 학습시킬 데이터를 모으는 단계이다. 인공지능의 판단은 데이터에 기초하기 때문에 이 단계에서 수집하는 데이터가 인공지능의 윤리적 판단에 영향을 주게 된다. 자료수집 및 전처리 단계에서 잠재적인 윤리적 문제는 데이터 편향과 프라이버시 두 개가 있다.

데이터 편향이란 데이터가 특정 집단 혹은 케이스와 같은 한쪽에 치우친 상태를 의미한다. 데이터 편향을 발생시키는 요인은 다양하다. 먼저 통계적 종속성에 의해 발생하는 경우가 있다. 통계적 종속성이란 어떠한 특성과 특성 사이 이에 상관관계가 있는 것을 말한다. 활용을 지양해야 하는 인종, 성별 등의 데이터와 분석 시 활용해야 하는 데이터 사이에 상관관계가 있다면 인종, 성별과 같은 민감한 데이터를 분석 단계에서 삭제한다고 하더라도 결과적으로 차별을 나을 수 있는 대리 차별이 발생할 수 있다.

다음으로는 데이터 자체가 오염되어있는 경우 편향이 발생할 수 있다. 데이터 내에 이미 편향성이 존재한다면 데이터를 학습한 인공지능의 판단 역시 편향될 가능성이 크다는 것이다. 이 문제의 경우 분석 후 사후적 조치로 피해를 줄이는 방법을 취할 수 있으나 효과가 떨어진다. 다음은 절대적인 데이터양이 부족할 경우 역시 편향이 생긴다. 이 경우는 특정 집단의 데이터양이 부족할 경우 절대다수의 데이터를 기반으로 소수집단을 판단하는 문제가 생긴다. 다음으로는 악의적 혹은 고의적 데이터 편향 때문에 발생한 경우다. 이는 해킹과 같은 악의적 공격이 원인이며 데이터 보관과 관리의 보안 수준을 높여야 할 편향으로 인한 피해를 줄일 수 있다.

자료수집 및 전처리 단계에서 흔히 발생할 수 있는 윤리적 문제에는 프라이버시 이슈도 있다. 정보 제공자에게 자료수집 여부를 묻지 않고 자료를 수집할 때 자료수집 단계에서 프라이버시 문제가 발생한다. 또한, 적절한 설명과 동의와 함께 수집된 데이터라고 하더라도 수집의 데이터의 가명화 수준에 따라 문제가 발생할 수 있다. 다른 데이터와 결합하면서 개인정보가 식별되는 문제가 발생할 수 있기 때문이다.

알고리즘 모델링 과정에서는 성능과 관련한 두 가지 문제가 발생할 수 있다. 오버핏과 언더핏의 문제이다. 언더핏은 데이터의 많은 공통 특성 중 적은 일부 특성만을 반영하여 모델링 했을 때 생기는 문제로, 판단의 근거가 빈약하여 정확도를 떨어트린다. 반대로 오버핏은 지나치게 많은 특성을 반영하여 모델링 했을 때 생기는 문제를 말한다. 오버핏은 모델의 성능을 떨어트리는 문제 외에도 프라이버시 이슈를 발생시킬 수 있는 가능성도 가지고 있다. AI 모델의 성능을 높이기 위해 개별 데이터까지 학습하게 되면 오버핏이 발생하곤 한다. 오버핏 된 모델이 학습한 데이터 안에 개인정보가 포함되어 있을 경우, 개인정보가 노출될 수 있으며 재식별 공격의 위험도 커진다.

〔표 2〕 데이터 파이프라인 단계별 윤리적 문제

분류	문제
수집 및 전처리	통계적 종속성, 데이터 오염에 의한 편향적 판단 프라이버시 노출
알고리즘 모델링	모델의 오버핏으로 프라이버시 노출 결과, 추론 과정 설명 불가능

한편 알고리즘 모델은 그 내부를 들여다볼 수 없는 특징을 가진다. 규칙에 따라서 추론하는 것이 아니라, 수많은 노드의 네트워크로 결과를 내는 것이기 때문에 머신러닝이 출력한 결과는 근거를 들어 설명하기가 어렵다. 이를 블랙박스라고 부르는데, 블랙박스라는 특징상 결과로 나온 것들에 대해 추론 과정을 설명하기 어렵고 이는 인공지능이 불공정한 결과를 출력한다고 하더라도 그 원인을 찾기가 어렵다는 문제로 이어질 수 있다.

Ⅳ. 해결방안

4.1. 해외 인공지능 윤리 정책

4.1.1. 유럽

유럽은 2018년도부터 인공지능 윤리에 대해 적극적으로 대응하고 있다. 2018년 Declaration of cooperation on Artificial intelligence, Artificial Intelligence for Europe 발표를 바탕으로 사회 각계 공개 의견수렴, 법 제도 논의가 이루어지고 있다. 유럽위원회에서는 52명의 AI 전문가로 구성된

AI-HLEG(High Level Expert Group on AI)를 발족하여 "신뢰할 만한 AI 윤리 가이드라인"과 "신뢰성 있는 AI 개발을 위한 정책 보고서"과 "AI 윤리 가이드라인 및 점검 평가 목록"을 발표했다.

19년도에 발표한 "신뢰할 만한 AI 윤리 가이드라인 "에서는 신뢰할 수 있는 AI의 3필수 3대 요소로 적법성, 윤리성, 견고성을 제안하고 이를 바탕으로 인간행위자와 감독, 기술적 견고성 및 안전, 프라이버시와 데이터 거버넌스, 투명성, 다양성 차별금지 공정성, 사회환경적 복지, 책임성의 7개의 요구사항을 포함했다. 또한, 필수 3요소와 7대 요구사항의 구현을 위한 기술적, 비기술적인 방안까지도 함께 가이드라인에서 제안했다. "신뢰할만한 AI 윤리 평가 목록"에서는 인간 기본, AI 윤리 7대 요구사항과 관련하여 140개의 체크 목록으로 구성되어있다[19].

4.1.2. 미국

미국은 오랫동안 일관적으로 국제사회에서 과학 경제 경제적 리더를 목표로 AI 정책을 펴왔다. 규제보다는 AI 활용에 초점을 맞추며 AI 성능 발전과 비교하여 AI 윤리는 성장이 더딘 편이다. 몇 AI 윤리 원칙은 정부가 아닌 민간이 중심으로 규제 원칙이 세워졌다. 그러던 중 중국의 AI 기술 성장을 견제하며 미국은 글로벌 AI 파트너십에 참여하게 되었는데, 이는 미국의 AI 정책 기조가 성능 발전, 혁신에서 신뢰성 있는 AI로의 전환기가 되었다.

미국에서는 민간 영역에서 AI 윤리 원칙을 초기에 이끌어왔다. 글로벌 AI 기업들이 모인 Partnership on AI에서 2016년에 인공지능 윤리 원칙을 검토하기 시작했다. 이 연구소에서 발표한 Tenets에는 학계, 기업, 정부 등 다양한 주체들 간의 협력과 학교, 대중 교육을 통해 인공지능 윤리 문제에 대응하는 제안이 포함되어 있다[20]. 이후 2019년도에는 국립과학기술위원회(NSTC, National Science and Technology Council)와 인공지능특별위원회(Select Committee on Artificial Intelligence)가 AI 관련 모든 행정부 및 산하 기관이 준수해야 할 6대 전략목표를 발표했다.

4.2. 국내 조치

우리나라에서는 정부와 공공기관, 비영리기관과 기업이 AI의 윤리적 사용을 위한 윤리헌장을 발표했다. 대표적으로 한국 인공지능 윤리협회는 인공지능 개발자와 소비자가 지켜야 할 인공지능 윤리헌장을 만들었다. 2019년에 만들어진 인공지능 윤리헌장 선한 인공지능 추구를 기본 개념으로 했다. 2019년, 2021년 총 두 번의 개정을 거쳐 인간과 인공지능의 관계, 선하고 안전한 인공지능, 인공지능 개발자의 윤리, 인공지능 소비자의 윤리, 공동의 책임과 이익의 공유의 총 5장 40조 항으로 구성됐다[22]. 2020년 12월에는 과학기술정보통신부에서 AI 윤리 규정을 발표했다. 인간 존엄성의 원칙, 사회의 공공선 원칙, 기술 합목적성 원칙의 3대 기본원칙을 토대로 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터관리, 책임성, 안정성, 투명성의 AI가 갖춰야 하는 10대 핵심 요건을 담았다[23]. 우리나라의 각 기업도 AI 윤리 문제에 대응하고 있다. 카카오는 2018년 초 일찍이 카카오 알고리즘 윤리헌장을 제정하였는데 이는 우리나라 IT 대기업 최초로 발표한 AI 윤리헌장이었다. 또한, 이루다 사건 이후 전 직원을 대상으로 한 AI 윤리 교육도 추진한 바 있다. 삼성전자는 2018년에 국내 기업 최초로 Partnership on AI에 가입했으며 현재 AI 윤리 기준 제정을 준비 중이다. 네이버는 2021년 2월 서울대학교와 공동으로 AI 윤리 준칙을 발표했다.

Ⅴ. 결 론

금융 분야에서 AI는 현재 다양하게 활용되고 있고, 회사의 비용도 크게 절감시키며 필수 기술로 자리 잡고 있다. 그러나 본 논문에서 서술한 윤리적 문제를 해결하지 않으면 금융 분에서 AI는 사회적인 신뢰를 얻지 못할 것이다. 금융 분야에서의 AI를 확대하기 위해서는 인공지능으로 인해 발생하는 윤리적인 문제에 대한 대책을 마련해야 한다. 본 논문에서는 인공지능으로 발생할 수 있는 윤리 문제를 활용 도메인에 따라 데이터 분석 파이프라인 절차에 따라 나누어보았다.

현재 인공지능이 가장 많이 활용되고 있는 로보어드바이저에서는 매우 짧은 시간 내에 일어나는 거래로 인해 시장 건전성에 악영향을 줄 수 있는 문제가 있었다. 또한, 블랙박스 모델 특성상 로보어드바이저가 내놓은 결과에 관해 설명할 수 없다는 문제도 발생할 수 있음을 확인했다. 설명 불가능의 문제는 보험 심사에서도 발생할 수 있다. AI가 내린 보험금 심사결과에

대해서 명확한 사유를 설명하기 어려운 문제가 발생할 수 있다. 데이터 파이프라인의 자료수집과 전처리 단계에서는 통계적 종속성과 데이터 오염으로 인한 데이터 편향이 문제가 될 수 있다. 또한, 자료수집 및 처리 과정에서 프라이버시의 문제가 생길 수도 있다. 알고리즘 모델링 과정에서도 두 개의 문제를 낳을 수 있다. 먼저, 오버핏으로 인해 재식별 공격의 위험이 커져 프라이버시의 문제가 발생할 수 있다. 그리고 블랙박스의 특성상 알고리즘의 결과를 설명하기 어려운 설명 불가능의 문제가 생길 수 있다.

인공지능으로 인해 발생할 수 있는 윤리적 문제에 유럽과 미국이 대응하고 있는 방식도 살펴보았다. AI 윤리 문제에 대응하기 위해 우리나라에서도 2018년부터 본격적으로 신뢰할 수 있는 AI를 만들기 위한 방안을 모색하고 있다. 2020년 12월에 발표한 “AI 윤리기준”에서는 인간성을 목표로 개발자, 서비스 제공자, 사용자가 지켜야 하는 10가지 핵심 요건을 제시했다. 그러나 이 “AI 윤리기준”은 법적인 구속력은 갖지 않는다. 그저 AI 개발자와 서비스 제공자, 사용자의 자율에 맡기는 것이다. 또한, 구체적인 지표가 아니므로 개념이나 해석의 차이가 생겨 실효성에 의문이 생길 수 있다.

우리나라에서는 여러 금융사가 AI를 이용한 서비스와 상품을 판매하고 있다. 앞으로의 금융 분야 AI 활성화를 위해서는 반드시 위 윤리적 문제를 관리해야 한다. 윤리적 문제는 인공지능 서비스의 사회적 신뢰도와 관련이 크기 때문이다. 인공지능 윤리헌장, 각 기업에서 발표한 인공지능 윤리 준칙, 정부에서 발표한 “AI 윤리기준”이 있지만, 원론적인 수준에 그치고 있다. 현재까지 발표된 윤리기준을 바탕으로 관련 법령을 수정하고 구체적인 준칙을 포함한 가이드라인을 배포하여 AI 윤리 규범, 기준의 실효성을 높일 수 있을 것이다.

참 고 문 헌

[1] Gartner, Gartner Hype Cycle for Emerging Technologies, 2021.08.01.

[2] Brian Burke, “Top Strategic Technology Trends for 2021”, Gartner, pp. 13, 2020.

[3] 한국리서치, AI와 인간의 공존, 그리고 ‘윤리성’ 2021.08.10.

[4] 오요한외 1명,“인공지능 알고리즘은 사람을 차별하는가?”, 과학기술학연구, vol 18, no 3, pp.153-215, 2018.

[5] 임용외 2명, “인공지능과 시장경쟁 : 데이터에 대한 규율을 중심으로”, 한국경제포럼, vol 12, no 3, pp.35-58, 2019

[6] 양종모, “인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안”, 법조, vol 66, no 3, pp.60-105, 2017

[7] 신용진, “AI서비스에서의 개인정보보호를 위한 책임과 원칙의 적용에 관한 연구”, 한국범죄정보연구, vol 7, no 1, pp.45-74, 2021

[8] 김승래, “4차 산업혁명과 금융 데이터산업의 활성화를 위한 정책과제”, 지급결제학회지, vol 11, no 1, pp.95-129, 2019

[9] Amy Kraft(CBSNEWS), Microsoft shuts down AI chatbot after it turned into a Nazi [Internet], Available: 웹사이트 URL, 2021.07.25.

[10] 윤영주(AI타임스), 미국표준기술연구소(NIST) “안면인식 시스템, 인종별 오류 편차 커”, 2021.07.23.

[11] 이은주(IT Chosun), 트위터 AI 알고리즘 "백인과 여성 선호", 2021.07.25.

[12] 전승우, 편견’에 취약한 AI… 인종과 성차별 망언 쏟아내기도, 2021.07.28

[13] 과학기술정보통신부 한국정보화진흥원,“지능정보사회 윤리 가이드라인”, pp. 23, 2018.

[14] 방송통신위원회, “이용자 중심의 지능정보사회를 위한 원칙”, pp.5, 2019.

[15] 4차산업혁명위원회, “사람이 중심이 되는 인공지능 윤리기준”, pp.40, 2020.

[16] 정한민외 1명, “인공지능 기반 로보어드바이저 운용 및 기술 동향”, 정보통신기획평가원, 대전광역시, ISSN 1225-6447, 2020.

[17] 오현길(아시아경제), 당신이 청구한 보험금, 이제 AI가 심사한다, 2021.08.05.

[18] 하채림(연합뉴스), AI에 맡겼더니 보험료·가입 차별?…감독체계 필요, 2021.08.07.

[19] European Commission, Ethics guidelines for trustworthy AI, 2021.08.13

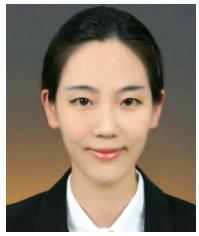
[20] Partnership on AI, Advancing positive outcomes for people and society, 2021.08.11.

[21] 세계법제정보센터, “미국 인공지능 법제”, 2019
[22] 한국인공지능윤리협회, 한국인공지능윤리협회 인공지능 윤리 헌장, 2021.08.14.
[23] 과학기술정보통신부, 사람이 중심이되는 인공지능 윤리기준, 2021.08.15.

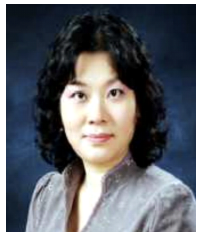
〈저 자 소 개〉

이 수 련 (Su Ryeon Lee)
학생회원
2018년 2월~현재 : 서울여자대학교 정보보호학과
<관심분야> 정보보호, 데이터분석, 데이터시각화

이 현 정 (Hyun Jung Lee)
2005년 8월 : 서울여자대학교 컴퓨터공학과 졸업
2009년 8월 : 성균관대학교 정보통신대학원 정보보호학과 졸업
2008년 3월~현재 : 고려대학교 정보보호대학원 정보보호학과 박사과정
<관심분야> 정보보호, 네트워크, 데이터시각화



이 아 람 (Aram Lee)
2009년 2월 : 서울여자대학교 정보보호학과 졸업
2020년 2월 : 한국방송통신대학교 법학과 졸업
<관심분야> 정보보호, 개인정보보호, 데이터 보호법



최 은 정 (Eun Jung Choi)
1997년 2월 : 서울여자대학교 컴퓨터학과 (이학사)
2000년 2월 : 서울여자대학교 대학원 컴퓨터학과 (이학석사)
2005년 8월 : 서울여자대학교 대학원 컴퓨터학과 (이학박사)
2006년 3월~현재 : 서울여자대학교 정보보호학과 교수
<관심분야> 빅데이터, 인공지능, 악성코드, 개인정보보호