ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2024 September: 30(3): 373~392 http://dx.doi.org/10.13088/jiis.2024.30.3.373

XAI 설명방식이 사용자의 인지된 이해도와 신뢰도에 미치는 영향: 조절초점의 역할

문상룡
국민대학교 비즈니스IT전문대학원
(sr.moon@kookmin.ac.kr)

안현철
국민대학교 비즈니스IT전문대학원
(hcahn@kookmin.ac.kr)

한윤
국민대학교 비즈니스IT전문대학원
(yoonhan@kookmin.ac.kr)

AI 기술의 발전에 따라 AI 모델이 생성한 결과를 이해하기 어려워지는 '블랙박스(Black-Box)' 문제가 중요한 이슈로 대두되고 있다. 이 문제를 해결하기 위한 설명 가능한 인공지능(XAI) 연구는 기술적 접근을 넘어서 사용자 중심적 접근으로 발전할 필요가 있다. 본 연구는 사용자 이해도를 높일 수 있는 XAI 설명방식을 제안하기 위해, 사용자의 개인 성향 특성인 '조절초점' 개념을 도입하였다. 600명의 일반 성인을 대상으로 반사실적(Counterfactual) 설명방식과 특징 중요도(Feature Importance) 설명방식을 비교하는 실험연구를 수행한 결과, 반사실적 방식이 사용자 이해도에서 유의미하게 높은 결과를 나타냈다. 또한, 조절초점 성향이 XAI 설명에 대한 사용자 이해도에 영향을 미치는 중요한 요인임이 확인되었다. 그리고 XAI 설명에 대한 사용자 이해도는 AI 시스템에 대한 사용자 신뢰도의 향상으로 이어진다는 사실도 확인하였다. 본 연구는 사용자 중심의 XAI 연구에 개인 특성을 고려한 확장된 이론적 프레임워크를 제시함으로써, AI 사용자에게 어떤 XAI 설명방식이 가장 적합한지에 대한 실질적인 가이드라인을 제공한다.

주제어 : 설명가능한 인공지능, 인지된 이해도, 조절초점, 신뢰

논문접수일 : 2024년 9월 20일 논문수정일 : 2024년 9월 27일 게재확정일 : 2024년 9월 27일
원고유형 : Fast Track 교신저자 : 한윤

1. 서론

인공지능(이하 AI) 기술은 금융, 의료, 교육, 마케팅, 미디어 등 다양한 분야에서 급속히 발전하고 있으며, 우리의 실생활에 혁신적인 변화를 가져오고 있다. 그러나 AI 기술이 고도화될수록 AI 모델의 성능은 향상되지만, 그 복잡성으로 인해 전문가조차도 AI 결과를 해석하기 어려운 '블랙 박스(Black Box)' 문제가 중요한 이슈로 떠오르고 있다. 이 문제는 AI 시스템이 뛰어난 성능에도 불구하고 중요한 의사결정 과정에서 활용되지 못하는 주요 장애 요인으로 작용하고 있다(Arrieta et al., 2020; Riveiro & Thill, 2021). AI가 모든 분야에서 널리 활용됨에 따라, AI 결과에 대한 설명 가능성에 대한 요구가 증가하고 있으며, 이는 AI 모델의 성능만큼이나 중요한 문제로 인식되고 있다. 특히, 유럽연합의 일반 데이터 보호 규정(GDPR)과 대한민국의 개인정보보호법과 같은 법률에서는 AI 결과에 대한 설명을 의무화하고 있으며, 이러한 규제는 앞으로 더욱 강화될 것으로 예상된다.

이러한 배경에서 AI 모델의 결과를 설명할 수 있는 기술로서 '설명 가능한 인공지능(Explainable AI, 이하 XAI)'은 AI 연구 분야에서 중요한 주제로 부상하고 있다. 그러나 지금까지의 XAI 연구는 주로 기술적 측면에 초점을 맞추었으며, 이로

인해 많은 기술이 개발되었지만 사용자가 실제로 이해할 수 있는 수준에는 미치지 못하고 있다. 오히려 다양한 XAI 기술 중에서 사용자가 최적의 기술을 선택해야 하는 추가적인 부담이 발생하고 있다(Arrieta et al., 2020). 이와 같은 한계를 극복하기 위해 사용자 중심의 관점에서 연구를 확장해야 한다는 주장이 제기되고 있다(Lu et al., 2020; Miller, 2019). 최근 사용자 중심의 XAI 연구에서는 XAI 설명방식의 효과성 연구(Warren, 2022; Ramon, 2021)와 인지심리학 등 다양한 분야의 연구 결과를 XAI 분야에 적용하는 등의 시도가 이루어지고 있다. 특히, 사용자의 개인 성향과 정보처리 과정에서 설명에 영향을 미치는 다양한 요인에 대한 연구가 주목받고 있다. 그러나 현재까지 이러한 요인들은 부분적으로만 적용되고 있으며(Herm, 2023; Riefle et al., 2024), 향후 연구에서는 이들 요인을 보다 포괄적으로 반영할 수 있는 방향으로 발전되어야 할 것이다.

본 연구는 XAI 설명의 효과에 영향을 미치는 요인 중 기존 연구에서 충분히 다루지 못한 사용자의 개인 성향을 고려하여, XAI 설명에 대한 사용자의 이해도를 향상시키는 것을 목표로 한다. 특히, Higgins et al. (1997)가 제안한 조절초점 이론(Regulatory Focus Theory)을 이론적 기반으로 하여, 사용자의 행동 동기 성향이 XAI 설명에 미치는 영향을 분석하고자 한다. 본 연구는 다음과 같은 연구 문제(Research Question)를 제시한다.

RQ: “개인의 동기 성향인 조절초점이 XAI 설명에 대한 사용자의 인지된 이해도에 어떤 영향을 미치며, XAI 설명에 대한 사용자의 이해도 향상은 AI 시스템에 대한 사용자 신뢰도 향상에 기여하는가?”

이 연구 문제는 XAI 알고리즘 개발에 초점을 맞추어 온 기존의 기술적 성과와 실제 사용자가 느끼는 설명에 대한 인지적 차이를 해소하는 데 기여할 수 있다. 또한, XAI 설명에 영향을 미치는 개인차 요인을 실증적으로 규명함으로써, XAI 설계에서 사용자 특성의 중요성을 보여줄 수 있을 것이다.

이 연구문제를 해결하기 위해, XAI 설명방식, 사용자의 조절초점, XAI 설명에 대한 사용자의 인지된 이해도, 그리고 AI에 대한 사용자 신뢰도 간의 관계를 규명하는 연구모델을 제시하고 실증 분석을 수행하였다. 600명의 성인을 대상으로 실험연구를 통하여 분석한 결과, XAI 설명방식에 따라 사용자 이해도가 유의미하게 차이가 있음을 확인하였으며, 사용자의 조절초점은 XAI 설명방식에 따른 사용자 이해도에 영향을 미치는 중요한 요인으로 나타났다. 본 연구는 사용자의 개인 특성에 맞춘 최적의 XAI 설명방식을 제공함으로써 사용자의 이해도를 높이는 데 기여할 수 있으며, 궁극적으로 AI에 대한 사용자의 신뢰도 향상에도 긍정적인 영향을 미칠 수 있음을 실증적으로 보여준다.

2. 이론적 배경

2.1. 인공지능 및 설명가능한 인공지능(XAI)

인공지능(AI) 기술은 기계학습과 심층학습의 발전에 힘입어 단순한 추론 문제에서부터 현실세계의 복잡한 비선형 문제를 해결할 수 있는 수준까지 지속적으로 발전하고 있다. 특히, 심층학습 기술은 대용량 학습 데이터 처리와 최적화 알고리즘 개발을 통해 금융, 의료, 교육, 마케팅 등 실생활의 다양한 분야에서 AI가 중요한 기술로 자리 잡게 하였고, 이는 인공지능 시대를 열게 된 주요 동력 중 하나로 평가된다. 그러나 AI 기술이

고도화됨에 따라 모델의 복잡성이 증가하면서, AI 모델의 결과를 해석하기 어려워지는 이른바 '블랙 박스(Black Box)' 문제가 대두되고 있다. 이는 AI 시스템의 뛰어난 성능에도 불구하고, 중요한 의사결정 과정에서 AI의 활용을 제한하는 주요 요인으로 작용하고 있다(Arrieta et al., 2020; Riveiro & Thill, 2021).

AI 결과에 대한 설명 필요성은 AI 기술의 확산과 함께 점점 더 중요하게 인식되고 있으며, 기술적 문제 뿐만 아니라 법적·윤리적 요구로도 발전하고 있다. 유럽연합(EU)은 2024년 AI 규제법을 통해 위험 등급에 따라 AI 기술을 차등 규제하고, AI 개발 기업에게 투명성 의무를 부여하고 있다. 대한민국 역시 2024년 3월 개인정보보호법 개정을 통해 AI를 활용한 자동화된 의사결정에 대한 설명 요구권을 구체화하였다. 이러한 규제 움직임은 전 세계적으로 확대될 가능성이 높으며, AI 결과 설명의 중요성은 더욱 커질 것으로 예상된다. 또한, AI 분야에서 중요한 화두 가운데 하나인 AI 공정성 문제를 예방하기 위한 핵심 과제로도 AI 설명이 강조되고 있으며(Ahn et al., 2020; Kim et al., 2023), 정밀의학, 자율주행, 금융 등 AI 결과가 중대한 영향을 미치는 분야에서는 특히 AI 결과에 대한 설명이 필수적인 요소로 인식되고 있다.

이러한 배경에서 '설명 가능한 인공지능(XAI)' 기술은 AI 연구의 중요한 주제로 부상하고 있다. XAI 기술은 크게 AI 모델내에 설명력을 구현하는 투명한 모델(Transparent Model)과 사후 설명 가능성 기법(Post-hoc Explainability Techniques)으로 별도의 XAI 설명 모델을 제공하는 방법으로 나눌 수 있다. 대표적인 알고리즘으로는 모델의 전체 예측에 각 특성이 얼마나 중요한 역할을 하는지 평가하는 글로벌 처리 방식인 특징 중요도(Feature Importance: FI) 알고리즘과 개별 예측에 대하여 모델의 예측 결과를 변경하기 위해 필요한 최소한의 특성 변화를 보여주는 로컬 처리 방식인 반사실적(Counterfactual: CF) 알고리즘이 있으며, 현재 최적의 XAI 기법을 선택하는 것이 어려울 정도로 다양한 알고리즘이 개발되고 있다(Arrieta et al., 2020; Mohseni et al., 2021; Alicioglu et al., 2022). 그러나 현재까지의 XAI 연구는 주로 AI 모델에 대한 설명을 자동으로 생성하는 알고리즘 개발에 중점을 두고 있어, 실제로 사용자가 이해할 수 있는 수준의 설명과는 상당한 차이를 보이고 있다(Ramon et al., 2021). 이러한 한계를 극복하기 위해, AI 개발자 중심의 기술적 연구에서 벗어나 사용자 중심의 XAI 방법론 개발로 연구 방향을 확장해야 한다는 주장이 제기되며, 최근에는 인지과학, 심리학, 마케팅 등 다양한 분야의 이론을 접목한 XAI 연구가 시도되고 있다(Lu et al., 2020; Miller, 2019).

2.2. 사용자 중심의 XAI 연구

사용자 중심의 XAI 연구는 XAI 설명의 필요성과 설명방식에 따른 사용자 인지 효과를 중심으로 진행되었다. 대부분의 연구는 XAI 설명이 사용자의 이해도와 신뢰도를 높이는 데 중요한 역할을 한다는 점을 보고하고 있다. Du et al.(2022)와 Panigutti et al.(2022)는 XAI를 적용한 의료 의사결정 지원 시스템이 의료진의 의사결정에 유의미한 영향을 미친다고 밝혔다. 또한, Hamm et al.(2023)은 위조 서명 탐지 실험에서 XAI 설명이 사용자의 인지된 설명 가능성, 유용성, 사용용이성에 긍정적인 영향을 미친다는 결과를 도출했다. 그러나 Lu et al.(2020)의 연구에서는 단순한 의사결정 트리로 제공된 설명이 오히려 설명이 없을

때보다 낮은 평가를 받았다. Kim et al.(2023)은 글로벌 및 로컬 방식과 시각화 기반 설명에서는 사용자 선호도가 높았지만, 텍스트 기반 설명에서는 텍스트량이 많아질수록 선호도가 감소한다는 점을 발견했다. 특히, 글로벌 방식보다 로컬 방식에 대해 사용자는 훨씬 긍정적으로 인식하는 경향이 나타났다.

XAI 설명방식에 따른 사용자 선호도와 이해도 비교 연구는 또 다른 중점 연구 주제이다. Warren et al.(2022)은 AI 시스템의 음주운전 판단에 대한 설명 과제에서 반사실적(CF) 설명방식이 사실적(factual) 방식보다 더 높은 사용자 만족도와 신뢰도를 제공한다는 결과를 도출했다. Xu et al. (2024)는 검색엔진 결과 설명에서 반사실적 방식이 사실적 설명보다 자동 평가 및 사용자 평가에서 우수한 결과를 나타냈으며, 반사실적 방식이 사용자가 원하는 결과에 도달하기 위한 직관적 설명을 제공함으로써 만족도와 신뢰도가 더 높았다고 강조하고 있다. 이러한 연구 결과는 반사실적 방식이 인간의 설명방식과 유사하다는 인지과학적 이론과 일치한다(Miller, 2019). Herm (2023)의 연구에서도 반사실적 방식은 특징 중요도(FI) 방식에 비해 과업 성과와 수행 시간에서 우수한 결과를 보였다. 그러나 Angerschmid et al.(2022)는 보험료 결정과 의료 진료 문제에서 XAI 설명이 AI 결정에 대한 사용자 신뢰도를 높이는 데 기여했지만, 반사실적 방식과 특징 중요도 방식 간에 사용자 신뢰도의 차이는 나타나지 않았다고 보고했다. Ramon et al.(2021)과 Riveiro et al.(2021)의 연구에서는 AI 모델의 결과가 긍정적인 경우에 특징 중요도 방식이 반사실적 방식보다 선호도가 높았으나, 결과가 부정적인 경우에는 반사실적 방식의 선호도가 상승한 것으로 나타났다. 이러한 연구 결과는 XAI 설명방식에 대한 사용자 이해도가 설명의 유형 뿐만 아니라 문제의 성격, 설명의 세부 요소 등 다양한 요인의 영향을 받는다는 것을 시사한다.

마지막으로, 사용자의 개인적 특성이 XAI 설명의 효과에 미치는 영향에 대한 연구는 XAI 분야에서 거의 다루지 않은 새롭게 주목받고 있는 주제이다. XAI 설명방식이 사용자의 객관적·주관적 이해도에 미치는 영향을 조사한 Riefle et al. (2024)의 실증 연구에서는 합리적 또는 직관적 인지 스타일에 따라 설명에 대한 객관적 이해도에 유의미한 영향을 미친다는 것을 발견했다. 하지만 주관적 이해도에서는 유의미한 차이가 나타나지 않았다. 이러한 연구는 사용자의 인지 스타일과 같은 개인 특성이 XAI 설명의 사용자 이해에 중요한 영향을 미치는 요인임을 실증적으로 보여주며, 향후 이와 같은 연구의 필요성을 인식하게 한다.

2.3. 조절초점 이론

Higgins et al. (1997)는 개인이 목표를 추구하는 방식에 따라 향상초점과 예방초점으로 구분한 조절초점 이론(Regulatory Focus Theory)을 제안하였다. 이 이론에 따르면, 사람들은 자신의 조절초점 성향에 맞는 방식으로 목표를 추구하며, 이 성향은 그들이 사용하는 동기 전략과 밀접하게 연관되어 있다. 향상초점 성향의 사람들은 현재보다 더 나은 상태를 달성하려 하며 긍정적인 결과와 이득에 주목하는 경향이 있으며, 예방초점 성향의 사람들은 안정성을 유지하고 부정적인 결과와 손실을 회피하려는 경향이 강하다.

조절 적합성 이론(Higgins, 2005)은 개인의 조절초점과 일치하는 조건에서 더 높은 참여 강도와 만족도, 그리고 과업 성과가 나타난다고 설명한다. 이는 개인의 판단, 의사결정, 태도 변화, 그리고

작업 수행에 영향을 미친다고 밝혀졌다. Lockwood et al.(2002)는 향상초점 성향자가 성공을 위한 전략을 제공하는 긍정적 롤모델에 더 큰 영향을 받으며, 예방초점 성향자는 실패를 피하기 위한 전략을 제공하는 부정적 롤모델에 의해 더 큰 동기부여를 받는다는 점을 실증적으로 입증했다. 이 연구는 개인의 조절초점에 맞는 전략이 그들의 동기부여에 중요한 영향을 미친다는 점을 보여주었다.

또한, Förster et al.(2005)는 조절초점과 글로벌-로컬 처리 스타일 간의 상호관계를 연구했다. 연구 결과에 따르면, 향상초점 성향자는 전체적인 패턴을 중시하는 글로벌 처리와 높은 적합성을 보이며, 예방초점 성향자는 구체적인 세부사항에 주목하는 로컬 처리와 더 높은 관련성을 나타냈다. 이러한 연구 결과는 조절초점 성향이 개인의 정보 처리 방식에 영향을 미친다는 것을 보여준다.

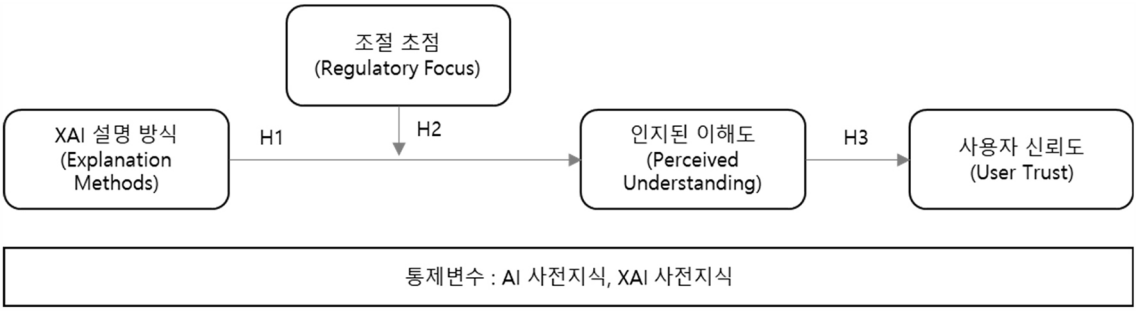
XAI 연구 분야에서는 아직 조절초점을 직접적으로 적용한 연구가 부족하지만, 인지 성향을 고려한 연구로서 Riefle et al.(2024)는 사용자의 합리적·직관적 인지 스타일에 따른 차이가 설명에 대한 사용자 이해도에 유의미한 영향을 미친다는 실증 연구 결과를 발표했다. 이는 사용자의 인지 스타일과 같은 개인적 특성이 XAI 설명의 효과에 중요한 역할을 한다는 것을 시사한다.

3. 연구 모델 및 가설 수립

본 연구는 <그림 1>과 같이 XAI 설명방식과 조절초점이 사용자의 인지된 이해도와 신뢰도에 미치는 영향을 규명하기 위한 연구모델을 제안한다. 이는 기존의 사용자 중심의 XAI 연구에서, XAI 설명방식에 따라 사용자가 인지하는 이해도와 신뢰도가 다르다는 연구 결과를 바탕으로 한다. 본 연구는 이러한 차이에 개인 특성으로서 Higgins et al. (1997)가 제안한 조절초점 이론을 적용하여, 변수들 간의 상호작용을 분석하는 연구모델을 설계하였다.

본 모델에서 조절초점(예방초점 vs. 향상초점)은 XAI 설명방식(반사실적 vs. 특징 중요도)이 XAI 설명에 대한 사용자의 인지된 이해도(Perceived Understanding)에 미치는 영향을 조절하는 조절변수로 가정하였다. 또한, XAI 설명방식에 의해 영향을 받은 사용자의 인지된 이해도는 AI에 대한 사용자 신뢰도(User Trust)에 직접적인 영향을 미치는 요인으로 설정하였다. 이와 함께, AI 사전지식과 XAI 사전지식을 통제변수로 설정하여 변수들 간의 관계를 분석하였다.

본 연구는 선행 연구를 토대로 XAI 설명방식이 사용자의 인지된 이해도에 미치는 영향을 분석하고자 가설을 수립하였다. Xu et al. (2024)은 검색



〈그림 1〉 연구 모델

엔진 결과에 대한 설명 방식에서 반사실적 방식이 사실적 방식보다 사용자 평가에서 더 우수한 결과를 보였으며, 이는 인간의 설명 방식이 반사실적 방식과 유사하다는 인지 과학 이론과 일치함을 확인하였다. 또한, Warren et al. (2022)은 AI의 음주운전 판단 설명 과제에서 반사실적 방식이 사실적 방식에 비해 사용자 주관적 평가에서 특히 유의미하게 높은 평가를 받았음을 보고하였다. 이러한 선행 연구 결과를 바탕으로, 본 연구에서는 사용자의 주관적 평가인 인지된 이해도 측면에서 반사실적 방식이 특징 중요도 방식보다 더 높은 평가를 받을 것으로 예상하여, 이를 토대로 다음과 같은 연구 가설을 제시한다.

H1: 반사실적 방식으로 XAI 설명을 제공할 때, 특징 중요도 방식으로 제공할 때 보다 사용자의 인지된 이해도가 더 높을 것이다.

Higgins et al.(1997)의 조절초점 이론에 따르면, 예방초점 성향을 지닌 사람들은 경계 및 회피 전략을 추구하며, 이러한 성향은 전반적인 특징보다는 세부적인 특징에 주의를 기울이는 경향이 있다. 반면, 열망과 성취를 중시하는 향상초점 성향은 전체적인 상황을 주목하는 경향이 있음을 발견하였다. Förster et al.(2005) 또한 머그잔의 가격을 이용한 실험을 통해, 예방초점 성향자는 로컬 처리에 더 적합하고, 향상초점 성향자는 글로벌 처리에 더 적합하다는 연구 결과를 제시하였다.

이와 같은 선행 연구들을 바탕으로, 본 연구에서는 XAI 설명 방식에서도 사용자의 조절초점에 따라 설명 방식에 대한 이해도 차이가 있을 것으로 예상한다. 즉, 예방초점 성향자는 부정적인 결과와 손실을 회피하는 구체적이고 직관적인 설명을 제공하는 로컬 처리 방식인 반사실적 설명에서 더 높은 이해도를 나타낼 것으로 예상된다. 반면, 향상초점 성향자는 AI 모델이 데이터를 어떻게 처리하는지에 대한 포괄적이고 전반적인 해석을 제공하는 글로벌 처리 방식인 특징 중요도 설명에서 더 높은 이해도를 보일 것으로 기대된다. 이러한 논리를 바탕으로 다음과 같은 연구 가설을 설정하였다.

H2: 사용자의 조절초점은 XAI 설명방식과 XAI 설명에 대한 사용자의 인지된 이해도 간의 관계에서 조절효과를 미칠 것이다.

H2a: 예방초점 성향의 경우, 반사실적 방식으로 XAI 설명을 제공할 때, 특징 중요도 방식으로 제공할 때 보다 사용자의 인지된 이해도가 더 높을 것이다.

H2b: 향상초점 성향의 경우, 특징 중요도 방식으로 XAI 설명을 제공할 때, 반사실적 방식으로 제공할 때 보다 사용자의 인지된 이해도가 더 높을 것이다.

Angerschmid et al.(2022)은 AI 결과에 대한 설명이 AI 시스템에 대한 사용자 신뢰도를 향상시키는 데 기여하며, XAI 가 사용자의 신뢰 형성에 중요한 요소임을 입증하였다. 또한, Warren et al.(2022)와 Larasati et al.(2020)의 연구는 XAI 설명 방식에 따른 사용자의 이해도와 신뢰도 사이에 밀접한 관계가 있음을 밝혔으며, 적절한 설명 방식이 사용자 신뢰도를 높일 수 있음을 실증적으로 제시하였다. 본 연구에서는 사용자의 조절초점에 부합하는 적절한 XAI 설명 방식이 사용자의 이해도를 향상시키고, 이러한 이해도의 향상이 AI 알고리즘에 대한 사용자 신뢰도를 증진시킬 것이라는 가정에 근거하여 다음과 같은 연구 가설을 수립하였다.

H3: XAI 설명에 대한 사용자의 인지된 이해도는 AI 알고리즘에 대한 사용자 신뢰도에 정(+)의 영향을 미칠 것이다.

4. 연구 방법

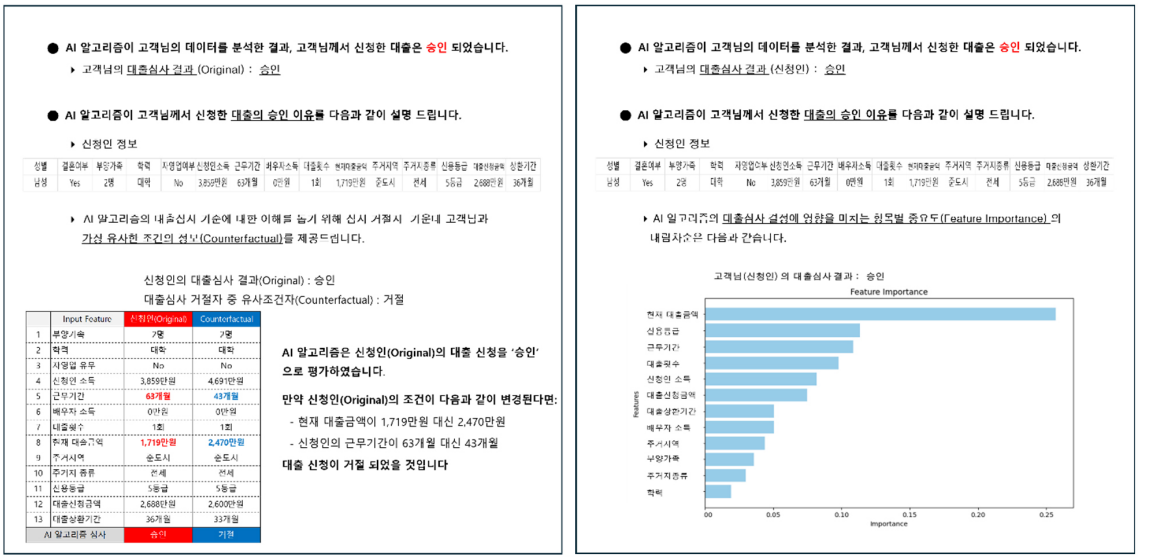
4.1. 실험 설계

본 연구는 XAI 설명방식과 사용자의 조절초점이 XAI 설명에 대한 사용자의 인지된 이해도에 미치는 영향, 그리고 인지된 이해도가 AI 모델에 대한 사용자의 신뢰도에 미치는 영향을 분석하였다. 실험은 AI 알고리즘이 대출 심사 업무를 수행하는 상황을 기반으로, 2(XAI 설명방식: 반사실적(CF) vs. 특징 중요도(FI)) × 2(조절초점: 예방초점 vs. 향상초점)의 집단 간 실험(between-subjects design)으로 설계되었다. XAI 설명방식으로는 대표적인 방식인 반사실적 방식과 특징 중요도 방식을 사용하였으며, 대출 심사 결과가 승인된 경우와 거절된 경우에 따른 차이가 실험 결과에 미치는 영향을 통제하기 위해 두 가지 경우를 모두 포함한 시나리오를 구성하였다.

설문 대상은 대출에 대한 기본적인 이해도를 가진 18세부터 70세까지의 일반 성인으로 설정하였으며, 응답자들은 예방초점과 향상초점에 관한 설문 항목을 통해 각각 예방초점, 향상초점, 중립의 세 유형으로 분류되었다. 이 중 예방초점과 향상초점에 속하는 응답자들을 대상으로 분석을 진행하였다. 각 응답자에게는 무작위로 대출 승인 또는 거절 시나리오 중 하나가 제시되었고, 이후 대출 심사 결과에 대한 설명은 반사실적 방식과 특징 중요도 방식이 무작위 순서로 제공되었다.

4.2. 실험 안내문

설문조사를 위하여 금융 대출 심사를 수행하는 AI 모델 시나리오를 개발하였다. 이 시나리오는 은행의 대출 심사 과정을 단순화하여 제시하였으며, AI 모델이 대출 승인 여부를 결정하는 상황을 두 가지 XAI 방식으로 설명하였다. 반사실적(CF) 방식은 AI 모델의 결정이 변경되기 위해 필요한 최소한의 입력 값의 변화를 구체적으로



〈그림 2〉 반사실적 설명방식(左)과 특징 중요도 설명방식(右) 안내문

제시하였으며, 특징 중요도(FI) 방식은 AI 모델의 결정에 영향을 미친 주요 특성들을 중요도 순서대로 막대 그래프로 시각화하여 제시하였다(안재현, 2020; 조수현, 신경식, 2022). 온라인 설문조사 과정에서는 AI 대출심사 결과를 설명하는 반사실적 방식과 특징 중요도 방식의 두가지 방식을 무작위 순서로 보여주었으며, XAI 설명방식에 대한 응답자의 이해를 높이기 위하여 설문 페이지별로 XAI 설명 결과를 참고자료로 보여주었다.

4.3. 측정 도구의 개발

본 연구의 연구모델을 검증하기 위하여 선행연구에서 실증된 문항을 바탕으로 타당성이 검증된 항목으로 측정항목을 구성하였으며, 일부 세부 항목에 대해서는 연구자가 보완, 추가하여 <표 1> 과 같이 측정도구를 개발하였다. 특히,

〈표 1〉 변수별 측정항목 및 관련문헌

변수	측정항목	항목 내용	관련문헌
조절초점(예방)	PVF1	나는 책임과 의무를 다하지 못할까봐 종종 걱정한다	최문규 (2016), Lockwood et al. (2002)
	PVF2	나는 미래에 내가 원치 않는 모습이 될까봐 자주 걱정한다.	
	PVF3	나는 나의 학업 또는 직무에서 목표를 달성하지 못할까봐 종종 걱정한다	
	PVF4	평소에 걱정하던 일들이 실제로 벌어지는 상황을 상상한다	
조절초점(향상)	PMF1	나는 희망과 소망을 어떻게 성취할지 종종 상상한다	
	PMF2	나는 미래의 바람직한 내 모습에 대해 종종 상상한다	
	PMF3	나는 학업 또는 직무를 어떻게 하면 성공적으로 달성할 것인가를 종종 생각한다	
	PMF4	평소에 바라던 일들이 실제로 일어나는 상황을 상상한다	
AI 사전지식	AIK1	나는 혼자서도 인공지능(AI) 제품이나 서비스를 편리하게 사용할 수 있다	Brown et al. (2005)
	AIK2	나는 필요하다면 혼자서 인공지능(AI) 제품이나 서비스를 쉽게 다룰 수 있다	
	AIK3	나는 주변의 도움 없이도 인공지능(AI) 제품이나 서비스를 사용할 수 있다	
XAI 사전지식	XAIK1	나는 설명가능한 인공지능(XAI)에 대한 충분한 지식을 가지고 있다	Peddibhotla (2006)
	XAIK2	나는 설명가능한 인공지능(XAI)이 어떤 방식으로 설명을 제공하는지 알고 있다	
	XAIK3	나는 설명가능한 인공지능(XAI)이 제공하는 정보를 잘 이해할 수 있다	
	XAIK4	나는 설명가능한 인공지능(XAI)에 대한 전문가라고 생각한다	
인지된 이해도	PU1	대출심사 AI 알고리즘의 심사 결과를 쉽게 이해할 수 있다	Riefle et al. (2024), Madsen et al. (2000), Hoffman et al. (2018)
	PU2	대출심사 AI 알고리즘의 작동 방식을 이해하고 있어서, 다음에 AI 를 사용할 때 어떤 일이 일어날지 알 수 있다	
	PU3	대출심사 AI 알고리즘이 어떤 과정으로 결과를 도출하는지 쉽게 파악할 수 있다	
	PU4	설명을 통해서, 대출심사 AI 알고리즘이 어떻게 작동하는지 이해할 수 있다	
사용자 신뢰도	UT1	나는 이 대출심사 AI 알고리즘을 신뢰하며, 잘 작동한다고 생각한다	Hoffman et al. (2018)
	UT2	대출심사 AI 알고리즘의 결과는 매우 예측 가능하다	
	UT3	대출심사 AI 알고리즘은 매우 믿을 만하며, 항상 정확한 결과를 제공할 것이라고 기대한다	
	UT4	이 대출심사 AI 알고리즘을 사용하면 올바른 답을 얻을 수 있다고 확신한다	
	UT5	이 대출심사 AI 알고리즘은 사람처럼 해당 작업을 잘 수행할 수 있다	

조절초점 측정항목은 4개의 대응되는 향상초점 항목과 예방초점 항목으로 구성하여 이를 기반으로 응답자를 향상초점자와 예방초점자로 구분할 수 있게 하였다. 모든 변수는 복수개의 측정항목들로 구성하여 신뢰성과 타당성을 확보하였으며, 모든 항목들은 7점 리커트(Likert) 척도로 측정하였다.

4.4. 자료의 수집

본 연구에서는 만 18세 이상 70세까지의 성인 남녀를 대상으로 설문조사를 시행하였다. 조사는 2024년 8월 19일부터 30일까지 전문 설문조사기관을 통하여 온라인 설문을 실시하였으며, AI 대출심사 결과가 승인된 경우에 대한 설문 300명과 거절된 경우에 대한 설문 300명을 포함해 총 600명을 목표로 진행하였다. 각 응답자에게는 두가지 XAI 설명방식인 반사실적 방식과 특징 중요도 방식의 설명 결과를 무작위 순서로 제시하고 응답자료를 수집하였다. 전문조사기관의 설문응답 품질 검토를 통과한 최종 600명으로부터 두가지 XAI 설명방식에 대한 응답자료를 수집함으로써 총 1,200건의 자료를 확보하였다.

분석은 SPSS Statistics 20으로 수행하였으며, 설문 응답자의 인구 통계학적 특성은 <표 2>와 같다. 응답자의 성별은 여성이 332명으로 55.3%, 남성이 268명으로 44.7%로 분포하였다. 나이는 18세이상부터 20대가 36.5%로 가장 많았고, 30대 31.8%, 40대 20.5% 순으로 나타났다. 학력은 대학교 재학 또는 졸업이 73.2%로 나타났고, 직업은 회사원이 49.3%로 가장 많았고 학생이 15.5%로 분포하였다.

〈표 2〉 응답자의 인구통계학적 특성

		빈도	백분율
성별	여성	332	55.3%
	남성	268	44.7%
나이	18 ~ 29	219	36.5%
	30 ~ 39	191	31.8%
	40 ~ 49	123	20.5%
	50 ~ 59	57	9.5%
	60 ~ 71	10	1.7%
학력	고등학교 재학/졸업	84	14.0%
	대학교 재학/졸업	439	73.2%
	대학원 재학/졸업	74	12.3%
	기타	3	0.5%
직업	학생	93	15.5%
	회사원	296	49.3%
	공무원	25	4.2%
	자영업	34	5.7%
	프리랜서	52	8.7%
	가정주부	44	7.3%
	은퇴/무직	34	5.7%
	기타	22	3.7%
합계		600	100.0%

5. 연구 결과

5.1. 측정 모델 평가

측정 모델의 집중타당도를 검증하기 위하여 탐색적 요인분석(EFA)을 수행하여 고유값을 기준으로 요인을 구분하고, 항목별로 요인 적재량을 확인하였다. 베리맥스(varimax) 회전방법을 이용한 주성분 요인분석 결과, <표 3>과 같이 측정도구는 6개의 요인으로 구분되고, 항목별 요인 적재량이 0.643 이상을 나타내는 것으로 확인되어 집중타당도가 확보되었다고 판단하였다. 내적 일관성 신뢰도 검증을 위하여 요인별 Cronbach's

〈표 3〉 집중타당도, 내적 일관성 신뢰도 평가 결과

항목	요인						공통성	고유값	신뢰도
	1	2	3	4	5	6			Cronbach's α
PVF1	-0.022	0.042	0.013	0.133	**0.808**	0.053	0.676	1.441	0.836
PVF2	0.033	0.012	0.034	0.085	**0.818**	-0.081	0.686		
PVF3	0.040	0.005	-0.002	0.056	**0.863**	0.084	0.756		
PVF4	0.026	0.004	-0.065	0.070	**0.763**	0.023	0.592		
PMF1	0.098	0.058	0.034	**0.856**	0.019	0.092	0.756	2.095	0.851
PMF2	0.073	0.090	0.026	**0.874**	0.046	0.039	0.782		
PMF3	0.106	0.068	0.119	**0.788**	0.127	0.165	0.694		
PMF4	0.072	0.132	-0.109	**0.704**	0.246	0.174	0.621		
AIK1	0.142	0.115	0.200	0.164	0.020	**0.863**	0.845	1.206	0.919
AIK2	0.090	0.132	0.166	0.161	0.038	**0.892**	0.877		
AIK3	0.125	0.140	0.208	0.130	0.028	**0.868**	0.849		
XAIK1	0.138	0.105	**0.890**	0.002	-0.017	0.170	0.852	2.643	0.886
XAIK2	0.119	0.135	**0.900**	0.049	-0.038	0.155	0.869		
XAIK3	0.107	0.243	**0.651**	0.139	0.027	0.343	0.632		
XAIK4	0.111	0.081	**0.841**	-0.052	0.001	0.056	0.733		
PU1	0.321	**0.752**	0.066	0.133	0.029	0.153	0.715	3.172	0.905
PU2	0.298	**0.808**	0.174	0.095	0.009	0.089	0.790		
PU3	0.274	**0.856**	0.143	0.082	0.014	0.089	0.843		
PU4	0.318	**0.777**	0.182	0.078	0.032	0.129	0.761		
UT1	**0.784**	0.322	0.076	0.134	0.007	0.107	0.754	7.611	0.917
UT2	**0.643**	0.467	0.075	0.107	0.023	0.135	0.668		
UT3	**0.861**	0.224	0.152	0.079	0.026	0.045	0.824		
UT4	**0.869**	0.204	0.116	0.078	0.023	0.066	0.821		
UT5	**0.825**	0.234	0.124	0.062	0.034	0.128	0.772		

alpha 계수를 확인한 결과 0.836에서 0.919로 나타나 매우 높은 수준의 신뢰성을 확인하였다.

5.2. 조절초점 유형 분류

신뢰성과 타당성이 확보된 예방초점 측정항목과 향상초점 측정항목을 기준으로 최문규(2016) 와 Lockwood et al.(2002)에서 사용한 분류기준을 적용하여 예방초점자와 향상초점자를 분류하였다. 구체적으로 예방초점 4개 항목의 합계 점수가 16점 이상이면서 향상초점 합계 점수보다 2점 이상 높은 응답자를 예방초점자로 선정하고, 향상초점 4개 항목의 합계 점수가 16점 이상이면서 예방초점 합계 점수보다 2점 이상 높은 응답자를 향상초점자로 분류하였다. 분류 결과, <표 4>와 같이 전체 피실험자 600명 가운데 예방초점자는 19.8% 인 119명, 향상초점자는 47.0% 인 282명, 중립유형 199명으로 판별되어, 예방초점자와 향상초점자 401명을 대상으로 분석을 수행하였다.

〈표 4〉 분석 대상 현황

		XAI 설명방식 (802건)	
		반사실적 (401건)	특징 중요도 (401건)
조절초점 (401명)	예방초점 (119명)	119건	119건
	향상초점 (282명)	282건	282건

5.3. 통제변수에 대한 사전 분석

XAI 설명에 대한 사용자의 인지된 이해도에 영향을 미칠 수 있는 통제 변수로 나이, AI 사전지식, XAI 사전지식을 설정하고, 이들이 인지된 이해도에 미치는 영향을 분석하였다. Pearson 상관관계 분석 결과, 나이는 인지된 이해도와 유의한 상관관계를 보이지 않았다($p = .096$). 반면, AI 사전지식과 인지된 이해도 간의 Pearson 상관계수는 .288, XAI 사전지식과 인지된 이해도 간의 Pearson 상관계수는 .333으로 유의하게($p<.001$) 나타나, 두 변수는 XAI 설명에 대한 사용자의 인지된 이해도와 정(+)의 상관관계가 있음이 확인되었다.

이와 같이 AI 사전지식과 XAI 사전지식은 사용자의 인지된 이해도에 유의한 영향을 미치는 요인으로 나타났으므로, XAI 설명방식과 조절초점이 XAI 설명에 대한 사용자의 인지된 이해도에 미치는 영향을 정확하게 분석하기 위해 두 변수의 영향을 통제하였다. 따라서 이후 분석은 공분산분석(ANCOVA)을 통해 통제된 상태에서 독립변수와 조절변수의 순수한 효과를 검증하였다.

5.4. XAI 설명방식이 사용자의 인지된 이해도에 미치는 영향

XAI 설명방식(반사실적 vs. 특징 중요도)이 사용자의 인지된 이해도에 미치는 영향을 분석하기 위해 공분산분석(ANCOVA)을 실시하였다. 종속변수는 인지된 이해도로 설정하였으며, AI 사전지식과 XAI 사전지식을 공변량으로 설정하였다. Levene의 검정 결과, 오차 분산의 동일성이 확인되었고($F(1, 800) = .072$, $p = .789$), 이는 두 집단 간의 분산이 동일하다는 가정이 충족되었음을 의미한다.

<표 5>의 ANCOVA 분석 결과, 공변량인 AI 사전지식($F(1, 798) = 27.294$, $p < .001$)과 XAI 사전지식($F(1, 798) = 53.600$, $p < .001$)은 인지된 이해도에 유의미한 영향을 미치는 것으로 나타났다. 독립변수의 주효과 분석 결과, XAI 설명방식은 인지된 이해도에 유의미한 영향을 미치는 것으로 나타났다($F(1, 798) = 5.473$, $p = .020$). XAI 설명방식간 평균차이를 분석한 결과, 반사실적

〈표 5〉 XAI 설명방식과 인지된 이해도의 ANCOVA 검정 결과

종속변수 : 인지된 이해도

소스	제 III 유형 제곱합	자유도	평균 제곱	F	유의 확률	부분 에타 제곱
AI 사전지식	34.742	1	34.742	27.294	0.000	0.033
XAI 사전지식	68.226	1	68.226	53.600	0.000	0.063
XAI 설명방식	6.967	1	6.967	5.473	0.020	0.007
오차	1015.755	798	1.273			

방식의 인지된 이해도의 평균값은 4.792(표준오차 = .056)로 특징 중요도 방식의 평균값 4.605(표준오차 = .056)보다 높게 나타나, 가설 H1은 지지되었다.

5.5. XAI 설명방식과 조절초점이 사용자의 인지된 이해도에 미치는 영향

연구가설 H2는 XAI 설명방식과 사용자의 조절초점이 XAI 설명에 대한 사용자의 인지된 이해도에 미치는 영향을 검증하는 것이었다. AI 사전지식과 XAI 사전지식을 통제변수로 설정하고 공분산분석(ANCOVA)을 수행하여 사용자의 인지된 이해도에 미치는 독립변수와 조절변수의 순수한 효과를 분석하였다. Levene 검정 결과, 집단 간 오차 분산은 동일한 것으로 나타나($F(3, 798) = 1.514$, $p = .209$), 공분산분석의 가정이 충족되었음을 확인하였다.

<표 6>의 분석 결과, AI 사전지식($F(1, 796) = 25.750$, $p < .001$)과 XAI 사전지식($F(1, 796) = 53.691$, $p < .001$)은 인지된 이해도에 유의한 영향을 미치는 공변량으로 나타났다. 독립변수인 XAI 설명방식 역시 사용자의 인지된 이해도에 유의미한 주효과를 보였으며($F(1, 796) = 7.821$, $p = .005$), CF 방식의 인지된 이해도($M = 4.803$)가 FI 방식($M = 4.559$)보다 높았다. 그러나, 조절초점에 따른 인지된 이해도의 차이와 XAI 설명방식과 조절초점 간의 상호작용 효과는 통계적으로 유의미하지 않은 것으로 나타났다.

<표 7>에 제시된 사후 검정(Bonferroni) 결과, 예방초점자에게 반사실적 설명방식($M = 4.831$)이 특징 중요도 설명방식($M = 4.446$)보다 유의미

〈표 6〉 XAI 설명방식, 사용자 조절초점과 인지된 이해도의 ANCOVA 검정 결과

종속변수 : 인지된 이해도

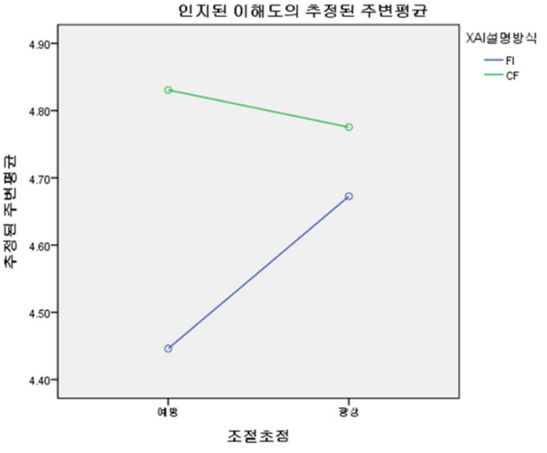
소스	제 III 유형 제곱합	자유도	평균 제곱	F	유의확률	부분 에타 제곱
AI 사전지식	32.712	1	32.712	25.750	0.000	0.031
XAI 사전지식	68.209	1	68.209	53.691	0.000	0.063
XAI 설명방식	9.936	1	9.936	7.821	0.005	0.010
조절초점	1.205	1	1.205	0.948	0.330	0.001
XAI 설명방식 * 조절초점	3.318	1	3.318	2.612	0.106	0.003
오차	1011.232	796	1.27			

〈표 7〉 Bonferroni 사후 검정 결과

종속변수 : 인지된 이해도

조절초점	XAI 설명방식		평균차	유의확률
	반사실적(CF)	특징 중요도(FI)		
예방	4.831	4.446	0.384	0.009
향상	4.775	4.673	0.103	0.279

하게 높은 인지된 이해도를 보였으며(p = .009), 이에 따라 가설 H2a는 지지되었다. 반면, <그림 3>을 통해서도 확인할 수 있듯이, 향상초점자에 대해서는 특징 중요도 설명방식(M = 4.673)이 예방초점자 경우보다는 높은 결과를 보였으나 반사실적 방식(M = 4.775)보다 낮게 나타나, 가설 H2b는 기각되었다.



〈그림 3〉 조절초점(예방초점 vs. 향상초점)의 조절효과 분석 결과

5.6. 사용자의 인지된 이해도가 사용자 신뢰도에 미치는 영향

사용자의 AI 사전지식과 XAI 사전지식을 통제한 상태에서, XAI 설명에 대한 사용자의 인지된 이해도가 AI 알고리즘에 대한 사용자 신뢰도에 미치는 영향을 분석하기 위해 다중회귀분석을 수행하였다. 분석 결과, 모형의 설명력은 R^2 = .424로 나타나, 독립변수들이 AI에 대한 사용자 신뢰도의 약 42.4%를 설명하는 것으로 확인되었다. 또한, Durbin-Watson 값이 2.094로 나타나 잔차의 독립성이 유지됨을 알 수 있었다.

회귀계수 분석 결과 <표 8>과 같이, 인지된 이해도가 AI 시스템에 대한 사용자 신뢰도에 가장 큰 영향을 미치는 변수로 나타났으며(β = .598, p < .001), 통계적으로 매우 유의미한 결과를 보였다. XAI 사전지식 또한 AI 신뢰도에 유의미한 영향을 미치는 것으로 나타났으나(β = .101, p = .001), 그 영향력은 인지된 이해도에 비해 상대적으로 낮았다. 반면, AI 사전지식(β = .033, p = .260)은 통계적으로 유의하지 않은 변수로 나타나, 사용자 신뢰도에 유의한 영향을 미치지 않는 것으로 확인되었다.

추가적으로 공선성 진단을 실시한 결과, 공차 값이 모두 .793 이상, VIF 값이 모두 1.261 이하로 나타나, 다중공선성의 문제가 없음을 확인하였다. 이러한 결과를 종합할 때, XAI 설명방식에 대한 사용자의 인지된 이해도는 AI 시스템에 대한 사용자 신뢰도에 가장 유의한 영향을 미치는 변수임이 입증되었으며, 이에 따라 연구가설 H3는 지지되었다.

〈표 8〉 인지된 이해도와 사용자 신뢰도간 다중회귀분석 결과

모형	비표준화 계수		표준화 계수	t	유의확률	공선성 통계량	
	B	표준오차	베타			공차	VIF
인지된 이해도	0.552	0.027	0.598	20.639	0.000	0.860	1.163
AI사전지식	0.032	0.028	0.033	1.127	0.260	0.818	1.223
XAI사전지식	0.084	0.025	0.101	3.344	0.001	0.793	1.261

종속변수 : 사용자 신뢰도

6. 결론 및 시사점

6.1. 연구 결과 요약 및 논의

본 연구는 설명 가능한 인공지능(XAI) 연구가 기술적 발전에도 불구하고 사용자 이해도를 충분히 높이지 못하는 한계를 인식하고, 사용자의 개인 성향인 조절초점을 반영한 사용자 중심의 XAI 연구모델을 제시하였다. 600명의 일반 성인을 대상으로 실험 연구를 수행한 결과, 다음과 같은 주요 사실을 확인하였다. 첫째, XAI 설명방식(반사실적 vs. 특징 중요도)에 따라 사용자 이해도에 유의미한 차이가 있음을 확인하였다. 둘째, 사용자의 조절초점 유형(예방초점 vs. 향상초점)에 따라서도 XAI 설명에 대한 사용자 이해도가 달라지며, 이는 개인 성향이 XAI 설명의 효과에 영향을 미치는 요인임을 증명하였다. 셋째, AI 및 XAI에 대한 사전 지식 수준 역시 XAI 설명의 효과에 영향을 미치는 요인임을 확인하였다. 마지막으로, XAI 설명에 대한 사용자 이해도의 향상은 AI 시스템에 대한 사용자 신뢰도의 향상으로 이어진다는 사실을 과학적으로 입증하였다.

본 연구는 다음과 같은 측면에서 중요한 이론적 및 실무적 시사점을 가진다. 이론적 측면에서 본 연구는 조절초점 이론을 XAI 연구에 통합함으로써, XAI의 효과성에 대한 확장된 이론적 프레임워크를 제공하였다. 이는 향후 사용자 중심적 XAI 연구의 발전에 기여할 수 있을 것이다. 실무적 측면에서, 본 연구는 사용자의 동기 성향을 고려한 XAI 시스템 설계 가이드라인 개발에 중요한 기여를 할 수 있으며, 이를 통해 XAI의 실제 효과성을 극대화하는 데 도움을 줄 수 있다. 또한, AI 사용자에게 어떤 XAI 설명방식이 가장 적합한지에 대한 구체적인 가이드라인을 제공할 수 있을 것이다. 정책적 측면에서는 AI의 설명 가능성에 대한 요구가 강화되는 상황에서, 본 연구는 효과적인 XAI 구현을 통해 신뢰성 있는 AI 정책 수립에 기여할 수 있을 것이다.

6.2. 연구의 한계 및 향후 연구 과제

사용자 중심의 XAI 연구는 아직 상대적으로 미흡한 연구 분야이다. 본 연구는 개인의 성향이 XAI 설명의 효과에 미치는 영향을 실증적으로 확인하며 기존 연구 결과들을 확장했으나, 설명의 효과에 영향을 미치는 요인은 매우 다양하므로, 이들을 포괄하는 추가 연구가 필요하다. 향후 연구 과제를 다음과 같이 제안할 수 있다. 첫째, 본 연구에서 적용한 조절초점 이외에도, AI 및 XAI에 대한 지식 수준 또는 해당 문제 상황에 대한 세부 조건 등 설명 효과에 영향을 미치는 다양한 요인에 대한 연구로 확대가 필요하다. 이들 요인 간의 관계를 설명할 수 있는 메커니즘에 대한 연구는 중요한 주제가 될 것이다. 둘째, 본 연구에서는 대표적인 XAI 설명방식인 반사실적 설명방식과 특징 중요도 설명방식을 비교분석하였으나, 각 방식 내에서도 설명의 수준과 세부 조건에 따라 사용자의 이해도가 달라질 수 있다. 이에 대한 연구도 의미 있는 주제가 될 것이다. 셋째, 본 연구에서 다룬 두 가지 XAI 설명방식 외에도 SHAP, LIME 등 다양한 XAI 알고리즘을 비교하는 연구가 필요하다. 이러한 연구를 통해 실무에서 사용자에게 최적의 XAI 설명을 제공할 수 있는 구체적인 지침을 마련할 수 있을 것이다.

참고문헌(References)

[국내 문헌]

개인정보보호위원회. (2024). 개인정보 보호법 시행령 개정안. Retrieved from https://www.pipc. go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9969

문건두, 김경재 (2023). Corporate bankruptcy prediction model using explainable AI-based feature selection. 지능정보연구, 29(1), 45-67. https://doi.org/10.13088/jiis.2023.29.2.241

박기경, 류강석, 박종원. (2015). 소비자의 조절초점, 상품구색의 크기, 그리고 선택모드가 금융상품 선택에 미치는 영향. 마케팅연구, 30(2), 123-140. https://doi.org/10.15830/kmr.2015.30.4.1

박경희, 서봉군, 박도형. (2022). 이미지 기반의 경험재, 텍스트 기반의 탐색재: 조절초점에 따른 제품 정보 구성 방식과 제품 유형의 일치 효과. 지능정보연구, 28(2), 45-68. https://doi.org/10.13088/jiis.2022.28.2.075

원종관, 홍태호, 배경일. (2021). 신용 데이터의 이미지 변환을 활용한 합성곱 신경망과 설명가능한 인공지능(XAI)을 이용한 개인신용평가. 정보시스템연구, 29(3), 75-92. https://doi.org/10.5859/KAIS.2021.30.4.203

정언용, 윤태섭, 여준상. (2013). 소비자와 브랜드 간 조절초점의 일치성이 신제품 평가에 미치는 영향. 한국심리학회지: 소비자·광고, 14(1), 45-67.

조수현, 신경식. (2022). 부도예측모형에서 도메인 지식을 통합한 반사실적 예시 기반 설명력 증진 방법. 지능정보연구, 29(1), 88-105. https://doi.org/10.13088/jiis.2022.28.2.307

최문규. (2016). 관여도, 조절초점에 따른 비주얼 표현기법의 차이에 대한 소비자 반응과 특이성 인식의 매개효과 연구 - 일러스트레이션 기법과 사진 기법의 효과 차이를 중심으로. 한국광고홍보학회, 27(3), 125-145. https://doi.org/10.16914/kjapr.2016.18.1.165

안재현. (2020). XAI 설명가능한 인공지능, 인공지능을 해부하다. 파주: 위키북스.

[국외 문헌]

Ahn, Y., Lin, Y. R. (2020). FairSight: Visual Analytics for Fairness in Decision Making. IEEE Transactions on Visualization and Computer Graphics, 26(1), 123-130. https://doi.org/10.1109/tvcg.2019.2934262

Alicioglu, G., Sun, B. (2022). A survey of visual analytics for Explainable Artificial Intelligence methods. Computers & Graphics, 97, 15-29. https://doi.org/10.1016/j.cag.2021.09.002

Angerschmid, A., Zhou, J., et al. (2022). Fairness and explanation in AI-informed decision making. Machine Learning and Knowledge Extraction, 4(3), 115-135. https://doi.org/10.3390/make4020026

Arrieta, A. B., Díaz-Rodríguez, N., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Atkinson, R. C., Shiffrin, R. M. (1968). Human Memory: A Proposed System and Its Control Processes. Psychology of Learning and Motivation, 2, 89-195. https://doi.org/10.1016/s0079-7421(08)60422-3

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic load within problems. Learning and Instruction, 16(4), 389-400. https://doi.org/10.1016/j.learninstruc.2006.09.001

Brown, S. A., Venkatesh, V. (2005). Model of adoption of technology in households: A baseline model

test and extension incorporating household life cycle. MIS Quarterly, 29(3), 399-426. https://doi.org/10.2307/25148690

Cierniak, G., Scheiter, K., Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? Computers in Human Behavior, 25(2), 315-324. https://doi.org/10.1016/j.chb.2008.12.020

Demajo, L. M., Vella, V., Dingli, A. (2020). Explainable AI for interpretable credit scoring. Computer Science & Information Technology (CS & IT), 12(3), 56-75. https://doi.org/10.5121/csit.2020.101516

Du, Y., Antoniadi, A. M., et al. (2022). The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations. Applied Sciences, 12(8), 1234-1250. https://doi.org/10.3390/app122010323

Fisher, A., Rudin, C., Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(1), 1-24.

Förster, J., Higgins, E. T. (2005). How global versus local perception fits regulatory focus. Psychological Science, 16(8), 631-636. https://doi.org/10.1111/j.1467-9280.2005.01586.x

Hadash, S., Willemsen, M. C., et al. (2022). Improving understandability of feature contributions in model-agnostic explainable AI tools. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1-10. https://doi.org/10.1145/3491102.3517650

Hamm, P., Klesel, M., et al. (2023). Explanation matters: An experimental study on explainable AI. Electronic Markets, 33(2), 287-301. https://doi.org/10.1007/s12525-023-00640-9

Herm, L. V. (2023). Impact of explainable AI on cognitive load: Insights from an empirical study. Thirty-first European Conference on Information Systems (ECIS 2023), 1-12.

Herm, L. V., Wanner, J., et al. (2021). I Don't Get IT, but IT seems Valid! The Connection between Explainability and Comprehensibility in (X)AI Research. Twenty-Ninth European Conference on Information Systems (ECIS 2021), 1-15.

Higgins, E. T. (1989). Continuities and discontinuities in self-regulatory and self-evaluative processes: A developmental theory relating self and affect. Journal of Personality, 57, 407-444. https://doi.org/10.1111/j.1467-6494.1989.tb00488.x

Higgins, E. T. (2005). Value from regulatory fit. Psychological Science, 16(11), 933-939. https://doi.org/10.1111/j.0963-7214.2005.00366.x

Higgins, E. T., Shah, J., Friedman, R. (1997). Emotional responses to goal attainment: Strength of regulatory focus as moderator. Journal of Personality & Social Psychology, 72(3), 515-525. https://doi.org/10.1037//0022-3514.72.3.515

Hoffman, R. R., Mueller, S. T., et al. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint, arXiv:1812.04608.

Kim, D., Song, Y., et al. (2023). How should the results of artificial intelligence be explained to users?-Research on consumer preferences in user-centered explainable artificial intelligence. Technological Forecasting and Social Change, 188, 122-134. https://doi.org/10.1016/j.techfore.2023.122343

Larasati, R., De Liddo, A., Motta, E. (2020). The effect of explanation styles on user's trust.

Oro. Retrieved from https://oro.open.ac.uk

Leppink, J., Paas, F., et al. (2013). Development of an instrument for measuring different types of cognitive load. Behavior Research Methods, 45(4), 1058-1072. https://doi.org/10.3758/s13428-013-0334-1

Lockwood, P., Jordan, C. H., Kunda, Z. (2002). Motivation by positive or negative role models: Regulatory focus determines who will best inspire us. Journal of Personality & Social Psychology, 83(4), 854-864. https://doi.org/10.1037/0022-3514.83.4.854

Lopes, P., Braga, C., et al. (2022). XAI systems evaluation: A review of human and computer-centered methods. Applied Science, 12(3), 300-320. https://doi.org/10.3390/app12199423

Lu, J., Lee, D., Kim, T. W., Danks, D. (2020). Good explanation for algorithmic transparency. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 93-100. https://doi.org/10.2139/ssrn.3503603

Madsen, M., Gregor, S. (2000). Measuring human-computer trust. 11th Australasian Conference on Information Systems, 1-11.

Mengqi, L., Shyam, S., et al. (2022). User Trust in Recommendation Systems: A comparison of Content-Based, Collaborative and Demographic Filtering. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1-8. https://doi.org/10.1145/3491102.3501936

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81-97. https://doi.org/10.1037/h0043158

Miller, T., Howe, P., Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv preprint, arXiv:1712.00547.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., et al. (2021). A multidisciplinary survey and framework for design and evaluation of Explainable AI systems. ACM Transactions on Interactive Intelligent Systems, 11(3), 123-145. https://doi.org/10.1145/3387166

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. Journal of Educational Psychology, 84(4), 429-434. https://doi.org/10.1037/0022-0663.84.4.429

Panigutti, C., Beretta, A., Giannotti, F., Pedreschi, D. (2022). Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1-12. https://doi.org/10.1145/3491102.3502104

Peddibhotla, N. B. (2006). Knowledge sharing in public document repositories. Retrieved from https://search.proquest.com

Ramon, Y., Farrokhnia, R. A., et al. (2021). Explainable AI for Psychological Profiling from Behavioral Data: An Application to Big Five Personality Predictions from Financial Transaction Records. Information, 12(4), 300-315. https://doi.org/10.3390/info12120518

Ramon, Y., Vermeire, T., et al. (2021). Understanding consumer preferences for explanations generated by XAI algorithms. arXiv preprint, arXiv:2107.02624.

Riefle, L., Hemmer, P., et al. (2024). The Role of Cognitive Styles for Explainable AI. Thirty-second European Conference on Information Systems (ECIS 2024), 1-13.

Riveiro, M., Thill, S. (2021). That's (not) the output I expected! On the role of end user expectations in creating explanations of AI systems. Artificial Intelligence, 289, 103-115. https://doi.org/10.1016/j.artint.2021.103507

Silva, A., Schrum, M., et al. (2022). Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. International Journal of Human-Computer Interaction, 38(2), 150-165. https://doi.org/10.1080/10447318.2022.2101698

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7

Warren, G., Keane, M. T., et al. (2022). Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. arXiv Preprint, arXiv:2207.01552.

Xu, Z., Lamba, H., Ai, Q., et al. (2024). CFE2: Counterfactual Editing for search result explanation. Proceedings of the 2024 ACM SIGIR ICTIR, 1-10. https://doi.org/10.1145/3664190.3672508

Abstract

The Impact of XAI Explanation Methods on Users' Perceived Understanding and Trust: The Role of Regulatory Focus

Sang-Ryong Moon* · Hyunchul Ahn* · Yoon Han**

As AI technology advances, the "black-box" problem—where AI model outcomes become increasingly difficult to interpret—has emerged as a critical issue. To address this challenge, explainable AI (XAI) research must evolve from a purely technical approach to a user-centered perspective. This study introduces the concept of regulatory focus, a personal trait, to propose XAI explanation methods that can enhance user understanding. A comparative experiment was conducted with 600 general adults to evaluate the Counterfactual and Feature Importance explanation methods. The results show that the Counterfactual explanation method significantly improves user understanding. Furthermore, the user's regulatory focus was identified as a significant factor influencing their understanding of XAI explanations. The study also confirmed that an increase in user understanding of XAI explanations leads to improved user trust in AI systems. By integrating individual traits into user-centered XAI research, this study offers an expanded theoretical framework and provides practical guidelines for determining which XAI explanation method is most appropriate for AI users.

Key Words : Explainable AI (XAI), Perceived Understanding, Regulatory Focus, Trust

Received : September 20, 2024 Revised : September 27, 2024 Accepted : September 27, 2024
Corresponding Author : Yoon Han

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Yoon Han
Graduate School of Business IT, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Korea
Tel: +82-2-910-5472, Fax: +82-2-910-4017, E-mail: yoonhan@kookmin.ac.kr

저 자 소 개



문상룡
현재 국민대학교 비즈니스IT전문대학원 박사과정에 재학 중이다. 부산대학교 컴퓨터공학과에서 공학사와 POSTECH 전자계산학과에서 공학석사를 취득하였다. 주요 관심분야는 빅데이터 및 인공지능, 인공지능의 평가 및 설명가능성(XAI) 등이다.



안현철
현재 국민대학교 비즈니스IT전문대학원 교수로 재직 중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심 분야는 금융 및 고객관계관리 분야의 인공지능 응용, 지능형 의사결정지원시스템, 정보시스템 수용과 관련한 행동 모형 등이다.



한윤
현재 국민대학교 비즈니스IT전문대학원에서 조교수로 재직 중이며, 싱가포르국립대학교(NUS)에서 박사 학위를 취득하였다. 계량경제 분석과 필드실험을 활용하여 디지털 미디어 환경에서 새로운 디지털 혁신이 소비자 행동에 미치는 영향에 관하여 폭넓게 연구하고 있다.