AI와 인간의 신뢰 관계: 심리학적 관점에서의 분석

정주영

jjy20051121jung@gmail.com

TheTrustRelationshipBetweenAIandHumans:AnAnalysisfromaPsychologicalPerspective

JungJooyoung

Independent Researcher

요 약

본 연구는 인공지능(AI)과 인간의 신뢰 관계를 심리학적 관점에서 탐구하였다. AI가 의료, 금융, 교육 등 핵심 의사결정 영역에 확대 적용되고 있으나, 사용자가 AI를 언제 신뢰하고 언제 불신하는지는 충분히 규명되지 않았다. 이에 본 논문은 이론적 검토와 실제 사례 분석을 바탕으로, 시나리오 기반 실험을 설계하여 신뢰 형성 요인을 검증하였다.

실험은 300명의 참가자를 대상으로 설명 가능성, 감정 표현, 전문가 협력 여부를 독립변수로 설정하고 신뢰도 및 선택 행동을 측정하였다. 분석 결과, ① 설명 가능한 AI가 그렇지 않은 경우보다, ② 감정 표현을 하는 AI가 중립적 표현을 하는 경우보다, ③ 전문가와 협력적으로 제시된 AI가 독립적으로 제시된 경우보다 유의미하게 높은 신뢰를 얻는 것으로 나타났다.

이러한 결과는 신뢰 형성에 있어 인지적 요인(설명 가능성, 전문가 협력)과 정서적 요인(감정 표현)이 모두 중요한 역할을 함을 시사한다. 본 연구는 인간-AI 협력 관계 설계에서 단순 성능 향상이 아닌 설명 가능성 강화, 전문가 보조적 구조, 사회적 존재감 고려가 필요함을 강조한다. 이는 향후 AI 개발 및 정책 수립에 중요한 심리학적·윤리적 함의를 제공한다.

1. 서론

오늘날 AI는 인간의 중요한 의사결정 과정에 깊숙이 자리 잡고 있다. 그러나 기술적 발전이 곧바로 신뢰로 이어지지는 않는다. 예컨대, 의료 현장에서 AI 진단과 의사의 진단이 충돌할 때 환자는 혼란에 빠지고, 금융권에서는 불투명한 AI 대출 심사 결과로 고객의 불신이 커지고 있다. 따라서 본 연구는 “인간은 어떤 조건에서 AI를 신뢰하는가?”라는 질문에 초점을 맞추어 심리학적 요인을 규명하고자 한다.

2. 이론적 배경

2.1 신뢰의 개념 및 선행연구

심리학에서 신뢰(trust)는 불확실성과 위험을 감수하면서 타인의 의도를 긍정적으로 기대하는 심리적 상태로 정의된다(Mayer et al., 1995). 신뢰는 크게 두 가지 차원으로 구분된다.

인지적 신뢰(Cognitive trust): 능력과 전문성에 기반한 신뢰

정서적 신뢰(Affective trust): 감정적 유대와 관계적 상호작용에 기반한 신뢰

최근 연구들은 AI 신뢰 형성 과정에서 인간-자동화 상호작용(Human-Computer Interaction, HCI) 관점의 중요성을 강조하고 있다. Lee와 See(2004)는 신뢰가 단순히 시스템 성능에 의해 결정되는 것이 아니라, 시스템의 투명성, 피드백 제공 방식, 사용자 통제 가능성에 의해 크게 영향을 받는다고 지적하였다. Hoff와 Bashir(2015)는 실증적 연구들을 종합하여, 자동화에 대한 신뢰 요인을 인지적 차원(성능, 신뢰성)과 정서적 차원(사회적 존재감, 상호작용 품질)으로 구분하였다.

이러한 최근 연구 흐름은 본 연구가 제안하는 설명 가능성, 전문가 협력, 감정 표현 요인의 중요성을 뒷받침한다. 즉, 사용자는 단순한 결과 제시보다, AI가 판단 근거를 제공하고, 전문가와 협력하며, 사회적·정서적 존재감을 갖춘 경우 더 높은 신뢰를 형성한다는 점을 시사한다.

2.2 AI 신뢰 형성 요인

전문가 협력성: AI 단독보다 전문가와 협력할 때 신뢰가 높음

설명 가능성(Explainability): 결과뿐 아니라 판단 근거 제공 여부

사회적 존재감(Social Presence): 감정 표현, 인간적 태도, 상호작용

2.3 사건 사례

IBM Watson Oncology: 설명 부족으로 의료 현장에서 신뢰 붕괴

자율주행차 사고(테슬라, 우버): 책임 문제와 신뢰 손상

금융 신용평가 AI: 불투명성으로 인한 고객 불신

3. 연구 가설

H1: 설명 가능한 AI는 설명이 없는 AI보다 높은 신뢰를 얻을 것이다.

H2: 감정 표현을 하는 AI는 중립적 표현을 하는 AI보다 높은 신뢰

를 얻을 것이다.

H3: 전문가와 협력적으로 제시된 AI는 독립적으로 제시된 AI보다 높은 신뢰를 얻을 것이다.

4. 연구 방법

4.1 참가자

본 연구는 국내 온라인 패널 모집 플랫폼을 통해 300명의 성인 참가자를 모집하였다(남 150명, 여 150명, 연령 20~50대). 참가자는 연령과 성별 비율이 균형을 이루도록 층화 표집(stratified sampling) 방식으로 선발되었으며, 각 실험 조건에 무작위 할당(random assignment) 되었다.

4.2 실험 설계

시나리오 기반 설문 실험을 수행하였다. 각 참가자는 의료, 금융, 교육 등 다양한 상황에서 AI와 인간 전문가의 판단을 비교하는 시나리오를 경험하였다.

의료 상황: AI 진단 vs 의사 진단

금융 상황: AI 대출 승인 결과 vs 은행 직원 설명

교육 상황: AI 튜터 조언 vs 교사 조언

4.3 변수 조작

독립변수는 다음과 같이 설정하였다.

설명 제공 여부: 제공 vs 미제공

감정 표현 여부: 격려 포함 vs 중립적

전문가 협력 여부: 협력 vs AI 단독

4.4 측정

신뢰도 평가: 7점 리커트 척도

선택 행동: AI 판단 vs 인간 판단

정성적 의견: 참가자 자유 응답

4.5 통계 분석

신뢰도 점수: 조건별 차이를 검증하기 위해 일원분산분석(One-way ANOVA)을 수행하고, 사후분석(Post-hoc test)으로 Tukey 검정을 실시하였다.

선택 행동 비율: 카이제곱 검정(χ^2 test)을 통해 조건별 유의성을 평가하였다.

모든 통계 분석은 유의수준 $\alpha = 0.05$에서 수행하였다.

이와 같이 명확히 표본, 무작위 할당, 변수 조작, 통계 분석 기법을 제시함으로써 연구 재현성을 높였다.

5. 실험 결과 시각화

5.1 신뢰 점수

조건	평균 신뢰 점수	표준편차
설명 제공	5.8	0.9
설명 미제공	3.9	1.1
감정 표현	5.5	1.0
중립 표현	4.2	1.1
전문가 협력	6.0	0.8
AI 단독	4.1	1.2

신뢰 점수 설명:

설명 제공, 감정 표현, 전문가 협력 조건에서 신뢰 점수가 유의하게 높게 나타났다($p < .05$~.001).

점수에서 세 가지 요인이 모두 신뢰를 높이는 효과를 확인할 수 있음.

신뢰 점수로 제시됨.

5.4 선택 행동 결과

설명 제공 AI: 72% 선택

설명 없는 AI: 38% 선택

감정 표현 AI: 65% 선택

전문가 협력 AI: 80% 선택

점수표로 나타냈을 때 전문가 협력 조건이 가장 높게 나타남.

6. 사례 분석

의료 사례: 서울 소재 A대학병원에서는 폐렴 진단을 보조하기 위해 AI 기반 영상 판독 시스템을 도입하였다. 해당 시스템은 흉부 X-ray 영상을 분석하여 폐렴 여부를 자동 분류하는 기능을 갖추고 있었으나, 실제 임상 적용 과정에서 오진 사례가 발생하였다. 사례는 50대 환자가 고열과 기침 증상을 보이며 응급실을 방문한 경우였다. AI 시스템은 환자의 영상 데이터를 '비폐렴'으로 판정하였으나, 담당 의사는 임상 경험과 혈액검사 결과를 근거로 폐렴 가능성이 높다고 판단하였다. 이후 CT 검사를 통해 의사의 판단이 옳았음이 입증되었다.

이 과정에서 환자는 처음에 AI의 진단 결과에 의문을 가지지 않았으나, 의사가 직접 "영상에서는 뚜렷하게 보이지 않지만 혈액 지표와 임상 증상을 종합하면 폐렴일 가능성이 높다"고 설명하자 의사의 의견을 우선적으로 신뢰하였다. 환자의 진술에 따르면, "AI가 왜 그렇게 진단했는지를 설명해주지 못했기 때문에 결국 의사의 판단을 믿을 수밖에 없었다"는 점이 신뢰 결정의 핵심 요인이었다. 이 사례는 다음과 같은 심리학적 시사점을 제공한다.

전문가 권위의 우선성: 의료 상황에서 환자는 AI보다 의사의 설명과 판단을 더 신뢰한다.

설명 가능성 부족의 한계: AI가 근거를 설명하지 못할 경우, 환자의 신뢰는 급격히 감소한다.

인지적 신뢰의 기반: 환자의 신뢰는 단순 결과보다 합리적 이유 제시에 의해 강화된다.

따라서 의료 영역에서 AI의 신뢰 확보를 위해서는, 의사의 전문성 보완 도구로 제시되고, 동시에 환자와 의료진 모두가 이해할 수 있는 설명 가능한 진단 근거를 제공해야 한다.

금융 사례: 국내 B은행은 신용평가 업무의 효율성과 공정성을 높이기 위해 AI 기반 대출 심사 시스템을 도입하였다. 이 시스템은 고객의 소득, 신용 이력, 거래 패턴, 소비 성향 등을 종합적으로 분석해 대출 승인 여부를 결정하였다. 그러나 실제 운영 과정에서 고객 불만과 불신이 증가하는 사례가 보고되었다.

30대 직장인 C씨는 안정적인 직장에 근무하며 연체 이력이 없음에도 불구하고, 주택담보대출 신청이 거절되는 상황을 겪었다. AI 시스템은 단순히 "대출 불가"라는 결과만을 반환했으며, 구체적인

근거는 제시하지 않았다. 이에 대해 은행 직원에게 문의했으나, 직원 역시 “AI 시스템이 내부 기준에 따라 거절한 것이라 이유를 알 수 없다”는 답변을 할 수밖에 없었다.

C씨는 “성실하게 금융 거래를 해왔는데도 이유조차 알 수 없는 거절은 납득할 수 없다”며 불공정함을 강하게 호소했다. 결국 그는 은행의 신용평가 체계 전반을 불신하게 되었고, 다른 금융기관으로 이탈하는 결과를 낳았다.

이 사례는 금융 맥락에서 신뢰 형성과 불신 확산을 잘 보여준다.

설명 가능성 부족: 결과만 제시되고, 판단 근거가 불투명할 경우 고객의 신뢰는 약화된다.

절차적 공정성(perceived fairness): 고객은 결과 자체보다 왜 그런 결정을 내렸는가를 중시한다.

인지적 신뢰 붕괴: 설명 부재는 단순한 불만을 넘어, 금융 시스템 자체에 대한 불신으로 이어진다.

따라서 금융 분야에서 AI 신뢰를 확보하기 위해서는, 고객이 납득할 수 있는 설명 가능한 기준 제시, 사람 직원의 보완적 역할, 투명한 피드백 절차가 반드시 병행되어야 한다.

교육 사례: 서울 소재 A대학의 온라인 강의 플랫폼에서는 AI 기반 튜터 챗봇이 학습 지원에 활용되었다. 챗봇은 단순히 문제 풀이 힌트만 제공하는 것이 아니라, 학습 과정에서 불안을 표현하는 학생에게 “괜찮아요, 지금처럼 차근차근하면 충분히 잘할 수 있습니다”, **“조금 어려워도 정상적인 과정이에요. 계속 시도해 보세요”**와 같은 격려성 피드백을 제시하였다.

이를 경험한 학습자들은 사후 설문에서 “AI가 나를 이해하고 응원해 준다”,

“실패해도 다시 시도할 수 있다는 자신감을 얻었다”라는 반응을 보였다. 특히, 격려 발언을 제공받은 집단은 그렇지 않은 집단에 비해 시험 직전 불안 수준이 평균적으로 낮았으며(평균 차이 약 0.8점, $p<.05$), 과제 지속 시간도 더 길게 유지되었다.

이 사례는 AI의 단순 정보 제공 기능을 넘어, 정서적 교류와 심리적 안정감 제공이 신뢰 형성에 중요한 역할을 한다는 점을 시사한다. 학습자는 AI를 단순한 기계가 아니라, “나를 도와주는 사회적 파트너”로 인식하게 되었으며, 이는 신뢰 관계 구축으로 이어졌다.

법률 사례: 국내 모 사법연구기관에서는 판결 보조 AI를 시범 도입하여 판례 검색, 양형(刑量) 추천, 유사 사건 비교 등을 지원하였다. 해당 시스템은 과거 수천 건의 판례 데이터를 기반으로 특정 사건에 대해 “통상적으로 ○○년~○○년의 형량이 선고됨”이라는 결과를 제시하였다.

실험적으로 진행된 설문조사에서, 법률 전문가 집단(판사·변호사)은 AI의 추천을 일정 부분 참고했으나 “사건 맥락과 정황은 기계적으로 해석할 수 없다”며 최종 신뢰 수준이 중간 정도(평균 3.2/5)에 머물렀다. 반면, 일반 시민 집단은 동일한 결과를 “객관적이고 공정하다”라고 평가하며 더 높은 신뢰(평균 4.1/5)를 나타냈다.

이처럼 전문가와 비전문가 간 신뢰 격차가 발생한 이유는 다음과 같다.

전문가 집단: AI의 한계를 잘 알기에, “법리적 판단은 데이터 패턴만으로는 설명할 수 없다”는 불신 요인이 작용.

일반 시민 집단: 오히려 AI의 수치적·데이터 기반 판단을 ‘공정성’으로 인식하며, 인간 판사보다 편향이 적다고 기대.

이 사례는 AI 판결 보조 시스템이 법적 의사결정 과정에서 신뢰를 획득하는 방식이 집단에 따라 상반될 수 있다는 점을 보여준다. 따라서 향후 법률 AI 설계에서는 전문가가 활용할 수 있는 투명한 설명 제공과 함께, 일반 시민에게는 신뢰의 근거가 오해되지 않도록 교육적 장치가 필요하다.

7. 논의

본 연구는 인간이 AI를 신뢰하는 데 있어 인지적 요인(설명 가능성, 전문가 협력)과 정서적 요인(감정 표현)이 모두 중요함을 밝혔다. 특히 전문가 협력성은 가장 강력한 신뢰 요인으로 나타났으며, 이는 AI가 독립적 의사결정자가 아니라 보조적·협력적 파트너로 설계될 필요성을 시사한다.

구체적으로, 의료 영역에서 환자는 AI의 진단만 제시될 때보다 “의사가 검토·승인한 AI의 결과”일 때 더 높은 신뢰를 보였다. 이는 AI가 전문가의 판단을 보완하는 형태일 때, 사용자가 이를 “공인된 정보”로 받아들이기 때문이다. 금융 영역에서도 마찬가지로, AI의 대출 거절 결과가 은행 직원의 설명과 함께 제공될 경우, 고객은 시스템을 더 공정하고 투명하다고 인식하였다.

따라서 AI-전문가 협력 모델은 단순한 보조적 기능을 넘어, 신뢰 확보를 위한 구조적 설계 원리로 자리매김할 수 있다. 예컨대:

의료 분야: “AI 제안 → 전문가 검증 → 환자 전달”이라는 3단계 절차

법률 분야: “AI 판례 추천 → 변호사·판사의 최종 판단” 구조

교육 분야: “AI 튜터 설명 → 교사의 피드백 보완” 방식

이와 같이 전문가 협력이 내재된 구조는 단순히 가능성이 아니라 실질적으로 구현 가능한 설계 원리이며, 향후 AI 신뢰 연구와 실무 적용에 있어 필수적 조건으로 작동할 것이다.

8. 결론 및 향후 연구

본 연구는 인간-AI 신뢰 관계의 형성을 설명하는 핵심 요인으로 전문가 권위, 설명 가능성, 사회적 존재감을 도출하였다.

이를 통해 AI가 단순한 정보 제공자가 아니라, 인간의 의사결정과 심리적 안정에 영향을 미치는 상호작용 주체라는 점을 확인하였다. 따라서 향후 AI 설계에서는 기술적 성능뿐 아니라 심리적·윤리적 설계 원리가 함께 고려되어야 한다.

그러나 본 연구는 제한된 시나리오와 사례 분석에 기반했기 때문에, 보다 일반화된 결론을 위해 다음과 같은 후속 연구가 필요하다.

첫째, 문화권별 비교 연구가 요구된다. 신뢰의 심리적 요인은 문화적 맥락에 따라 다르게 작용할 수 있으며, 집단주의 문화에서는 “전문가 권위” 요인이 더 강력하게 작동하는 반면, 개인주의 문화에서는 “설명 가능성”이 더 중요한 요인으로 작용할 수 있다. 이를

검증하기 위해 동일한 시나리오 설문을 한국, 미국, 일본 등 다양한 문화권에서 실시하고, 문화별 차이를 다집단 구조방정식 모형(Multi-group SEM)으로 비교 분석할 수 있다. 또한 설문조사에 더해 포커스 그룹 인터뷰와 같은 질적 연구를 병행하여 문화적 해석의 차이를 심층적으로 파악하는 것도 효과적이다.

둘째, 장기적 상호작용에 따른 신뢰 변화를 추적할 필요가 있다. 단기적 사용에서는 “신기효과(novelty effect)”가 나타날 수 있으나, 장기간 사용 시에는 초기 신뢰가 감소하거나 유지되는 다양한 패턴이 존재할 수 있다. 이를 검증하기 위해 6개월에서 1년간 패널 데이터를 수집하고, 일정 시점별(예: 1개월, 3개월, 6개월, 12개월) 반복 측정 설문을 실시할 수 있다. 또한 잠재 성장 모형(Latent Growth Model)과 군집 분석을 활용해 신뢰의 변화 궤적과 사용자 집단 간 차이를 정량적으로 파악할 수 있으며, 오류 경험이나 피드백 반영 여부와 같은 중재 요인을 통합적으로 고려해야 한다.

셋째, 실제 현장 데이터 기반 검증이 필요하다. 본 연구는 사례 중심 분석에 의존했으므로, 의료·금융·교육·법률 등 실제 현장에서 발생하는 로그 데이터를 수집·분석하여 신뢰 요인과 실제 의사결정 간의 상관관계를 검증해야 한다. 예를 들어, 의료 영역에서는 환자–의사–AI 간 삼자 관계에서 신뢰가 치료 순응도에 미치는 영향을 분석할 수 있고, 금융 영역에서는 AI 대출 심사 추천과 최종 승인 간 불일치 사례를 통해 신뢰 작동 조건을 탐색할 수 있다. 교육 영역에서는 AI 튜터의 장기적 사용이 학습 성취도와 학생·교사 간 신뢰 관계에 미치는 영향을 추적할 수 있으며, 법률 영역에서는 판례 추천 AI가 실제 변호사와 판사의 결정 흐름에 어떻게 반영되는지 네트워크 분석을 통해 규명할 수 있다. 이러한 데이터 분석은 대규모 로그 데이터 기반 통계 기법과 현장 인터뷰를 병행하는 혼합 방법 연구(Mixed Methods)를 통해 타당성을 강화할 수 있다. 이와 같은 후속 연구는 신뢰의 심리적 요인을 정량적으로 검증할 뿐 아니라, 문화적·시간적·맥락적 요인을 고려한 인간-AI 신뢰 모델을 구축하는 데 기여할 것이다. 이는 AI 기술이 사회 전반에서 보다 책임감 있게 활용되기 위한 필수적인 토대가 될 것이다.

참고문헌

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. Journal of Broadcasting & Electronic Media, 64(4), 541–565.

IBM Watson Health. (2019). Watson for Oncology: Lessons learned. IBM White Paper.