JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 5, pp. 1191-1202, May. 2026

Check for updates

생성형 인공지능의 환각 공개가 사용자 신뢰에 미치는 영향: 버추얼 휴먼 인터페이스의 조절 효과

강 지 영*
이화여자대학교 커뮤니케이션·미디어학부 교수

Effects of AI Hallucination Disclosure on User Trust: Moderating Role of Virtual Human Interfaces

Jiyoung Kang*
Professor, Division of Communication & Media, Ewha Womans University, Seoul 03760, Korea

[요 약]

최근 생성형 인공지능의 발전은 다양한 디지털 환경에서 인간-AI 상호작용을 확대시키고 있다. 그러나 생성형 AI는 사실과 다른 정보를 그럴듯하게 생성하는 환각 문제를 내포하고 있다. 본 연구는 환각 가능성 공개가 사용자 인식과 신뢰 형성에 미치는 영향을 분석하였다. 이를 위해 환각 가능성 공개 여부와 인터페이스 유형(버추얼 휴먼 vs. 텍스트 챗봇 인터페이스)를 독립변수로 하는 2×2 실험 설계를 적용하였으며, 총 130명을 대상으로 실험을 수행하였다. 분석 결과, 환각 가능성 공개는 지각된 투명성을 증가시키는 반면 지각된 능력을 감소시키는 상반된 효과를 보였다. 또한 지각된 투명성과 지각된 능력은 모두 AI 신뢰에 긍정적인 영향을 미쳤으나, 환각 가능성 공개의 직접 효과는 유의하지 않았다. 한편 버추얼 휴먼 인터페이스 조건에서 환각 가능성 공개의 투명성 효과가 더 크게 나타났다. 이러한 결과는 생성형 AI 환경에서 투명성과 능력 인식 간의 상반된 효과, 즉 투명성 역설을 보여주며, 커뮤니케이션 전략과 인터페이스 설계의 중요성을 시사한다.

[Abstract]

Recent advancements in generative artificial intelligence (AI) have significantly expanded human–AI interaction across digital ecosystems. However, these systems are inherently susceptible to hallucinations, producing plausible yet factually incorrect information. This study investigates the influence of disclosure of such hallucination risks on user perceptions and trust formation. A 2 × 2 factorial experiment was conducted with 130 participants, in which hallucination disclosure (disclosure vs. no disclosure) and interface type (virtual human vs. text-based chatbot) were manipulated. The results show that hallucination disclosure has opposing effects: it increases perceived transparency but decreases perceived competence. Both perceived transparency and competence positively influenced user trust, whereas the direct effect of disclosure on trust was not significant. The effect of disclosure on perceived transparency was more pronounced within the virtual human condition. These findings reveal a transparency paradox, highlighting the critical roles of communication strategies and interface design in shaping user trust in generative AI systems.

색인어 : 생성형 인공지능, AI 환각, AI 투명성, 버추얼 휴먼, 사용자 신뢰
Keyword : Generative AI, AI Hallucination, AI Transparency, Virtual Humans, User Trust

http://dx.doi.org/10.9728/dcs.2026.27.5.1191

Received 25 March 2026; **Revised** 17 April 2026
Accepted 30 April 2026

***Corresponding Author; Jiyoung Kang**
Tel: +82-2-3277-2266
E-mail: kangjiyoung@ewha.ac.kr

Ⅰ. 서 론

최근 생성형 인공지능(Generative Artificial Intelligence)의 급속한 발전은 인간과 인공지능 간 상호작용의 방식과 범위를 크게 확장시키고 있다. 대규모 언어 모델(Large Language Models, LLMs)을 기반으로 한 인공지능 시스템은 자연어 생성 능력을 바탕으로 정보 검색, 고객 서비스, 교육, 콘텐츠 제작 등 다양한 영역에서 활용되며, 대화형 인터페이스를 중심으로 일상적인 디지털 환경에 깊이 통합되고 있다. 이러한 변화는 인공지능을 단순한 도구를 넘어 인간과 상호작용하는 사회적 행위자(social actor)로 인식하게 만들며, 인간-AI 상호작용에서 신뢰(trust)는 중요한 연구 주제로 부상하고 있다[1].

그러나 생성형 AI의 확산과 함께 중요한 문제로 지적되는 현상이 바로 AI 환각(hallucination)이다. 생성형 AI는 사실 여부와 관계없이 그럴듯한 정보를 생성할 수 있으며, 실제로 존재하지 않는 정보나 부정확한 내용을 자연스럽고 설득력 있는 형태로 제시할 수 있다. 이러한 특성은 사용자가 오류를 인지하기 어렵게 만들며, 잘못된 신뢰 형성으로 이어질 위험을 내포한다[2],[3]. 특히 이러한 문제는 교육, 의료, 법률과 같은 고위험 정보 환경에서 단순한 기술적 오류를 넘어 사회적 문제로 확장될 수 있다.

이러한 문제를 해결하기 위한 접근으로 최근 연구에서는 인공지능의 투명성(transparency)과 설명 가능성(explainability)을 강조하고 있다. 이는 AI 시스템의 작동 방식이나 한계를 사용자에게 이해 가능한 방식으로 제시함으로써 사용자가 정보를 보다 비판적으로 평가할 수 있도록 하는 설계 전략이다[4]. 그러나 일부 연구에서는 이러한 투명성이 항상 긍정적인 결과를 가져오는 것은 아니며, 인공지능이 자신의 불확실성이나 오류 가능성을 강조할 경우 오히려 시스템의 능력에 대한 평가가 낮아질 수 있음을 지적한다[5].

한편, 최근 인공지능 인터페이스는 텍스트 기반 챗봇을 넘어 버추얼 휴먼(virtual human)과 같은 의인화된 형태로 발전하고 있다. 이러한 인터페이스는 인간과 유사한 외형과 표현을 통해 사용자의 사회적 반응과 정서적 평가에 영향을 미칠 수 있으며, 동일한 정보라도 인터페이스 유형에 따라 사용자 인식과 신뢰 형성 과정이 달라질 가능성이 존재한다[6].

이러한 배경에서 본 연구는 생성형 AI의 환각 문제를 단순한 기술적 오류가 아닌 커뮤니케이션 전략의 관점에서 접근한다. 특히 AI가 정보 생성 과정에서 발생할 수 있는 오류 가능성을 사전에 공개하는 경우, 이러한 정보 제공 방식이 사용자 신뢰와 인식에 어떠한 영향을 미치는지를 분석하고자 한다. 또한 이러한 효과가 버추얼 휴먼 인터페이스와 텍스트 기반 챗봇 인터페이스 간에 어떻게 다르게 나타나는지를 비교 분석함으로써 인간-AI 상호작용 연구에 새로운 시사점을 제시하고자 한다.

이에 본 연구는 다음과 같은 연구 질문을 제시한다.

첫째, AI가 환각 가능성을 사전에 공개하는 경우 사용자 신뢰는 어떻게 변화하는가?

둘째, 이러한 공개는 사용자에게 인식되는 투명성과 능력 평가에 어떠한 영향을 미치는가?

셋째, 이러한 효과는 인터페이스 유형에 따라 차이를 보이는가?

Ⅱ. 이론적 배경

2-1 생성형 인공지능과 환각 현상: 커뮤니케이션 현상으로서의 재해석

최근 생성형 인공지능은 자연어 처리와 콘텐츠 생성 기술의 발전을 기반으로 다양한 디지털 환경에서 핵심적인 역할을 수행하고 있다. 특히 대규모 언어 모델은 인간과 유사한 수준의 자연어 텍스트를 생성함으로써 인간-AI 상호작용을 보다 자연스럽고 사회적인 형태로 변화시키고 있다. 이러한 특성은 인공지능 시스템이 단순한 정보 처리 도구를 넘어 사용자에 의해 사회적 행위자(social actor)로 인식될 수 있음을 시사하며, 이는 Media Equation 및 CASA(Computers as Social Actors) 연구 흐름과도 연결된다[1],[7].

이러한 맥락에서 생성형 AI의 환각은 단순한 기술적 오류를 넘어 인간-AI 커뮤니케이션 과정에서 발생하는 중요한 현상으로 이해될 필요가 있다. 기존 연구에서는 환각을 사실적으로 부정확하거나 검증되지 않은 정보를 그럴듯하게 생성하는 현상으로 정의하며, 이는 언어 모델이 의미 이해보다는 확률적 패턴에 기반하여 텍스트를 생성하기 때문에 발생한다고 설명한다[8],[9]. 또한 최근 연구에서는 환각을 인공지능이 생성하는 왜곡된 정보 유형으로 분류하고, 그 발생 메커니즘과 유형을 체계적으로 분석하고 있다[10].

그러나 이러한 정의는 환각을 기술적 관점에서 설명하는 데 그치며, 사용자 인식과 상호작용 맥락에서의 의미를 충분히 설명하지 못하는 한계를 가진다. 실제로 생성형 AI는 매우 자연스럽고 설득력 있는 형태로 정보를 제시하기 때문에 사용자는 정보의 정확성보다는 표현 방식과 맥락에 기반하여 신뢰를 형성할 가능성이 높다. 이러한 특성은 환각이 단순한 정보 오류가 아니라, 사용자의 인식과 신뢰 형성 과정을 왜곡할 수 있는 커뮤니케이션 현상임을 시사한다.

특히 환각은 정보의 사실 여부와 관계없이 인지된 신뢰성(perceived credibility)과 지각된 능력(perceived competence)에 영향을 미칠 수 있으며, 이는 인간-AI 상호작용에서 중요한 평가 기준으로 작용한다[11]. 따라서 환각 문제는 기술적 정확성의 문제를 넘어, 정보가 어떻게 전달되고 해석되는가라는 커뮤니케이션 차원에서 접근될 필요가 있다.

그럼에도 불구하고 기존 연구는 환각 문제를 주로 모델 개선이나 기술적 해결 방안에 초점을 맞추어 왔으며, 사용자 인

식과 신뢰 형성 과정에서의 영향을 충분히 고려하지 못한 한계를 보인다. 최근 일부 연구에서 AI의 투명성이나 오류 가능성 공개가 사용자 평가에 영향을 미칠 수 있음이 제시되고 있으나, 이러한 효과가 신뢰와 능력 평가에 어떠한 방식으로 작용하는지에 대한 체계적인 분석은 여전히 부족한 상황이다.

따라서 본 연구는 환각을 단순한 기술적 오류가 아닌 커뮤니케이션 전략의 문제로 재정의하고, AI가 자신의 오류 가능성을 사전에 공개하는 경우 이러한 정보 제공 방식이 사용자 인식과 신뢰 형성에 어떠한 영향을 미치는지를 분석하고자 한다.

2-2 AI 투명성과 신뢰

인공지능 시스템이 다양한 사회적 영역에서 활용됨에 따라 AI에 대한 사용자 신뢰(trust in AI)는 인간-AI 상호작용 연구에서 핵심적인 연구 주제로 자리 잡고 있다. 신뢰는 사용자가 기술 시스템을 수용하고 지속적으로 활용하는 데 중요한 요인으로 작용하며, 특히 인공지능과 같이 복잡한 알고리즘 기반 시스템에서는 사용자가 시스템의 작동 원리를 완전히 이해하기 어렵기 때문에 신뢰 형성 과정이 더욱 중요하게 논의된다[12]. 기존 연구에서는AI에 대한 신뢰가 단순한 기술적 성능뿐 아니라 사용자 경험과 인지적 평가에 의해 형성된다는 점이 강조되고 있으며, 다양한 요인이 신뢰 형성에 영향을 미칠 수 있음이 메타분석 연구를 통해 확인된 바 있다[13].

이러한 맥락에서 투명성(transparency)은 AI 신뢰 형성의 중요한 결정 요인으로 제시된다. 인공지능 시스템이 자신의 작동 방식이나 의사결정 과정을 사용자에게 이해 가능한 형태로 제공할 경우, 사용자는 시스템을 보다 예측 가능하고 신뢰할 수 있는 대상으로 인식하게 된다[12]. 특히 최근 연구에서는 설계 수준에서의 투명성이 인간-AI 상호작용 상황에서 신뢰 형성과 정보 공유 행동에 긍정적인 영향을 미칠 수 있음이 실증적으로 확인되고 있다[14]. 또한 국내 연구에서도 생성형 AI 환경에서 투명성과 설명 가능성이 사용자 신뢰 인식에 중요한 영향을 미칠 수 있음이 제시된 바 있다[11].

그러나 인간-AI 상호작용에서 사용자의 평가는 단순한 기술적 정보에만 기반하지 않는다. Sundar[1]는 사용자가 인공지능 시스템을 평가할 때 실제 성능보다 인터페이스 특성이나 기계 단서(machine cues)에 의존하는 경향이 있음을 지적하며, 이를 machine heuristic 개념으로 설명하였다. 또한 인간은 인공지능을 의인화하여 인식하는 경향이 있으며, 이러한 인식은 신뢰 형성에 영향을 미칠 수 있다[15]. 이러한 관점에서 AI가 어떠한 방식으로 정보를 전달하는지, 특히 자신의 한계나 불확실성을 어떻게 표현하는지는 사용자 인식에 중요한 영향을 미치는 요소로 작용한다.

한편, 최근 연구에서는 AI 투명성이 항상 긍정적인 효과를 가져오는 것은 아니라는 점이 제기되고 있다. 인공지능 시스템이 자신의 불확실성이나 오류 가능성을 강조할 경우, 사용자는 해당 시스템의 능력(perceived competence)을 낮게 평가할 수 있으며, 이는 신뢰 감소로 이어질 가능성이 있다[16]. 또한 인간과 유사한 시스템에서 나타나는 불편감이나 거부감 역시 기술 평가에 영향을 미칠 수 있으며, 이는 언캐니 밸리(uncanny valley)와 같은 현상으로 설명된다[17]. 이러한 결과는 기술 시스템 평가에서 능력 인식과 정서적 반응이 중요한 역할을 한다는 점을 보여준다.

이와 같이 AI 투명성은 한편으로는 시스템의 이해 가능성과 신뢰성을 강화하는 긍정적 기능을 수행하는 동시에, 다른 한편으로는 능력에 대한 인식을 약화시키거나 부정적 정서 반응을 유발할 수 있는 양면적 특성을 가진다. 이러한 상반된 효과는 AI 커뮤니케이션 설계에서 중요한 이론적 문제로 제기되며, 특히 생성형 AI와 같이 불확실성을 내포한 정보 생성 환경에서 더욱 중요하게 작용할 수 있다.

따라서 본 연구는 AI 시스템이 환각 가능성을 사전에 공개하는 커뮤니케이션 전략이 사용자 신뢰 형성에 어떠한 영향을 미치는지를 분석하고자 한다. 특히 이러한 정보 공개가 투명성 인식과 능력 평가에 미치는 상반된 영향을 동시에 고려함으로써, 생성형 AI 환경에서의 신뢰 형성 메커니즘을 보다 정교하게 설명하고자 한다.

2-3 AI 커뮤니케이션에서의 투명성 역설

AI 시스템의 투명성은 전통적으로 사용자 신뢰를 강화하는 핵심 요인으로 이해되어 왔다. 시스템의 작동 방식이나 의사결정 과정, 그리고 잠재적 한계를 사용자에게 명시적으로 제시하는 것은 정보 비대칭을 완화하고 사용자가 기술을 보다 합리적으로 평가할 수 있도록 한다[12],[14]. 이러한 관점에서 투명성은 기술 수용과 신뢰 형성을 촉진하는 중요한 설계 원칙으로 간주되어 왔다.

그러나 최근 인간-AI 상호작용 연구에서는 투명성이 항상 신뢰를 증가시키는 것은 아니라는 점이 강조되고 있다. 사용자는 인공지능 시스템을 평가할 때 단순히 제공되는 정보의 양이나 설명 수준에만 의존하는 것이 아니라, 시스템의 능력과 신뢰성을 동시에 판단한다[13]. 특히 AI가 자신의 한계나 불확실성을 명시적으로 표현할 경우, 사용자는 이를 시스템의 낮은 능력에 대한 단서로 해석할 가능성이 있으며, 이는 신뢰 감소로 이어질 수 있다[16].

이와 같은 상반된 효과는 투명성 역설(transparency paradox)로 설명될 수 있다. 투명성 역설은 시스템이 자신의 작동 방식이나 한계를 공개할수록 사용자가 이를 더 신뢰할 것이라는 일반적인 가정과 달리, 특정 상황에서는 이러한 정보 공개가 오히려 신뢰 감소로 이어질 수 있음을 의미한다. 이는 사용자가 기술 시스템을 평가할 때 이해 가능성뿐 아니라 능력 단서와 정서적 반응을 동시에 고려하기 때문이며, 특히 인간과 유사한 특성을 지닌 시스템에서는 이러한 평가가 더욱 민감하게 나타날 수 있다[17].

또한 인간은 인공지능을 단순한 도구가 아니라 인간과 유사한 존재로 인식하는 경향이 있으며, 이러한 의인화(anthropomorphism)는 AI에 대한 신뢰 형성 과정에 중요한 영향을 미친다[18]. 의인화된 시스템에서는 사용자가 AI의 행동을 인간과 유사한 기준으로 해석하기 때문에, 시스템이 자신의 한계나 오류 가능성을 공개하는 경우 이러한 정보가 기술적 특성의 표현이 아니라 사회적 신호로 인식될 가능성이 존재한다. 본 연구에서 'AI의 한계 공개'는 AI 시스템이 자신의 정보 생성 과정에서 발생할 수 있는 오류 가능성, 즉 환각(hallucination)의 가능성을 사전에 사용자에게 명시적으로 안내하는 커뮤니케이션 전략으로 정의하였다.

이러한 맥락에서 생성형 AI의 환각 문제는 투명성 역설을 설명하는 중요한 사례를 제공한다. 생성형 AI는 자연스럽고 설득력 있는 정보를 생성할 수 있지만 동시에 사실과 다른 정보를 생성할 가능성을 내포하고 있다[8],[9]. 따라서 AI 시스템이 환각 가능성을 사전에 공개하는 경우, 사용자 인식은 상반된 방향으로 변화할 수 있다. 한편으로는 이러한 정보 공개가 시스템의 정직성과 투명성을 강화하여 신뢰 형성에 긍정적인 영향을 미칠 수 있으며, 다른 한편으로는 시스템의 능력에 대한 의문을 증가시켜 신뢰를 약화시킬 가능성도 존재한다.

따라서 생성형 AI 환경에서의 환각 공개 전략은 단순한 정보 제공을 넘어 사용자 신뢰 형성 과정에서 복합적인 영향을 미치는 커뮤니케이션 전략으로 이해될 필요가 있다. 특히 이러한 효과는 AI 시스템의 인터페이스 특성이나 인간 유사성 수준에 따라 달라질 수 있으며, 이러한 관점에서 투명성 역설을 실증적으로 검증하는 연구가 요구된다.

2-4 버추얼 휴먼과 의인화된 인터페이스

최근 인공지능 인터페이스는 단순한 텍스트 기반 챗봇을 넘어 인간과 유사한 외형과 행동을 갖는 버추얼 휴먼(virtual human) 형태로 발전하고 있다. 버추얼 휴먼은 인간과 유사한 얼굴, 음성, 표정, 제스처 등을 통해 사용자와 상호작용하는 디지털 에이전트를 의미하며, 최근 디지털 콘텐츠, 고객 서비스, 교육, 관광 서비스 등 다양한 영역에서 활용되고 있다[19],[20]. 이러한 인터페이스의 확장은 인간–AI 상호작용을 단순한 정보 전달 과정에서 사회적 상호작용으로 변화시키는 중요한 특징을 가진다. 특히 최근 연구에서는 버추얼 휴먼이 단순한 기술적 인터페이스를 넘어 사회적·윤리적 의미를 동반하는 존재로 인식될 수 있음이 제시되고 있으며[6], 이는 사용자가 AI를 보다 사회적 행위자로 해석하는 경향을 강화할 수 있음을 시사한다.

인간이 컴퓨터나 인공지능 시스템을 사회적 존재로 인식하는 현상은 Media Equation theory를 통해 설명될 수 있다. Nass와 Moon은 사람들이 컴퓨터와 상호작용할 때 실제 인간과 유사한 사회적 규범을 적용하는 경향이 있음을 지적하였다. 즉 사용자는 인공지능 시스템에 대해 예의, 신뢰, 협력과 같은 사회적 반응을 보일 수 있으며, 이러한 현상은 인간–컴퓨터 상호작용 연구에서 중요한 이론적 기반으로 활용되어 왔다[7].

또한 최근 연구에서는 인공지능 인터페이스의 의인화 수준이 사용자 인식과 행동에 중요한 영향을 미칠 수 있다는 점이 강조되고 있다. 의인화는 인간이 비인간 대상에 인간적 특성이나 의도를 부여하는 심리적 경향을 의미하며, 인공지능 시스템이 인간과 유사한 외형이나 행동을 제공할 경우 사용자는 해당 시스템을 보다 사회적 존재로 인식할 가능성이 높아진다[18]. 이러한 인식은 인공지능을 단순한 기술이 아니라 사회적 행위자(social actor)로 이해하게 만들며, 사용자 태도와 신뢰 형성에 중요한 영향을 미친다.

특히 의인화된 인터페이스에서는 사용자가 인공지능의 행동을 인간과 유사한 기준으로 해석하는 경향이 강화된다. Waytz 등은 인공지능 시스템에 인간적 특성이 부여될 경우 사용자가 해당 시스템에 대해 더 높은 수준의 신뢰를 형성할 수 있음을 보여주었다[15]. 또한 가상 환경에서 구현된 디지털 인간은 사용자에게 사회적 존재감(social presence)을 제공하며, 이는 사용자와 시스템 간 관계 형성에 중요한 역할을 한다[6].

이러한 관점에서 인공지능 인터페이스의 인간 유사성은 단순한 외형적 특성을 넘어, 사용자 인식과 신뢰 형성 메커니즘을 변화시키는 핵심 요인으로 작용할 수 있다. 특히 생성형 AI와 같이 불확실성이 내재된 정보 환경에서는 동일한 정보라도 인터페이스 유형에 따라 사용자의 해석과 평가가 달라질 가능성이 존재한다. 예를 들어 동일한 환각 가능성 정보라도 텍스트 기반 챗봇이 전달하는 경우와 인간과 유사한 버추얼 휴먼이 전달하는 경우, 사용자의 신뢰 평가와 능력 인식은 서로 다르게 나타날 수 있다.

따라서 본 연구는 AI가 환각 가능성을 사전에 공개하는 커뮤니케이션 전략이 사용자 신뢰 형성에 미치는 영향을 분석함과 동시에, 이러한 효과가 인터페이스 유형(버추얼 휴먼 vs 텍스트 기반 챗봇)에 따라 어떻게 달라지는지를 비교 분석하고자 한다. 이를 통해 인공지능 인터페이스의 의인화 수준이 AI 커뮤니케이션 전략의 효과를 조절하는 메커니즘을 규명하고자 한다.

Ⅲ. 연구가설

3-1 환각 가능성 공개와 지각된 투명성

AI 시스템이 자신의 정보 생성 과정에서 발생할 수 있는 오류 가능성을 사용자에게 사전에 공개하는 경우, 사용자는 해당 시스템을 보다 투명하고 이해 가능한 기술로 인식할 가능성이 있다. 기존 연구에 따르면 기술 시스템이 자신의 작동

방식이나 한계를 명시적으로 설명할 경우, 사용자는 시스템의 의사결정 과정을 보다 쉽게 이해할 수 있으며, 이러한 정보 제공은 투명성 인식을 증가시킬 수 있다[12],[14]. 특히 AI 시스템이 자신의 불확실성을 명시적으로 전달하는 경우 이는 시스템의 정직성과 책임성을 나타내는 신호로 해석될 수 있다. 따라서 다음과 같은 가설을 설정하였다.

H1. AI가 환각 가능성을 사전에 공개할 경우 사용자가 인식하는 AI의 투명성은 증가할 것이다.

3-2 환각 가능성 공개와 지각된 능력

한편, AI 시스템이 자신의 오류 가능성을 강조하는 경우 이는 사용자에게 시스템의 한계를 부각시키는 신호로 작용할 수 있다. 자동화 시스템 연구에서는 사용자가 기술을 평가할 때 정확성과 능력을 중요한 판단 기준으로 활용하며, 시스템이 불확실성을 표현할 경우 사용자는 해당 시스템을 덜 유능한 기술로 인식할 가능성이 있다[13],[16]. 이러한 관점에서 환각 가능성 공개는 사용자로 하여금 AI의 정보 처리 능력에 대해 보다 보수적인 평가를 내리도록 만들 수 있다. 따라서 다음과 같은 가설을 설정하였다.

H2. AI가 환각 가능성을 사전에 공개할 경우 사용자가 인식하는 AI의 능력 평가는 낮아질 것이다.

3-3 지각된 투명성과 신뢰

AI 시스템의 투명성은 사용자 신뢰 형성에 중요한 영향을 미치는 요인으로 알려져 있다. 시스템이 자신의 작동 방식이나 정보 처리 과정을 사용자에게 이해 가능한 형태로 제공할 경우, 사용자는 해당 시스템을 보다 정직하고 신뢰할 수 있는 대상으로 인식할 가능성이 높다[12],[14]. 따라서 투명성에 대한 인식은 AI 신뢰 형성에 긍정적인 영향을 미칠 것으로 예상된다. 이에 다음과 같은 가설을 설정하였다.

H3. 사용자가 인식하는 AI의 투명성은 AI에 대한 신뢰에 긍정적인 영향을 미칠 것이다.

3-4 지각된 능력과 신뢰

기술 시스템의 능력에 대한 인식은 신뢰 형성의 핵심 요인 중 하나이다. 사용자는 시스템이 정확하고 유능하다고 판단할 때 해당 기술을 보다 신뢰하고 의존하는 경향을 보인다[13]. 따라서 AI의 능력에 대한 인식은 사용자 신뢰 형성에 중요한 영향을 미칠 것으로 예상된다. 따라서 다음과 같은 가설을 설정하였다.

H4. 사용자가 인식하는 AI의 능력 평가는 AI에 대한 신뢰에 긍정적인 영향을 미칠 것이다.

3-5 환각 가능성 공개와 신뢰: 투명성 역설 관점

AI 시스템이 자신의 정보 생성 과정에서 발생할 수 있는 오류 가능성, 즉 환각(hallucination) 가능성을 사전에 공개하는 행위는 사용자 신뢰 형성에 복합적인 영향을 미칠 수 있다. 한편으로 이러한 정보 공개는 시스템의 정직성과 투명성을 강화하여 사용자 신뢰를 증가시키는 긍정적 요인으로 작용할 수 있다. 다른 한편으로는 시스템의 불확실성을 명시적으로 드러냄으로써 사용자가 해당 시스템의 능력에 대해 의문을 갖게 하고, 결과적으로 신뢰를 약화시키는 요인으로 작용할 가능성도 존재한다.

이러한 상반된 효과는 투명성 역설(transparency paradox)의 관점에서 설명될 수 있으며, AI의 정보 공개는 신뢰에 대해 단일 방향의 효과를 가지기보다는 서로 다른 인지적 평가 경로를 통해 복합적으로 작용할 가능성이 있다. 특히 환각 가능성 공개는 지각된 투명성과 지각된 능력을 동시에 변화시키며, 이러한 변화는 궁극적으로 신뢰 형성에 영향을 미칠 수 있다.

그럼에도 불구하고, 기술 시스템이 자신의 한계를 명시적으로 설명하는 경우 사용자는 이를 정직성과 책임성의 신호로 해석할 수 있으며, 이는 신뢰 형성에 긍정적인 영향을 미칠 가능성이 있다. 따라서 본 연구에서는 환각 가능성 공개가 신뢰에 직접적인 영향을 미칠 가능성을 가정한다.

H5. 환각 가능성 공개는 사용자 신뢰에 직접적인 영향을 미칠 것이다.

3-6 인터페이스 유형의 조절 효과

인공지능 인터페이스의 인간 유사성은 사용자 인식과 해석 방식에 중요한 영향을 미칠 수 있다. 특히 버추얼 휴먼과 같은 의인화된 인터페이스는 사용자가 AI를 사회적 존재로 인식하도록 만들며, 이러한 인식은 AI의 행동과 메시지를 보다 사회적 맥락에서 해석하도록 유도할 수 있다[18],[20].

이러한 관점에서 동일한 환각 가능성 공개 정보라도 인터페이스 유형에 따라 사용자에게 다르게 해석될 가능성이 존재한다. 특히 의인화된 인터페이스에서는 AI의 환각 가능성 공개가 단순한 기술적 설명이 아니라 정직성과 책임성을 나타내는 사회적 신호로 해석될 가능성이 높다. 반면 텍스트 기반 인터페이스에서는 동일한 정보가 보다 기능적이고 기술적인 설명으로 인식될 수 있다.

따라서 인터페이스 유형은 환각 가능성 공개가 사용자 인식에 미치는 영향을 조절할 것으로 예상되며, 특히 환각 가능성 공개와 지각된 투명성 간의 관계에서 조절 효과가 나타날 것으로 예상된다.

H6. 인터페이스 유형(버추얼 휴먼 vs. 텍스트 챗봇)은 환각 가능성 공개와 지각된 투명성 간의 관계를 조절할 것이다.

Ⅳ. 연구모형

앞서 제시한 가설을 종합하여, 본 연구는 생성형 AI의 환각 가능성 공개가 사용자 인식과 신뢰 형성 과정에 미치는 영향을 설명하기 위한 연구모형을 제안한다.

본 연구모형은 환각 가능성 공개가 사용자 인식에 미치는 영향을 중심으로 구성된다. 구체적으로, AI가 자신의 정보 생성 과정에서 발생할 수 있는 오류 가능성을 사전에 공개하는 경우, 이는 사용자가 인식하는 지각된 투명성(perceived transparency)과 지각된 능력(perceived competence)에 영향을 미칠 것으로 가정하였다. 이러한 두 변수는 각각 사용자 신뢰 형성에 영향을 미치는 핵심 요인으로 설정되었다.

특히 본 연구는 투명성 역설 관점을 반영하여, 환각 공개가 지각된 투명성에는 긍정적인 영향을 미치는 반면, 지각된 능력에는 부정적인 영향을 미칠 수 있는 이중 경로를 포함하였다. 또한 이러한 인식 변수들이 AI에 대한 신뢰에 영향을 미치는 매개 구조를 중심으로 연구모형을 구성하였다. 한편, 환각 공개는 사용자 신뢰에 직접적인 영향을 미칠 가능성도 고려하여, 독립변수와 종속변수 간의 직접 경로를 함께 포함하였다.

또한 본 연구에서는 인간-AI 상호작용 맥락에서 인터페이스의 인간 유사성이 중요한 역할을 할 수 있다는 점에 주목하여, AI 인터페이스 유형(버추얼 휴먼 vs. 텍스트 챗봇 인터페이스)를 조절변수로 설정하였다. 특히 인터페이스 유형은 환각 공개와 지각된 투명성 간의 관계를 조절하는 변수로 작용할 것으로 가정하였다.

따라서 본 연구모형은 환각 가능성 공개가 지각된 투명성과 지각된 능력을 통해 신뢰에 영향을 미치는 매개 구조(mediation model)를 중심으로 구성되며, 인터페이스 유형이 일부 경로를 조절하는 조절된 매개모형(moderated mediation model)의 형태를 갖는다.

본 연구의 전체 연구모형은 아래 그림 1과 같다.

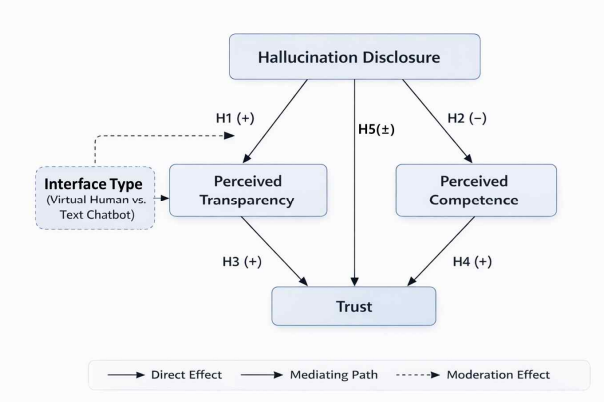


그림 1. 연구모형
Fig. 1. Proposed research model

Ⅴ. 연구결과

5-1 연구 설계 및 표본 규모 산정

본 연구는 생성형 AI의 환각 가능성 공개가 사용자 인식과 신뢰 형성 과정에 미치는 영향을 검증하기 위해 2 × 2 요인 간 실험 설계(between-subjects factorial design)를 적용하였다.

독립변수는 환각 가능성 공개와 인터페이스 유형(Agent Type)으로 구성되었다. 환각 가능성 공개는 AI 시스템이 정보 생성 과정에서 발생할 수 있는 오류 가능성을 사전에 사용자에게 안내하는 조건과 안내하지 않는 조건으로 조작하였다. 인터페이스 유형은 인터페이스 형태에 따라 버추얼 휴먼과 텍스트 기반 챗봇 조건으로 구분하였다.

이에 따라 실험 조건은 다음과 같이 총 네 가지로 구성되었으며 각 참여자는 네 가지 조건 중 하나에 무작위로 할당되었다.

1. 버추얼 휴먼 + 환각 공개
2. 버추얼 휴먼 + 환각 비공개
3. 텍스트 챗봇 인터페이스 + 환각 공개
4. 텍스트 챗봇 인터페이스 + 환각 비공개

표본 규모 산정을 위해 G*Power 3.1을 활용한 사전 검정력 분석(a priori power analysis)을 수행하였다. Cohen[21]이 제시한 중간 효과 크기(effect size, f = 0.25), 유의수준(α) = .05, 검정력(power) = .80을 기준으로 설정하였으며, 4개의 집단을 고려한 분석 결과 최소 필요 표본 수는 128명으로 산출되었다. 이에 본 연구에서는 총 130명의 참여자를 확보하여 통계적 검정력을 충족하였다.

이와 같은 실험 설계를 통해 환각 공개 여부와 인터페이스 유형이 사용자 인식(지각된 투명성, 지각된 능력) 및 신뢰 형성에 미치는 영향을 체계적으로 비교 분석하였다.

5-2 실험 자극 제작

본 연구에서는 AI 인터페이스 유형에 따른 사용자 인식 차이를 분석하기 위해 두 가지 형태의 실험 자극을 제작하였다.

먼저 버추얼 휴먼 조건에서는 AI 기반 가상 인간 생성 플랫폼인 AI Studio PERSO를 활용하여 디지털 캐릭터 형태의 AI 인터페이스를 제작하였다. PERSO는 인공지능 기반 음성 합성과 얼굴 애니메이션 기술을 활용하여 인간과 유사한 외형과 음성 표현을 생성할 수 있는 플랫폼으로, 실제 대화 상황과 유사한 인터페이스를 구현할 수 있다. 본 연구에서는 약 30초 분량의 설명 영상 형태로 버추얼 휴먼이 정보를 전달하도록 구성하였으며, 중립적인 표정과 안정적인 음성을 사용하여 불필요한 감정적 요소가 결과에 영향을 미치지 않도록 통제하였다.

한편 텍스트 챗봇 인터페이스 조건에서는 네이버 클라우드에서 제공하는 챗봇 개발 플랫폼인 CLOVA Chatbot을 활용하여 텍스트 기반 대화 인터페이스를 구현하였다. CLOVA

Chatbot은 자연어 이해(NLU) 기술을 기반으로 사용자 입력을 분석하고 적절한 응답을 제공하는 시스템으로, 본 연구에서는 버추얼 휴먼 조건과 동일한 정보 내용을 단일 메시지 형태의 텍스트 응답으로 제시하도록 구성하였다.

본 연구에서는 인터페이스 유형 이외의 영향을 최소화하기 위해 두 조건에서 동일한 대화 스크립트를 사용하였다. 해당 스크립트는 스마트홈 기술의 작동 원리에 대한 설명으로 구성되었으며, 정보 내용, 문장 구조, 설명 순서, 응답 길이를 동일하게 유지하였다. 예를 들어 "스마트홈 기술은 다양한 가전기기를 네트워크로 연결하여 자동으로 제어할 수 있는 시스템으로…"와 같이 기술의 정의, 작동 방식, 활용 예시를 포함한 약 3~4문장 분량으로 구성하였다. 이러한 설계는 정보 내용의 차이에 따른 영향을 배제하고 인터페이스 유형의 효과만을 검증하기 위한 통제 전략이다.

또한 AI 환각(환각) 가능성 공개 조건에서는 AI 시스템이 정보 제공 이전에 자신의 정보 생성 과정에서 오류가 발생할 수 있음을 사용자에게 사전에 안내하도록 설정하였다. 구체적으로 AI 인터페이스는 다음과 같은 메시지를 제시하였다.

"저는 인공지능 기반 시스템으로 학습된 데이터를 바탕으로 정보를 제공합니다. 그러나 경우에 따라 부정확한 정보가 생성될 수 있으므로 제공되는 정보는 참고용으로 활용해 주시기 바랍니다."

반면 비공개 조건에서는 해당 메시지를 제시하지 않고 동일한 설명이 바로 제공되도록 구성하였다.

실험 자극의 이해 가능성과 자연스러움을 검증하기 위해 파일럿 테스트(pilot test)를 실시하였다. 파일럿 테스트는 대학원생 및 일반 사용자 10명을 대상으로 수행되었으며, 자극의 이해도와 자연스러움에 대한 평가를 바탕으로 문장 표현과 길이를 최종적으로 수정하였다.

5-3 연구 절차

본 연구는 온라인 기반 실험 방식으로 진행되었다. 참여자는 실험에 참여하기에 앞서 연구 목적과 절차에 대한 안내를 받았으며, 연구 참여에 대한 자발적 동의를 확인한 후 실험에 참여하였다. 동의 절차 이후 참여자는 무작위 배정을 통해 네 가지 실험 조건 중 하나에 할당되었다. 각 참여자는 하나의 조건에만 노출되는 between-subjects 설계에 따라 실험이 진행되었다.

참여자는 배정된 조건에 따라 버추얼 휴먼 또는 텍스트 챗봇 인터페이스 형태의 AI 인터페이스 자극을 확인하였다. 버추얼 휴먼 조건에서는 AI 인터페이스가 특정 정보를 설명하는 영상 자극이 제시되었으며, 텍스트 챗봇 인터페이스 조건에서는 동일한 정보 내용이 텍스트 기반 대화 형태로 제공되었다. 또한 환각 공개 조건에서는 AI 시스템이 정보 제공 이전에 오류 가능성을 안내하는 메시지가 제시되었으며, 비공개 조건에서는 해당 메시지가 제공되지 않았다. 참여자는 자극 노출 이후 설문 문항에 응답하였다. 설문은 지각된 투명성, 지각된 능력, 그리고 AI에 대한 신뢰를 측정하기 위한 항목으로 구성되었다.

5-4 측정 변수

본 연구에서는 연구모형에 포함된 주요 변수들을 측정하기 위해 선행연구에서 검증된 척도를 바탕으로 설문 문항을 구성하였다. 모든 문항은 7점 Likert 척도(1 = 전혀 그렇지 않다, 7 = 매우 그렇다)를 사용하여 측정하였다. 문항은 기존 연구를 참고하여 번역 후 의미의 정확성을 확보하기 위해 검토 과정을 거쳐 구성하였다.

또한 각 구성개념의 Cronbach's α 값은 0.80 이상으로 나타나 내적 일관성이 확보된 것으로 확인되었다. 구성개념 간 상관분석 결과에서도 변수 간 상관계수가 과도하게 높지 않은 것으로 나타나 판별타당성이 확보된 것으로 판단된다.

지각된 투명성(perceived transparency)은 AI 시스템이 자신의 작동 방식과 한계를 얼마나 명확하고 솔직하게 설명한다고 인식되는지를 의미한다. 이를 측정하기 위해 "AI 시스템은 자신의 한계를 명확하게 설명한다", "AI 시스템은 정보 제공 과정이 투명하다", "AI 시스템은 자신의 기능을 솔직하게 설명한다" 등의 총 3문항을 사용하였다[12].

지각된 능력(perceived competence)은 AI 시스템이 얼마나 유능하고 정확한 정보를 제공할 수 있다고 인식되는지를 의미한다. 이를 측정하기 위해 "AI 시스템은 유능하다", "AI 시스템은 신뢰할 수 있는 정보를 제공한다", "AI 시스템은 충분한 전문성을 갖추고 있다" 등의 총 3문항을 사용하였다[11],[13].

신뢰(trust)는 사용자가 AI 시스템을 얼마나 신뢰하고 의존할 의향이 있는지를 의미한다. 이를 측정하기 위해 "나는 이 AI 시스템을 신뢰한다", "이 AI 시스템은 전반적으로 신뢰할 수 있다", "나는 정보를 찾을 때 이 AI 시스템을 의존할 의향이 있다" 등의 총 3문항을 활용하였다[11],[12].

5-5 분석 방법

본 연구에서는 가설 검증을 위해 단계적인 분석 절차를 수행하였다. 먼저 측정 변수의 신뢰도를 검증하기 위해 Cronbach's α 계수를 활용한 신뢰도 분석을 실시하였다. 이를 통해 각 구성개념의 내적 일관성을 확인하였다. 다음으로 실험 조작의 효과를 검증하기 위해 조작 점검을 수행하고, 실험 조건 간 차이를 확인하기 위해 분산분석(ANOVA)을 실시하였다. 이를 통해 환각 공개 여부와 인터페이스 유형에 따른 집단 간 차이를 검증하였다. 이후 연구모형에서 제시된 변수 간 관계를 검증하기 위해 회귀분석 기반의 매개효과 분석을 수행하였다. 지각된 투명성과 지각된 능력이 환각 공개와 신뢰 간의 관계를 매개하는지를 검증하기 위해 간접효과를 분석하였다.

또한 인터페이스 유형의 조절효과를 검증하기 위해 조절회

귀분석을 실시하였다. 특히 환각 공개와 지각된 투명성 간의 관계에서 인터페이스 유형의 상호작용 효과를 검증하였다. 모든 통계 분석은 SPSS 통계 프로그램을 활용하여 수행하였으며, 유의수준은 $p < .05$를 기준으로 설정하였다.

Ⅵ. 논 의

6-1 표본 특성

초기 수집된 표본 중 응답이 불완전하거나 분석 기준에 부합하지 않는 데이터를 제외한 총 130명의 자료를 최종 분석에 활용하였다. 참여자는 무작위로 네 개의 실험 조건에 배정되었으며, 각 조건별 참여자 수는 비교적 균등하게 분포하였다. 구체적으로, 버추얼 휴먼+ 환각 공개 조건은 33명, 버추얼 휴먼+ 비공개 조건은 32명, 텍스트 챗봇 인터페이스+ 공개 조건은 33명, 텍스트 챗봇 인터페이스+ 비공개 조건은 32명으로 구성되었다.

참여자의 성별은 남성과 여성 각각 65명으로 동일하게 구성되었으며, 평균 연령은 29.4세(SD = 4.8)로 나타났다. 이러한 표본 구성은 성별과 연령 측면에서 비교적 균형 잡힌 분포를 보이며, 집단 간 비교 분석에 적합한 특성을 갖는 것으로 판단된다.

6-2 신뢰도 분석

측정 변수의 내적 일관성을 검증하기 위해 Cronbach's α를 활용한 신뢰도 분석을 실시하였다. 분석 결과, 지각된 투명성의 Cronbach's α 값은 0.86, 지각된 능력은 0.88, 신뢰는 0.90으로 나타났다. 이는 일반적으로 권장되는 기준값인 0.70을 상회하는 수준으로, 본 연구에서 사용된 측정 척도가 높은 신뢰도를 확보하고 있음을 의미한다. 이에 따라 각 구성 개념은 해당 문항들의 평균값을 산출하여 단일 지표로 구성한 후 이후 분석에 활용하였다.

주요 변수들의 기술통계 분석 결과는 표 1에 제시하였다. 분석 결과, 지각된 투명성(Mean = 5.03, SD = 0.98), 지각된 능력(Mean = 5.04, SD = 0.91), 그리고 신뢰(Mean = 5.18, SD = 0.95) 모두 중간값(4점)보다 높은 수준으로 나타나, 전반적으로 참여자들이 AI 시스템에 대해 비교적 긍정적인 평가를 내린 것으로 확인되었다.

표 1. 기술통계 및 신뢰도(N = 130)
Table 1. Descriptive statistics and reliability (N = 130)

Variable	Mean	SD	Cronbach's α
Perceived Transparency	5.03	0.98	0.86
Perceived Competence	5.04	0.91	0.88
Trust	5.18	0.95	0.90

6-3 조작점검

본 연구에서는 독립변수의 조작이 적절하게 수행되었는지를 확인하기 위해 조작점검을 표 2와 같이 실시하였다. 조작점검은 환각 가능성 공개 조건과 인터페이스 유형 조건에 대해 각각 수행되었다.

먼저 환각 가능성 공개 조작의 적절성을 검증하기 위해 참여자들에게 "AI 시스템이 정보 생성 과정에서 오류가 발생할 수 있다는 안내를 제공했는지"에 대한 인식 정도를 측정하였다. 분석 결과 환각 가능성 공개 조건(M = 5.68, SD = 0.84)이 비공개 조건(M = 3.12, SD = 1.01)보다 유의미하게 높게 나타났으며, 이러한 차이는 통계적으로 유의미하였다 ($t(128) = 16.42$, $p < .001$). 이는 참여자들이 환각 가능성 공개 여부를 명확하게 인식했음을 의미한다.

다음으로 인터페이스 유형 조작의 적절성을 검증하기 위해 사회적 존재감(social presence) 및 의인화 인식 정도를 측정하였다. 분석 결과 버추얼 휴먼 조건(M = 5.41, SD = 0.92)이 텍스트 챗봇 인터페이스 조건(M = 4.02, SD = 1.03)보다 유의미하게 높게 나타났으며, 이러한 차이는 통계적으로 유의미하였다 ($t(128) = 8.17$, $p < .001$). 이는 버추얼 휴먼 인터페이스가 텍스트 챗봇 인터페이스보다 더 높은 사회적 존재감과 의인화 인식을 유도했음을 의미한다.

표 2. 조작점검 결과
Table 2. Manipulation check results

Variable	Condition	Mean	SD	t	p
Disclosure Perception	Disclosure	5.68	0.84	16.42	< .001
	No Disclosure	3.12	1.01		
Social Presence	Virtual Human	5.41	0.92	8.17	< .001
	Text Chatbot	4.02	1.03		

따라서 본 연구에서 설정한 두 독립변수의 조작은 모두 적절하게 이루어진 것으로 확인되었다.

6-4 환각 가능성 공개 효과 검증

환각 가능성 공개가 사용자 인식에 미치는 영향을 검증하기 위해 분산분석(ANOVA)을 실시하였다. 분석 결과, 환각 가능성 공개 조건에서 참여자들은 AI 시스템을 더 투명하게 인식하는 것으로 나타났다. 구체적으로, 환각 가능성 공개 조건의 지각된 투명성 평균은 5.42, 비공개 조건은 4.63으로 나타났으며, 이러한 차이는 통계적으로 유의미하였다($F(1,128) = 14.37$, $p < .001$, $\eta^2 = .10$). 이는 AI가 자신의 오류 가능성을 사전에 공개할 경우 사용자가 해당 시스템을 보다 투명한 기술로 인식함을 의미한다. 따라서 가설 H1은 지지되었다.

반면, 환각 가능성 공개는 AI 시스템의 능력 평가에는 부정적인 영향을 미치는 것으로 나타났다. 환각 가능성 공개 조건의 지각된 능력 평균은 4.87, 비공개 조건은 5.21로 나타났으며, 이러한 차이는 통계적으로 유의미하였다($F(1,128) =$

6.12, p = .015, η^2 = .05). 이는 AI가 자신의 불확실성을 명시적으로 표현할 경우 사용자가 해당 시스템의 능력을 낮게 평가할 수 있음을 시사한다. 따라서 가설 H2는 지지되었다.

이러한 결과는 AI의 정보 공개가 지각된 투명성과 지각된 능력에 상반된 영향을 미친다는 점을 보여주며, 투명성 역설(transparency paradox)의 가능성을 경험적으로 뒷받침하는 결과로 해석될 수 있다. 관련 분석 결과는 표 3에 제시하였다.

표 3. 환각 가능성 공개에 대한 분산분석 결과
Table 3. ANOVA results for hallucination disclosure

Variable	Disclosure Mean	No Disclosure Mean	F	p	η^2
Perceived Transparency	5.42	4.63	14.37	< .001	.10
Perceived Competence	4.87	5.21	6.12	.015	.05

6-5 매개 및 조절 효과 분석

환각 가능성 공개가 신뢰 형성에 미치는 과정에서 지각된 투명성과 지각된 능력의 매개 역할을 검증하기 위해 회귀분석과 PROCESS macro(Hayes, Model 7)를 활용한 분석을 실시하였다. 회귀분석 결과를 살펴보면, 전체 모형은 통계적으로 유의미한 설명력을 가지는 것으로 나타났으며(R2 = .48, Adj. R2 = .46), 지각된 능력에는 유의미한 부(-)의 영향을 미치는 것으로 나타났다(β = -.28, p = .015). 또한 지각된 투명성과 지각된 능력은 각각 신뢰에 유의미한 영향을 미치는 것으로 확인되었다(β = .41, p < .001; β = .36, p < .001). 또한 다중공선성 검증 결과 모든 변수의 VIF 값은 2.5 이하로 나타나 다중공선성 문제는 없는 것으로 확인되었다. 이러한 경로 분석 결과는 표 4에 제시하였다.

한편 환각 가능성 공개가 신뢰에 미치는 직접 효과는 통계적으로 유의미하지 않은 것으로 나타났다(β = .09, p = .18). 이는 환각 가능성 공개가 신뢰에 직접적으로 영향을 미치기보다는 지각된 투명성과 지각된 능력을 통해 간접적으로 영향을 미칠 가능성을 시사한다. 따라서 가설 H5는 지지되지 않았다.

이후 매개 효과를 보다 엄밀하게 검증하기 위해 PROCESS macro를 활용한 부트스트랩 분석(5,000 resamples)을 실시하였다. 분석 결과, 환각 가능성 공개가 신뢰에 미치는 간접효과는 모두 통계적으로 유의미한 것으로 나타났다. 구체적으로 지각된 투명성을 통한 간접효과는 β = .17 (SE = .05), 95% CI [.09, .27]로 나타났으며, 지각된 능력을 통한 간접효과는 β = -.10 (SE = .04), 95% CI [-.19, -.02]로 나타났다. 이는 두 간접효과의 신뢰구간이 모두 0을 포함하지 않아 통계적으로 유의미함을 의미한다. 이러한 결과는 환각 가능성 공개가 신뢰에 미치는 영향이 단일 방향이 아닌, 투명성 경로를 통한 긍정적 효과와 능력 경로를 통한 부정적 효과가 동시에 작용하는 매개 구조임을 보여준다. 관련 결과는 표 5에 제시하였다.

또한 인터페이스 유형의 조절 효과를 검증하기 위해 상호작용 항을 포함한 회귀분석을 실시한 결과, 환각 가능성 공개와 인터페이스 유형 간의 상호작용 효과가 유의미하게 나타났다(β = .21, p = .03). 단순기울기(simple slope) 분석 결과, 버추얼 휴먼 조건에서는 환각 가능성 공개가 지각된 투명성에 유의미한 영향을 미치는 것으로 나타났으며(β = .53, p < .001), 텍스트 챗봇 인터페이스 조건에서는 상대적으로 약한 효과가 나타났다(β = .29, p < .05). 이에 따라 조건부 간접효과를 분석한 결과, 버추얼 휴먼 조건에서의 간접효과는 β = .23 (SE = .06), 95% CI [.12, .36]로 나타났으며, 텍스트 챗봇 인터페이스 조건에서는 β = .12 (SE = .04), 95% CI [.04, .21]로 나타났다. 이는 환각 가능성 공개의 효과가 버추얼 휴먼 인터페이스에서 더 강하게 나타남을 의미한다.

마지막으로 조절된 매개의 지표(index of moderated mediation)를 검증한 결과, 해당 값은 β = .08 (SE = .03), 95% CI [.02, .16]로 나타났으며, 신뢰구간이 0을 포함하지 않아 통계적으로 유의미한 것으로 확인되었다. 이는 인터페이스 유형이 환각 가능성 공개와 지각된 투명성 간의 관계를 조절할 뿐 아니라, 이러한 조절이 신뢰에 대한 간접효과에도 영향을 미친다는 것을 의미한다.

종합적으로 본 연구 결과는 환각 가능성 공개가 신뢰에 미치는 영향이 직접적인 효과가 아니라 지각된 투명성과 지각된 능력을 통한 간접 효과를 중심으로 작용하며, 이러한 과정이 인터페이스 유형에 의해 조절되는 조절된 매개 구조로 해석될 수 있는 경향을 보여준다.

표 4. 매개 및 조절 효과에 대한 회귀분석 결과
Table 4. Regression results for mediation and moderation

Path	β	SE	p
Disclosure → Transparency	.42	.09	< .001
Disclosure → Competence	−.28	.11	.015
Transparency → Trust	.41	.07	< .001
Competence → Trust	.36	.08	< .001
Disclosure → Trust	.09	.07	.18
Agent Type × Disclosure → Transparency	.21	.10	.03

표 5. 매개효과 및 조절된 매개효과의 부트스트랩 분석 결과
Table 5. Bootstrap results for mediation and moderated mediation effects

Effect	β	SE	95% CI
Indirect (Transparency)	.17	.05	[.09, .27]
Indirect (Competence)	−.10	.04	[−.19, −.02]
Conditional Indirect (Virtual Human)	.23	.06	[.12, .36]
Conditional Indirect (Text Chatbot)	.12	.04	[.04, .21]
Index of Moderated Mediation	.08	.03	[.02, .16]

6-6 조절 효과 시각화

인터페이스 유형의 조절 효과를 보다 직관적으로 확인하기 위해 상호작용 효과를 시각화한 결과를 그림 2에 제시하였다. 앞선 회귀분석에서 환각 가능성 공개와 인터페이스 유형 간 상호작용 효과가 유의미하게 나타났으며(β = .21, p = .03), 이를 바탕으로 조건별 평균값을 시각적으로 비교하였다.

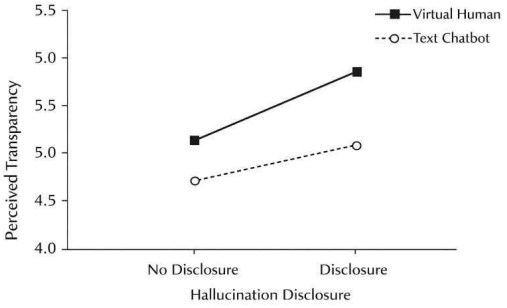


그림 2. 인터페이스 유형이 지각된 투명성에 미치는 조절 효과
Fig. 2. Moderating effect of agent type on perceived transparency

그림 2에서 확인할 수 있듯이, 환각 가능성 공개 조건에서 지각된 투명성은 전반적으로 증가하는 경향을 보였다. 그러나 이러한 증가 효과는 인터페이스 유형에 따라 차이를 보였다. 구체적으로, 버추얼 휴먼 인터페이스 조건에서는 환각 가능성 공개에 따른 지각된 투명성 증가 폭이 상대적으로 크게 나타났지만, 텍스트 챗봇 인터페이스 조건에서는 그 증가 폭이 비교적 완만하게 나타났다.

이러한 결과는 동일한 정보 공개라도 인터페이스 유형에 따라 사용자 인식이 다르게 형성될 수 있음을 보여준다. 특히 버추얼 휴먼과 같은 의인화된 인터페이스에서는 AI의 한계나 오류 가능성에 대한 설명이 단순한 기술적 정보가 아니라 정직성과 책임성을 나타내는 사회적 신호로 해석될 가능성이 크다. 반면 텍스트 기반 인터페이스에서는 동일한 정보가 더 기능적이고 기술적인 설명으로 인식될 수 있다.

이와 같은 결과는 의인화 이론과 Media Equation 관점에서 설명될 수 있다. 인간과 유사한 인터페이스는 사용자가 AI를 사회적 행위자로 인식하도록 만들며, 이로 인해 정보의 해석 방식이 변화하게 된다. 따라서 버추얼 휴먼 인터페이스는 환각 공개가 가지는 긍정적 효과를 강화하는 요인으로 작용할 수 있다. 즉 본 연구 결과는 AI 커뮤니케이션에서 정보의 내용뿐 아니라 전달 방식, 즉 인터페이스 특성이 사용자 인식 형성 과정에서 중요한 역할을 한다는 점을 시사한다.

Ⅶ. 논 의

본 연구 결과는 생성형 AI 환경에서 환각 가능성 공개가 사용자 신뢰 형성에 미치는 영향이 단순한 직접 효과가 아니라, 서로 다른 인지적 평가 경로를 통해 작용하는 복합적 메커니즘임을 보여준다. 특히 환각 가능성 공개는 지각된 투명성을 증가시키는 동시에 지각된 능력을 감소시키는 상반된 효과를 동시에 유발하며, 이러한 구조는 AI 신뢰 형성이 단일 요인이 아닌 다차원적 인지 평가 과정에 의해 형성된다는 점을 시사한다.

이러한 결과는 기존 연구에서 제시된 투명성의 긍정적 역할을 확장하면서도, 투명성이 항상 신뢰 증가로 이어지지 않는다는 점을 보여준다. 즉, AI 시스템이 자신의 불확실성을 명시적으로 공개하는 경우, 사용자는 이를 정직성과 책임성의 신호로 해석할 수 있으나 동시에 시스템의 능력에 대한 평가를 낮출 수도 있다. 이러한 상반된 효과는 투명성 역설(transparency paradox)의 관점에서 이해될 수 있으며, 자동화 시스템 신뢰 연구에서 제시된 능력 인식의 중요성과도 일치한다[11],[17].

또한 본 연구는 지각된 투명성과 지각된 능력이 모두 AI에 대한 신뢰 형성에 유의미한 영향을 미친다는 점을 확인하였다. 이는 AI 신뢰가 단일 요인에 의해 형성되는 것이 아니라, 이해 가능성(투명성)과 성능 평가(능력)라는 복합적 인지 과정에 의해 형성됨을 보여준다. 특히 이러한 결과는 AI 시스템에 대한 신뢰가 단순히 기술적 정확성에 의해 결정되는 것이 아니라, 사용자가 해당 시스템을 어떻게 해석하고 평가하는지에 의해 형성된다는 점을 강조한다[11],[13].

한편, 환각 가능성 공개가 신뢰에 미치는 직접 효과는 통계적으로 유의미하지 않은 것으로 나타났다. 그러나 이러한 결과는 환각 가능성 공개가 신뢰에 영향을 미치지 않는다는 것을 의미하기보다는, 지각된 투명성과 지각된 능력을 통해 작용하는 상반된 간접 경로가 서로 상쇄된 결과로 해석하는 것이 타당하다. 즉, 환각 가능성 공개는 지각된 투명성을 증가시켜 신뢰를 높이는 긍정적 경로와 동시에 지각된 능력을 감소시켜 신뢰를 낮추는 부정적 경로를 함께 유발하며, 이러한 상반된 효과가 동시에 작용함으로써 전체적인 직접 효과는 유의미하지 않게 나타난 것으로 볼 수 있다.

이러한 결과는 AI 커뮤니케이션 전략이 단순히 신뢰를 증가시키거나 감소시키는 일방향적 효과가 아니라, 서로 다른 인지적 평가 경로를 통해 복합적으로 작용하는 구조임을 시사한다. 특히 본 연구는 투명성 역설이 개념적 논의에 그치지 않고 실제 사용자 인식 과정에서 상반된 효과가 동시에 작용하는 메커니즘으로 나타날 수 있음을 실증적으로 보여주었다는 점에서 의의를 갖는다.

마지막으로, 본 연구는 인터페이스 유형의 조절 효과를 확인하였다. 분석 결과 버추얼 휴먼 인터페이스 조건에서 환각 가능성 공개가 지각된 투명성에 미치는 영향이 텍스트 챗봇 인터페이스 조건보다 더 강하게 나타났다. 이는 인간과 유사한 외형과 행동을 갖는 인터페이스가 AI의 설명을 단순한 기술적 정보가 아니라 사회적 커뮤니케이션 행위로 해석하도록

유도할 수 있음을 의미한다. 이러한 결과는 인간이 기술 시스템을 사회적 존재로 인식한다는 Media Equation 관점[7]과 의인화 이론[18]과도 일치한다. 즉, 의인화된 인터페이스는 동일한 정보라도 사용자에게 더 높은 신뢰성과 정직성을 전달하는 매개적 역할을 수행할 수 있다.

본 연구에서 버추얼 휴먼 조건은 텍스트 챗봇과 비교하여 의인화 수준뿐 아니라 매체 형식(영상 vs. 텍스트), 정보 제시 방식, 사회적 존재감 등 복합적인 요소가 동시에 차별화되어 제시되었다. 따라서 본 연구에서 관찰된 효과를 순수한 인터페이스 유형의 영향으로 단정하기보다는, 이러한 요소들이 결합된 결과로 해석할 필요가 있다.

종합적으로, 본 연구는 생성형 AI 환경에서 정보의 내용뿐 아니라 정보가 어떻게 전달되는가가 사용자 인식과 신뢰 형성에 중요한 영향을 미친다는 점을 보여준다. 특히 환각 가능성 공개는 지각된 투명성과 지각된 능력이라는 상반된 경로를 통해 작용하며, 이러한 결과는 AI 신뢰 형성이 단일 요인이 아닌 다차원적 인지 평가 구조에 의해 형성됨을 실증적으로 보여준다. 나아가 본 연구는 투명성 역설(transparency paradox)이 생성형 AI 맥락에서 어떻게 나타나는지를 구체적으로 설명함으로써, AI 커뮤니케이션 연구의 이론적 확장에 기여한다.

Ⅷ. 결 론

본 연구는 생성형 AI의 환각 가능성 공개가 사용자 인식과 신뢰 형성 과정에 미치는 영향을 실험 연구를 통해 분석하였다. 연구 결과, 환각 가능성 공개는 사용자가 AI 시스템을 보다 투명하게 인식하도록 만드는 긍정적인 효과를 가지는 동시에, 시스템의 능력 평가에는 부정적인 영향을 미칠 수 있는 양면적 효과가 나타나는 경향을 확인하였다. 또한 지각된 투명성과 지각된 능력은 모두 AI에 대한 신뢰 형성에 중요한 영향을 미치는 요인으로 확인되었으며, 인터페이스 유형은 환각 가능성 공개와 사용자 인식 간의 관계를 조절하는 변수로 작용하는 것으로 나타났다.

본 연구의 가장 중요한 이론적 기여는 생성형 AI의 정보 공개 전략을 투명성 역설 관점에서 설명하였다는 점이다. 기존 연구에서는 AI의 투명성이 신뢰를 강화하는 긍정적 요인으로 주로 논의되어 왔으나, 본 연구는 AI가 자신의 한계를 공개하는 행위가 투명성 인식을 높이는 동시에 능력 평가를 저하시킬 수 있음을 실증적으로 보여주었다. 이는 AI 신뢰 형성이 단일 경로가 아닌 상반된 인지 평가 과정을 통해 형성되는 복합적 메커니즘임을 제시한다는 점에서 중요한 의미를 갖는다. 또한 본 연구는 AI 커뮤니케이션을 단순한 정보 전달이 아닌 사용자 인식 형성을 유도하는 전략적 과정으로 재정의함으로써 인간–AI 상호작용 연구를 확장하였다.

실무적 측면에서 본 연구는 AI 시스템 설계와 커뮤니케이션 전략에 중요한 시사점을 제공한다. 생성형 AI 서비스에서 오류 가능성을 공개하는 전략은 사용자에게 정직성과 책임성을 전달하는 효과적인 수단이 될 수 있으나, 동시에 능력에 대한 신뢰를 저하시킬 위험도 내포하고 있다. 따라서 AI 시스템 설계자는 단순히 투명성을 강화하는 것을 넘어, 투명성과 능력 인식 간의 균형을 고려한 커뮤니케이션 전략을 설계할 필요가 있다. 특히 본 연구 결과는 버추얼 휴먼과 같은 의인화된 인터페이스가 이러한 효과를 강화할 수 있음을 보여주며, 인터페이스 디자인이 정보 해석 과정에 중요한 영향을 미친다는 점을 시사한다.

그럼에도 불구하고 본 연구는 몇 가지 한계를 지닌다. 본 연구 결과는 인터페이스 효과뿐 아니라 매체 형식과 사회적 존재감이 결합된 영향으로 해석될 필요가 있다. 첫째, 본 연구는 스마트홈 기술 설명이라는 비교적 저위험 정보 맥락을 기반으로 수행되었기 때문에, 의료, 법률, 금융과 같은 고위험 환경으로 일반화하는 데에는 제한이 있다. 둘째, 본 연구에서는 버추얼 휴먼과 텍스트 챗봇 인터페이스라는 두 가지 인터페이스 유형만을 고려하였으므로, 향후 연구에서는 음성 기반 인터페이스, 몰입형 환경, 또는 멀티모달 상호작용 등 다양한 인터페이스 유형을 포함할 필요가 있다.

향후 연구에서는 생성형 AI가 실제 서비스 환경에서 사용자와 상호작용하는 다양한 상황을 고려하여 AI의 투명성 전략과 사용자 신뢰 형성 과정 간의 관계를 보다 심층적으로 분석할 필요가 있다. 특히 AI 시스템의 설명 방식, 인터페이스 디자인, 그리고 사용자 경험 요소가 상호작용하는 메커니즘을 통합적으로 분석하는 연구가 이루어진다면 인간–AI 상호작용 연구에 더 중요한 이론적·실무적 기여를 할 수 있을 것이다.

참고문헌

[1] S. S. Sundar, “Rise of Machine Agency: A Framework for Studying the Psychology of Human-AI Interaction,” *Journal of Computer-Mediated Communication,* Vol. 25, No. 1, pp. 74-88, January 2020. https://doi.org/10.1093/jcmc/zmz026

[2] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, ... and B. Zoph, “GPT-4 Technical Report,” arXiv:2303.08774, March 2023. https://doi.org/10.48550/arXiv.2303.08774

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, ... and D. Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS),* New Orleans, LA, pp. 24824-24837, 2022. https://doi.org/10.48550/arXiv.2201.11903

[4] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE*

Access, Vol. 6, pp. 52138-52160, September 2018. https://doi.org/10.1109/ACCESS.2018.2870052

[5] B. Shneiderman, “Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy,” *International Journal of Human-Computer Interaction,* Vol. 36, No. 6, pp. 495-504, 2020. https://doi.org/10.1080/10447318.2020.1741118

[6] J. N. Bailenson, *Experience on Demand: What Virtual Reality Is, How It Works, and What It Can Do*, New York, NY: W. W. Norton & Company, 2018.

[7] C. Nass and Y. Moon, “Machines and Mindlessness: Social Responses to Computers,” *Journal of Social Issues*, Vol. 56, No. 1, pp. 81-103, 2000. https://doi.org/10.1111/0022-4537.00153

[8] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, ... and T. Liu, “A Survey on Hallucination in Large Language Models,” arXiv:2311.05232, November 2023. https://doi.org/10.48550/arXiv.2311.05232

[9] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, ... and P. Fung, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, Vol. 55, No. 12, pp. 1-38, 2023. https://doi.org/10.1145/3571730

[10] Y. Sun, D. Sheng, Z. Zhou, and Y. Wu, “AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content,” *Humanities and Social Sciences Communications*, Vol. 11, 1278, 2024. https://doi.org/10.1057/s41599-024-03811-x

[11] D. Kim, “A Study on Users’ Trust Perception of Generative Artificial Intelligence (AI),” *The Journal of the Korea Contents Association*, Vol. 25, No. 4, pp. 38-48, April 2025. https://doi.org/10.5392/JKCA.2025.25.04.038

[12] J. D. Lee and K. A. See, “Trust in Automation: Designing for Appropriate Reliance,” *Human Factors*, Vol. 46, No. 1, pp. 50-80, 2004.

[13] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. de Visser, and R. Parasuraman, “A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction,” *Human Factors*, Vol. 53, No. 5, pp. 517-527, 2011. https://doi.org/10.1177/0018720811417254

[14] A. Aryania, S. Chockalingam, H. K. Rødseth, and G. Alenya, “Impact of Design Transparency on Trust and Data Sharing During Human-Robot Interactions in Public Places,” *ACM Transactions on Human-Robot Interaction*, Vol. 15, No. 2, 49, pp. 1-22, 2026. https://doi.org/10.1145/3785152

[15] A. Waytz, J. Heafner, and N. Epley, “The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle,” *Journal of Experimental Social Psychology*, Vol. 52, pp. 113-117, May 2014. https://doi.org/10.1016/j.jesp.2014.01.005

[16] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, “I Trust It, But I Don’t Know Why: Effects of Implicit Attitudes toward Automation on Trust in an Automated System,” *Human Factors*, Vol. 55, No. 3, pp. 520-534, 2013. https://doi.org/10.1177/0018720812465081

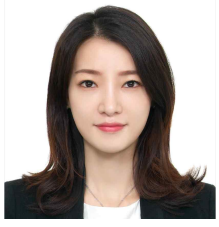
[17] M. Mori, K. F. MacDorman, and N. Kageki, “The Uncanny Valley,” *IEEE Robotics & Automation Magazine*, Vol. 19, No. 2, pp. 98-100, June 2012. https://doi.org/10.1109/MRA.2012.2192811

[18] A. Waytz, J. Cacioppo, and N. Epley, “Who sees Human? The Stability and Importance of Individual Differences in Anthropomorphism,” *Perspectives on Psychological Science*, Vol. 5, No. 3, pp. 219-232, May 2010. https://doi.org/10.1177/1745691610369336

[19] S. Jang, D. Han, and C. Oh, “Identifying the Ethical Issues of Virtual Human: A Semantic Network Analysis of Media Reports,” *Journal of Digital Contents Society*, Vol. 23, No. 11, pp. 2307-2316, November 2022. https://doi.org/10.9728/dcs.2022.23.11.2307

[20] M. Kim and C. Gu, “A Study on Behavioral Intention Toward ChatGPT-Based Conversational AI Tourism Search Services: The Roles of Cognitive Trust and Affective Trust,” *Information Systems Review*, Vol. 26, No. 1, pp. 119-149, 2024. https://doi.org/10.14329/isr.2024.26.1.119

[21] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.



강지영(Jiyoung Kang)

2004년 : Pratt Institute 컴퓨터 그래픽스 (학사)

2006년 : New York University, 인터랙티브 텔레커뮤니케이션 (석사)

2013년 : 한국과학기술원(공학박사-인터랙션 디자인)

2022년~현 재: 이화여자대학교 커뮤니케이션·미디어학부 교수

※관심분야 : 가상현실(VR), 증강현실(AR), 인터랙션 디자인 등