한국심리학회지: 일반
Korean Journal of Psychology: General
2026, Vol. 45, No. 3, 253-276
http://dx.doi.org/10.22257/kjp.2026.6.45.3.253

OPEN ACCESS

AI 출처표지가 얼굴 정서 평정에 미치는 영향: 내집단 편향과 AI 의인화의 조절 효과*

태 진 이　　최 원 일†

광주과학기술원 인문사회과학부

생성형 인공지능(Generative artificial intelligence)의 발전으로 실제와 구분하기 어려운 딥페이크 자극이 증가하면서, 컴퓨터로 생성된 얼굴에 대한 사람들의 인식이 중요한 연구 주제로 부상하고 있다. 본 연구는 동일한 얼굴에 대한 정서적 반응이 출처표지(인간 vs AI), 얼굴의 인종(내집단 vs 외집단), 그리고 개인의 AI 태도에 따라 어떻게 달라지는지를 검증하였다. 한국인 대학생 50명이 실험에 참여하였으며, 참가자들은 '인간' 또는 'AI'라는 출처표지를 확인한 후 한국인(내집단) 또는 백인(외집단)의 긍정(웃음), 부정(분노), 중립 표정 사진을 보고 정서를 평정하였다. 또한 참가자의 특성을 측정하기 위해 AI에 대한 일반적 태도와 AI 의인화 정도를 묻는 설문에 응답하였다. 실제 자극은 모두 실제 배우가 촬영한 사진이었으나, 출처표지만 조작하여 제시하였다. 분석 결과, 정서 유형에 따른 주효과가 관찰되어 긍정 표정이 중립 표정에 비해 더 긍정적으로, 부정 표정은 중립 표정에 비해 더 부정적으로 평가되었으며, 내집단 얼굴은 외집단에 비해, 인간 표지 조건은 AI 표지 조건에 비해 더 긍정적으로 평가되었다. 흥미롭게도, 출처표지와 정서의 상호작용 효과는 내집단 얼굴에서만 나타났다. 구체적으로 내집단의 긍정 표정에서 인간으로 표지된 사진이 AI로 표지된 사진보다 더 긍정적으로 평가되었다. 이러한 출처표지 효과는 AI 의인화 경향이 높은 참가자에게서는 더 적게 나타났다. 본 연구는 제시된 사진의 출처표지의 효과가 집단 정체성이나 AI에 대한 태도에 따라 다를 수 있음을 보여줌으로써, AI에 대한 인식이 정서 처리에 미치는 영향을 통합적으로 이해하고자 했다는 점에서 의의가 있다.

주요어 : 딥페이크, 정서 지각, 출처표지 효과, 타인종 효과, 개인차 연구

* 본 논문은 대한민국 교육부(NRF-2020S1A3A2A02103899) 지원에 의해 수행되었습니다.

† 교신저자: 최원일, 광주과학기술원 인문사회과학부, 광주광역시 북구 첨단과기로 123
Email: wichoi@gist.ac.kr

Copyright © 2026, The Korean Psychological Association. This is an open-access article distributed under the terms of the Creative Commons Attribution –NonCommercial Licenses(https://creativecommons.org/licenses by-nc/4.0) which permits unrestricted non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

생성형 인공지능(Generative artificial intelligence)의 기하급수적인 발전으로 컴퓨터가 생성한 글, 사진, 영상 등 다양한 정보들이 우리의 삶에 들어왔다. 과거에는 실제 사람 사진과 컴퓨터로 생성된 사진을 구분하는 일이 어렵지 않았다. 왜냐하면 인간은 기본적으로 다른 자극과 달리 사람 얼굴에 예민한 특성이 있을 뿐만 아니라(Birmingham & Kingstone, 2009; Haan et al., 2002; Vuilleumier et al., 2001), 컴퓨터로 생성된 얼굴 자극들이 실제 사람 사진에 비해 표정이나 이목구비가 부자연스러운 특성을 보였기 때문이다. 하지만 최근 인공지능 기술의 비약적 발전에 힘입어 각종 소프트웨어를 통해 실제와 가까운 얼굴 사진을 만들 수 있게 됨에 따라 사람이 등장해야 할 광고나 영상 콘텐츠에 컴퓨터로 생성된 인간과 매우 유사한 비인간이 등장하는 사례가 점차 늘어나고 있다. 그러다보니 과거에는 실제 사람 안내원이 민원인을 만나 응대했다면, 최근에는 사람들이 컴퓨터 화면에 나오는 AI 상담원을 보고 자신의 고민이나 민원 사항을 전달하거나, 실제 면접관이 아니라 AI 면접관이 구직자와 면접하는 기업의 사례도 늘어나는 추세이다. 이러한 시대적 흐름에 발맞춰 최근 심리학 연구에서는 컴퓨터로 생성된 자극을 처리하는 인지적 처리 기제를 이해하기 위한 경험적 연구도 증가하고 있다(Eiserbeck et al., 2023; Pearson et al., 2026).

컴퓨터로 생성된 얼굴에 대한 사람들의 인식을 이해하기 위해서는 먼저 이러한 얼굴이 실제 사람 얼굴과 유사하게 처리되는지를 확인할 필요가 있다. 만약 컴퓨터로 생성된 얼굴이 지각적 수준에서부터 실제 얼굴과 다르게 처리된다면, 이 얼굴에 대한 정서적 평가나 사회적 판단 또한 달라질 수 있기 때문이다. 이러한 맥락에서 최근의 얼굴 지각 연구에서는 참가자가 컴퓨터로 생성된 얼굴을 볼 때, 실제 사람의 얼굴을 볼 때와 유사한 정보처리가 일어나는지 확인하려는 방편으로 얼굴처리 과정에서 독특하게 나타나는 현상들을 활용하고 있다.

사람들이 다른 자극과 구별되게 얼굴 사진을 처리한다는 것을 보여주는 대표적인 현상으로는 얼굴역전효과(face inversion effect; Yin, 1969)가 있다. 얼굴역전효과란 사람들이 다른 사물과 비교해서 뒤집어진 얼굴 사진을 인식하는데 더 큰 어려움을 보이는 현상을 의미한다. 컴퓨터로 생성된 얼굴 사진이 실제 얼굴과 유사하게 처리되는지를 살펴보고자 했던 연구들은 컴퓨터로 생성된 얼굴 사진에서도 얼굴역전효과가 유의미하게 나타나는지를 알아보았다(Balas & Pacella, 2015; Carlson et al., 2012; Kätsyri, 2018; Matheson & McMullen, 2011; Matheson et al., 2012). 일부 연구는 컴퓨터로 생성된 얼굴도 실제 얼굴과 마찬가지로 얼굴역전효과를 보인다는 결과를 보고했지만(Balas & Pacella, 2015; Kätsyri, 2018; Matheson & McMullen, 2011; Matheson et al., 2012), 다른 연구에서는 컴퓨터로 생성된 얼굴에서 얼굴역전효과가 나타나지 않는다는 결과(Carlson et al., 2012)를 보고하기도 하였다. 상반된 결과에 대해 메타 분석을 통해 두 얼굴의 역전효과 크기를 비교한 Miller 등(2023)의 연구는 사람들이 컴퓨터로 생성된 얼굴을 봤을 때도 실제 얼굴과 마찬가지로 얼굴역전효과를 보이기는 하지만, 효과크기가 상대적으로 낮을 뿐이라고 설명하였다.

컴퓨터로 생성된 얼굴에서 관찰되는 얼굴역전효과는 사람들이 컴퓨터로 생성된 얼굴을 지각적 처리 수준에서 실제 얼굴과 유사한 방

식으로 처리함을 시사한다. 그러나 사람의 얼굴은 단순히 그 사람의 정체성을 나타내는 것을 넘어 그 사람의 심리적 상태를 표현하는 기능도 가지고 있다. 사람들은 상대의 표정을 통해 상대의 내면 상태를 추론하고, 이를 활용하여 어떤 행동을 취할지 결정하게 된다. 그렇다면 컴퓨터로 생성된 얼굴이 지각적 수준에서 실제 얼굴과 유사하게 처리된다는 사실이 정서적 반응에서도 동일한 양상을 보인다는 것을 의미하는가? 다시 말해 사람들은 컴퓨터로 생성된 얼굴이 나타내는 표정에 대해서도 실제 사람 얼굴 표정과 유사한 정서적 반응을 보이는 것일까?

컴퓨터로 생성된 얼굴 사진과 실제 사람의 사진에서 표현된 정서 처리의 차이를 살펴본 연구들은 대체로 정서명명과제를 사용했다. 정서명명과제란 참가자들이 제시되는 사진을 보고, 그 사진이 표현하고 있는 정서의 명칭을 고르는 과제이다. 얼굴 출처에 따른 정서명명과제 수행의 차이를 살펴본 연구(Afzal et al., 2009; Baltrušaitis et al., 2010; Costantini et al., 2004; 2005; Fabri et al., 2004; Kätsyri et al., 2003; Moser et al., 2007)에 따르면 컴퓨터로 생성된 얼굴은 실제 배우 사진과 비교해서 피부의 주름(Miller et al., 2022), 동공의 변화(Demos et al., 2008), 그리고 얼굴색의 변화(Peromaa & Olkkonen, 2019)와 같이 정서 변화 탐지에 단서가 될 수 있는 정보가 부족하기 때문에 사람들이 실제 사람이 표현하는 정서를 탐지할 때보다 컴퓨터로 생성된 얼굴에 표현된 정서를 탐지하기 어려워한다고 설명한다.

그러나 정서명명과제의 결과만을 가지고는 사람들이 컴퓨터로 생성된 정서 자극을 통해 경험하는 정서 강도를 파악하기는 어렵다. 그 이유는 정서명명과제는 참가자가 제시된 자극에서 표현하고 있는 정서를 정확하게 맞췄는지에만 초점을 맞추고 있기 때문이다. 그러나 정서평정과제는 참가자들이 제시된 사진에 표현된 정서를 맞추기보다, 사진을 보고 상대가 표현하고 있는 정서를 어느 정도의 강도로 느끼는지를 측정하기 위해 목적을 가지고 있다. 그러다보니 해당 과제의 참가자들은 제시된 사진을 보고 사진을 봤을 때 자신이 느꼈던 정서가(긍정 혹은 부정) 혹은 각성가(차분 혹은 흥분)의 정도를 평가하라는 지시를 받게 된다. 따라서 평정과제를 사용하게 되면 컴퓨터로 생성된 얼굴 자극에 대한 사용자들의 정서 경험을 측정할 수 있다는 장점을 가지게 되는 것이다.

하지만 컴퓨터로 생성된 얼굴과 실제 얼굴에 대한 사람들의 정서 평정을 비교한 선행연구들은 일관되지 않은 결과를 보고하고 있다. 제시되는 사진을 보고 사진에서 느껴지는 정서의 강도를 평정하게 한 일부 연구는 사람들이 실제 배우의 사진보다 컴퓨터로 생성된 얼굴을 봤을 때 더 약하게 정서를 경험했다고 지각한다는 결과를 보고했으나(Kegel et al., 2020; Mühlberger et al., 2009), 다른 연구는 반대의 결과를 보고하기도 했다(Kätsyri et al., 2020). 비록 기존 연구들이 컴퓨터가 생성한 얼굴과 실제 배우가 촬영한 사진이라는 물리적 속성이 지나치게 다른 자극을 사용했기 때문에 두 자극에 대한 정서적 판단이 달라졌을 가능성도 있다. 하지만, 무엇보다 먼저 고려해야 하는 부분은 참가자들이 실험 자극으로 제시된 컴퓨터로 생성된 얼굴에 대해 가지는 인지적 판단이다. 즉, 어떤 대상에 대한 정서적 평정은 단순히 그 대상에 대한 지각적인 특성 외에도 이에 대한 기대와 신념과 같은 상위 인지적 요소에 영향을 받을 수 있다. 그러므

로 사람들이 자신이 보는 대상에 대해 부정적 인식이 있다면, 그 대상이 전달하는 메시지의 특성과 무관하게 그 대상에 대한 인식이 반영되어 결과적으로 그 메시지를 부정적으로 해석할 수 있을 것이다.

이러한 맥락에서 Eiserbeck 등(2023)은 동일한 시각 자극일지라도 출처에 대한 표지가 정서 평정에 어떠한 영향을 미치는지 실증적으로 검증하였다. 이들은 얼굴 사진을 보고 느껴지는 정서를 평가하는 상황에서 특정한 정서 표정을 짓고 있는 배우의 얼굴이 실제 사람인지 아니면 AI로 생성된 얼굴인지에 따라 사람들의 평정 반응이 달라지는지를 확인하기 위해 연구를 진행하였다. 참가자들은 긍정(행복), 부정(분노), 그리고 중립 얼굴 사진을 실험 자극으로 보게 되고, 사진에서 느껴지는 정서를 1점(매우 부정적)에서 101점(매우 긍정적) 사이로 평가하였다. 이 연구에서 가장 중요한 부분은 참가자가 평가해야 할 사진이 나오기 전에 사진의 표지가 제시되었다는 점이었다. 실험에 사용된 자극들은 모두 전문 배우의 얼굴을 촬영한 사진들이었지만, 이 연구에서는 실제 사람의 얼굴과 컴퓨터로 생성된 얼굴에 대한 참가자들의 인식 차이를 살펴보고자 자극의 출처 변인을 조작하여 ‘실제(Real)’ 혹은 ‘거짓(Fake)’이라는 표지를 제시하였다. 실제 조건은 다음에 보이는 사진이 실제 사람의 얼굴이라고 알려준 조건이고, 거짓 조건은 컴퓨터로 생성된 얼굴이라고 가르쳐준 조건이었다. 평정 과제의 반응을 분석한 결과를 보면 표지 조건에 따른 효과가 관찰되었다. 즉, 사람들은 사진에서 표현하고 있는 정서 유형과 관계없이 ‘거짓’ 표지로 제시된 사진을 ‘실제’ 표지 조건에 비해 더 부정적으로 인식하는 경향이 있었다. 출처 표지의 효과는 사진에서 표현하고 있는 정서의 종류에 따라 다르게 나타나기도 했다. 부정적인 정서를 표현하고 있는 사진을 보는 상황에서 사람들은 출처 표지와 관계없이 일관되게 평정하는 경향이 관찰되었던 반면에, 긍정적인 정서 사진을 보는 상황에서는 거짓 표지 조건의 얼굴 사진을 실제 표지 조건의 얼굴보다 덜 긍정적으로 평가하였다. 이러한 결과는 동일한 얼굴 자극이라도 그 얼굴의 출처가 어디인지에 대한 인식이 정서 평정에 영향을 줄 수 있음을 보여준다.

Eiserbeck 등(2023)은 출처표지가 정서 평정에 영향을 미치는 이유는 사람들이 평정 시, 자극의 물리적 특성뿐만 아니라 출처와 같은 상위 인지 정보를 함께 고려하기 때문이라고 설명한다(Bar, 2004). 그런데 이때 실제 참가자가 처리해야 하는 정보의 특성 외에 다른 인지 정보의 영향은 정서 유형에 따라 달라질 수 있다고 연구자는 설명한다. 부정 정서, 특히 분노는 잠재적 위협 신호로서 맥락과 무관하게 우선 처리되므로 출처표지의 영향을 덜 받는다. 그러나 긍정 정서는 표정에서 느껴지는 진정성 평가가 중요하게 여겨지는 정서이다 보니, 부정 정서에 비해 민감하게 반응하게 된다. 다시 말해, 화난 얼굴은 진짜 사람의 얼굴이든 AI가 만든 얼굴이든 위협적이지만, 웃는 얼굴은 진짜인지 가짜인지가 정서 평가에 중요한 영향을 미치게 되는 것이다. 이러한 발견은 동일한 물리적 자극이라도 출처 인식이 정서 평정에 영향을 미치며, 특히 긍정 정서의 경우 표현된 정서의 진정성이 핵심적 역할을 함을 시사한다.

그러나 Eiserbeck 등(2023)의 연구만으로 컴퓨터로 생성된 얼굴에 대한 사람들의 인식이 정서 평정에 미치는 영향을 완전히 이해하기

에는 한계점이 존재한다. 이 연구에서는 제시된 사진이 실제 사람을 찍은 것인지 아니면 컴퓨터로 생성된 것인지를 구별하기 위해서 실제(Real)와 거짓(Fake)이라는 표지를 사용하였다. 사람들이 사진의 출처를 나타내는 용어인 '거짓(Fake)'를 보았을 때, 단순히 이것이 실제 사람이 아닌, 컴퓨터로 생성된 얼굴임을 나타내기 위한 목적으로 쓰인 용어라고 여기고 사진에서 표현하는 정서의 강도를 평가했을 수 있다. 하지만 'Fake'라는 단어와 연합되어 온 부정적인 정서(Xue et al., 2024)가 사진의 정서 판단에 영향을 줬을 가능성도 존재한다. 그러므로 이 연구에서 관찰된 결과가 주어진 얼굴이 컴퓨터가 생성한 것이라는 인식 때문인지, 아니면 컴퓨터로 생성된 얼굴을 지칭하는 용어인 'Fake'가 나타내는 부정적 정서의 영향 때문인지 구분하기 어려울 수 있다. 따라서 본 연구에서는 단순히 사진의 출처를 나타내는 명칭(인간 vs AI)을 활용하여 컴퓨터로 생성된 얼굴에 대한 사람들의 인식이 해당 자극의 정서 평정에 미치는 영향을 재검증하고자 한다.

한편, Eiserbeck 등(2023)을 포함한 기존 연구들은 주로 실험 참가자들과 인종이 같은 내집단 얼굴 사진을 실험 자극으로 사용하였다. 그러나 얼굴 지각 연구에 따르면 사람들은 내집단 얼굴과 외집단 얼굴을 다르게 처리한다는 것이 잘 알려져 있다. 구체적으로, 사람들은 자신과 다른 인종의 얼굴(외집단)을 자신과 같은 인종의 얼굴(내집단)에 비해 더 부정확하게 재인하는 특성을 보이는데, 이러한 현상을 타인종 효과(other race effect; ORE)라고 부른다(McKone et al., 2019; Meissner & Brigham, 2001). 인종에 따른 얼굴 지각의 차이는 얼굴에 표현된 정서에 대한 지각에도 영향을 미친다(Craig et al., 2012; Hugenberg, 2005). 예를 들어, Hugenberg(2005)의 연구에서는 컴퓨터로 생성된 백인과 흑인의 다양한 표정 자극을 사용하여 사진에 등장하는 배우의 인종에 따른 정서 탐지 패턴의 차이를 살펴보았다. 실험 결과, 백인 참가자들은 자신과 같은 인종인 백인이 웃고 있는 사진을 빠르게 탐지하는 패턴을 보였고, 자신과 다른 인종인 흑인의 사진에 대해서는 부정적인 정서 사진을 빠르게 탐지하는 패턴이 관찰되었다. 이에 대해 저자는 내집단에 대한 긍정적인 평가와 외집단에 대한 부정적인 평가가 특정 인종 사진에 대한 정서 평가에 다른 영향을 주게 된 것이라 설명한다. 다시 말해, 사람들이 내집단에 대해 긍정적인 평가를 하게 되면, 긍정 정서가 활성화되고, 자연스럽게 내집단 구성원의 웃는 얼굴을 빠르게 탐지하게 되고, 외집단에 대한 부정적인 평가와 이를 통한 부정 정서의 활성화는 외집단 구성원의 부정적인 정서 표현을 빠르게 탐지하게 만드는 것이다. 만약 사람들이 내집단과 외집단에 대해 서로 다른 정서적 평가를 내리고 있다면, Eiserbeck 등이 보고한 출처표지 효과도 제시된 얼굴의 인종에 따라 달라질 가능성이 있다.

안타깝게도 출처표지 효과가 제시되는 배우의 인종적 특성과 얼굴에 표현된 정서 유형에 따라 어떻게 달라지는지를 직접 검증한 선행 연구는 존재하지 않는다. 따라서 본 연구는 탐색적 접근을 통해 이러한 삼원상호작용의 가능성을 검증하고자 한다. 그럼에도 불구하고 기존 이론들은 세 변인이 독립적으로 작용하기보다는 상호작용할 가능성을 시사한다. 첫째, 얼굴 처리 연구에 따르면 사람들은 내집단 얼굴에 대해서는 세밀하고 개별화된 처리를 하는 반면, 외집단 얼굴에 대해서는 범

주적이고 피상적인 처리를 하는 경향이 있다(Hugenberg et al., 2010). 이러한 결과는 출처표지와 같은 맥락 정보가 정서 평정에 미치는 영향을 조절할 수 있음을 시사한다. 예를 들어, 내집단 얼굴을 처리할 때는 출처표지를 포함한 다양한 맥락 정보가 통합되어 정서 평정에 반영될 수 있지만, 외집단 얼굴을 처리할 때는 표정이라는 핵심 특징에만 주목하여 출처표지의 영향이 약화될 수 있다. 둘째, Eiserbeck 등(2023)의 연구에서 출처표지 효과가 긍정 표정에서 두드러지게 나타났고, Hugenberg(2005)의 연구에서도 내집단에 대한 긍정 편향이 특히 긍정 표정 처리에서 관찰되었으므로 출처표지와 인종의 상호작용이 표현 정서의 유형에 따라 다시 조절될 가능성을 시사한다. 따라서 본 연구는 이러한 이론적 가능성을 바탕으로 세 변인 간의 삼원상호작용을 검증하고자 한다.

더 나아가 인공지능 기술에 기반하여 생성된 얼굴에 대한 정서 평정은 AI에 대해 개인이 가지고 있는 태도에 따라 달라질 가능성이 있다. 예를 들어, AI에 대해 긍정적인 태도를 가진 사람들은 AI가 제공하는 정보를 더 잘 수용하는 경향을 보일 수 있다. 반면, AI가 인간을 위협하는 대상이라는 인식을 가진 사람들은 AI가 제공하는 정보를 잘 받아들이지 않거나 AI로 생성된 자극을 부정적으로 평가할 수 있다. 최근 Pearson 등(2026)의 연구에서는 참가자들에게 화면에 제시되는 사진이 실제 사진인지 아니면 AI가 생성한 사진인지를 판단하도록 하였다. 이때 참가자들의 판단을 돕기 위해 AI 혹은 사람이 제공한 조언이 사진 옆에 제시되었고, 참가자들은 AI에 대한 태도를 측정하는 설문(General Attitude toward Artificial Intelligence Scale; GAAIS; Schepman & Rodway, 2020)에도 응답하였다. 분석 결과, AI에 대해 긍정적인 태도를 가진 사람일수록 실제 사람과 AI로 생성된 사진을 구분하는 정확도가 감소했으며, 이러한 패턴은 AI의 조언이 제공된 조건에서만 유의미하게 관찰되었다. 이러한 결과는 AI에 대한 태도가 AI에 기반하여 생성된 자극에 대한 처리 과정에 영향을 미칠 수 있음을 시사한다. 따라서 본 연구에서는 참가자들의 AI에 대한 태도를 측정하여, 출처표지 효과가 개인의 AI 태도에 따라 어떻게 조절되는지를 확인하고자 한다.

그러나 AI에 대한 태도는 AI에 대한 유용성과 신뢰성을 바탕으로 한 태도뿐만 아니라, AI를 의도와 감정을 가진 인간과 유사한 사회적 행위자로 지각하는 의인화 경향에 따라서도 달라질 수 있다. 선행 연구에 따르면 AI를 인간과 유사하게 지각하는 사람일수록 자율주행 자동차에 대한 신뢰도가 높고(Waytz et al., 2014), AI와의 상호작용에서 즐거움을 느끼며(Kim et al., 2019), AI와의 심리적 거리를 가깝게 느끼는 경향이 있었다(Li & Sung, 2021). 이러한 설명을 고려할 때, 정서 평정 과제와 같이 자극에 대한 즉각적 정서 반응을 측정하는 과제의 경우, 참가자가 AI에 대해 가지고 있는 일반적인 태도뿐만 아니라 AI를 얼마나 사람과 유사한 대상으로 인식하는지가 반응에 영향을 미칠 수 있다. 예를 들어, AI를 사회적 행위자로 지각하는 사람들은 AI로 생성된 얼굴도 사회적 의미를 가진 대상으로 생각할 수 있으며, 이 경우 출처표지의 영향이 약화될 수 있다. 따라서 본 연구에서는 AI에 대한 일반적 태도와 AI 의인화 경향을 모두 측정하여, 어떤 개인차 변인이 출처표지 효과를 조절하는 핵심 요인인지를 탐색하고자 한다.

종합하면, 컴퓨터로 생성된 얼굴에 대한 정

서 평정은 단일 요인이 아닌 여러 요인의 복합적 상호작용에 의해 결정될 가능성이 높다. 기존 연구들은 주로 출처표지(Eiserbeck et al., 2023), 얼굴 인종(Hugenberg, 2005), 또는 개인의 AI 태도를 각각 독립적으로 검증하였으나, 실제 정서 평정 상황에서는 이러한 요인들이 동시에 작용하며 서로 영향을 주고받을 수 있다. 예를 들어, 출처표지 효과는 내집단 얼굴에서는 나타날 수 있지만 외집단 얼굴에서는 나타나지 않을 수 있으며, 이러한 패턴이 개인의 AI 태도에 따라 다시 조절될 수 있다. 따라서 세 변인(출처 정보, 인종, 정서가)의 삼원상호작용을 통합적으로 검증하고, 개인차 변인의 조절 효과를 함께 고려하는 것이 컴퓨터로 생성된 얼굴에 대한 정서 반응의 심도 깊은 이해를 가능하게 할 것이다.

이상의 논의를 바탕으로 본 연구는 세 가지 연구 목적을 설정하였다. 첫째, 얼굴 출처를 나타내는 중립적 명칭(인간 vs AI)을 사용하여 Eiserbeck 등(2023)의 결과를 재검증하고자 한다. 만약 기계에 의해 생성된 얼굴 사진에 대한 부정적인 사회적 통념이 표지와 무관하게 존재한다면, AI로 표기된 긍정적인 얼굴을 상대적으로 덜 긍정적으로 평가하는 양상이 관찰될 것으로 예상된다. 둘째, 기존 연구가 내집단 얼굴만을 사용한 것과 달리, 본 연구에서는 외집단 얼굴도 포함함으로써 출처표지 효과가 얼굴의 인종에 따라 달라지는지를 탐색적으로 살펴보고자 한다. 이를 통해 내집단과 외집단 얼굴에 대한 서로 다른 정서적 평가가 출처효과와 어떻게 상호작용하는지 확인할 수 있을 것이다. 마지막으로, AI에 대한 개인의 태도가 출처표지 효과를 조절하는지를 확인하고자 한다. 개인이 AI에 대해 가지고 있는 태도는 AI로 생성된 자극에 대한 평가에 영향을 미칠 수 있으므로, 이러한 개인차 변인을 함께 고려하는 것은 컴퓨터로 생성된 얼굴 자극에 대한 개인의 심리적 반응을 심층적으로 이해하는 데 필수적이다. 결론적으로 본 연구는 출처 정보(인간 vs AI), 인종(내집단 vs 외집단), 정서가(긍정, 부정, 중립)의 삼원상호작용을 탐색적으로 검증하고, 개인의 AI 태도가 이러한 효과를 어떻게 조절하는지 분석함으로써, 컴퓨터로 생성된 얼굴에 대한 정서 평정이 다양한 요인에 의해 달라질 수 있음을 실험적으로 검증하고자 한다. 특히 이전 연구들에서 이와 같은 삼원상호작용 효과를 종합적으로 살펴보지 않았다는 점을 볼 때, 본 연구는 출처표지에 따른 다양한 종류의 얼굴 자극에 대한 정서 평정 과정을 통합적으로 이해할 수 있는 실험적 증거를 제공한다는 점에서 의의가 있다.

방 법

연구 대상

본 연구는 광주과학기술원에 재학 중인 학생들을 대상으로 진행되었다. 참가자들은 실험에 참여하기 전 기관에서 승인받은 연구 동의서(20230919-HR-73-03-02)를 읽었고, 자발적으로 연구 참여 의사를 밝힌 학생들만이 이후 연구에 참여할 수 있었다[1]. 표본 크기는 선행

1) 출처표지의 순수한 효과를 검증하기 위해 자극의 실제 출처에 대한 완전한 정보를 사전에 제공하지 않았으며, 참가자 간 정보 공유 방지를 위해 즉각적인 사후설명을 실시하지 않았다. 참가자들은 동의서를 통해 연구에 대한 설명을 언제든 요청할 수 있었으나, 해당 요청은 없었다.

연구를 바탕으로 한 사전 검정력 분석을 통해 결정하였다. Eiserbeck 등(2023)에서 보고된 출처표지와 정서의 상호작용 효과로부터 계산된 효과크기(Cohen's f = 0.15)를 바탕으로 G*Power 3.1(Faul et al., 2007)을 사용하여 분석하였다. 본 연구에서는 선형 혼합 효과 모형을 사용하였으나, G*Power 3.1에서는 해당 모형에 대한 검정력 계산을 직접 제공하지 않으므로, 동일한 독립변인 효과를 분석할 수 있는 반복측정 ANOVA 모형을 사용하여 근사적으로 계산하였다. 분석 결과, α = .05, power = .80 기준으로 최소 41명의 참가자가 필요한 것으로 나타났다. 따라서 본 실험에서는 불성실한 참가자의 반응으로 인해 일부 참가자의 데이터가 제거될 수 있을 것을 고려하여 총 50명(평균:21.5세, 표준편차: 6.1)의 참가자를 모집하였다. 전체 참가자 중 남성 참가자의 수는 26명이었고, 여성 참가자의 수는 24명이었다.

실험 도구

본 실험은 E-prime3(Psychology Software Tools, Pittsburgh)로 만들어졌고, 참가자들은 24인치 모니터를 통해 실험 자극을 보았으며, 키보드를 이용해서 자극에 대해 반응하라는 지시를 받았다.

실험 자극

실험 자극으로 정서를 표현하고 있는 얼굴 자극을 사용하였다. 특히 본 연구에서는 사진의 출처에 따른 정서 평가 양상이 배우의 인종에 따라 달라지는지를 살펴보고자 했으므로, 서로 다른 두 인종의 정서 얼굴 자극을 실험 자극으로 선정하였다. 첫 번째 인종은 실험 참가자들과 인종이 같은 한국인 사진으로 총 120명의 배우 사진이 실험 자극으로 사용되었다. 현재까지 배포된 정서 얼굴 사진 세트 중에서 120명의 서로 다른 배우의 사진을 포함하고 있는 세트는 없었기 때문에 세 가지 정서 얼굴 자극 세트인 YFACE(Chung et al., 2019), KOFEE(Lee et al., 2013), ChaeLee(Lee et al., 2013)에서 실험 자극을 선정하였다. 두 번째로 실험 자극으로 제시된 배우의 인종은 백인으로 백인의 정서 얼굴 사진은 기존 연구들에서 많이 사용되는 KDEF(Karolinska Directed Emotional Face, Lundqvist et al., 1998)와 FACE (Ebner et al., 2010)에서 선택되었고, 한국인과 마찬가지로 120명의 서로 다른 백인 얼굴 사진이 선택되었다. 이들 데이터베이스는 모두 각 정서 범주(긍정, 부정, 중립)에 대한 인식 타당도가 검증된 표준화된 자극 세트였다.

본 연구에서는 참가자들이 같은 자극을 반복해서 보지 않고, 다양한 자극을 한 번씩 볼 수 있도록 실험을 설계했기 때문에 각 데이터베이스에서 가능한 많은 자극을 포함하는 것을 원칙으로 하였다. 이에 따라 본 실험에서 사용된 얼굴 자극 세트 중 대부분의 사진이 실험 자극으로 사용되었으며, 최소 기준에 따라 자극의 수가 조정되었다. KOFEE, ChaeLee, KDEF 세트에서는 남성과 여성 배우의 성비를 맞추기 위해 기술적 품질(예: 조명, 해상도)이 상대적으로 낮은 일부 사진을 제외하였고, FACE에서는 모든 젊은 성인 배우의 사진을 사용하고, 중년 성인 배우 중 외관상 젊은 성인과 유사한 연령대로 보이는 9명의 사진을 추가로 사용하였다. 최종적으로 각 배우는 긍정(웃음), 부정(분노), 중립 표정을 짓고 있는 3장의 사진으로 제시되었으며, 총 720장의 사

진(한국인 360장, 백인 360장)이 실험 자극으로 사용되었다.

설문 도구

AI 자극에 대한 개인의 인식은 일반적으로 인공지능이라는 새로운 기술에 대한 태도에 따라 달라질 가능성이 있다. 이에 따라서 본 연구에서는 인공지능에 대해 참가자가 가지고 있는 태도를 측정할 수 있는 설문조사를 두 가지 진행했다. 첫 번째 척도는 인공지능에 대한 일반적인 태도를 측정하는 척도인 ATTARI-12이다. ATTARI-12 척도는 Stein 등(2024)이 개발한 척도로 12개의 문항(예: AI는 세상을 더 나은 곳으로 만들 것이라 생각한다)으로 구성되어 있으며 인공지능에 대한 인지적, 정서적, 행동적 태도를 측정하기 위해 고안되었다. 전체 점수가 높을수록 인공지능에 대해 긍정적인 태도를 가지고 있다고 해석하고, 전체 점수가 낮을수록 인공지능에 대해 부정적인 태도를 가지고 있다고 해석된다. 원저자가 보고한 내적 합치도는 .93이었고, 본 연구의 내적 합치도는 .858였다.

두 번째는 개인이 인공지능을 의인화하는 정도를 측정하기 위해 Pelau 등(2021)의 연구에서 사용한 문항 중 AI에 대한 의인화 정도를 측정하는 데 사용된 19문항(예: AI는 고유의 선호도와 기분을 가지고 있다)을 선택하였다. Palau 등(2021)이 보고한 내적 신뢰도는 0.96이었고, 본 연구의 내적 신뢰도는 .875이었다. ATTARI-12와 마찬가지로 총점을 통해 개인의 태도를 유추하는데, 전체 점수가 높을수록 응답자가 인공지능이 인간과 유사한 속성을 가진다고 생각한다고 볼 수 있다.

실험 절차

실험이 시작되면 화면 중앙에 고정점(+)이 450ms 동안 나타난다. 고정점이 사라진 자리에는 뒤이어 등장할 사진의 출처를 나타내는 표지(예: 사람 혹은 AI)가 500ms 동안 제시된다. 단어가 사라진 후 100ms 동안의 짧은 공백이 있다가, 연이어 참가자가 사진에서 느껴지는 정서를 판단해야 할 사진이 나타난다. 참가자들은 사진을 보고 자신이 느끼는 정서를 1점에서 9점까지로 평가하여 키보드 오른쪽 숫자 버튼을 통해 응답하라는 지시를 받게 된다. 1점에 가까울수록 부정적인 정서를 느

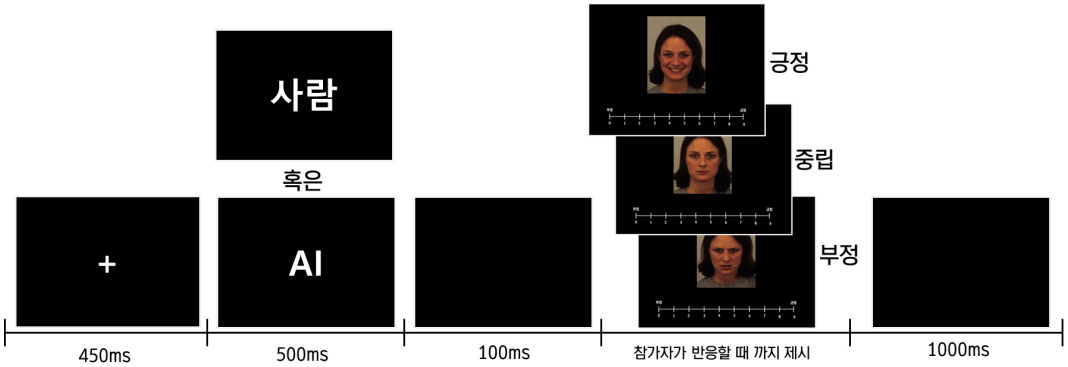


그림 1. 정서 평정 실험 절차

꼈음을 의미하고, 긍정적인 정서를 느끼면 9점에 가까운 점수를 매겨야 했다. 사진에 대한 반응이 끝난 후에는 검은 화면이 1000ms 간 제시되는 것으로 한 시행이 구성된다(그림 1 참조).

본 실험에는 240명의 배우의 긍정・중립・부정 사진이 모두 포함되어서 총 720장의 사진이 사용된다. 실험 참가자들은 240장의 사진을 보게 되는데, 이 과정에서 개별 참가자가 한 배우가 표현하는 하나의 정서 사진만을 보도록 연구를 설계했기 때문에 역균형화 절차를 통해 3개의 리스트를 제작하였고, 개별 사진은 두 개의 출처 표지 조건에 제시되어야 하므로 최종적으로 총 6개의 리스트가 제작되었다. 참가자들은 6개 리스트 중 하나에 무선적으로 할당되었다. 본 평정 실험을 완료한 참가자들은 인공지능에 대한 태도를 묻는 두 가지 설문에 응답하고 실험이 종료되었다. 실험 소요 시간은 대략 20분 내외였다.

분석 방법

본 연구는 각 참가자가 240개의 서로 다른 얼굴 사진을 평가하는 반복 측정 설계로 진행되었다. 이에 참가자와 자극 변수를 무선 효과(random effects)로 모델링할 수 있는 선형 혼합 효과 모형(linear mixed effects model; LMM)을 사용하여 자료를 분석하였다. 선형 혼합 효과 모형은 전통적인 반복측정 분산분석과 달리 개별 참가자 및 개별 자극의 변산성을 동시에 통제할 수 있는 장점이 있다(Baayen et al., 2008). 따라서 분석은 R 프로그램(version 4.4.1, R Core Team, 2024)의 lme4 패키지(Bates et al., 2015)를 사용하여 수행되었다. 고정 변인으로는 사진에 표현된 정서가(긍정, 부정, 중립), 배우의 인종(한국인, 백인), 사진의 출처 정보(AI, 사람), 3개의 이원 상호작용 (정서가 * 인종, 정서가 * 사진의 출처, 인종 *출처 정보), 세 변인의 삼원 상호작용 변수가 포함되었다. 무선 절편으로는 참가자와 사진 자극의 고유번호를 포함하였다. 무선 효과 모형은 최대 무선 효과 모형부터 단계적으로 모형을 축소하는 방식을 통해 모형 수렴과 특이적합(singular fit)의 문제가 없는 모형을 찾았다. 결론적으로 최종 모형은 참가자 수준에서 정서와 인종의 상호작용을, 사진 자극 수준에서 정서를 무선 기울기로 선정되었다. 세 가지의 고정 변인인 정서가, 배우의 인종, 사진의 출처 정보는 범주형 변수로 요인화 시켰고, 세 가지 변인은 모두 더미 코딩(dummy coding)을 실시하여 분석에 투입하였다. 정서가는 3수준 변인(긍정, 부정, 중립)이므로 중립을 기준 집단으로 설정하고 2개의 더미 변수(긍정 vs 중립, 부정 vs 중립)를 생성하였다. 배우 인종과 사진 출처표지는 각각 2수준 변인으로 백인을 기준 집단으로 한 1개의 더미 변수(한국인 vs 백인)와 AI를 기준 집단으로 한 1개의 더미 변수(인간 vs AI)를 생성하였다. 따라서 모델에서 추정된 각 회귀 계수(Estimate)는 설정된 기준 집단 대비 각 조건의 상대적인 차이를 의미한다. 예를 들어, ‘정서: 긍정’의 회귀 계수는 중립 표정 대비 긍정 표정의 평정치 차이를 나타내며, ‘정서: 부정’의 회귀 계수는 중립 표정 대비 부정 표정의 평정치 차이를 나타낸다. 마지막으로 고정 변인의 효과에 대한 유의 확률은 lmerTest 패키지(Kuznetsova et al., 2017)를 이용하여 계산되었다. 삼원상호작용이 유의한 경우, 본 연구의 주요 목적인 출처표지, 인종, 정서의 상호작용 양상을 이해하기 위한 계획 비교로서 단순효과 분석을 수행하

였다. 구체적으로 내집단과 외집단 얼굴에서 각각 출처표지 × 정서 상호작용을 검증하였으며, 이원상호작용이 유의한 경우 각 정서 조건에서 출처표지 효과를 검증하였다. 또한 AI에 대한 일반적 태도 및 AI 의인화 경향이 출처표지 효과를 조절하는지 확인하기 위해 상관분석 및 회귀분석을 수행하였다.

결 과

출처 종류에 따른 인종과 정서에 따른 정서가 평정의 차이 비교

표 1은 사진 출처 종류와 배우의 인종과 정서에 따른 정서 평정가의 기술통계가 정리되어 있다. 참가자들이 사진을 보고 평정한 정서가 수치를 종속변인으로 하여 고정 변인의 효과를 확인한 분석 결과는 표 2에 제시되었다.

분석 결과를 보면, 기본적으로 참가자들은 배우의 정서에 맞게 사진에 대한 정서를 평가했는데 중립 얼굴에 비해 분노 표정은 더욱 부정적으로 인식하고, 웃고 있는 표정은 더 긍정적으로 인식하였다. 또한 참가자들은 사진에 등장하는 배우가 백인일 때와 비교하여 자신과 같은 한국인일 때 더 긍정적으로 평가했고, 등장하는 사진의 출처가 AI라고 했을 때보다 실제 사람이라고 명시되었을 때 이 사진을 더욱 긍정적인 정서를 담고 있다고 인식하였다. 상호작용 분석 결과에서는 중립 정서 대비 부정과 긍정 정서에 대한 평정 효과가 배우의 인종에 따라 다르게 나타났다.

정서와 인종 간의 상호작용을 상세히 검토하기 위해 각 정서 수준별로 인종에 따른 단순 주효과 분석을 진행하였다. 분석 결과, 중립 조건에서는 백인 배우에 비해 한국인 배우의 점수가 유의미하게 높게 나타났다 (β = 0.145, p = .002). 그러나 분노 조건(ϕ = .494)과 행복 조건(ϕ = .626)에서는 인종 간의 유의미한 차이가 관찰되지 않았다.

흥미롭게도 정서, 배우 인종, 사진 출처 간의 삼원 상호작용이 관찰되었다. 삼원 상호작용을 규명하기 위해 배우 인종별로 단순 이원 상호작용을 분석하였다. 분석 결과, 외집단인 백인 배우 사진을 본 조건에서는 정서와 사진 출처 간의 상호작용이 관찰되지 않았으나 (ϕ=.987), 내집단인 한국인 사진을 본 조건에서는 정서와 출처 간의 유의미한 상호작용이 관찰되었다(ϕ=.006). 추가적으로 한국인 조건 내에서 각 정서 수준별 출처표지 효과를 분석한 결과, 중립 조건에서는 사진 출처에 따른 차이가 유의미하지 않았다(β=-0.01, p=.815). 그러나 웃고 있는 사진이 제시된 조건에서는

표 1. 사진 출처, 배우의 인종, 그리고 배우의 정서에 따른 정서가 평균 및 표준편차

	사진 출처: 사람			사진 출처: AI		
	부정	중립	긍정	부정	중립	긍정
한국인	2.65 (1.55)	4.54 (1.03)	6.69 (1.42)	2.53 (1.54)	4.48 (0.98)	6.42 (1.65)
백인	2.7 (1.51)	4.44 (1.03)	6.58 (1.47)	2.62 (1.43)	4.34 (1.09)	6.47 (1.59)

표 2. 정서, 배우 인종, 사진 출처에 따른 정서가 평정의 선형 혼합 효과 분석 결과

		Estimate	*SE*	*t*	*p*
(절편)		4.306	0.077	55.995	<.001
정서: 분노		**-1.779**	**0.125**	**-14.261**	**<.001**
정서: 행복		**2.280**	**0.139**	**16.359**	**<.001**
배우 인종: 한국인		**0.193**	**0.057**	**3.361**	**<.001**
사진 출처: 사람		**0.106**	**0.043**	**2.446**	**.014**
정서:분노 * 배우 인종: 한국인		**-0.259**	**0.098**	**-2.647**	**.009**
정서:행복 * 배우 인종: 한국인		**-0.202**	**0.086**	**-2.355**	**.019**
정서:분노 * 사진 출처: 사람		0.010	0.061	0.155	.877
정서:행복 * 사진 출처: 사람		0.003	0.061	0.047	.962
배우 인종:한국인 * 사진 출처: 사람		-0.096	0.061	-1.564	.118
정서: 분노 * 배우 인종: 한국인 * 사진 출처: 사람		0.114	0.087	1.311	.190
정서: 행복 * 배우 인종: 한국인 * 사진 출처: 사람		**0.189**	**0.087**	**2.174**	**.030**
무선 효과		*Variance*	*SD*		
참가자	(절편)	0.232	0.481		
	정서:분노	0.554	0.744		
	정서:행복	0.799	0.894		
	배우 인종: 한국인	0.037	0.193		
	정서: 분노 * 배우 인종: 한국인	0.031	0.178		
	정서: 행복 * 배우 인종: 한국인	0.025	0.16		
사진의 고유번호	(절편)	0.041	0.203		
	정서: 분노	0.312	0.559		
	정서: 행복	0.187	0.432		

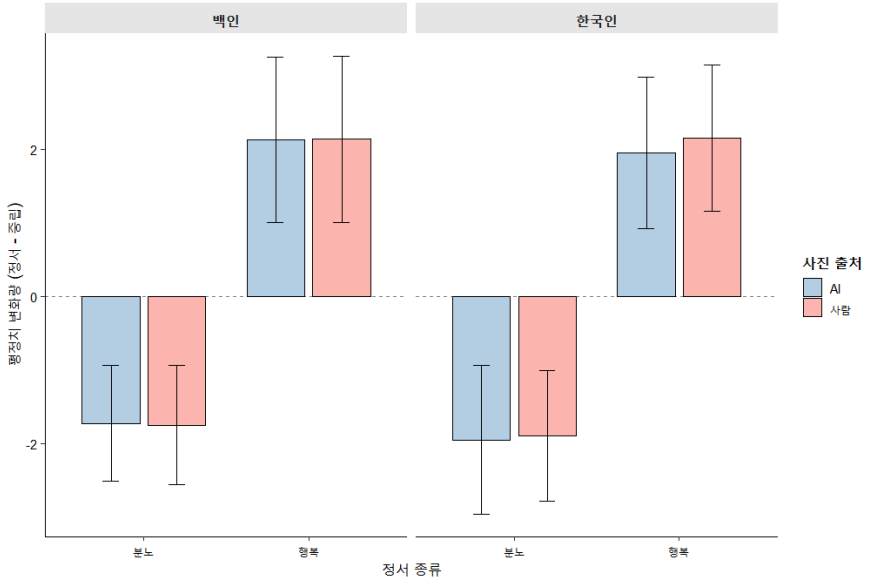


그림 2. 배우 인종과 사진 출처에 따른 정서 조건별 평정치 변화량 비교

AI 출처표지가 있던 조건에 비해 사람들이 실제 사람의 얼굴이라는 출처표지가 있었을 때, 사진을 더 긍정적으로 평정하였다(β=-0.201, p<.001).

인공지능에 관한 일반적 태도 및 의인화 경향과 긍정 내집단 얼굴의 출처표지 효과크기의 관계

본 분석에서는 한국인의 웃는 얼굴에서 관찰되었던 출처표지 효과, 즉 AI 표지 조건에 비해 인간 표지 조건에서 웃고 있는 얼굴을 더 긍정적으로 평가했다는 결과가 개인의 인공지능에 대해서 가지고 있던 태도나 의인화 정도에 따라 어떻게 달라지는지를 확인하고자 하였다. 이를 위해 '긍정 내집단 얼굴의 출처표지 효과크기'를 {(인간 표지 웃는 얼굴의 정서 평정치 - 인간 표지 중립 얼굴의 정서 평정치) - (AI 표지 웃는 얼굴의 정서 평정치 - AI 표지 중립 얼굴의 정서 평정치)}의 식을 통해 계산하였다. 그리고 이 출처표지 효과크기와 참가자들이 응답한 설문 척도 점수 간의 상관 분석을 진행하였다. 상관 분석 결과, 출처표지 효과크기는 인공지능에 대한 의인화 정도를 묻는 설문인 의인화 척도와 유의미한 부적 상관이 관찰되었다(표 3 참조). 다시 말해, 인공지능을 사람과 비슷한 대상으로 생각할수록, 인간과 AI 얼굴에서 느끼는 긍정 정서의 정도 차이가 줄어드는 경향이 있었다.

AI에 대한 태도를 가늠할 수 있는 두 개의 척도에 대한 점수와 출처표지 효과크기의 종합적으로 살펴보기 위해 출처표지 효과크기 점수를 종속변인으로, ATTARI-12 점수와 의인화 척도 점수를 독립변인으로 하는 다중회귀분석을 실시하였다(표 4 참조). 회귀계수의 안정성을 검증하기 위해 10,000회 부트스트래핑을 실시하여 95% 신뢰구간을 산출하였다. 분석 결과, 상관 분석과 마찬가지로 의인화 척도가 긍정 내집단 얼굴의 출처표지 효과크기와 유의미한 관계를 보여주었다(β =-.283, 95% CI [-0.016, -0.01], p < .05). 이는 의인화 척도 점수가 한 단위 증가하면 긍정 내집단 얼굴의 출처표지 효과크기가 0.283만큼 줄어든다는 것을 나타낸다(그림 3 참조). 상관 분석 결과와 마찬가지로 ATTARI-12점수는 출처표지 효과크기와 통계적으로 유의미한 관계를 보이지 않았다.

표 3. ATTARI-12, 의인화 척도, 긍정 내집단 얼굴의 출처표기 효과크기의 평균, 표준편차 및 상관계수

	1. ATTARI-12	2. 의인화척도	3. 긍정 내집단 얼굴의 출처표지 효과크기
M	3.12	4.21	1.53
SD	0.84	0.72	0.59
1	-		
2	-.011	-	
3	-.175	-.281*	-

*p < .01

표 4. 긍정 내집단 얼굴의 출처표지 효과크기와 AI 태도 설문 결과의 관계에 대한 다중회귀분석 결과

변수	비표준화 계수		표준화 계수	*t*	*p*
	B	*SE*	*β*		
(절편)	1.167	0.465		2.512	0.015
ATTARI-12	-0.008	0.006	-0.178	-1.297	0.201
의인화척도	-0.008	0.004	-0.283	-2.059	0.045

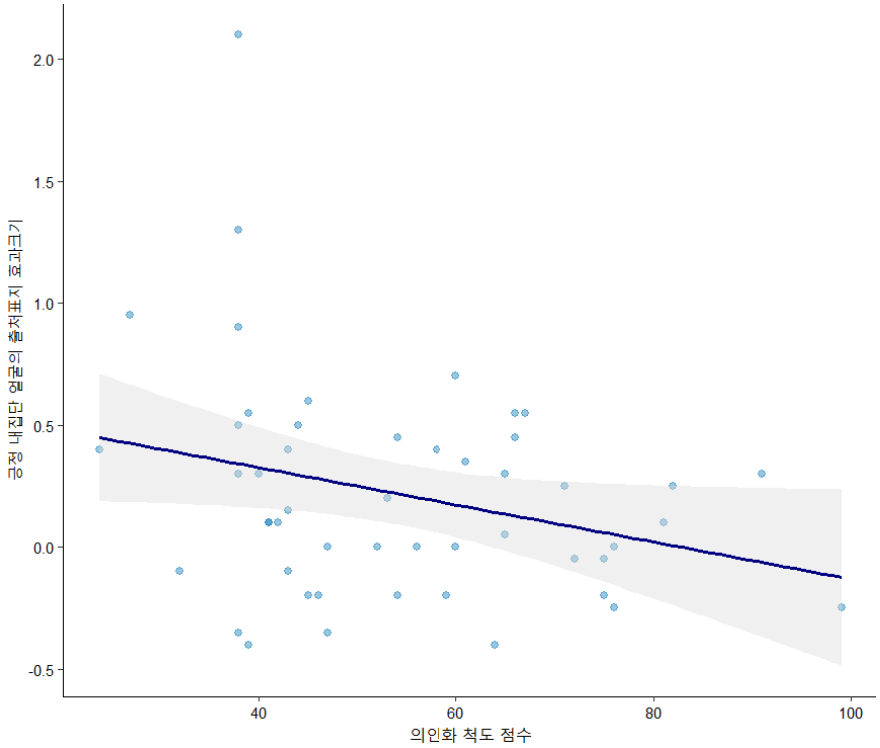


그림 3. 의인화 점수에 따른 긍정 내집단 얼굴의 출처표지 효과크기의 회귀 분석 결과

논 의

본 연구의 목적은 제시되는 사진의 출처표지(인간 vs AI)와 사진에 등장하는 배우의 인종에 따라 참가자의 정서 판단이 달라지는지, 그리고 이러한 차이가 개인이 AI에 대해 가진 태도에 따라 조절되는지를 확인하는 것이었다. 참가자들은 출처(AI 혹은 인간)를 확인한 후 다양한 정서 표정 사진을 보고, 사진에서 느껴지는 정서를 평가하는 과제를 수행하였다. 실험 결과, 사람들은 화난 표정을 가장 부정적으로 평가했고, 무표정에서는 중립적 반응을 보였으며, 웃는 표정에는 긍정적으로 반응했다. 또한 내집단 얼굴(한국인)이나 인간으로

출처가 표시된 사진이 외집단 얼굴(백인)이나 AI로 명시된 사진에 비해 더 긍정적으로 평가되었다. 상호작용 효과도 관찰되었는데, 참가자들은 백인의 무표정한 얼굴을 볼 때 한국인 얼굴보다 더 부정적으로 평가하는 특성을 보였다. 뿐만 아니라, 출처표지와 정서의 상호작용 효과는 사진에 등장하는 배우의 인종에 따라 달라졌다. 구체적으로, 외집단(백인) 사진에서는 출처표지와 정서의 상호작용이 관찰되지 않았지만, 내집단(한국인) 사진에서는 유의미한 상호작용이 나타났다. 즉, 한국인의 웃는 표정을 볼 때 AI 표지 사진이 사람 표지의 사진보다 더 부정적으로 평가되었다. 내집단의 긍정적 표정에 대한 출처표지 효과크기는 개별 참가자들이 AI를 인간과 유사한 대상으로 생각하는지를 측정하는 의인화 척도 점수와 관련이 있기도 했다. 보다 구체적으로, AI를 실제 사람처럼 느낀다고 보고한 사람(의인화 점수가 높은 사람)일수록 AI로 생성된 웃는 한국인 사진을 인간의 사진만큼 긍정적으로 평가했으며, 그 결과 출처표지조건 간 차이가 작아졌다.

Eiserbeck 등(2023)의 연구는 출처를 나타내는 표지로 '실제(Real)'와 '거짓(Fake)'을 사용했는데, '거짓(Fake)'라는 용어가 단순히 자극의 특성을 나타낼 뿐만 아니라 그 자체로 부정적인 함의를 지닐 수 있다는 한계가 있었다. 이러한 한계를 보완하기 위해 본 연구에서는 단순히 사진의 출처를 표현하는 보다 중립적인 용어인 '인간'과 'AI'를 출처표지로 사용하였다. 본 연구는 중립적 출처표지를 사용했음에도 불구하고 Eiserbeck 등(2023)과 유사한 결과를 관찰하였다. 이러한 결과는 단순히 출처표지 용어의 부정적 함의 때문이 아니라, 사람들이 AI 또는 컴퓨터로 생성된 자극 자체에 대해 부정적 인식이 있을 가능성을 시사한다. 또 한 가지 주목할 점은 출처표지 효과가 긍정 표정에서만 나타났다는 것이다. 서론에서 제시한 바와 같이, Eiserbeck 등(2023)은 긍정 정서가 진정성 평가에 민감하고, 부정 정서는 위협 탐지를 위해 맥락 독립적으로 처리된다는 정서가 의존적 처리 가설을 제시하였다. 본 연구에서도 기존 연구와 마찬가지로 분노와 중립 얼굴에서는 출처표지 효과가 나타나지 않고, 오직 긍정 표정에서만 유의미한 출처표지 효과가 관찰되었다. 이는 긍정 정서 평가 상황에서 표정을 짓는 대상의 진정성 여부가 핵심적 요소임을 재확인시켜주는 결과로 보여진다.

AI에 대한 사람들의 태도를 살펴본 기존 연구를 보면, 사람들이 AI를 긍정적인 대상으로 여기기보다 부정적인 대상으로 생각한다는 결과를 보고하고 있다(Grassini & Koivisto, 2024; Liang & Lee, 2017; Murtarelli et al., 2023; Persson et al., 2021; Schepman & Roday, 2020). 특히 2025년 스탠퍼드 대학교의 스탠퍼드 인간중심 인공지능연구소(Human-Centered Artificial Intelligence, HAI)의 연례보고서(AI Index 25, Maslej et al., 2025)를 보면 32개국에 거주 중인 23,685명을 대상으로 진행한 인터뷰에서 '인공지능을 활용한 제품과 서비스가 나를 불안하게 만든다' 라는 진술에 대해 대표적인 생성형 인공지능 프로그램인 챗지피티(ChatGPT)가 등장하기 전인 2022년에는 39%였던 데 반해 2024년에는 50%로 10% 이상 증가한 것으로 나타났다. 사람들이 AI에 대한 부정적 인식을 가지게 된 이유에 대해 연구자들은 AI로 인해 일자리가 줄어들고, 사람 간의 관계가 나빠질 수 있다는 염려 때문이라고 설명한다(Park & Yoon, 2025; Persson et al., 2021; Schepman &

Rodway, 2020).

그러나 AI에 대한 인식은 모든 개인에서 동일하게 나타나는 것은 아니었다. 본 연구에서는 AI에 대한 개인차 변인으로 일반적 태도와 의인화 경향을 함께 측정하였는데, 출처표지 효과는 의인화 경향에 의해서만 조절되었고 AI에 대한 일반적 태도와는 유의미한 관계를 보이지 않았다. 이처럼 AI에 대한 의인화 정도가 AI에 대한 인식에 영향을 미친다는 결과는 기존 연구 결과들과도 일치한다. 그렇다면 왜 참가자가 AI에 대해서 가지고 있는 일반적인 태도와 AI를 의인화하여 인식하고 있는 정도가 서로 다른 영향을 주고 있는 것일까?

ATTARI-12로 측정되는 일반적 AI 태도는 AI에 대한 명시적 평가를 반영하는 인지적·정서적 성향을 의미한다(Stein et al., 2024). 그러다 보니 해당 설문은 AI에 대해 참가자가 가지고 있는 의식적 판단에 관련된 의견을 묻는 문항들로 구성되어 있다. 반면, 의인화는 AI를 단순한 도구가 아닌 의도나 감정을 가진 행위자로 지각하는 경향을 의미하며(Epley et al., 2007), 사회적 지각과 관계 형성에 관여하는 인지 과정을 반영한다. 그 결과 주어진 자극이 표현하는 정서가 무엇인지를 판단하는 인지 반응을 요구하는 정서 평정 과제에서는 참가자가 가지고 있는 AI에 대한 일반적 태도보다는 이 대상에 대한 의인화 정도가 더 직접적인 영향을 준다고 볼 수 있다. 즉, AI 의인화 경향이 높은 사람들은 AI로 생성된 얼굴도 사람과 유사한 사회적 행위자의 표현으로 처리하므로, 출처표지('AI' 대 '인간')가 제공하는 범주적 차이를 덜 두드러지게 지각하게 된다. 반면 의인화 경향이 낮은 사람들은 AI 산출물을 명확히 비인간 범주로 분류하므로, 동일한 웃는 얼굴이라도 출처에 따라 질적으로 다르게 처리하는 것이다. 이와 달리 일반적 AI 태도는 이 대상에 대한 일반적인 태도를 반영하므로 특정 얼굴 자극에 대한 즉각적 정서 반응에는 직접적으로 영향을 미치지 못했을 가능성이 있다. 예를 들어, 한 참가자가 'AI는 장점보다 단점이 더 많다고 생각한다'는 이유로 AI에 부정적 태도를 가질 수 있지만, 이것이 AI로 만들어진 웃는 얼굴 사진을 평가하는 순간의 정서 반응까지 자동적으로 조절하기는 어렵기 때문이다. 따라서 본 연구 결과는 AI 생성 자극에 대한 정서적 반응이 AI에 대한 일반적인 태도보다는 사회적 행위자로 지각하는 정도에 따라 유의미하게 달라질 수 있음을 시사한다.

추가로 본 연구를 통해 발견한 흥미로운 결과는 출처표지 효과가 내집단 얼굴에서만 유의미하게 나타나고 외집단 얼굴에서는 관찰되지 않았다는 점이다. 이러한 패턴은 내집단과 외집단 정보의 처리 방식 차이로 설명할 수 있다. 선행 연구에 따르면, 사람들은 친숙한 내집단 구성원에 대해서는 정보를 세분화하고 구체적인 표상의 형태로 처리하는 반면, 친숙도가 낮은 외집단 구성원에 대해서는 포괄적이고 추상적인 범주적 표상을 형성하는 경향이 있고 한다(Ostrom et al., 1993). 세분화된 표상은 대상에 대한 풍부하고 상세한 정보를 포함하며, 이는 맥락 정보(예: 출처정보)를 함께 처리하고 통합하는 것을 가능하게 한다. 반면 추상적이고 범주적인 표상은 핵심적 특징(예: 얼굴의 전형적 특성, 표정)에만 초점을 맞추고 부가적인 맥락 정보는 덜 정교하게 처리하게 만든다. 이러한 관점에서 볼 때, 한국인 참가자들이 내집단인 한국인 얼굴을 볼 때는 세분화된 표상을 형성하며, 이 과정에서 출처표지 정보('AI' 또는 '인간')도 함께 부호화되고 정

서 판단 과정에 통합적으로 처리했을 가능성이 있다. 그에 따라 출처표지가 정서 평정에 유의미한 영향을 미치게 된 것으로 보인다. 반면, 외집단인 백인 얼굴에 대해서는 상대적으로 추상적이고 범주적인 처리가 이루어졌을 것이고, 출처표지와 같은 세부 정보는 덜 정교하게 처리되거나 평가 과정에서 충분히 고려되지 않았을 것이다. 그러다 보니 외집단 얼굴에서는 출처표지가 정서 판단에 미치는 영향이 미약하게 나타나게 된 것이다. 이러한 설명이 타당하다면, 참가자들은 정서 평정 과정에서 제시된 내집단 얼굴의 출처정보를 외집단 얼굴의 출처정보에 비해 더 정확하게 기억할 것으로 예측된다. 그러나 본 연구에서는 출처정보에 대한 기억 검사를 실시하지 않았으므로, 후속 연구에서는 내집단과 외집단에 따른 맥락정보 기억의 차이를 직접적으로 비교하여 본 해석을 검증할 필요가 있다.

본 연구에서 관찰된 내집단-외집단 차이 해석 시 한 가지 방법론적 한계를 고려해야 한다. 본 연구에서는 한국인 얼굴과 백인 얼굴이 서로 다른 데이터베이스에서 선별되었으므로(한국인: YFACE, KOFEE, ChaeLee; 백인: KDEF, FACE), 데이터베이스 간 체계적 차이(예: 촬영 환경, 조명, 표정 강도, 매력도)가 존재할 가능성이 있다. 비록 분석 시 자극의 차이가 가져올 수 있는 무선 효과를 포함함으로써 개별 자극 간 변산성이 가지는 영향을 통계적으로 통제하였다. 그러나 각 데이터베이스의 자극 간의 물리적 속성을 통제하는 절차는 거치지 않았으므로 추후 연구에서는 이런 지각적 속성도 철저히 통제한 연구를 수행할 필요성도 있다.

종합하면, 본 연구는 제시된 얼굴 자극에 대한 정서 반응이 단순히 자극의 물리적 특성만으로 결정되는 것이 아니라, 자극의 출처에 대한 인식, 자극 대상과 관찰자의 사회적 관계(내집단 vs 외집단), 그리고 개인의 AI 태도라는 다층적 요인들의 상호작용에 기반하여 형성됨을 보여주었다. 이는 앞으로 AI가 생성한 콘텐츠에 대한 인간의 정서적 반응을 이해하기 위해서는 자극 특성, 사회적 맥락, 개인차를 통합적으로 고려해야 함을 시사한다.

하지만 본 연구만을 가지고 AI라는 출처표지에 대한 사람들의 인식을 포괄적으로 이해하는 데에 몇 가지 한계점이 존재하는 것은 사실이다. 첫째, 본 연구는 Eiserbeck 등(2023)이 사용한 'Real' vs 'Fake' 표지가 가진 명시적 부정성을 보완하기 위해 '인간' vs 'AI' 표지를 사용하였으나, 'AI'라는 용어 자체의 정서가를 사전에 측정하지 않았다. 따라서 본 연구에서 관찰된 출처표지 효과가 'AI'라는 용어 자체가 가진 부정적 정서가 때문인지, 아니면 AI 또는 컴퓨터로 생성된 자극에 대한 일반적인 부정적 인식 때문인지를 명확히 구분하기 어렵다. 서론에서 제시한 바와 같이 'Fake'에 비해 'AI'가 더 중립적이라는 가정은 직관적으로 타당해 보이지만, 이를 뒷받침하는 경험적 증거는 제공하지 못했다는 한계가 존재한다. 후속 연구에서는 출처표지 용어(예: 'AI', '인간', '실제', '가짜' 등)에 대한 정서가를 사전 평정을 통해 측정하여, 용어 자체의 정서가를 통제한 상태에서 출처 인식의 순수한 효과를 검증해 보고자 한다.

둘째, 본 연구는 Eiserbeck 등(2023)과 마찬가지로 실제 배우가 촬영한 사진에 출처표지만 조작하여 실험을 진행하였다. 이러한 접근은 출처표지의 순수한 효과를 검증하기 위해 자극의 물리적 특성을 통제했다는 장점이 있으나, 실제 AI가 생성한 얼굴 자극을 사용했을

때도 동일한 패턴이 나타나는지는 확인하지 못했다는 한계가 있다. 실제 AI 생성 얼굴은 미묘한 부자연스러움(uncanny valley effect; Di Natale et al., 2023)이나 독특한 시각적 특징을 가질 수 있으며, 이러한 물리적 특성이 출처표지 효과와 상호작용할 가능성이 있다. 따라서 후속 연구에서는 실제 AI로 생성된 얼굴 사진을 사용하여 본 연구에서 관찰된 출처표지 효과가 재현되는지, 그리고 물리적 특성과 출처표지가 어떻게 상호작용하는지를 검증할 것을 제안한다.

셋째, 본 연구에서는 AI에 대한 일반적 태도와 AI 의인화 경향만을 측정하였으나, AI 생성 자극에 대한 반응에 영향을 미칠 수 있는 다른 개인차 변인들은 고려하지 못했다. 선행 연구에 따르면 개인의 성격 특성(예: 개방성, 신경증), 기술 수용성, 혁신성 등이 신기술에 대한 태도와 행동에 영향을 미치는 것으로 알려져 있다(Choung et al., 2022; Hering et al., 2025; Liang et al., 2025). 특히 개방성이 높은 사람은 새로운 기술에 대해 더 긍정적이고 수용적인 태도를 보이므로, AI 생성 자극에 대한 평가에서도 차이를 보일 가능성이 있다. 따라서 후속 연구에서는 성격 특성(예: Big Five, HEXACO)이나 기술 수용 관련 변인들을 추가로 측정하여, AI 생성 콘텐츠에 대한 정서적 반응을 예측하는 다층적 요인들을 체계적으로 규명할 필요가 있다.

마지막으로, 본 연구는 통제된 환경에서 촬영된 정적인 얼굴 표정 사진을 사용하였다. 그러나 일상생활에서 사람들이 접하는 정서 표현은 훨씬 더 역동적이고 맥락이 풍부하다. 최근 정서 연구들은 실험실에서 통제된 자극 외에도 자연스러운 상황에서 촬영된 사진이나 정서의 변화를 동적으로 확인할 수 있는 동영상 자극을 사용하여, 자극 특성에 따른 정서 지각의 차이를 살펴보고 있다(Chen et al., 2024; Kim et al., 2025; Krumhuber et al., 2017). 특히 동영상 자극은 표정의 변화 과정, 움직임의 자연스러움 등 추가적인 정보를 제공하므로, AI 생성 여부를 탐지하거나 진정성을 판단하는 데 중요한 단서가 될 수 있다. 따라서 보다 다양한 유형의 정서 자극(예: 자연스러운 표정, 동영상)을 활용하여 출처표지 효과가 자극의 생태학적 타당성에 따라 어떻게 달라지는지를 검증한다면, 실험실 결과를 실제 상황에 일반화할 수 있는 생태학적 타당성을 높일 뿐만 아니라, 일상에서 접하는 AI 생성 콘텐츠(예: 딥페이크 영상, AI 아바타)에 대한 사람들의 반응을 예측하는 데 실용적 시사점을 제공할 것이다.

참고문헌

Afzal, S., Sezgin, T. M., Gao, Y., & Robinson, P. (2009). Perception of emotional expressions in different representations using facial feature points. Proceedings of the 3rd International *Conference on Affective Computing and Intelligent Interaction and Workshops*, 1-6. https://doi.org/10.1109/ACII.2009.5349549

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language,* 59(4), 390-412.. https://doi.org/10.1016/j.jml.2007.12.005

Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in Human Behavior, 52*, 331-337.

https://doi.org/10.1016/j.chb.2015.06.018

Baltrušaitis T., Riek, L. D., & Robinson, P. (2010). Synthesizing expressions using facial feature point tracking: How emotion is conveyed. *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments,* 27-31. https://doi.org/10.1145/1877826.1877835

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 617-629. https://doi.org/10.1038/nrn1476

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48. https://doi.org/10.18637/jss.v067.i01

Birmingham, E., & Kingstone, A. (2009). Human social attention: A new look at past, present, and future investigations. *Annals of the New York Academy of Sciences, 1156*(1), 118-140. https://doi.org/10.1111/j.1749-6632.2009.04468.x

Carlson, C. A., Gronlund, S. D., Weatherford, D. R., & Carlson, M. A. (2012). Processing differences between feature based facial composites and photos of real faces. *Applied Cognitive Psychology, 26*(4), 525-540. https://doi.org/10.1002/acp.2824

Chen, P., Zhang, C., Li, B., Tong, L., Wang, L., Ma, S., ... & Yan, B. (2024). An fMRI dataset in response to large-scale short natural dynamic facial expression videos. *Scientific Data, 11*(1), 1247. https://doi.org/10.1038/s41597-024-04088-0

Choung, H., David, P., & Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. International *Journal of Human–Computer Interaction*, 39(9), 1727-1739. https://doi.org/10.1080/10447318.2022.2050543

Chung, K. M., Kim, S., Jung, W. H., & Kim, Y. (2019). Development and validation of the Yonsei face database (YFace DB). *Frontiers in Psychology*, *10*, 2626. https://doi.org/10.3389/fpsyg.2019.02626

Costantini, E., Pianesi, F., & Cosi, P. (2004). Evaluation of synthetic faces: Human recognition of emotional facial displays. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 276-287. https://doi.org/10.3389/fpsyg.2019.02626

Costantini, E., Pianesi, F., & Prete, M. (2005). Recognising emotions in human and synthetic faces: The role of the upper and lower parts of the face. *Proceedings of the 10th International Conference on Intelligent User Interfaces,* 20-27. https://doi.org/10.1145/1040830.1040846

Craig, B. M., Mallan, K. M., & Lipp, O. V. (2012). The effect of poser race on the happy categorization advantage depends on stimulus type, set size, and presentation duration. *Emotion, 12*(6), 1303-1314. https://doi.org/10.1037/a0028622

Demos, K. E., Kelley, W. M., Ryan, S. L., Davis, F. C., & Whalen, P. J. (2008). Human amygdala sensitivity to the pupil size of others. *Cerebral Cortex, 18*(12), 2729-2734. https://doi.org/10.1093/cercor/bhn034

Di Natale, A. F., Simonetti, M. E., La Rocca, S., & Bricolo, E. (2023). Uncanny valley effect: A qualitative synthesis of empirical research to assess the suitability of using virtual faces in psychological research. *Computers in Human*

Behavior Reports, 10, 100288. https://doi.org/10.1016/j.chbr.2023.100288

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods, 42*(1), 351-362. https://doi.org/10.3758/brm.42.1.351

Eiserbeck, A., Maier, M., Baum, J., & Abdel Rahman, R. (2023). Deepfake smiles matter less—the psychological and neural impact of presumed AI-generated faces. *Scientific Reports, 13*(1), 16111. https://doi.org/10.1038/s41598-023-42802-x

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

Fabri, M., Moore, D., & Hobbs, D. (2004). Mediating the expression of emotion in educational collaborative virtual environments: An experimental study. *Virtual Reality, 7*(2), 66-81. https://doi.org/10.1007/s10055-003-0116-7

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191. https://doi.org/10.3758/BF03193146

Grassini, S., & Koivisto, M. (2024). Understanding how personality traits, experiences, and attitudes shape negative bias toward AI-generated artworks. *Scientific Reports, 14*(1), 4113. https://doi.org/10.1038/s41598-024-54294-4

Haan, M. D., Pascalis, O., & Johnson, M. H. (2002). Specialization of neural mechanisms underlying face recognition in human infants. *Journal of Cognitive Neuroscience, 14*(2), 199-209. https://doi.org/10.1162/089892902317236849

Hering, M., Zabel, S., Neef, N., & Otto, S. (2025, May). Predicting individual AI adoption: Personal technological innovativeness on par with AI attitude. In *International Conference on Human-Computer Interaction* (pp. 294-303). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-94171-9_28

Hugenberg, K. (2005). Social categorization and the perception of facial affect: Target race moderates the response latency advantage for happy faces. *Emotion, 5*(3), 267-276. https://doi.org/10.1037/1528-3542.5.3.267

Kätsyri, J. (2018). Those virtual people all look the same to me: Computer-rendered faces elicit a higher false alarm rate than real human faces in a recognition memory task. *Frontiers in Psychology, 9*, 1362. https://doi.org/10.3389/fpsyg.2018.01362

Kegel, L. C., Brugger, P., Frühholz, S., Grunwald, T., Hilfiker, P., Kohnen, O., Loertscher, M. L., Mersch, D., Rey, A., Sollfrank, T., Steiger, B. K., Sternagel, J., Weber, M., & Jokeit, H. (2020). Dynamic human and avatar facial expressions elicit differential brain responses. *Social Cognitive and Affective Neuroscience, 15*(3), 303-317.

https://doi.org/10.1093/scan/nsaa039

Kim, A., Cho, M., Ahn, J., & Sung, Y. (2019). Effects of gender and relationship type on the response to artificial intelligence. *Cyberpsychology, Behavior, and Social Networking, 22*(4), 249-253. https://doi.org/10.1089/cyber.2018.0581

Kim, H., Bian, Y., & Krumhuber, E. G. (2025). A review of 25 spontaneous and dynamic facial expression databases of basic emotions. *Affective Science, 6*(2), 380-394. https://doi.org/10.1007/s42761-024-00289-3

Krumhuber, E. G., Skora, L., Küster, D., & Fou, L. (2017). A review of dynamic datasets for facial expression research. *Emotion Review, 9*(3), 280-292. https://doi.org/10.1177/1754073916670022

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1-26. https://doi.org/10.18637/jss.v082.i13

Lee, K. U., Kim, J., Yeon, B., Kim, S. H., & Chae, J. H. (2013). Development and standardization of extended ChaeLee Korean facial expressions of emotions. *Psychiatry Investigation, 10*(2), 155-163. https://doi.org/10.4306/pi.2013.10.2.155

Lee, S. B., Koo, S. J., Song, Y. Y., Lee, M. K., Jeong, Y. J., Kwon, C., ... & An, S. K. (2013). Theory of mind as a mediator of reasoning and facial emotion recognition: findings from 200 healthy people. *Psychiatry Investigation*, 11(2), 105-111. https://doi.org/10.4306/pi.2014.11.2.105

Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User-AI assistant interactions. *Computers in Human Behavior, 118*, 106680. https://doi.org/10.1016/j.chb.2021.106680

Liang, H., Mao, X., & Reiss, M. J. (2025). Personality traits measured by the HEXACO personality inventory and the dark triad predict university students' attitudes and misconduct behaviors related to generative artificial intelligence. *Scientific Reports, 15*(1), 41787. https://doi.org/10.1038/s41598-025-25744-4

Liang, Y., & Lee, S. A. (2017). Fear of autonomous robots and artificial intelligence: Evidence from national representative data with probability sampling. *International Journal of Social Robotics, 9*(3), 379-384. https://doi.org/10.1007/s12369-017-0401-3

Lundqvist, D., Flykt, A., & Öhman, A. (1998). Karolinska Directed Emotional Faces (KDEF) [Database record]. APA PsycTests. https://doi.org/10.1037/t27732-000

Maslej, N., et al. (2025). The AI Index 2025 Annual Report. Stanford: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf

Matheson, H. E., & McMullen, P. A. (2011). A computer-generated face database with ratings on realism, masculinity, race, and stereotypy. *Behavior Research Methods, 43*(1), 224-228. https://doi.org/10.3758/s13428-010-0029-9

Matheson, H. E., Bilsbury, T. G., & McMullen, P. A. (2012). Second-order relational face processing is applied to faces of different race and photographic contrast. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale, 66*(1), 51-62. https://doi.org/10.1037/a0026062

McKone, E., Wan, L., Pidcock, M., Crookes, K., Reynolds, K., Dawel, A., ... & Fiorentini, C. (2019). A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact. *Scientific Reports, 9*(1), 12820. https://doi.org/10.1038/s41598-019-49202-0

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3-35. https://doi.org/10.1037/1076-8971.7.1.3

Miller, E. J., Foo, Y. Z., Mewton, P., & Dawel, A. (2023). How do people respond to computer-generated versus human faces? A systematic review and meta-analyses. *Computers in Human Behavior Reports, 10*, 100283. https://doi.org/10.1016/j.chbr.2023.100283

Miller, E. J., Krumhuber, E. G., & Dawel, A. (2022). Observers perceive the Duchenne marker as signaling only intensity for sad expressions, not genuine emotion. *Emotion, 22*(5), 907-919. https://doi.org/10.1037/emo0000772.supp

Moser, E., Derntl, B., Robinson, S., Fink, B., Gur, R. C., & Grammer, K. (2007). Amygdala activation at 3T in response to human and avatar facial expressions of emotions. *Journal of Neuroscience Methods, 161*(1), 126-133. https://doi.org/10.1016/j.jneumeth.2006.10.016

Mühlberger, A., Wieser, M. J., Herrmann, M. J., Weyers, P., Tröger, C., & Pauli, P. (2009). Early cortical processing of natural and artificial emotional faces differs between lower and higher socially anxious persons. *Journal of Neural Transmission, 116*(6), 735-746. https://doi.org/10.1007/s00702-008-0108-6

Murtarelli G, Collina C, Romenti S (2023) "Hi! How can I help you today?": Investigating the quality of chatbots-millennials relationship within the fashion industry. *The TQM Journal 35*(3):719-733. https://doi.org/10.1108/tqm-01-2022-0010

Ostrom, T. M., Carpenter, S. L., Sedikides, C., & Li, F. (1993). Differential processing of in-group and out-group information. *Journal of Personality and Social Psychology, 64*(1), 21-34. https://doi.org/10.1037//0022-3514.64.1.21

Park, K., & Yoon, H. Y. (2025). AI algorithm transparency, pipelines for trust not prisms: mitigating general negative attitudes and enhancing trust toward AI. *Humanities and Social Sciences Communications*, 12(1), 1160. https://doi.org/10.1057/s41599-025-05116-z

Pearson, J., Dror, I. E., Jayes, E., Whordley, G. R., Mason, G., & Nightingale, S. (2026). Examining human reliance on artificial intelligence in decision making. *Scientific Reports*. https://doi.org/10.1038/s41598-026-34983-y

Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics

in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior, 122*, 106855. https://doi.org/10.1016/j.chb.2021.106855

Peromaa, T., & Olkkonen, M. (2019). Red color facilitates the detection of facial anger —but how much? *PLoS One, 14*(4), 1-13. https://doi.org/10.1371/journal.pone.0215610

Persson A, Laaksoharju M, Koga H (2021) We mostly think alike: Individual differences in attitude towards AI in Sweden and Japan. *The Review of Socionetwork Strategies, 15*(1), 123-142. https://doi.org/10.1007/s12626-021-00071-y

R Core Team (2024). R: A Language and Environment for Statistical Computing (Version 4.4) [Computer Software]. https://cran.r-project.org

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports, 1*, 100014. https://doi.org/10.1016/j.chbr.2020.100014

Stein, J. P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards AI: measurement and associations with personality. *Scientific Reports, 14*(1), 2909. https://doi.org/10.1038/s41598-024-53335-2

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron, 30*(3), 829-841. https://doi.org/10.3389/conf.fnhum.2011.207.00312

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113-117. https://doi.org/10.1016/j.jesp.2014.01.005

Xue, H., Zhang, J., Shen, C., & Wojcieszak, M. (2024). The majority of fact-checking labels in the United States are intense and this decreases engagement intention. *Human Communication Research, 50*(4), 530-544. https://doi.org/10.1093/hcr/hqae007

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141-145. https://doi.org/10.3410/f.717989861.793473051

1차원고접수 : 2026. 03. 25
최종게재결정 : 2026. 04. 20

한국심리학회지: 일반
Korean Journal of Psychology: General
2026, Vol. 45, No. 3, 253-276
http://dx.doi.org/10.22257/kjp.2026.6.45.3.253

The Effect of AI Labels on Emotional Ratings of Faces: Moderating Roles of In-Group Bias and AI Anthropomorphism

Jini Tae **Wonil Choi**

School of Humanities and Social Sciences, Gwangju Institute of Science and Technology

With the advancement of Generative AI, the proliferation of deepfake stimuli that are difficult to distinguish from reality has made understanding how people perceive computer-generated faces an important research topic. This study examined how emotional responses to identical faces differ depending on labels (human vs. AI), facial race (in-group vs. out-group), and individual attitudes toward AI. Fifty Korean university students participated in the experiment. Participants were presented with either a "human" or "AI" source label, followed by photos of Korean (in-group) or Caucasian (out-group) faces displaying positive (happy), negative (angry), or neutral expressions, which they then rated for emotional valence. General attitudes toward AI and AI anthropomorphism tendency were also measured. All stimuli were photographs of actual actors, with only the source labels manipulated. Results showed main effects of emotion type, with positive expressions rated more positively and negative expressions rated more negatively compared to neutral expressions. In-group faces were rated more positively than out-group faces, and human-labeled photos were rated more positively than AI-labeled photos. Interestingly, the interaction effect between the label and the emotion variables was observed only for in-group faces. Specifically, for in-group positive expressions, human-labeled photos were rated more positively than AI-labeled photos. This label effect was attenuated among participants with high AI anthropomorphism tendency. This study contributes to an integrated understanding of how perceptions of AI influence emotional processing by demonstrating that the effects of labels can vary depending on group identity and attitudes toward AI.

Key words : Deepfake, Emotion Perception, Label Effect, Other-Race Effect, Individual Differences