한국심리학회지 : 문화 및 사회문제
Korean Journal of Culture and Social Issues
2026, Vol. 32, No. 3, 329~357.
http://dx.doi.org/10.20406/kjcs.2026.8.32.3.329

OPEN ACCESS

로봇학대는 가능한가?: 인간유사성, 쌍도덕, 도덕적 기반의 역할*

최 민[1] 정 은 경†

본 연구는 인간과 유사한 외양을 지닌 로봇이 늘어나는 상황에서, 로봇의 인간유사성이 로봇을 해한 가해자에 대한 책임 판단(도덕적, 법적)에 어떤 영향을 미치는지 살펴보았다. 쌍도덕 이론을 바탕으로 로봇에 대한 지각된 행위능력과 경험능력을 매개변수로, 도덕적 기반을 조절변수로 설정하였다. 예비연구를 통해 인간유사성 수준이 다른 네 종류의 로봇 이미지를 선정한 뒤, 각 조건에 무작위로 배정된 성인 400명이 로봇이 주인에 의해 신체적 손상을 입는 시나리오를 읽고, 관련 변인을 평가하였다. 분석 결과, 로봇의 인간유사성은 도덕적, 법적 책임 판단 모두에 정적 영향을 미쳤다. 병렬 매개 분석에서 도덕적 판단은 지각된 행위능력만, 법적 판단은 지각된 경험능력만을 통해서 간접효과가 나타나, 동일한 피해 상황에서도 판단 유형에 따라 서로 다른 마음지각 차원이 작동함이 드러났다. 조절된 매개 분석에서는 도덕적 판단에서만 도덕적 기반이 두 경로를 유의하게 조절했으며, 법적 판단에서는 조절효과가 나타나지 않았다. 이러한 결과는 로봇의 외양이 행위와 경험이라는 두 차원의 마음지각을 유발하고, 도덕 판단은 개인의 도덕적 가치에 따라 유연하게 달라지는 반면 법적 판단은 상대적으로 안정적인 구조를 보일 수 있음을 시사한다.

주요어 : 인간유사성, 도덕적 판단, 법적 판단, 마음지각, 쌍도덕이론, 도덕적 기반 이론

* 이 논문은 2023년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2023S1A5A2A01078593).

1) 강원대학교(춘천) 심리학과 석사과정
† 교신저자 : 정은경, 강원대학교, 강원특별자치도 춘천시 강원대학교길 1 강원대학교 사회과학관 409호
E-mail : ekchung@kangwon.ac.kr

Copyright ©2026, The Korean Psychological Association of Culture and Social Issues
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://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.

2021년 11월, 우리나라에서는 처음으로 로봇 학대가 논란으로 등장했다. 당시 한 대선 후보가 로봇산업 전시회에서 로봇이 넘어졌을 때의 복원능력을 확인하기 위해 로봇을 밀어 뒤집어 넘어뜨렸는데, 이 장면이 영상으로 퍼지자 사람들의 반응은 극명히 엇갈렸다. 일부 언론과 네티즌들은 “로봇학대다”, “감정이입 능력에 문제가 있다”라고 비판하였으며, 반대 입장에서는 “감정을 느낄 수 없는 로봇에게 학대라는 개념 자체가 무의미하다”, “기계 성능 테스트는 원래 그런 방식으로 이루어진다”와 같은 반응으로 로봇학대라는 개념 자체가 불가능하다고 지적하였다. 이러한 논란은 “사족 보행 로봇을 걷어차는 것은 도덕적으로 잘못됐을까”(한국일보, 2021)라는 기사로 이어졌다. 이 기사에서는 ‘로봇학대’ 논란의 반응이 엇갈린 이유로 로봇의 도덕적 지위를 다르게 본다는 점을 언급했다. ‘로봇학대’를 비판한 이들은 로봇의 도덕적 지위를 부정하는 반면, 반대 입장에서는 로봇을 도덕적으로 대해야 할 의무가 있다는 점을 인정한다고 말했다. 또한, “로봇박사, ‘로봇 학대 논란, 좋은 화두 던졌다’”(노컷뉴스, 2021)라는 기사에서는 인간이 로봇에게 도덕적 지위를 부여하고, 그에 따라 감정적이고 도덕적인 판단을 하는 현상이 실제로 존재함을 제시하였다.

상기한 사례는 로봇기술 발전이 사회에 미치는 영향이 단순히 인간을 편리하게 한다는 개발 목적에만 머무르지 않을 것임을 시사한다. 특히 인공지능의 급속한 발전으로 최근에는 로봇과 인공지능이 결합되는 현상이 나타나고 있다. 실제로 엔비디아의 CEO 젠슨 황은 2025년에 열린 소비자 가전 전시회(CES)에서 AI의 궁극적 미래는 로봇, 자율주행차와 같은 물리적 실체를 갖춘 ‘피지컬(physical) AI’의 시대가 될 것이라고 하였다(조선일보, 2025). 인공지능이 장착된 로봇은 더욱 인간과 유사한 기능을 수행할 것이고, 이는 로봇이 인간과 어떤 관계에 있는지에 대한 규명과 연구, 나아가 법률적 판단의 필요성을 촉발할 것이다.

AI 및 로봇과 같은 4차산업 기술의 결과물과 관련된 윤리 문제는 크게 세 개 범주로 나뉜다(이상돈, 정채연, 2017). 첫 번째는 생산 프로그래밍의 윤리적 지침으로 지금까지 4차 산업 관련 윤리 연구 및 정책이 집중해 온 주제이다. 두 번째는 사고에서의 법적 책임 문제이다. 예를 들면 운전자 없는 자율주행자동차 사고에서 책임의 주체는 누구인가 하는 것으로 현재 활발한 연구(예, 이혜원, 정은경, 2021; Awad et al., 2018)와 정책 개발이 이루어지고 있다. 마지막은 기술의 주체적 책임과 권리에 대한 것으로, 예를 들면, 로봇에게 권리와 책임이 존재하는지 여부에 대한 것이다. 인간의 기본권 중 하나인 신체의 자유에 대한 권리가 로봇에게도 적용이 되는지, 그리고 신체의 자유에 대한 권리의 침해인 상해나 학대가 로봇에게도 가능한지 등에 대한 문제라고 할 수 있다.

실제로 초기의 로봇 관련 연구는 대부분 로봇의 기능성과 기술력, 그리고 인간에게 제공할 수 있는 실용성에 집중되었다. 그러나 최근에는 로봇이 의료, 교육 등 인간과 밀접하게 상호작용하는 영역에서 사용되면서 로봇에 대한 사람들의 인식과 윤리적 평가, 나아가 책임 귀인 판단에 대한 논의가 활발히 이루어지고 있다. 특히 로봇의 외양이 인간과 유사할수록(human-likeness), 인간은 로봇을 기계라기보다는 사회적 존재로 인식하고(Kanda et al., 2004; Schombs et al., 2023), 그 로봇에게 감정, 의도, 자율성과 같은 인간적인 특성을 부여

하는 경향을 보이는 것으로 보고되고 있다(Mathur & Reichling, 2016). 이러한 연구들은 사람들이 로봇에게 인간과 같은 윤리적, 도덕적 판단과 지각을 보일 가능성이 있으며, 특히 로봇의 외양이 어떠한가에 도덕적 판단이 상당히 영향을 받을 가능성을 시사한다.

본 연구는 로봇 외양의 인간유사성이 로봇에 대한 도덕적 판단에 어떤 영향을 주는지 알아보고자 하였다. AI나 로봇의 외양이 도덕적 판단에 미치는 영향을 알아본 연구들이 존재하기는 하나(예, 신홍임, 2021; 윤은진, 정은경, 2025; 최윤빈, 장대익, 2022; Laakasuo et al., 2021; Malle et al., 2016; Schombs et al., 2023; Zhang et al., 2024), 이들은 대부분 AI나 로봇의 책임성에 대해 다루었으며, '피해자로서의 로봇', 즉 로봇의 권리에 대해 다룬 실증 연구는 비교적 적은 상황이다. 이에 본 연구에서는 도덕적 판단 중 권리침해 판단에 로봇 외양의 인간유사성이 미치는 영향을 살펴보고자 한다. 아울러 로봇 외양의 인간유사성과 권리침해 판단 간의 관계를 매개하는 변인으로써 선행연구에서 제시된 쌍도덕 이론(Theory of Dyadic Morality; TDM, Gray et al., 2007)의 역할을 알아보고자 한다. 마지막으로 이러한 매개모델이 개인이 가지고 있는 도덕적 가치에 따라 달라지는지를 도덕적 기반이론((Moral Foundations Theory, Haidt & Joseph, 2004; Haidt & Graham, 2007)이 제시한 개념들을 통해 확인하고자 한다.

로봇 외양의 인간유사성과 도덕적 판단

로봇 외양의 인간 유사성(Human- Likeness, HL)은 로봇의 물리적 외양이 인간과 얼마나 닮았는지를 나타낸다(Broadbent et al., 2013; Kanda et al., 2004). 로봇 관련 연구에서는 로봇 외양의 인간 유사성과 의인화(anthropomorphsim) 개념을 분리할 필요가 있는데, 이는 의인화라는 용어 역시 대상을 인간처럼 만든다는 느낌을 주기 때문이다. 의인화는 비인간 존재에 대해 감정, 의도, 자율성과 같은 인간적 속성을 추론적으로 부여하는 것을 의미하는 것으로(Epley et al., 2007; Waytz et al., 2010), 로봇뿐 아니라 동물, 자연 등에 대해서도 의인화가 이루어질 수 있다. 즉 만약 로봇에 대해 사람들이 인간적 속성을 부여하고 느낀다면 이는 로봇 의인화가 일어났다고 할 수 있다.

선행연구들은 로봇의 외양이 사람들의 사회적 반응을 유발하는 핵심 요인임을 반복적으로 보고해 왔다. 특히 인간 유사성이 높은 로봇은 단지 기능적 파트너를 넘어서, 상호작용의 주체로 인식되며 사회적 관계 속에서 특정한 역할을 수행할 수 있는 존재로 지각되기도 한다(곽소나, 2014; Prematilake & Li, 2024; Tay et al., 2014). 실제로 Broadbent 등(2013)은 의료로봇의 외양이 환자들의 수용성과 기대 수준에 영향을 미칠 수 있음을 시사하였다. 또한 인간 유사성이 높을수록 로봇에 대한 정서적 반응, 신뢰, 친근감, 사회적 수용성이 높아진다는 결과들도 존재한다(Fong et al., 2003; Goetz et al., 2003; Powers & Kiesler, 2006; Złotowski et al., 2015). 이러한 연구들은 해당 로봇에 대한 도덕적 판단의 가능성을 증가시키고, 로봇의 외양은 단순한 디자인 요소를 넘어서, 인간-로봇 상호작용의 성격 자체를 근본적으로 바꾸는 요인으로 작용함을 시사한다.

실제로 로봇 외양의 인간유사성이 도덕적 판단에 영향을 준다는 결과들은 꾸준히 보고되고 있다. Malle 등(2016)은 인간 외양을 많이 닮은 로봇이 실수하거나 실패했을 때, 사용자

들이 그 행위나 실패에 대해 더 많은 도덕적 기대와 책임을 부여하려는 경향을 보인다고 보고하였고, De Graaf와 Allouch (2013)는 로봇이 인간처럼 보일수록 그 로봇이 행한 결과에 대해 도덕적 해석을 더 많이 시도한다고 밝혔다. Strait 등(2017) 역시 비휴머노이드에 비해 휴머노이드에 대해 더 강한 도덕적 비난을 한다고 하였으며, Liberman-Pincu 등(2022)은 로봇의 신체구조, 윤곽, 색채 등의 시각적 요소를 세분화하여 분석한 결과, 이러한 외적 요소들이 사람들의 도덕적 인식에 영향을 준다고 보고하였다. 최윤빈과 장대익(2022) 역시 로봇의 외양이 도덕적 책임 지각에 영향을 준다는 것을 밝혔다.

현재까지 로봇에 대한 도덕 판단 연구는 대부분 로봇의 행위나 오류에 대한 책임 귀인에 집중되어 있으며, 로봇을 능동적 행위자(agent)로 설정한 상황에 집중되어 왔다. 로봇이 피해자, 특히 신체적 해를 입는 수동적 존재로 제시되는 맥락에서 로봇 외양의 인간유사성이 가해자에 대한 도덕적 판단에 미치는 영향을 실증적으로 다룬 연구는 아직까지 발견되지 않는다. 그러나 Darling 등(2020)의 연구에서 사람들은 16개 학대 행동에 대해 피해자가 사람일 때나 로봇일 때나 동일한 수준으로 비도덕적인 것으로 평가하는 것으로 나타나 로봇에 대한 학대라는 개념이 가능함을 보여주었다. Wieringa 등(2024)의 연구에서도 로봇에게 해를 가하는 사람을 그렇지 않은 사람보다 덜 친사회적인 것으로 평가하는 것으로 나타나기도 하였다.

로봇 외양의 인간유사성 효과는 범죄심리학 및 사회인지 분야에서 발견되는 피해자의 외양이 제 3자의 도덕적 판단에 영향을 미친다는 연구들을 통해 추론될 수 있다. 예를 들어, Maeder 등(2015)은 피해자의 외모가 배심원의 유죄 판단과 증언 신뢰도에 영향을 미칠 수 있음을 발견하였으며, Gulati 등(2024) 역시 온라인 환경에서도 개인의 매력도가 도덕적 신뢰도나 능력 판단에 영향을 미친다고 보고하였다. 폭력 범죄 시나리오에서도 피해자의 매력도가 높을수록 관찰자가 피해자를 더 신뢰하고 동정심을 느끼며, 가해자에게 더 큰 책임을 부여하는 경향이 있는 것으로 나타난다(Gerdes et al., 1988; Hall et al., 2024). 이처럼 다양한 심리학 분야에서 피해자의 외양이 도덕 판단에 미치는 영향은 반복적으로 검증되어 왔으며, 이러한 결과들은 로봇의 외양이 권리침해 상황에서의 판단에도 영향을 미칠 가능성을 추론케 한다. 아울러 로봇의 인간유사성이 높을수록 의인화가 증가한다는 선행연구들(Kanda et al., 2004; Mathur & Reichling, 2016; Schombs et al., 2023)을 고려해 볼 때, 로봇 외양이 인간과 닮을수록 가해자에 대한 책임판단이 강해질 것으로 예상된다. 다만, 본 연구에서는 도덕적 판단과 법적 판단이 서로 다르다는 선행연구(Schleim et al., 2011)를 고려하여 두 판단 유형을 분리하여 측정하고자 한다. 이에 본 연구에서는 다음과 같은 가설을 설정하였다.

가설 1: 로봇외양의 인간유사성은 가해자에 대한 책임 판단에 정적 영향을 줄 것이다.

가설 1-1: 로봇외양의 인간유사성은 가해자에 대한 도덕적 판단에 정적 영향을 줄 것이다.

가설 1-2: 로봇외양의 인간유사성은 가해자에 대한 법적 판단에 정적 영향을 줄 것이다.

쌍도덕의 매개역할

쌍도덕 이론은 도덕적 판단은 위해(harm)와 관련된 인지적 틀을 통해 이루어진다는 이론으로(Gray et al., 2012), 도덕 판단은 기본적으로 도덕적 행위자와 도덕적 경험자라는 한 쌍의 구조로 구성된다고 주장한다. 도덕적 행위자(moral agent)는 의도적이고 자율적인 결정을 내릴 수 있는 존재, 즉 도덕적 책임을 질 수 있는 대상을 말하고, 도덕적 경험자(moral patient)란 감정, 고통, 정서적 반응을 경험할 수 있는 존재, 즉 도덕적으로 보호받아야 할 대상을 말한다. Gray 등 (2012)은 우리가 어떤 상황을 도덕적으로 판단하려면 이 두 요소가 모두 인식되어야 한다고 하였다. 예를 들면, 행위능력이 큰 가해자가 경험능력이 큰 피해자에게 위해를 가한다면 사람들은 더욱 비도덕적으로 판단할 것이다(Schein & Gray, 2018).

도덕적 행위자와 도덕적 경험자라는 구조는 마음지각이론(Mind Perception Theory, Gray et al., 2007)을 기반으로 한다. 마음지각이론에 의하면, 사람들은 인간을 비롯해 다양한 존재들을 평가할 때 행위능력(agency)과 경험능력(experience)이라는 두 차원으로 이루어진 마음지각을 사용하며, 이에 따라 사회적, 도덕적 판단을 내린다. 행위능력은 계획하고 생각하며 행동할 수 있는 능력을, 경험능력은 감정을 느끼고 고통과 즐거움을 경험할 수 있는 능력을 의미한다. 이 연구에 따르면, 로봇과 같은 존재는 어떤 행동을 할 수 있는 행위능력은 높게 지각되지만 감정을 느낄 수 있는 경험능력은 낮게 지각되는 것으로 나타난다. 이 두 차원은 쌍도덕 구조의 기반을 형성하며, 인간은 이 두 차원을 통해 누가 혹은 무엇이 도덕적 주체 혹은 도덕적 보호의 대상이 될 수 있는지 판단하게 된다. 다시 말해 감정을 느낄 수 있다고 여겨지는 대상 즉, 경험능력이 큰 대상은 도덕적 보호의 대상으로 판단하고 자율성과 의도를 지닌 대상 즉, 행위능력이 큰 대상은 도덕적 주체로 판단하여 더 많은 도덕적 책임을 부여한다.

이처럼 마음지각은 도덕적 판단의 심리적 메커니즘으로 작동하며, 최근 연구들은 이러한 구조가 인간뿐 아니라 로봇과 같은 비인간 존재에게도 동일하게 적용됨을 밝히고 있다. Tanibe 등(2017)은 로봇에게 도움을 주는 경험이 해당 로봇이 감정을 지닌 존재라는 인식을 강화하고, 이로 인해 로봇을 도덕적으로 보호해야 한다는 판단이 유발될 수 있다고 보고하였다. 또한, Yam 등(2021)은 사람들이 감정을 지닌 것으로 지각되는 서비스 로봇에 대해 더 긍정적인 태도를 보이며, 해당 로봇이 서비스를 실패했을 때에도 보다 관대하게 반응한다고 보고하였다. 특히, 로봇이 기쁨이나 슬픔과 같은 감정을 경험할 수 있다고 인식될 경우, 사용자들은 해당 로봇을 더 인간적인 존재로 간주하고, 도덕적 판단이나 용서의 대상이 될 수 있다고 평가하였다. Koban와 Wieringa(2024)는 사람들이 로봇이 피해를 입는 장면을 목격할 때, 해당 로봇이 감정을 지닌 존재로 인식되는 경향이 있으며, 이러한 마음지각은 로봇을 도덕적 보호의 대상으로 간주하려는 판단으로 이어질 수 있음을 주장하였다. Seo 등(2015)은 실제 물리적 로봇과 화면 속 가상 로봇에 대한 감정이입 수준을 비교한 실험에서, 물리적으로 실체를 지닌 로봇에 대해 더 높은 감정이입과 마음 지각이 나타나는 경향을 발견하였다. 이러한 결과는 로봇의 인간과 유사한 물리적 실체를 지닐수록 인간의 마음 지각을 더 강하게 유도함을 시사한다.

인간을 닮은 로봇의 외양이 마음지각과 도덕적 판단에 영향을 미친다는 것을 제시한 연구들도 존재한다. Riek 등(2009)은 인간 유사성이 높은 로봇에 대해 더 많은 감정이입이 발생된다고 보고하였으며, Kahn 등(2012)은 아이들이 로봇과 사회적 관계를 맺을 수 있다고 보며, 로봇이 인간처럼 보이거나 행동할 경우 아이들이 감정적 유대를 느끼고 도덕적 관심을 가지는 경향이 있다고 밝혔다. 또한 최윤빈과 장대익(2022)은 AI의 의인화가 도덕적 책임 판단에 미치는 영향을 쌍도덕 이론 관점에서 분석한 실험연구를 통해, 의인화 수준이 높은 AI에 대해 사람들이 더 높은 경험능력을 지각하고, 이러한 지각이 AI실패에 대한 AI의 도덕적 책임 판단을 유의하게 매개함을 확인하였다. 특히, 지각된 경험능력의 경우, 모든 실험에서 AI의인화와 AI의 도덕적 책임 간의 관계를 정적으로 매개한 반면, 지각된 행위능력의 매개 효과는 일부 실험에서만 관찰되었다. 이러한 연구들은 로봇의 외양이 지각된 경험능력과 지각된 행위능력이라는 두 매개변수를 통해 도덕적 판단에 영향을 미칠 수 있음을 시사한다.

상기한 연구들을 종합해 보면, 로봇의 외양과 행위능력 및 경험능력간의 관계는 도덕적 상황 맥락과 종속변인 등에 따라 유의도 여부는 다소 다른 결과를 보이지만 변인들 간에는 정적인 관계가 있는 것으로 나타나고 있다. 이에 본 연구에서는 다음과 같은 가설을 설정하였다.

가설 2: 로봇외양의 인간유사성과 가해자에 대한 책임 판단 간의 관계를 쌍도덕이 정적으로 매개할 것이다.

가설 2-1: 로봇외양의 인간유사성과 가해자에 대한 도덕적 판단 간의 관계를 지각된 행위능력이 정적으로 매개할 것이다.

가설 2-2: 로봇외양의 인간유사성과 가해자에 대한 도덕적 판단 간의 관계를 지각된 경험능력이 정적으로 매개할 것이다.

가설 2-3: 로봇외양의 인간유사성과 가해자에 대한 법적 판단 간의 관계를 지각된 행위능력이 정적으로 매개할 것이다.

가설 2-4: 로봇외양의 인간유사성과 가해자에 대한 법적 판단 간의 관계를 지각된 경험능력이 정적으로 매개할 것이다.

도덕적 기반의 조절된 매개효과

도덕적 기반 이론(Moral Foundations Theory; MFT)은 기존 도덕 이론이 지나치게 서구의 가치만을 반영하였다는 한계를 극복하기 위해 제시된 이론으로, 비서구권 문화에서 나타나는 도덕적 가치를 포함한 5가지 도덕적 판단 기반을 제시하였다(Haidt & Joseph, 2004). 5개 도덕적 기반은 위해(harm/care), 공평성(fairness/reciprocity), 내집단(ingroup/ loyalty), 권위(authority/respect), 순수함(purity/sanctity)으로, 도덕적 기반 각각은 문화, 개인 성향, 정치적 태도 등에 따라 강조되는 정도가 달라지며, 동일한 행위에 대해서도 사람마다 도덕적 반응이 달라지는 이유를 설명해 준다(Graham et al., 2011). 예를 들면, 간통에 대한 한국인의 판단을 도덕적 기반 이론을 통해 살펴본 연구(정은경, 손영우, 2011)에서는 간통에 대해 진보와 보수가 확연히 다른 도덕적 기반을 사용하여 판단하는 것으로 나타났다. 다만, 한국인을 대상으로 한 타당화 연구(정은경 등, 2016)에서는 한국에서의 도덕적 기반은 5가지가 아니라 위해, 공평성, 규범존중 총 3가지의 기반이

가장 안정적이고 설득력 있는 구조인 것으로 나타났다.

선행연구에 의하면 도덕적 기반은 다양한 변인들 간의 관계에서 조절 역할을 하는 것으로 보고된다. 예를 들면 Clifford와 Jerit(2013)은 정치적 메시지에 대한 수용 태도가 개인의 도덕적 기반에 따라 달라진다는 점을 실증하였는데, 특히 위해와 공평성을 중시하는 사람일수록 보편적 인권 이슈에 민감하게 반응하는 경향을 보였다. 또한 Nilsson 등(2020)은 기부 행동을 예측하는 데 있어 개인의 도덕적 기반이 중요한 조절변인으로 작용한다는 점을 밝혔으며, 특히 위해 기반이 강한 사람들은 고통받는 타인에 대한 정보를 접했을 때 기부 의향이 더 크게 증가하는 것으로 나타났다.

이처럼 사람마다 어떤 도덕적 기반을 중시하느냐에 따라 동일한 자극에 대한 해석과 평가가 달라질 수 있다. 피해 로봇의 행위능력, 경험능력 지각이 인간 가해자에 대한 책임 판단으로 이어지는 정도 역시 개인이 중요시하는 도덕적 가치에 따라 달라질 수 있다. 예를 들어, 위해를 중요한 도덕적 가치라고 생각하는 사람은 로봇에게 가해진 손상 자체에 집중하여 인간 가해자의 행위를 더욱 심각하게 인식할 수 있다. 공평성을 중시하는 사람은 로봇이 인간처럼 감정이 있다고 판단할 경우, 로봇을 해친 행위에 대해 공평하지 않다고 생각하여 더 높은 수준의 책임을 부여할 수도 있지만 반대로 로봇을 인간과 동일하게 여기는 것을 불공평하다고 느껴 인간 책임을 낮게 볼 수도 있다. 또한 규범존중을 중요시하는 사람은 사회적 규범이나 도덕 질서 위반 여부에 주목하여, 학대라는 반규범적인 행위를 더 비난할 가능성이 있다.

AI나 로봇 외양과 도덕적 판단에서 도덕적 기반의 조절효과를 직접적으로 살펴본 연구는 발견되지 않는다. 다만 의인화와 관련된 것은 아니나 디지털치료와 세무업무라는 두 영역에서의 AI의 서비스 실패와 인간의 서비스 실패에 대한 책임판단을 비교한 연구(성주현, 정은경, 2025)에서 두 변인의 관계를 도덕적 기반 중 일부 가치기반이 조절하는 것으로 나타났다. 상기 연구에서는 3 요인으로 구성된 도덕적 기반 중 공평성과 규범존중의 조절효과가 보고되었는데, 두 변인 모두가 낮을수록 인간이냐 AI이냐에 따른 도덕적 판단이 크게 달라지는 것으로 나타났다. 이러한 결과는 선행연구와 다른데, 기존 연구들은 공평성이라는 개인중심 가치와 규범존중이라는 공동체중심 가치가 보통은 서로 다른 방향으로 영향을 주는 것으로 나타났기 때문이다. 이에 저자들은 AI냐 인간이냐가 도덕적 판단에 미치는 영향은 개별 가치가 아니라 도덕적 가치 자체를 얼마나 중요하게 생각하고 있는지에 영향을 받는 것 같다고 결론지었다. 실제로 Ogba 등(2021)의 연구에서도 개별 요인이 아니라 도덕적 기반 전체 점수가 조절변수로 사용되기도 하였다. 이에 본 연구에서도 도덕적 기반 각각에 대한 분석은 이루어질 예정이나 가설은 도덕적 기반의 전체 강도를 조절변인으로 사용할 예정이며, 이에 다음과 같이 가설을 설정하였다.

가설 3: 로봇외양의 인간유사성이 쌍도덕을 거쳐 가해자에 대한 책임 판단으로 영향을 미치는 매개효과는 도덕적 기반 강도에 의해 조절될 것이다.

가설 3-1: 로봇외양의 인간유사성이 지각된 행위능력을 거쳐 가해자에 대한 도덕적 판단으로 영향을 미치는 매개효과는 도덕적 기반

강도에 의해 조절될 것이다.

가설 3-2: 로봇외양의 인간유사성이 지각된 경험능력을 거쳐 가해자에 대한 도덕적 판단으로 영향을 미치는 매개효과는 도덕적 기반 강도에 의해 조절될 것이다.

가설 3-3: 로봇외양의 인간유사성이 지각된 행위능력을 거쳐 가해자에 대한 법적 판단으로 영향을 미치는 매개효과는 도덕적 기반 강도에 의해 조절될 것이다.

가설 3-4: 로봇외양의 인간유사성이 지각된 경험능력을 거쳐 가해자에 대한 법적 판단으로 영향을 미치는 매개효과는 도덕적 기반 강도에 의해 조절될 것이다.

예비 연구

예비연구는 본연구에서 사용할 로봇 이미지를 개발하기 위해 실시되었다.

방 법

연구 대상

본 연구는 온라인 설문조사 업체를 통해 실시되었으며 로봇이미지에 익숙한 20대와 30대를 대상으로 선정하였다. 총 180명(남자 90명, 여자 90명)의 피험자가 설문조사에 응했다. 참가자들의 평균 연령은 30.15세(*SD* = 5.28)이었다.

연구 절차

인간 외양 유사 정도에 따라 인간과 가장 닮지 않은 로봇 이미지부터 인간과 가장 닮은 로봇 이미지까지 6개의 이미지가 생성되었다. 그림 2에 제시된 6단계의 로봇 이미지를 사용해 연구를 실시하였다. 참가자들은 6개 로봇 이미지에 무작위로 배정되었고, 각 조건에 30명씩 할당되었다. 참가자들은 배정된 로봇 이미지를 보면서 로봇의 인간유사성과 의인화 정도를 묻는 질문들에 응답하고 마지막에 인구통계학적 특성을 묻는 질문들에 응답하였다.

연구 도구

로봇 이미지

로봇 외양을 다룬 연구들은 대부분 얼굴을 사용하여 로봇 외양의 인간화 효과를 다루었다(예, Mathur & Reichling, 2016; Müller et al., 2021). 그러나 본 연구에서는 로봇을 인격체로 느끼는지를 다루므로 얼굴만이 아닌 로봇 전체의 외양을 통해 인간과 유사함 정도가 드

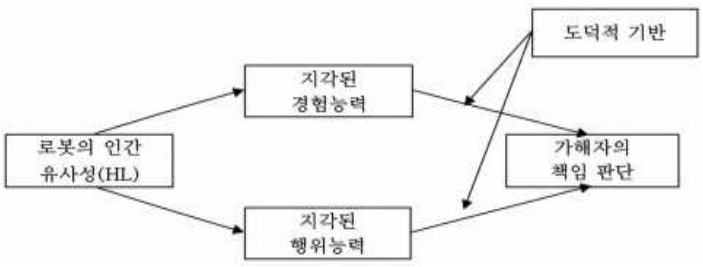
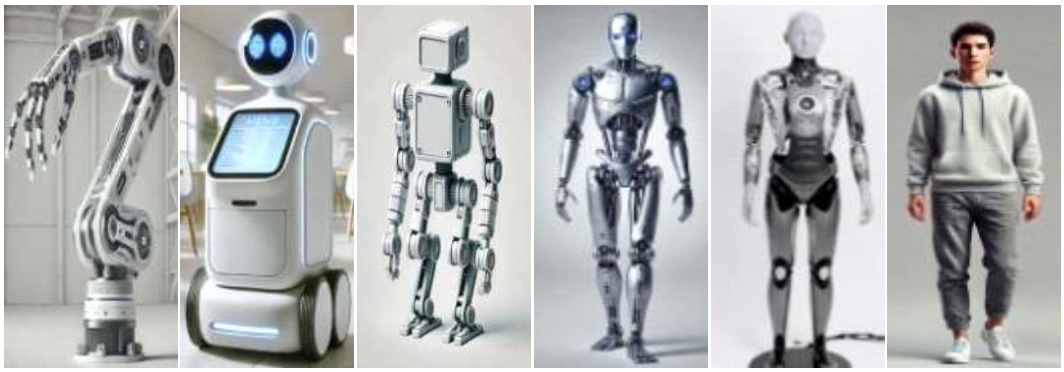


그림 1. 연구모형



1단계 2단계 3단계 4단계 5단계 6단계

그림 2. 연구에 사용된 로봇 이미지

러날 수 있도록 총 6개의 이미지를 제작하였다. 또한 로봇의 외관 연구에서 이미지를 생성할 때 종종 연속이미지를 생성하는 모핑 기법(morphing technique)을 이용하기도 하나 사람 얼굴에 모핑 기법을 사용하면 얼굴의 눈과 같은 부분에 비정상적인 특징이 생겨 이미지에 대한 불쾌감이 형성된다는 선행연구(Palomaki et al., 2018)를 참고하여 Palomaki 등(2018)과 Laakasuo 등(2021)이 사용한 논모핑(non-morphed) 이미지, 즉 사진과 같은 단일 정지 이미지를 사용하였다. 로봇 외양을 통해 인간 유사성 정도를 느낄 수 있도록 가장 낮은 1단계는 신체의 일부만 존재하는 산업용 로봇을, 가장 높은 6단계는 인간 모습을 한 로봇을 설정하였으며, 양 극단 사이의 이미지는 Palomaki 등(2018)과 Laakasuo 등(2021)이 사용한 혼다의 아시모(Asimo), 영화 '아이, 로봇'의 소니(Sonny)와 함께 2022년 엔지니어드아츠社의 '아메카' 등 기존 로봇 이미지를 참조하여 4개를 제작하였다. 로봇 이미지 제작을 위해 기존 로봇 이미지와 함께 생성형 AI(ChatGPT)도 사용하였는데, "인간 외양 유사 정도에 따라 인간과 가장 닮지 않은 로봇 이미지부터 인간과 가장 닮은 로봇 이미지까지 6단계의 이미지를 만들어줘"라는 명령어에 의해 생성된 로봇 이미지들과 기존 로봇 이미지와 유사하게 제작하도록 한 이미지 등이 개발되었다. 상기한 바와 같이 1단계는 산업용 로봇팔을 생성형 AI로 구현하였으며, 가장 인간스러운 로봇 모습은 6단계로 "20대-30대 정도의 운동복을 입은 성인 남성의 이미지"를 생성하도록 하여 그 중 채도나 분위기 등이 다른 이미지와 유사한 것을 선정하였다. 5단계는 인간의 얼굴을 가장 근접하게 구현한 로봇으로 평가받는 실제 '아메카'의 이미지를 선정하였으며, 4단계는 아메카의 이미지를 기본으로 하되 신체가 좀 더 로봇같고 이목구비가 불완전한 이미지를 생성하도록 하였다. 3단계는 사지는 존재하되 이목구비는 없는 로봇을 생성하였으며, 2단계는 사지가 없으나 인간같은 느낌을 주는 로봇 이미지를 생성하도록 하였다. 단계 별 로봇 이미지 개발 및 선정 기준을 정리하면 다음과 같다.

- 1단계: (인간 신체의 일부만 유사) 공장에서 사용하는 산업용 로봇팔 이미지

- 2단계: (인간 신체의 일부가 누락) 팔 등 몸체 일부는 없으나 전체적으로 인간 같은 느낌을 주는 이미지
- 3단계: (인간 신체는 있으나 얼굴이 모호) 전체적으로 인간 신체를 갖추었으나 얼굴의 이목구비가 없어서 로봇으로 확연하게 인식되는 이미지
- 4단계: (인간 신체와 로봇 같은 이목구비) 영화 '아이·로봇'에 나오는 소니와 유사하게 인간을 닮았으나 얼굴표정이 없고 금속성 로봇 이미지
- 5단계: (인간 신체와 인간과 유사한 이목구비) 인간을 닮았고 인간과 훨씬 유사한 얼굴을 보유한 로봇 이미지
- 6단계: (인간 이미지) 운동복을 입은 인간 남성 전신 이미지

인간 유사성(Human-likeness, HL)

제시된 로봇의 인간 유사성에 대한 측정은 Mathur와 Reichling(2016)이 사용한 2문항을 번안해 7점 척도(1점: 전혀 아니다, 7점: 매우 그렇다)로 측정하였다. 측정문항은 다음과 같다. "이 로봇은 얼마나 사람같습니까?", "이 로봇은 얼마나 기계같습니까?(역코딩). 본 척도 점수가 높을수록 로봇의 인간유사성이 높은 것을 의미한다. 본 척도의 Spearman-Brown 계수는 .734였다.

의인화

제시된 로봇의 의인화 수준을 측정하기 위해 동물, 자연, 기술과 같은 다양한 존재자들에 대한 의인화 척도를 개발한 선행연구(Waytz et al., 2010)에서 사용되었던 설문 문항 중, 기술에 대한 의인화 척도를 다루는 네 가지 설문을 번안한 최윤빈과 장대익(2022)의 설문 문항을 사용하였다. 설문 문항은 4문항이고 7점 척도(1점: 전혀 아니다, 7점: 매우 그렇다)이다. 설문 문항의 예는 "이 로봇이 얼마나 의도를 가질 수 있는 것처럼 보이십니까?, "이 로봇이 얼마나 감정을 경험할 수 있는 것처럼 보이십니까?"이며 본 척도의 Cronbach's Alpha는 .818였다.

결 과

로봇의 인간 유사성(HL)

로봇 이미지에 따른 로봇의 인간 유사성 평균이 차이가 있는지 알아보기 위해 일원배치 분산분석(one-way ANOVA)을 실시하였다(표 1). 그 결과, $F(5,174) = 6.215$, $p < .001$으로 로봇 이미지에 따른 HL 평균에 차이가 있는 것으로 나타났다. 이미지간 비교를 위해 실시한 Bonferroni 보정을 적용한 대응별 비교 결과, 6단계는 나머지 모든 단계보다 HL이 유의미하게 높은 것으로 나타났으며, 5단계는 6단계 및 3단계와, 4단계는 6단계와, 3단계는 6단계 및 5단계와 유의한 차이가 나타났다. 2단계와 1단계는 유의한 차이가 없었다.

로봇의 의인화 수준

로봇 이미지에 따른 의인화 수준 평균이 차이가 있는지 알아보기 위해 일원배치 분산분석(one-way ANOVA)을 실시하였다(표 2). 그 결과, $F(5,174) = 5.737$, $p < .001$로 로봇 이미지에 따른 AI의 의인화 수준 평균이 차이가 있는 것으로 나타났다. 이미지간 비교를 위해 실시한 Bonferroni 보정을 적용한 대응별 비교

표 1. 로봇 이미지 간의 인간 유사성 차이 검정

로봇 이미지	인간 유사성			
	평균	표준편차	*F*	Boferroni
1단계	2.250	1.223		
2단계	2.400	1.020		
3단계	1.967	1.090	6.215***	1,2,3,4,5<6
4단계	2.400	1.148		3<5
5단계	2.700	.867		
6단계	3.500	1.520		

주. $^{***}p < .001$

표 2. 로봇 이미지 간의 의인화 수준 차이 검정

로봇 이미지	의인화			
	평균	표준편차	*F*	Boferroni
1단계	2.633	1.176		
2단계	3.300	1.168		1,2+,3<6
3단계	2.583	1.260		1,3<5
4단계	3.575	1.130	5.737***	1,3<4
5단계	3.408	.816		1<3
6단계	3.792	1.237		

주. $^{***}p$ <.001, + p <.10

결과, 6단계는 1단계, 3단계와는 유의한, 2단계와는 경계선 수준(p=.09)의 차이를 보였으며, 5단계는, 1단계, 3단계와, 4단계는 1단계, 3단계와, 3단계는 1단계를 제외한 모든 집단과, 2단계는 1단계, 3단계와는 유의하고 6단계와는 경계선 수준의 차이를 보였다.

상기 결과들을 종합해 보면, 먼저 공업용 로봇팔인 1단계보다 이목구비를 제외한 인간 신체를 갖춘 3단계가 더 기계같다는 응답이 일관적으로 나왔다. 결과적으로, 1단계보다는 3단계가 나머지 단계들과 더 차이를 보였기에 HL이 가장 낮은 단계로 3단계 이미지를 사용하고자 하였다. 또한 2단계의 HL은 4단계와 동일할 정도로 높게 나왔는데, 이는 2단계가 나머지 단계들 이미지에 비해 귀여운 이미지로 표현되었기 때문인 것으로 보인다. 이러한 결과와 아기 스키마 특성을 가진 귀여운 디자인의 로봇 이미지가 긍정적인 반응을 이끌어

내고 사람들에게 더 매력적으로 다가갈 수 있다는 많은 선행연구들(Lv et al., 2021; Yu et al., 2022; Nishiwaki et al., 2017; Aragon et al., 2015; Lieber-Milo, 2022; Sherman & Hadit, 2011)을 고려하여 2단계도 배제되었다. 나머지 3~6단계에서는 다른 집단간 차이는 유의하였으나 3-4단계, 4-5단계 간 차이는 방향성은 적절하지만 유의하지는 않은 것으로 나타났다. 그러나 본 연구에서는 로봇 이미지 단계라는 집단변수를 독립변인으로 사용하는 것이 아니라 선행연구에 따라(최윤빈, 장대익, 2022) HL 점수라는 연속변수를 독립변인으로 사용하기 때문에 집단간 차이가 아주 분명할 필요는 없다. 이에 본 연구에서는 최종적으로 3 ~ 6 단계 이미지 4개를 HL 자극으로 선정하였다.

본 연구

본연구는 예비연구에서 개발된 4개의 이미지를 사용하여 로봇의 인간 유사성 정도가 도덕적 판단에 영향을 미치는지와 쌍도덕의 매개효과 및 도덕적 기반의 매개된 조절효과를 알아보고자 하였다.

방 법

연구 대상

본 연구는 온라인 설문조사 업체를 통해 실시되었으며 20대부터 50대까지의 연령과 성별이 균등하게 분포되도록 참가자를 모집하였다. 참가자들의 평균 연령은 39.73세(SD = 10.81)이었다.

연구 절차

참가자들은 4개의 로봇 외양의 인간유사성 조건에 무작위로 배정되었으며, 각 조건에 100명씩 할당되었다. 참가자들은 로봇 이미지와 함께 로봇이라고 명시된 로봇 권리침해 시나리오를 제시받았다. 이후 연구 척도들과 인구통계 특성 질문들에 응답하였다.

연구 도구

로봇 가해 시나리오

로봇 권리침해 시나리오는 지난 2021년 대선시 이슈가 되었던 '로봇학대' 논란을 참고하여 제작되었다. 로봇이 인간에 의해 가벼운 몸체 손상을 당하는 시나리오이고 내용은 아래와 같다.

> 로봇 '키보'의 주인은 40대 남성 스티브입니다. 매우 스트레스가 많던 어느 날, 스티브는 상사로부터 꾸중을 듣고 나서 화풀이로 '키보'를 발로 찼습니다. '키보'는 넘어져서 팔이 부러졌습니다.

로봇 외양의 인간 유사성(HL)

4개 로봇 이미지 단계별 차이 분석 결과, 3단계와 4단계간 차이를 제외한 모든 조합에서 유의한 HL 점수 차이를 보였다. 본 연구에서는 의인화가 아닌 인간 유사성을 측정하였으며, 선행연구에 따라(최윤빈, 장대익, 2022) 로봇 외양의 인간 유사성을 로봇 이미지 집단별로 점수를 부여하는 방식이 아니라 이미지에 대해 참가자들이 평가한 HL 점수 자체를 인간 유사성 척도의 점수로 사용하였다. 로봇의 인간 유사성 문항은 예비연구에서 사용한

것과 같다. 본 척도의 Spearman-Brown 계수는 .724이었다.

쌍도덕: 행위능력과 경험능력

참가자들이 로봇의 행위능력과 경험능력을 얼마나 지각하는지는 Yam 등(2021)의 척도를 사용하여 측정하였으며, 지각된 행위능력과 경험능력 각 4문항씩 제시되었다. 참가자들은 “이 로봇은 다른 존재와 소통할 능력이 있다”(행위능력), “이 로봇은 고통을 느낄 수 있다”(경험능력)와 같은 문항들에 7점 척도(1점: 전혀 아니다, 7점: 매우 그렇다)로 응답하였다. 행위능력의 Cronbach’s Alpha는 .836, 경험능력의 Cronbach’s Alpha는 .940였다.

가해자의 책임판단(도덕적, 법적)

가해자의 책임판단은 가해자 스티브의 도덕적 책임, 법적 책임이 어느 정도인지 묻는 질문으로 측정되었다. 특히 법적 책임판단은 실제적 전문적 판단이 아닌 일반인이 가해자에게 법적 책임을 부여하려는 정도라고 정의된다. 도덕적 판단과 법적 판단은 선행연구들(성주현, 정은경, 2025; 안정용, 성용준, 2021; 장혜지, 조은경, 2025)에서 사용된 사회적, 법적 책임 문항들을 참고하여 본 연구에 적합하도록 문항을 재구성하였다. 참가자들은 도덕적 판단 문항 2문항(“스티브의 행동은 사회적으로 비난받을 만하다”, “스티브는 나쁘다”)과 법적 책임판단 3문항(“스티브의 행동은 법적으로 처벌받을 만하다”, “스티브는 상해죄로 처벌받을 만하다”, “스티브는 학대죄로 처벌받을 만하다”)에 대해 7점 척도(1점: 전혀 아니다, 7점: 매우 그렇다)로 응답하였다. Cronbach’s Alpha는 전체 문항에 대해서는 .892, 도덕적 판단에 대해서는 .732, 법적 책임판단에 대해서는 .943으로 나타났다.

본 연구에서는 도덕적 판단과 법적 책임판단이 개념적으로 다를 수 있기에 두 개념이 구분된 개념인지를 검증하기 위해 탐색적 요인분석과 확인적 요인분석을 실시하였다. 탐색적 요인분석 결과, 고유값이 1을 초과하는 요인은 1개로 나타났다. 요인적재량은 법적 책임판단 문항들은 동일 요인에 매우 높은 적재량(.88∼.95)을 보인 반면, 도덕적 판단 문항은 상대적으로 낮고 분산된 적재 양상을 보였다. 이러한 결과는 전체 문항이 단일 요인으로도 설명될 수 있음을 시사한다. 그러나 개념적으로는 두 변인이 분명히 구분되므로, 2요인 모델로 확인적 요인분석을 실시하였다. 그 결과, 도덕적 판단과 법적 판단을 구분하는 2요인 모형은 전반적으로 양호한 적합도를 나타냈다($\chi^2(4) = 51.64$, $p < .001$; CFI = .975; TLI = .937; SRMR = .035). 다만 RMSEA는 .154로 높게 나타났으며, 이는 두 요인의 분리가 충분히 명확하지 않을 가능성을 시사한다. 비록 자유도가 매우 낮은(df=4) 소규모 모형에서 과대추정되는 경향이 있다는 점을 고려할 필요가 있으나 본 결과만으로 두 구성개념이 완전히 독립된 차원이라고 단정하기에는 한계가 있다.. 구체적 요인적재량을 살펴보면, 도덕적 판단 문항은 .73∼.80의 적재량을 보였으며, 법적 책임판단 문항은 .87∼.95로 매우 높은 적재량을 나타냈다. 또한 두 요인 간 상관은 .70으로 나타나, 두 구성개념이 상당히 관련되어 있으면서도 완전히 동일하지는 않은 것으로 나타났다.

상기 결과를 종합하면, 본 척도는 단일 요인으로도 설명 가능하지만, 이론적, 개념적으로 구분되는 도덕적 판단과 법적 책임판단이라는 두 차원으로 분리하여 분석할 수 있는

가능성도 제시되었다. 따라서 본 연구에서는 두 판단 유형을 완전히 독립된 구성개념으로 전제하기보다는, 상호 밀접하게 관련되어 있으면서도 부분적으로 구분될 수 있는 판단 차원으로 해석하고 도덕적 판단과 법적 책임판단을 구분하여 분석을 실시하였다.

통제 변인

본 연구에서는 로봇 기술에 대한 익숙함과 유용성에 대한 믿음이 도덕적 판단에 영향을 미칠 것이라고 예상하였다. 이에 “귀하는 로봇 기술에 대해 얼마나 잘 알고 있습니까?”(1점: 전혀 모른다, 7점: 매우 잘 안다)와 “귀하는 로봇이 인류에 유익하다고 생각하십니까?”(1점: 전혀 아니다, 7점: 매우 그렇다)라는 문항을 제시해 본 연구 문제의 모든 분석에서 통제변인으로 사용하였다.

자료분석

수집한 자료는 모두 SPSS 29와 SPSS process macro 4.2를 사용하여 분석되었다. 주요 연구 변인들의 평균, 표준편차, 왜도, 첨도를 확인하기 위해 기술 통계 분석을 수행하였다. 또한 주요 연구 변인들의 상관관계를 확인하기 위해 상관관계 분석을 수행하였다. 로봇의 인간 유사성과 도덕적 책임판단 간의 관계를 살펴보기 위해, 통제변인을 1단계에, 인간 유사성을 2단계에 투입한 위계적 회귀분석을 실시하였다. 쌍도덕의 매개효과를 확인하기 위해 PROCESS 모델 4번을 통한 다중병렬매개모형 분석을 실시하였다. 마지막으로 도덕적 기반의 조절효과 및 조절된 병렬매개효과를 확인하기 위하여 PROCESS 모델 59번을 실시하였다.

결 과

본 연구의 주요 연구 변인들의 기술통계량과 상관관계를 표 3에 제시하였다. 모든 연구 변인들의 왜도(-0.75 ~ 0.82)와 첨도(-.68~1.54)가 정규분포 가정을 충족하는 것을 확인하였다(Curran et al., 1996).

로봇의 인간 외양성과 가해자의 도덕적 책임판단 간의 관계

로봇기술지식과 로봇인류유익에 대한 응답을 통제변수로 설정한 위계적 회귀분석을 실시하였다. 먼저 도덕적 판단에 대한 분석 결과, 회귀모형은 유의한 것으로 나타났다(F = 4.632, p < .01). 회귀모형의 설명력은 3.4%(R^2 = .034)였으며,. 로봇의 HL이 도덕적 책임판단에 정적인 영향(B = .147, p < .001)을 미치는 것으로 나타났다. 다음으로 법적 책임판단에 대한 분석 결과, 회귀모형은유의한 것으로 나타났다(F = 11.880, p < .001). 회귀모형의 설명력은 8.3%(R^2 = .083)였으며, 로봇의 HL이 법적 책임판단에 정적인 영향(B = .190, p < .001)을 미치는 것으로 나타났다. 따라서 위의 결과를 통해 가설1-1, 1-2가 모두 지지되었다.

쌍도덕의 매개효과

먼저 도덕적 판단에 대한 쌍도덕의 매개효과 분석 결과(표 4), 로봇의 HL은 경험능력(B = .403, p < .001)과 행위능력(B = .366, p < .001) 모두에 유의한 정적 영향을 미치는 것으로 나타났다. 이는 로봇의 외양이 인간과 유사할수록 사람들이 로봇의 경험능력과 행위능력을 더 높게 지각함을 의미한다. 반면, 로봇

표 3. 연구 변수 간의 평균, 표준 편차 및 상관관계(*N*=400)

	1	2	3	4	5	6	7	8	9
1	-								
2	.474***	-							
3	.441***	.519***	-						
4	.159**	.218***	.253***	-					
5	.065	.250***	-.013	.439***	-				
6	.024	.209***	-.017	.396***	.748***	-			
7	.243***	.224***	.456***	.242***	.045	.061	-		
8	.225***	.288***	.256***	.197***	.060	.135**	.083	-	
9	.059	.144**	-.150**	.005	.145**	.167***	-.117*	.114*	-
M	3.05	4.01	2.45	3.39	4.15	4.08	3.27	3.65	5.08
SD	1.52	1.25	1.38	.88	.93	1.00	1.27	1.29	1.10
왜도	.45	-.28	.82	-.53	-.47	-.61	.25	-.19	-.75
첨도	-.38	-.29	-.05	.30	.47	.83	-.30	-.68	1.54

주. $^{*}p$ <.05, $^{**}p$ <.01, $^{***}p$ <.001; 1= 로봇의 HL, 2= 행위능력, 3= 경험능력, 4= 규범존중, 5= 위해, 6= 공평성, 7= 도덕적 책임판단, 8= 로봇기술지식, 9= 로봇인류유익

의 HL이 도덕적 판단에 미치는 직접효과는 유의하지 않았다(B = .068, *ns.*). 매개변수들이 종속변수에 미치는 영향을 살펴보면, 로봇의 경험능력은 도덕적 판단에 유의한 영향을 미치지 않았으나(B = .001, *ns.*), 로봇의 행위능력은 도덕적 판단에 유의한 정적 영향을 미쳤다 (B = .141, p < .01). 즉, 로봇의 행위능력이 크다고 인식될수록 가해자에게 더 큰 도덕적 책임을 부여하는 것으로 나타났다. 이는 로봇의 HL이 도덕적 판단에 직접적인 영향을 주기보다는, 행위능력이라는 매개변수를 통해 간접적으로 영향을 주는 것을 의미한다. 간접효과의 유의성을 알아보기 위하여 95% 신뢰구간 내 부트스트래핑(bootstraping) 분석을 실시하였다 (표본 수 5,000개의 반복 검증). 분석 결과는 표 4에 제시하였다. 로봇의 HL과 도덕적 판단의 관계에서 행위능력의 매개효과는 95%의 신뢰구간 [.012 ~ .270]로 신뢰구간 내에 '0'이 포함되어 있지 않아 유의한 간접효과가 있는 것으로 확인되었다. 반면에 경험능력의 매개효과는 유의하지 않은 것으로 나타났다. 따라서 가설 2-1은 지지, 가설 2-2는 기각되었다.

다음으로, 법적 책임판단에 대한 쌍도덕의 매개효과 분석 결과(표 4), 로봇의 HL이 두 매개변인에 미치는 영향은 위와 동일하다. 로봇의 HL이 법적 책임판단에 미치는 직접효과 역시 위와 동일하게 유의하지 않았다(B = .031, *ns.*). 매개변수들이 종속변수에 미치는 영향을 살펴보면, 로봇의 경험능력은 법적 책임

표 4. 쌍도덕의 병렬매개 효과 검증 (*N*=400)

모델	종속변수	독립변수	*B*	*SE*	*t*	LLCI	ULCI	R^2	*F*
도덕적 판단	경험능력	로봇의 HL	.403***	.041	9.806	.322	.484		
		로봇 인류 유익	-.186***	.056	-3.341	-.296	-.077	.195	96.158***
		로봇 기술 지식	.199***	.049	4.076	.103	.295		
	행위능력	로봇의 HL	.366***	.036	10.112	.295	.437		
		로봇 인류 유익	.171***	.049	3.490	.075	.267	.281	51.600***
		로봇 기술 지식	.166***	.043	3.876	.082	.251		
	도덕적 판단	로봇의 HL	.068	.049	1.401	-.027	.164		
		경험능력	.001	.058	0.018	-.113	.115		
		행위능력	.141*	.066	2.151	.012	.270	.117	6.444***
		로봇 인류 유익	.050	.061	0.827	-.069	.169		
		로봇 기술 지식	-.104*	.052	-2.011	-.206	-.002		
법적 판단	경험능력	로봇의 HL	.357***	.041	8.670	.276	.438		
		로봇 인류 유익	-.186***	.056	-3.341	-.296	-.077	.242	42.088***
		로봇 기술 지식	.199***	.049	4.076	.103	.295		
	행위능력	로봇의 HL	.366***	.036	10.112	.295	.437		
		로봇 인류 유익	.171***	.049	3.490	.075	.267	.281	51.600***
		로봇 기술 지식	.166***	.043	3.876	.082	.251		
	법적 판단	로봇의 HL	.031	.047	0.658	-.062	.124		
		경험능력	.525***	.056	9.306	.414	.636		
		행위능력	-.121	.064	-1.897	-.247	.004	.287	19.687***
		로봇 인류 유익	-.122*	.059	-2.072	-.238	-.006		
		로봇 기술 지식	-.029	.051	-0.571	-.128	.070		

*p <.05, **p <.01, ***p <.001

판단에 유의한 정적효과를 미쳤으나(B = .525, p < .001), 행위능력은 유의한 영향을 미치지 않았다(B = -.121, *ns.*). 즉, 로봇의 경험능력이 크다고 인식될수록 가해자에게 더 큰 법적 책임을 부여하는 것으로 나타났다. 간접효과의 유의성을 알아보기 위하여 95% 신뢰구간 내 부트스트래핑(bootstraping) 분석을 실시하였다(표본 수 5,000개의 반복 검증). 분석 결과는 표 4에 제시하였다. 로봇의 HL과 법적 책임판단의 관계에서 경험능력의 매개효과는 95%의 신뢰구간 [.414 ~ .636]으로 신뢰구간 내에 '0'이 포함되어 있지 않아 유의한 간접효과가 있

는 것으로 확인되었다. 반면에 행위능력의 매개효과는 유의하지 않은 것으로 나타났다. 따라서 가설 2-3은 기각, 가설 2-4는 지지되었다.

쌍도덕과 도덕적 기반의 조절된 매개효과

먼저 도덕적 판단에 대한 도덕적 기반의 조절된 매개효과를 분석하였다(표 5). Muller 등(2005)이 정립한 조절된 매개의 필수 조건에는 개별 매개효과의 유의성은 포함되지 않으므로, 본 연구에서는 두 개의 매개변인을 포함하여 조절된 병렬매개 분석을 실시하였으며, 그 결과는 표 5에 제시하였다. 분석 결과는 다음과 같다. 로봇의 HL은 경험능력(B = .357, p < .001)과 행위능력(B = .366, p < .001) 모두에 유의한 정적 영향을 미치는 것으로 나타났다. 종속변수인 도덕적 판단에 대해서는, 경험능력의 주효과와(B = -.929, p < 0.1), 행위능력의 주효과(B = 1.023, p < .001)는 모두 유의한 것으로 나타났다. 또한, 경험능력과 도덕적 기반의 상호작용 효과(B = .246, p < .001)와 행위능력과 도덕적 기반의 상호작용 효과(B = -.233, p < .001)는 모두 통계적으로 유의하였다. 이는 경험능력과 행위능력이 도덕적 판단

표 5. 로봇의 HL과 도덕적 판단간의 관계에서 쌍도덕과 도덕적 기반의 조절된 매개효과

종속변수	변수	*B*	*S.E.*	*t*	LLCI	ULCI
경험능력	(상수)	1.577	.337	4.677***	.914	2.240
	로봇의 HL	.357	.041	8.670***	.276	.438
	로봇 기술 지식	.199	.049	4.076***	.103	.295
	로봇 인류 유익	-.186	.056	-3.341***	-.296	-.077
행위능력	(상수)	1.419	.297	4.787***	.836	2.002
	로봇의 HL	.366	.036	10.112***	.295	.437
	로봇 기술 지식	.166	.043	3.876***	.082	.251
	로봇 인류 유익	.171	.049	3.490***	.075	.267
도덕적 판단	(상수)	1.218	.948	1.285	-.645	3.082
	로봇의 HL	.068	.049	1.401	-.027	.164
	경험능력 (A)	-.929	.290	-3.207**	-1.499	-.360
	행위능력 (B)	1.023	.264	3.878***	.504	1.541
	도덕적 기반 (C)	.642	.242	2.648**	.165	1.118
	A × C	.246	.072	3.419***	.104	.387
	B × C	-.233	.068	-3.428***	-.366	-.099
	로봇 기술 지식	-.104	.052	-2.011*	-.206	-.002
	로봇 인류 유익	.050	.061	.827	-.069	.169

*p <.05, **p <.01, ***p <.001

에 미치는 영향이 개인의 도덕적 기반 수준에 따라 달라짐을 의미한다. 마지막으로, 조절된 매개효과의 유의성을 검증한 결과, 경험능력을 통한 간접효과의 조절된 매개지수(Index = .0877, 95% CI [.033, .144])와 행위능력을 통한 간접효과의 조절된 매개지수(Index = -.0852, 95% CI [-.144, -.034]) 모두 신뢰구간에 0이 포함되지 않아 유의한 것으로 나타났다. 보다 구체적으로 조건부 간접효과를 살펴본 결과, 경험능력의 간접효과는 도덕적 기반 수준이 낮을 때(B = -.0607, 95% CI [-.127, .001]), 평균 수준에서는 유의하지 않았으며(B = .0110, 95% CI [-.031, .050]), 높은 수준에서는 유의한 정적 효과(B = .0668, 95% CI [.016, .120])를 보였다. 이는 도덕적 기반 수준이 증가함에 따라 경험능력을 통한 간접효과가 정적 방향으로 변화함을 의미한다. 한편, 행위능력의 간접효과는 도덕적 기반 수준이 낮을 때 유의한 정적 효과(B = .111, 95% CI [.051, .174])를 보였으나, 평균 수준과(B = .041, 95% CI [-.014, .094]), 높은 수준에서는(B = -.013, 95% CI [-.086, .053]) 모두 유의하지 않았다. 이러한 결과는 행위능력을 통한 간접효과가 도덕적 기반이 낮은 집단에서만 유의하게 나타남을 시사한다.

다음으로, 법적 판단에 대한 도덕적 기반의

표 6. 도덕적 판단에 대한 쌍도덕의 조건부 간접효과

		B	*S.E.*	95% 신뢰구간	
				LLCI	ULCI
경험능력	-1*SD*	-.061	.032	-.127	.001
	Mean	.011	.020	-.031	.050
	+1*SD*	.067	.026	.016	.120
행위능력	-1*SD*	.111	.032	.051	.174
	Mean	.041	.028	-.014	.094
	+1*SD*	-.013	.036	-.090	.053

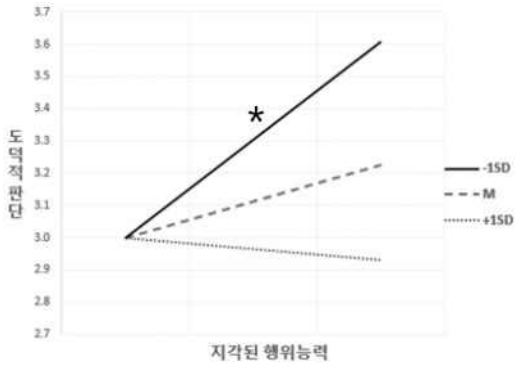


그림 3. 행위능력과 도덕적 기반의 상호작용효과

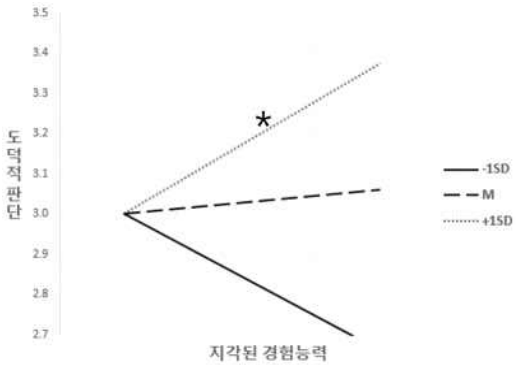


그림 4. 경험능력과 도덕적 기반의 상호작용효과

조절된 매개효과를 분석하였다. 로봇의 HL이 쌍도덕에 미치는 영향은 위와 동일하므로, 종속변수인 법적 책임판단에 대한 결과를 살펴보면, 로봇의 HL이 법적 책임판단에 미치는 직접효과는 유의하지 않은 것으로 나타났다(B = .031, *ns.*). 또한 매개변수의 직접효과를 살펴보면, 경험능력(B = .110, *ns.*)과 행위능력(B = .336, *ns.*) 모두 법적 책임판단에 유의한 영향을 미치지 않는 것으로 나타났다. 조절효과를 검증한 결과, 경험능력 × 도덕적 기반(B = .109, *ns.*) 상호작용은 유의하지 않았으며, 행위능력 × 도덕적 기반(B = -.121, p = .069) 역시 통계적으로 유의하지 않은 것으로 나타났다. 구체적으로 조건부 간접효과를 살펴보면, 경험능력을 통한 간접효과는 도덕적 기반 수준이 낮을 때(B = .160, 95% CI [.087, .243]), 평균일 때(B = .192, 95% CI [.136, .253]), 높을 때(B = .217, 95% CI [.150, .285]) 모두에서 통계적으로 유의한 것으로 나타났다. 그러나 조절된 매개효과의 지수(Index of moderated mediation)는 유의하지 않아(B = .039, 95% CI [-.022, .094]), 해당 간접효과는 도덕적 기반 수준에 따라 유의하게 달라지지 않는 것으로 확인되었다. 반면, 행위능력을 통한 간접효과는 도덕적 기반 수준이 낮을 때와 평균 수준에서는 유의하지 않았으며, 높은 수준에서만 유의한 부적 효과(B = -.078, 95% CI [-.141, -.012])가 나타났다. 그러나 조절된 매개효과의 지수는 역시 유의하지 않았다(B = -.044, 95% CI [-.091, .003]). 따라서 법적 책임판단에 대해서는 로봇의 HL이 경험능력을 통해 일관되게 간접효과를 가지는 것으로 나타났으나, 이러한 간접효과가 도덕적 기반 수준에 따라 유의하게 달라지는 조절된 매개효과는 확인되지 않았다. 이에 따라 가설 3-1, 3-2는 지지되었으나 가설 3-3, 3-4는 기각되었다.

논 의

본 연구는 로봇이 권리와 책임의 주체가 될 수 있는지를 실증적 접근을 통해 알아보고,

표 7. 가설검증 결과

가설	세부 가설	내용	결과
H1	H1-1	인간유사성 → 도덕적 판단	**지지**
	H1-2	인간유사성 → 법적 판단	**지지**
H2	H2-1	인간유사성 → 행위능력 → 도덕적 판단	**지지**
	H2-2	인간유사성 → 경험능력 → 도덕적 판단	기각
	H2-3	인간유사성 → 행위능력 → 법적 판단	기각
	H2-4	인간유사성 → 경험능력 → 법적 판단	**지지**
H3	H3-1	행위능력 → 도덕적 판단 (조절된 매개)	**지지**
	H3-2	경험능력 → 도덕적 판단 (조절된 매개)	**지지**
	H3-3	행위능력 → 법적 판단 (조절된 매개)	기각
	H3-4	경험능력 → 법적 판단 (조절된 매개)	기각

이러한 판단에 영향을 주는 변인들의 역할을 살펴보고자 하였다. 구체적으로, 본 연구에서는 로봇의 외양이 인간과 유사한 정도가 해당 로봇을 손상시킨 가해자의 책임 판단(도덕적 판단, 법적 판단)에 영향을 미칠 것이라고 가정하였으며, 이 과정에서 쌍도덕과 도덕적 기반이 매개 및 조절의 역할을 하는지 살펴보았다. 본 연구의 주요 결과는 다음과 같다.

먼저, 로봇의 인간 유사성은 가해자에 대한 도덕 판단과 법적 판단 모두에 유의한 정적 영향을 주는 것으로 나타났다. 이러한 결과는 유사한 선행 연구들의 결과(신홍임, 2021, 최윤빈, 장대익, 2022; Malle et al., 2016)와 일관된 것으로, 로봇의 외양이 인간같을수록 사람들은 로봇 훼손의 비윤리성을 더 크게 인식한다는 것을 의미한다. 심지어 본 연구에서는 (인간 유사성과 관계없이) 가해자의 상해죄와 학대죄 처벌에 대해 참여자의 약 41%가 '보통이다' 이상으로 동의하는 것으로 나타났다. 물론 이러한 결과가 로봇의 권리를 인정해야 할 직접적 근거가 될 수는 없으나 적어도 기술이 발달하여 로봇이 인간의 외모와 매우 유사해진다면 사람들은 로봇의 생존권을 강하게 침해하는 것으로 보이는 사람에 대해서는 심지어 법적 처벌을 요구할 가능성도 배제할 수는 없다. 다만 이러한 결과는 로봇이 아닌 단순 사물에 대한 '화풀이 기물파손'에 대한 부정적 반응과 구별할 수 없으므로 추후에는 다양한 '사물'조건이 포함된 연구를 통해서 결과를 다시 확인할 필요가 있다.

둘째, 로봇의 인간 유사성이 가해자에 대한 판단을 쌍도덕이 매개하는지를 살펴본 결과, 가해자에 대한 판단이 도덕적 판단인지, 법적 판단인지에 따라 서로 다른 매개경로가 형성되는 것으로 나타났다. 먼저, 도덕적 판단에 대한 분석 결과, 로봇의 인간유사성은 경험능력과 행위능력 모두에 유의한 정적 영향을 미쳤으나, 실제로 도덕적 판단에 영향을 미친 것은 행위능력만이었다. 즉, 로봇이 인간과 유사할수록 사람들은 해당 로봇을 더 '행위할 수 있는 존재'로 인식하게 되며, 이러한 행위능력 지각이 가해자에 대한 도덕적 책임 판단을 증가시키는 것으로 나타났다. 반면 경험능력은 도덕적 판단에 유의한 영향을 미치지 않았다. 흥미롭게도, 법적 판단에서는 정반대의 양상이 발견되었는데, 로봇의 인간유사성은 동일하게 경험능력과 행위능력을 모두 증가시켰지만, 실제 법적 책임 판단에 영향을 미친 것은 경험능력이었다. 즉, 로봇이 인간과 유사할수록 로봇은 '느끼고 고통을 경험할 수 있는 존재'로 지각되며, 이러한 경험능력 지각이 가해자에 대한 법적 책임 판단을 강화한다는 것이다. 행위능력은 법적 책임 판단에 유의한 영향을 미치지 않았다. 도덕적 판단과 법적 판단 모두에서 인간유사성이 미치는 직접효과는 유의하지 않았으며, 쌍도덕을 통한 간접효과만 유의한 것으로 나타났다.

상기 결과는 몇 가지 중요한 이론적 시사점을 보여준다. 우선 두 판단 유형의 매개모델 모두에서 인간유사성의 직접효과는 유의하지 않았고 반드시 마음 지각을 통해서만 가해자 판단에 영향을 미쳤다는 점이다. 이는 로봇의 외형적 인간유사성보다는 기능적 혹은 정서적인 속성이 어떻게 지각되는지가 사람들의 윤리적 태도를 형성하는데 더 결정적일 수 있다는 것을 의미한다. 또한 본 연구는 판단 유형에 따라 쌍도덕의 역할이 서로 달라진다는 것을 최초로 밝혔는데, 이러한 결과는 우선, 기존 Gray 등(2007, 2012)이 제시한 쌍도덕 이론을 지지하는 결과이다. 구체적으로, 사람들이

판단 유형에 따라 행위능력과 경험능력을 사용하는 정도가 달라진다는 점은 Gray 등(2007, 2012)이 행위능력은 주로 책임과 비난과 관련되어 있으며, 경험능력은 도덕적 지위, 즉 보호받을 권리와 자격과 관련되어 있다고 제안한 것과 일관된 결과이다. 즉 인간과 유사할수록 사람들은 로봇을 단순한 사물이 아닌 주체적인 판단과 의도를 가질 수 있는 존재(행위능력)로 인식하고 이러한 존재의 권리를 훼손한 것에 대해 도덕적 비난을 강화할 수 있다. 반대로 법적 판단에 대해서는 인간과 유사할수록 고통을 겪는 피해자로 보는 도덕적 피동체(moral patient)로 지각하는 것이 강해지고 이것이 법적 판단으로 이어질 수 있다. 아울러 본 결과는 법적 판단의 기제와 도덕적 판단의 기제가 질적으로 다를 수 있다는 것을 시사하는데, 도덕적 판단과 법적 판단이 서로 다르다는 fMRI 연구(Schleim et al., 2011) 등이 존재하기는 하나 두 판단 유형에 대한 직접적인 비교 연구는 생각보다 부족한 상태이므로 추가적인 연구가 요구된다.

한 가지 주목할 점은 로봇에 대한 인간의 인식 변화이다. 표 5의 결과를 살펴보면, 인간유사성은 행위능력과 경험능력지각에 비슷한 수준의 영향력을 보이고 있다. 이러한 결과는 행위능력과 경험능력 차원에 따라 다양한 대상의 위치를 연구한 고전적 연구(Gray et al., 2007) 결과와는 상당히 달라진 것이다. 당시 연구에서 로봇은 행위능력은 중간정도이나 경험능력은 거의 없는 존재로 지각되어 책임은 물을 수 있으나 보호의 대상은 아닌 것으로 해석되었다. 이러한 변화는 로봇이 과거보다 훨씬 인간과 가까이 존재한다는 시대적 변화와 함께 본 연구에서 로봇이 손상을 입은 것으로 기술되어 제시 맥락이 달랐던 점이 영향을 미친 것으로 보인다. 최근 국내에서도 인공지능에 대한 태도를 체계적으로 측정하는 척도가 타당화되는 등(이선영 등, 2025) 로봇과 AI에 대한 인식을 정밀하게 포착하려는 흐름이 나타나고 있다. AI나 로봇에 대한 쌍도덕 지각에 대한 선행 연구에서는 연구에 따라 때로는 행위능력이(Kahn et al., 2012; Malle et al., 2015), 때로는 경험능력이(Gray & Wegner, 2009; Gray & Wegner, 2012) 도덕적 판단과 관련되는 것으로 나타났는데, 본 연구는 로봇이 피해자로 기술된 만큼 과거에 꾸준히 높게 지각되던 행위능력과 함께 경험능력 역시 높게 지각되었을 수 있다. 따라서 추후 연구에서는 다양한 맥락을 사용하여 쌍도덕 지각과 도덕적, 법적 판단간의 관계를 살펴볼 필요가 있겠다.

마지막으로, 도덕적 기반의 조절된 매개효과 역시 판단 유형에 따라 다르게 나타났다. 먼저 법적 판단에 대해서는 조절된 매개효과는 모두 유의하지 않아 법적 판단은 개인의 도덕적 기반에 덜 의존하는 것으로 나타났다. 반면, 인간 유사성이 도덕적 판단에 미치는 매개 경로는 개인의 도덕적 기반의 강도에 따라 달라지는 것으로 나타났으며, 이는 선행연구(성주현, 정은경, 2025)와 일관된 결과이다. 조절된 매개효과는 경험능력과 행위능력 경로 모두에서 나타났는데, 구체적으로 경험능력 경로에서의 조절된 매개효과는 도덕적 기반이 높은 집단에서만 유의했던 반면, 행위능력 경로에서의 조절된 매개효과는 도덕적 기반이 낮은 집단에서만 유의한 것으로 나타났다. 이는 도덕적 판단에서 도덕적 기반이 높은 사람일수록 경험능력 경로가, 낮은 사람일수록 행위능력 경로가 더 중요하다는 것을 의미한다. 이러한 결과는 쌍도덕 이론의 확장적 해석 가

능성을 시사한다. 기존 쌍도덕 연구는 도덕적 판단이 행위자와 피해자의 이원적 구조로 이루어진다고 보았으나, 본 연구는 개인의 도덕적 성향에 따라 이 두 요소의 상대적 중요성이 달라질 수 있음을 보여준다. 즉, 동일한 로봇이라도 어떤 개인은 행위능력 중심으로, 다른 개인은 경험능력 중심으로 도덕적 판단을 내린다는 것이다. 이는 쌍도덕의 실제 작동양상과 관련된 것으로, 추후 쌍도덕의 추가적 조절변인들을 지속적으로 탐색할 필요가 있다.

본문에는 제시되지 않았으나 도덕적 기반은 개인기반(위해, 공정성)과 공동체기반(규범존중) 가치로 나뉘어질 수 있으므로 하위요소에 따라 결과가 달라지는지를 추가적으로 확인하였다. 그 결과, 행위능력 경로의 결과는 세 개 기반이 모두 조절된 매개 효과가 유의하여 전체 결과와 유사한 결과를 보인 반면, 경험능력 경로는 규범존중에서만 유의한 조절된 매개효과를 보였다. 즉 규범존중이 높은 사람일수록 경험능력이 도덕적 판단에 미치는 영향이 유의하게 증가하였으며, 경험능력을 통한 간접효과 역시 강화되는 것으로 나타났다. 이는 규범존중이 공동체기반 가치임을 고려해볼 때 다소 의외의 결과이다. 다만, 위해와 공정성이 개인 피해에 민감한 가치라는 점을 고려해 볼 때, 위해와 공정성에 민감한 사람들은 기본적으로 타자에게 해를 가하는 행위를 원래 강하게 비난하는 성향을 지니기 때문에 경험능력의 차이가 도덕적 판단의 추가적 변동을 설명할 여지가 상대적으로 적었을 수 있다. 추후 연구에서는 이러한 가능성을 검증할 수 있는 설계가 필요할 것으로 보인다.

본 연구의 학문적, 실무적 시사점은 다음과 같다. 첫째, 본 연구는 인간에 의해 손상된 로봇의 외양, 특히 인간 유사성이 그 가해자에 대한 도덕적, 법적 판단에 영향을 미친다는 것을 실증적으로 밝힌 최초의 연구이다. 특히 기존 연구들은 AI나 로봇의 도덕적 책임성 여부를 주로 '가해자'로서의 역할에 초점을 맞춘 반면, 본 연구는 '피해자'로서의 로봇이라는 관점을 도입했다는 점에서 의의가 있다. 본 연구 결과는 인간의 외모가 도덕적, 법적 책임 판단에 영향을 미친다는 수많은 선행 연구들과 맥을 같이 하는 것으로, 앞으로 인간과 함께 생활하는 로봇 디자인 등을 개발할 때 로봇 외양의 중요성에 대한 시사점을 줄 수 있다. 둘째, 본 연구는 로봇관련 도덕적, 법적 판단에 행위능력과 경험능력이라는 쌍도덕이 영향을 미친다는 것을 다시 한번 확인하였다. 특히 판단의 유형에 따라 쌍도덕의 매개역할이 달라졌다는 것은 단순한 도덕적 판단과 법적 판단에 작동하는 인간의 심리적 특성이 다를 수 있음을 보여주는 것이다. 이러한 결과는 추후 도덕적 판단 연구 정교화에 기여할 수 있을 것으로 보인다. 마지막으로 도덕적 기반이 법적 판단에는 유의한 영향을 미치지 못하나 도덕적 판단에는 영향을 미친다는 점, 그리고 도덕적 기반이 낮은 사람들에게는 행위능력이, 높은 사람들에게는 경험능력이 중요한 판단기준으로 작동한다는 점을 제시한 것도 본 연구의 중요한 발견점이다. 이러한 결과는 로봇 관련 정책 등을 대중이 수용할 수 있도록 하는데 활용될 수 있을 것이다.

본 연구의 한계 및 향후 연구를 위한 제언은 다음과 같다. 첫째, 본 연구는 로봇의 외양을 정지 이미지 기반으로 제시하였으며, 실제 상호작용이 없는 상황에서 판단을 측정하였다. 그러나 인간-로봇 상호작용은 동적이고 상황 맥락에 민감하게 반응하기 때문에, 실제 서비스 현장이나 실시간 상호작용 상황에서는 다

른 반응이 나타날 수 있다. 향후에는 음성, 제스처, 반응성 등을 포함한 실시간 상호작용 맥락에서 인간 유사성의 효과를 분석하는 것이 필요하다. 또한 본 연구에서는 각 이미지를 개별 점수로 사용하지 않고 이미지와 관계없이 HL점수를 인간 유사성으로 사용하였기에 예비 연구에서 일부 이미지 간 HL 점수 차이가 유의하지 않았아도 본 연구에서 사용하였다. 그러나 이미지 간 차이의 불분명함이 결과의 명료성에 영향을 미쳤을 수도 있으므로 추후 연구에서는 좀 더 확실하게 구분되는 로봇 이미지를 사용하여 결과를 확인할 필요가 있다. 둘째, 로봇에 대한 판단은 단순히 인간과의 유사성 뿐 아니라 귀여움이나 친숙함 등과 같은 기타 특성이 영향을 미칠 수 있다(Sun et al., 2025). 따라서 추후 연구에서는 귀여움 등과 같은 인상 속성을 측정하여 통제할 필요가 있다. 셋째, 본 연구의 초점이 로봇의 외양에 맞춰져 있어서 본 연구에서는 고려하지 못하였으나 가해자의 '화풀이와 손상'에 대한 도덕적 판단은 로봇이 아닌 '물건'에 대해서도 발생할 수 있다. 대상에 대한 인간유사성을 측정하였으나 이것이 과연 로봇으로써의 인간 유사성인지 아니면 단순한 물건의 인간 유사성인지는 구분하지 못하였다. 아울러 소유권의 문제도 추가로 고려되어야 하는데, 본 연구에서는 자신의 소유 로봇으로 한정하였으나 타인의 로봇을 훼손하는 것에 대해서는 또 다른 판단이 내려질 수 있다. 따라서 추후 연구에서는 로봇이 아닌 물건, 로봇, 그리고 인간 모두를 포함하는 포괄적 디자인, 그리고 소유권 조건을 적용해 도덕적 판단을 다시 살펴볼 필요가 있다. 넷째, 본 연구에서는 법적 판단을 묻는 문항에 '상해죄'와 '학대죄'에 해당하는지를 묻는 문항이 포함되어 있었는데, 엄밀히 말하면 이것은 태도보다는 법적 범주를 묻는 문항으로 해석될 수도 있으므로 추후에는 이를 분리 측정하여 연구결과를 다시 확인할 필요가 있다. 다섯째, 로봇과 관련된 판단에는 정서적 불쾌감이 영향을 미치는 것으로 알려져 있는데(Gray & Wegner, 2012; Mathur & Reichling, 2016), 본 연구에서는 로봇 이미지에 대한 정서적 불쾌감은 고려하지 못하였다. 따라서 추후 연구에는 이를 고려한 설계가 필요하다. 마지막으로, 본 연구는 한국인을 대상으로 한 단일 문화권 표본에 기반하고 있어 문화 간 일반화에는 제한이 있다. 도덕 판단은 문화적 규범과 윤리 기준에 따라 큰 차이를 보일 수 있으므로, 향후에는 다양한 문화권 샘플을 대상으로 결과의 재현 가능성과 문화적 특수성을 비교 분석할 필요가 있다.

결론적으로, 본 연구는 로봇의 인간유사성이 도덕적 판단에 영향을 미치며, 이러한 책임 판단에 미치는 영향이 단일한 경로가 아니라, 판단의 유형(도덕적 vs 법적)에 따라 서로 다른 심리적 경로를 통해 작동함을 최초로 밝혔다는데 그 의의가 있다. 추후 연구에서 인간-로봇 상호작용 맥락에서 책임 귀속을 이해하기 위해서는 도덕적 판단과 법적 판단을 구분하여 접근할 필요가 있으며, 본 연구에서 제시된 도덕적 기반과 같은 다양한 조절변인의 역할도 정교하게 규명되어야 할 것이다.

참고문헌

곽소나 (2014). The Impact of the Robot Appearance Types on Social Interaction with a Robot and Service Evaluation of a Robot. 디

자인학연구, 27(2), 81-93.

박희원 (2021, 11월 20일). 로봇박사 “이재명 로봇 학대 논란, 좋은 화두 던졌다”. *노컷뉴스*. https://www.nocutnews.co.kr/news/5660328

성주현, 정은경 (2025). AI서비스 실패에 대한 책임지각: 도덕적 기반의 조절효과. 한국심리학회지: 문화 및 사회문제, 31(2), 245-268.

신홍임 (2021). 불쾌한 골짜기: 로봇 속성의 의인화, 마음지각 및 도덕적 처우의 관계. 감성과학, 24(4), 3-16.

안정용, 성용준 (2021). 의인화가 인공지능의 윤리적 책임 평가에 미치는 영향: 지각된 자유의지의 매개효과를 중심으로. 한국심리학회지: 소비자・광고, 22(2), 155-177.

윤은진, 정은경 (2025). Human-Likeness and Moral Responsibility in AI: The Roles of Dyadic Morality and Emotional Unpleasantness. 미래사회, 16(3), 218-239.

이상돈, 정채연 (2017). 자율주행자동차의 윤리화의 과제와 전망. IT와 법연구, 15, 281-325.

이선영, 김효미, 안현의 (2025). 한국판 인공지능에 대한 태도 척도(ATTARI-12)의 타당화: 긍・부정 문항 방법효과에 따른 쌍요인 모형 검증. 한국심리학회지: 문화 및 사회문제, 31(2), 269-292.

이정구 (2025, 11월 13일). 젠슨 황 CEO “이젠 피지컬 AI 시대”라는데... 피지컬 AI란 무엇?. *조선일보*. https://www.chosun.com/economy/weeklybiz/2025/11/13/POGIJSMM5VEU7ITALLNHFPTVYY/

이혜원, 정은경 (2021). 자율주행 자동차 사고 상황에 대한 한국인의 윤리적 판단. 한국심리학회지: 일반, 40(1), 105-129.

임지연 (2021, 11월 6일). 사족 보행 로봇을 걷어차는 것은 도덕적으로 잘못됐을까. *한국일보*. https://www.hankookilbo.com/news/article/A2021110517130003017

장혜지, 조은경 (2025). 범죄유형에 따라 음주가 가해자와 피해자에 대한 비난 판단에 미치는 영향. 한국심리학회지: 법, 16(1), 89-109.

정은경, 박상혁, 이수란, 손영우 (2016). 한국인을 대상으로 한 도덕적 기반 질문지 적용에 대한 연구. 한국심리학회지: 사회 및 성격, 30(3), 47-61.

정은경, 손영우 (2011). 진보와 보수의 도덕적 가치 판단의 차이: 간통죄를 중심으로. 한국심리학회지: 일반, 30(3), 727-741.

최윤빈, 장대익 (2022). AI의 의사결정에 대한 도덕판단에서 의인화가 미치는 영향 - 쌍도덕 이론을 중심으로 -. 인지과학, 33(4), 169-203.

Aragón, O. R., Clark, M. S., Dyer, R. L., & Bargh, J. A. (2015). Dimorphous expressions of positive emotion: Displays of both care and aggression in response to cute stimuli. Psychological science, 26(3), 259-273.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59-64.

Broadbent, E., Kumar, V., Li, X., Sollers 3rd, J., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with display screens: A robot with a more humanlike face display is perceived to have more mind and a

better personality. *PLOS ONE, 8*(8), e72589.

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics, 75*(3), 659-671.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16.

Darling, K., Nandy, P., & Breazeal, C. (2020). Empathic concern and the effect of stories in human-robot interaction. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 770-775.

De Graaf, M. M., & Allouch, S. B. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics and Autonomous Systems, 61*(12), 1476-1486.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864-886.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*(3-4), 143-166.

Gerdes, E. P., Dammann, E. J., & Heilig, K. E. (1988). Perceptions of rape victims and assailants: Effects of physical attractiveness, acquaintance, and subject gender. *Sex Roles, 19*, 141-153.

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2003)*(pp. 55-60). IEEE.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology, 101*(2), 366.

Gray, K., Gray, H. M., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*(3), 505-520.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125-130.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101-124.

Gulati, A., Martínez-Garcia, M., Fernández, D., Lozano, M. A., Lepri, B., & Oliver, N. (2024). What is beautiful is still good: The attractiveness halo effect in the era of beauty filters. *Royal Society Open Science, 11*(11), 240882

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98-116.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133*(4),

55-66.

Hall, M., Debowska, A., & Hales, G. K. (2024). The effect of victim attractiveness and type of abuse suffered on attributions of victim blame and credibility in intimate partner violence: A vignette-based online experiment. *Violence Against Women, 30*(9), 2396-2416.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... & Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology, 48*(2), 303.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary. E. H., Reichert. L. A., Freier. G. N., & Severson, R. L. (2012, March). Do people hold a humanoid robot morally accountable for the harm it causes?. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*(pp. 33-40).

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction, 19*(1-2), 61-84.

Koban, K., & Wieringa, M. S. (2024). Think once, think again: Exploring the harm-made mind effect through dual-process theory. *International Journal of Social Robotics, 16*(11), 2189-2203.

Laakasuo, M., Palomaki, J., & Kobis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics, 13*(7), 1679-1688.

Liberman-Pincu, E., Parmet, Y., & Oron-Gilad, T. (2023). Judging a socially assistive robot by its cover: the effect of body structure, outline, and color on users' perception. *ACM Transactions on Human-Robot Interaction, 12*(2), 1-26.

Lieber-Milo, S. (2022). Pink purchasing: Interrogating the soft power of Japan's kawaii consumption. *Journal of Consumer Culture, 22*(3), 747-761.

Lv, X., Liu, Y., Luo, J., Liu, Y., & Li, C. (2021). Does a cute artificial intelligence assistant soften the blow? The impact of cuteness on customer tolerance of assistant service failure. *Annals of Tourism Research, 87*, 103114.

Maeder, E. M., Yamamoto, S., & Saliba, P. (2015). The influence of defendant race and victim physical attractiveness on juror decision-making in a sexual assault trial. *Psychology, Crime & Law, 21*(1), 62-79.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*(pp. 117-124).

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*(pp. 125-132). IEEE.

Mathur, M. B., & Reichling, D. B. (2016).

Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition, 146*, 22-32.

Müller, B. C., Gao, X., Nijssen, S. R., & Damen, T. G. (2021). I, robot: How human appearance and mind attribution relate to the perceived danger of robots. *International Journal of Social Robotics, 13*(4), 691-701.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*(6), 852.

Nishiwaki, Y., Furukawa, M., Yoshikawa, S., Karatas, N., & Okada, M. (2017, March). iBones: A weak robot to construct mutual coordination for handing out tissue packs. *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction,* 413-413.

Nilsson, A., Erlandsson, A., & Vastfjall, D. (2020). Moral foundations theory and the psychology of charitable giving. *European Journal of Personality, 34*(3), 431-447.

Ogba, K. T., Onyishi, I. E., & Chukwuorji, J. C. (2021). Self-disclosure in intimate relationships: Moderating role of online moral foundations. *Current Psychology, 40*(10), 4785-4790.

Palomaki, J., Kunnari, A., Drosinou, M., Koverola, M., Lehtonen, N., Halonen, J., Koverola, M., & Laakasuo, M. (2018). Evaluating the replicability of the uncanny valley effect. *Heliyon, 4*(11), e00939.

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. *In Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI '06),* 218-225.

Premathilake, G. W., & Li, H. (2024). Delving into emotions: A study of social robots' use in hotels. *Proceedings of the 32nd European Conference on Information Systems (ECIS)* (Paper 1997).

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009, March). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction*(pp. 245-246).

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32-70.

Schleim, S., Spranger, T. M., Erk, S., & Walter, H. (2011). From moral to legal judgment: The influence of normative context in lawyers and other academics. *Social Cognitive and Affective Neuroscience, 6*(1), 48-57.

Schombs, S., Klein, J., & Roesler, E. (2023). Feeling with a robot—the role of anthropomorphism by design and the tendency to anthropomorphize in human-robot interaction. *Frontiers in Robotics and AI, 10*, 1149601.

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., & Young, J. E. (2015). Poor thing! Would you feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* 125-132.

Sherman, G. D., & Haidt, J. (2011). Cuteness and

disgust: The humanizing and dehumanizing effects of emotion. *Emotion Review, 3*(3), 245-251.

Strait, M. K., Floerke, V. A., Ju, W., Maddox, K., Remedios, J. D., Jung, M. F., & Urry, H. L. (2017). Understanding the uncanny: Both atypical features and category ambiguity provoke aversion toward humanlike robots. Frontiers in psychology, 8, 1366.

Sun, X., Cheng, Y., & Zhang, M. (2025). The influence of cuteness types and sound characteristics of in-vehicle robots on user trust and experience in automatic driving. *Kybernetes, 55*(7), 3194-3215.

Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLOS ONE, 12*(7), e0180952.

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior, 38*, 75-84.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219-232.

Wieringa, M. S., Müller, B. C. N., Bijlstra, G., & Bosse, T. (2024). Robots are both anthropomorphized and dehumanized when harmed intentionally. *Communications Psychology, 2*(1), 72.

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2021). Robots at work: People prefer－and forgive－service robots with perceived feelings. *Journal of Applied Psychology, 106*(10), 1557-1575.

Yu, X., Xu, Z., Song, Y., & Liu, X. (2022). The cuter, the better? The impact of cuteness on intention to purchase AI voice assistants: A moderated serial-mediation model. Frontiers in Psychology, 13, 1036848.

Zhang, W., Slade, E. L., & Pantano, E. (2024). Humanlike service robots: A systematic literature review and research agenda. *Psychology & Marketing, 41*(12), 3157-3181.

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human-robot interaction. *International Journal of Social Robotics, 7*(3), 347-360.

논문 투고일 : 2026. 02. 21
1 차 심사일 : 2026. 04. 22
게재 확정일 : 2026. 06. 06

한국심리학회지 : 문화 및 사회문제
Korean Journal of Culture and Social Issues
2026, Vol. 32, No. 3, 329～357.

Can Robots Be Abused? Exploring the Roles of Human-Like Appearance, Dyadic Morality, and Moral Foundations

Min Choi **Eun Kyoung Chung**

Department of Psychology, Kangwon National University

This study examined how the human-likeness of robots influences judgments of responsibility (moral and legal) toward a perpetrator who harms a robot, in a context where humanlike robots are increasingly prevalent. Based on the theory of dyadic morality, perceived agency and experience were specified as mediators, and moral foundations were included as a moderator. Through a pilot study, four robot images differing in levels of human-likeness were selected. A total of 400 adults were randomly assigned to one of the conditions and asked to read a scenario in which a robot was physically damaged by its owner, after which they evaluated the relevant variables. The results showed that robot human-likeness had a positive effect on both moral and legal responsibility judgments. Parallel mediation analysis revealed that moral judgment was mediated only by perceived agency, whereas legal judgment was mediated only by perceived experience, indicating that different dimensions of mind perception operate depending on the type of judgment even in the same victim scenario. Moderated mediation analysis further showed that moral foundations significantly moderated both pathways only in moral judgment, while no moderation effect was found in legal judgment. These findings suggest that robot appearance elicits mind perception along the dimensions of agency and experience, and that moral judgments vary flexibly depending on individuals' moral values, whereas legal judgments tend to follow a relatively more stable structure.

Key words : Human-likeness, Moral responsibility, Mind perception, Dyadic morality, Moral foundation theory