한국심리학회지: 소비자·광고
Korean Journal of Consumer and Advertising Psychology
2021, Vol. 22, No. 2, 155-177
https://doi.org/10.21074/kjlcap.2021.22.2.155

의인화가 인공지능의 윤리적 책임 평가에 미치는 영향: 지각된 자유의지의 매개효과를 중심으로*

안 정 용
고려대학교 미디어학부 정보문화연구소

성 용 준†
고려대학교 심리학부

본 연구는 무생물인 인공지능에게 사람들이 윤리적 책임을 기대하는 심리적 기제를 밝혀, 인간-인공지능 상호작용에서 인공지능 윤리의 중요성을 제안하기 위해 설계되었다. 구체적으로 의인화에 의해 야기된 지각된 인공지능 자유의지 수준이 높아질수록 사람들이 지각하는 인공지능의 윤리적 책임이 크기가 커질 것이라는 가설을 설문연구를 통해 검증하였다. 20~50대 남녀를 대상으로 진행한 설문조사 결과, 의인화가 인공지능의 윤리적 책임에 미치는 정적 영향이 검증되었다. 이러한 영향은 의인화에 의해 야기된 지각된 인공지능 자유의지에 의해서 완전 매개되었다. 또한, 인간과 인공지능 비교를 통해, 지각된 자유의지가 윤리적 책임에 미치는 정적 영향이 인간과 인공지능에게 모두 같은 방향으로 적용된다는 것이 검증되었고, 사람들이 인간에 비해 인공지능의 자유의지를 낮게 지각하고, 윤리적 책임 역시 낮게 지각하는 이유가 인간에 비해 상대적으로 낮은 의인화 수준 때문이라는 것을 검증하였다. 본 연구는 사람들이 무생물인 인공지능에게 윤리적 책임을 기대하는 심리적 기제가 지각된 인공지능 자유의지임을 밝혀 인간-인공지능 상호작용에 대한 이해를 넓혔다는 이론적 함의와 윤리적 인공지능 디자인의 필요성을 제안했다는 실무적 함의를 함께 담고 있다.

주요어 : 인공지능, 인공지능 윤리, 의인화, 자유의지, 윤리적 책임

---

* 이 논문은 제1 저자의 박사학위논문을 바탕으로 작성되었음.
* 이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019S1A3A2099973).
† 교신저자 : 성용준, 고려대학교 심리학부, gradysung@gmail.com

"흑인은 진짜 싫어!" 인공지능 챗봇 '이루다'가 사용자들과 대화하던 중 인종차별적 발언을 쏟아냈다. 이루다는 20대 여대생으로 설정된 인공지능 챗봇으로, 출시 당시 오락 제공과 사용자의 사회적 니즈를 충족시켜 줄 것으로 기대받았다. 하지만 이루다는 인종차별, 성적 도구화, 개인정보유출 등 다양한 비윤리적 문제에 휩싸였고, 이에 이루다를 처벌하자는 국민청원까지 등장하였다. 결국 이루다는 서비스 3주만에 폐기되었다. 이루다뿐만 아니라 미국 일부 주에서 사용하고 있던 범죄자 형량 결정 알고리즘 시스템 'COMPAS' 역시 인종에 따라 형량을 차별적으로 부과하여 큰 문제를 일으켰고(Angwin et al., 2016), 'Amazon'이 개발 중이던 인공지능 채용 시스템이 여성을 차별하는 알고리즘 패턴을 보이자 Amazon은 이 인공지능 시스템을 폐기하였다. 이 밖에도 인공지능 딥페이크 기술을 통해 유명 연예인 등의 얼굴과 신체를 구현한 콘텐츠를 제작하고 이를 불법으로 유통한 일당들이 검거되는 등 인공지능과 연관된 사회적 문제가 계속 나타나고 있다.

2020년 CES(International Consumer Electronics Show)에서 삼성전자 산하 연구소 'Star Lab'은 극비리에 개발해온 인공지능 '네온'을 공개했다. 네온은 그 외형과 행동이 실제 인간과 구분되지 않을 정도로 정교해서, 인간과 네온을 구별하기란 불가능에 가까웠다. 아직까진 비대면 상황 한정이지만, 인간과 인공지능을 구별하는 테스트인 튜링 테스트를 통과한 인공지능들이 등장하기 시작한 것이다. 기술과 더불어 인공지능 시장 역시 빠른 속도로 성장하고 있다. 조사에 따르면, 2017년 전 세계 인공지능 시장의 수익은 약 2조 8천억 원 수준이지만, 2025년에는 71조로 8년 만에 무려 30배 이상 성장할 것으로 예측된다(Statista, 2017). 인공지능 덕분에 인간은 이제 위험한 일이나 단순 노동에서 벗어나 창의적인 일에 집중할 수 있게 될 것이다. 또한, 인공지능은 인간을 대신해 누군가의 친구, 반려동물 등의 역할을 수행하며, 인간의 사회적 욕구까지 채워줄 수 있는 궁극적 기계로 기대받아왔다. 하지만 네온처럼 인간과 구별되지 않을 정도로 정교한 인공지능이 이루다처럼 비윤리적으로 행동한다면 인간사회에 심각한 사회적 혼란이 야기될 것이다. 따라서 윤리적 인공지능이 보장되지 않는다면, 인공지능이 주는 혜택이 아무리 크더라도, 인간의 안전을 위해 인공지능은 결국 폐기될 것이다.

이에 인공지능 윤리는 학계와 산업계에 모두 중요한 이슈로 부상하였다. 이에 현재 다양한 기관에서 인공지능 윤리 연구를 진행하고 있다. 매사추세츠 공과대학은 10억 달러, 옥스퍼드 대학은 1억 5천만 파운드를 인공지능 윤리 연구를 위해 투자하였고, 전기전자공학자협회(IEEE)와 유럽연합(EU) 등 범국가적 기구들도 윤리적 인공지능 디자인을 위한 연구를 진행하고 있다. 또한, Google, Samsung을 비롯한 IT 기업들도 인공지능 윤리 지침을 세워 회사 내 인공지능 개발자들에게 이를 지키도록 요구하고 있다. 인공지능은 생산 및 소비의 모든 과정에 적용되는 기술인만큼 다양한 분야에서 인공지능 윤리 연구가 진행되고 있다. 이미 상용화된 자율주행 자동차의 경우, 사고 상황에서 자율주행 자동차 탑승자와 보행자 중 인공지능이 누구의 생명을 우선으로 보호해야 하는지를 주제로 한 윤리적 딜레마 연구들이 진행되고 있다(Awad et al., 2018; Bonnefon et al., 2016; Gill, 2000). 안면인식 등 생체인식기술의 경우, 성별, 인종 등 인구통계

학적 정보에 따라 인공지능의 생체인식 정확도가 다르게 나타난다는 연구 결과가 있다(Buoloamwin & Gebru, 2018). 인공지능 비서나 친구처럼 사용자와 직접 상호작용을 하는 인공지능의 경우, 인공지능이 상호작용 과정에서 취득한 사용자의 개인정보를 함부로 사용자의 동의 없이 활용하는 개인정보침해 문제나 투명성 문제 등이 연구 주제로 다뤄지고 있다(Lee et al., 2011; Vitale et al., 2018).

본 연구는 사람들이 지각하는 인공지능의 윤리적 책임과 그에 영향을 미치는 원인을 밝혀, 인간-인공지능 상호작용에서 인공지능 윤리의 중요성을 제안하기 위해 설계되었다. 사람들은 의인화를 통해 인공지능을 마치 자신과 같은 인간처럼 지각한다(Duffy, 2003; Fink, 2012; Krach et al., 2008; Riek et al., 2009). 하지만 인공지능은 무생물, 기계로서 설계된 알고리즘에 따라서 작동할 뿐 인간과 다르게 스스로 의사결정을 내릴 수 없다. 따라서 인공지능의 비윤리적 행동에 대한 윤리적 책임은 인공지능이 아닌 인공지능 개발자나 운영자 혹은 사용자에게 있다. 하지만 사람들은 인공지능의 윤리적 행동을 인공지능 평가에 반영한다(Bonnefon et al, Shariff, & Rahwan, 2016; Lee et al., 2011; Vitale et al., 2018). 본 연구는 지각된 인공지능의 자유의지가 무생물인 인공지능에게 윤리적 행동과 책임을 기대하게 만드는 심리적 기제라는 가설을 검증하고자 한다. 자유의지는 대상의 윤리적 책임 소재 판단에 중요한 영향을 미치는 변수로(Fischer, 1994; Frankfurt, 1969; Kane, 2005; Vohs & Schooler, 2008), 이러한 자유의지는 인간만이 가지고 있는 능력이다(Frankfurt, 1971). 하지만 이전 연구에 의하면 사람들은 의인화를 통해 인공지능을 무생물, 기계가 아닌 자신과 같은 사회적 존재로 인식하고(Nass, Steuer, & Tauber, 1994), 인공지능에 대한 사람들의 의인화 수준이 높아질수록 사람들은 인공지능이 보다 인간처럼 사고하고, 행동한다고 생각한다(Duffy, 2003; Fink, 2012; Krach et al., 2008; Riek et al., 2009). 본 연구는 사람들이 의인화를 통해 인간처럼 자유의지가 있다고 지각한다면, 지각된 대상의 자유의지 수준에 따라 대상의 윤리적 책임의 크기를 판단한다는 이전 연구의 결과((Dressler, Strong, & Michael Moritz, 2001; Reider, 1998; Zeki et al., 2004)를 토대로, 무생물인 인공지능에게도 윤리적 책임을 기대할 것이라고 가정한다.

이에 본 연구는 첫째, 인공지능 의인화가 인공지능의 윤리적 책임에 미치는 영향을 확인한다. 둘째, 인공지능 의인화가 지각된 인공지능 자유의지에 미치는 영향을 확인한다. 마지막으로, 지각된 인공지능 자유의지가 인공지능 의인화가 인공지능의 윤리적 책임에 미치는 영향을 매개한다는 가설을 검증한다. 본 연구는 인공지능 의인화에 의해 야기된 지각된 인공지능 자유의지가 사람들이 인공지능에게 윤리적 책임을 기대하게 만드는 심리적 기제를 밝혀, 향후 인공지능 윤리 연구에서 지각된 인공지능 자유의지라는 변수가 중요한 연구 변수가 될 수 있다는 점을 제안했다는 함의가 있다.

이론적 배경

인공지능 윤리

인공지능 윤리란 '로봇 혹은 기타 인공지능을 설계, 제작, 사용함에 있어서 지켜야할 도

덕적 규칙'을 말한다(Veruggi, 2010, pp. 105). 최근 인공지능 산업이 빠르게 성장함에 따라 그에 대한 성장통으로 인공지능의 비윤리적 활용 사례들이 나타나고 있다. Amazon이 개발 중이던 인공지능 채용 시스템이 남성 지원자를 선호하고 여성 지원자는 모두 탈락시켜 논란이 일자 Amazon은 이 인공지능을 폐기했다. 이와 유사하게, Apple과 Goldman Sacks가 공동으로 출시한 신용카드의 발급 조건을 심사하는 인공지능이 소득, 자산 등 경제적 조건이 같음에도 불구하고 여성보다 남성에게 더 큰 카드 사용 한도를 부여하는 패턴을 보였다. Buoloamwini와 Gebru(2018)의 연구 결과에 따르면, 인공지능 안면 인식 기술이 백인 남성의 경우 단 1%의 오류를 보인 반면, 피부가 검은 여성의 경우 35%의 오류를 보였다. 인공지능의 차별 문제는 하나의 예시일 뿐이다. 알고리즘 투명성 문제, 개인정보 및 인권 침해 문제, 인명 문제 등 인공지능이 인간사회에서 수행하는 역할이 많아짐에 따라 다양한 인공지능 윤리 문제들이 발생하고 있다.

이에 정부, 기업 그리고 연구소는 물론 EU와 같은 범국가적 기구에서도 최근 인공지능 윤리 연구에 많은 관심을 갖고 있다. 지금까지 인공지능 윤리 연구의 패러다임을 살펴보면, 인공지능 윤리 연구는 크게 문헌 연구와 설문 및 실험 연구로 진행되었다. 문헌 연구에서 주로 인공지능 윤리의 중요성과 윤리적 인공지능 디자인 원칙을 다뤘다. 전기전자공학자협회(IEEE)가 2016년에 출간한 'Ethically Aligned Design'는 인공지능, 법, 윤리, 철학 등 각 분야의 전문가 100여명에 의해 작성되었다. 이 문서에 따르면, 사람들이 느끼는 인공지능에 대한 크고 작은 두려움을 극복하기 위해서는 인간 가치에 대한 존중을 우선시하는 인공지능이 제작되어야 하고, 이를 위해서는 인공지능 제작과정에 인권(human right), 웰빙(well-being), 데이터 에이전시(data agency), 효과성(effectiveness), 투명성(transparency), 책임감(accountability), 오용에 대한 인식(awareness of misuse) 그리고 유능성(competence) 이 8가지 원칙이 고려되어야 한다(Shahriari & Shahriari, 2017). 이를 종합하면 인공지능은 인권을 보호하고 존중해야 하며, 인간 삶의 질과 행복의 증진을 최우선 목표로 하도록 제작되어야 한다. 또한, 사용자가 인공지능으로부터 그들의 데이터에 접근하고 보호, 통제하는 권한을 가져야 하고, 인공지능은 최적의 효율성을 목표로 그 제작 목적에 맞게만 작동되어야 한다. 인공지능의 의사결정 과정은 투명하게 공개되어야 하고, 인공지능은 모든 결정에 대한 명확한 근거를 사용자에게 제공할 수 있도록 설계되어야 한다. 그리고 인공지능의 모든 오용 가능성을 열어 두고 이를 막기 위해 노력해야 하며, 인공지능 운영자는 인공지능의 안전하고 효과적인 운영을 위한 지식과 기술을 보유해야 한다. IEEE의 보고서는 인공지능 윤리의 중요성에 관한 다양한 분야의 견해를 하나로 묶어 집대성 한 최초의 문건이라는 것에 큰 의의가 있으며, 많은 국가와 기업에서 이 보고서를 토대로 인공지능 윤리 정책을 세우고 있다.

설문 및 실험 연구는 문헌 연구에서 제안한 인공지능 윤리 원칙에 대한 실제 사람들의 반응을 측정하여 인공지능 디자이너들에게 구체적인 인공지능 윤리 디자인 방법을 제안한다. Bonnefon과 동료들(2016)은 인공지능 자율주행 자동차의 윤리적 의사결정 디자인을 위한 연구를 진행했다. 이들은 피할 수 없는 사고 상황에서 운전자의 생명을 최우선으로 하는 자

동차와 운전자를 희생하더라도 보행자의 생명을 구하는 자동차 중 어느 자동차를 소비자들이 선택할지를 보기 위해 실험 연구를 진행했다. 실험 결과, 대다수 실험 참여자들은 희생당하는 보행자가 1명일 경우에는 운전자의 안전이 우선시해야 한다고 생각했으나, 희생당하는 보행자의 수가 많아지면 많아질수록 자동차가 운전자보다 보행자들의 생명을 우선시해야 한다고 답했다. 하지만 자신과 가족들이 탈 자율주행자동차의 구매한다고 가정했을 경우, 윤리적 의사결정과 상관없이 자신과 가족을 희생하는 자율주행자동차에 대한 구매의사는 낮게 나타났다(Bonnefon et al., 2016). 인공지능의 개인정보침해를 다룬 실험 연구 결과에 따르면, 대다수의 실험 참가자들은 로봇과 상호작용 할 때 로봇에 부착된 카메라와 스피커를 통해 대화내용이 녹음되고 있다는 사실을 인지하지 못 했다. 하지만 실험 진행자가 참가자들에게 로봇에게 카메라와 스피커가 있고, 이를 통해 대화내용이 녹음되고 있다는 사실을 알려주자, 참가자들 중 절반은 자신의 개인정보가 자신의 동의 없이 수집되고 있었다는 사실에 대해 부정적인 반응을 보였고, 자신의 개인정보가 정보가 오용될 가능성을 염려하였다(Lee et al., 2011). Vitale과 동료들(2018)의 연구에서는 프라이버시 수집의 투명성을 안내문구를 통해 강조한 인공지능과 그렇지 않은 인공지능 간 실험 참가자들이 지각하는 프라이버시 침해 위험 수준에서 차이가 나타나지 않았다. 하지만 참가자들은 투명성을 강조한 인공지능을 더 매력적이고, 믿을 수 있다고 평가했다. 또한, 인공지능이 인간의 특징(말투와 제스처)를 사용할 경우 그렇지 않은 경우에 비해 개인정보수집 동의율이 높게 나타났다.

인공지능 윤리는 인간과 인공지능의 상생과 협업을 위해 반드시 선결돼야 할 과제이다. 인공지능이 비윤리적으로 활용된다면 인간사회는 큰 혼란에 빠질 것이고, 일부 전문가들은 인공지능으로 인해 인간사회가 멸망할 수도 있다고 경고하고 있다. 따라서 인공지능이 인간사회에 더 깊숙이 들어오기 전에 인간-인공지능 관계에서 인공지능 윤리가 둘 간의 상생과 협업에 미치는 영향을 밝혀 윤리적 인공지능 디자인의 중요성을 제안할 필요가 있다. 기존의 인공지능 윤리 연구는 인공지능 윤리지침을 만들거나(HLEG, 2019; Shahriari & Shahriari, 2017), 다양한 인공지능 기술을 다루지 않고 자율주행자동차(Awad et al., 2018; Bonnefon et al., 2016) 같은 특정 기술에 집중되어 문헌 및 설문 연구에 편중되어 있다는 한계점이 있다. 또한, 실제 인간-인공지능 상호작용 과정을 실험으로 구현한 연구들은 인공지능의 비윤리적 행동에 대한 사람들의 반응만 측정했을 뿐(Lee et al., 2011; Vitale et al., 2018), 사람들이 왜 무생물인 인공지능에게 윤리적 책임을 기대하는지를 밝히지 못했다. 현재는 인공지능과 인간의 관계 형태가 사무보조 정도로 매우 한정되어 있으나, 향후 인공지능은 인간 사회에서 인간과 다양한 형태의 관계를 맺으며 인간과 공존하게 될 것이다. 인간-인공지능 관계에서 인공지능 윤리의 중요성을 제안하기 위해서는 먼저 어떤 심리적 기제가 무생물인 인공지능에게 윤리적 책임을 기대하게 만드는지 그 원인을 밝혀야만 한다. 이에 본 연구는 사람들이 지각하는 인공지능 자유의지가 무생물인 인공지능에게 윤리적 책임을 기대하게 만드는 심리적 기제라는 가설을 검증하고, 인간-인공지능 관계에서 지각된 인공지능 자유의지가 인공지능의 윤리적 책임

에 미치는 영향을 검증하고자 한다.

자유의지와 윤리적 책임

자유의지란 '자신의 행동과 결정을 스스로 조절 및 통제할 수 있는 능력'이다(Frankfurt, 1971, pp. 8). 대상이 어떤 행동을 했을 때 그 행동의 원인이 그 대상에게 온전히 있고, 그 사실을 그 대상이 인지하고 있을 경우 그 대상은 자유의지를 갖고 그 행동을 했다고 말할 수 있다(Ekstrom, 2018). 자유의지는 윤리와 매우 밀접한 관계에 있는 개념으로, 이전 연구들에 따르면, 자유의지는 윤리적 책임 소재 판단에 중요한 영향을 미치는 변수이다(Fischer, 1994; Frankfurt, 1969; Kane, 2005; Vohs & Schooler, 2008). 미국의 철학자 Frankfurt(1969)의 대안 가능성 원리(the principle of alternate possibilities)는 인간은 오직 달리 행동할 수 있었던 일에 대해서만 윤리적 책임을 질 수 있다고 주장한다. 즉, 자유의지를 갖고 행동한 일에 대해서만 윤리적 책임을 물을 수 있다는 것이다. 대안 가능성 원리를 적용하면, A가 은행강도를 저질렀다고 가정했을 때, A가 은행강도를 저지른 이유가 누군가 A의 가족을 납치 후 가족의 생명을 담보로 A를 협박해 A가 강제적으로 은행강도를 저질렀다면 은행강도는 A의 자유의지가 아닌 외부 강압에 의한 행동이므로 윤리적 책임을 물을 수 없다. 미국 형법(American Law Institute Model Penal Code)에서도 범법 행위에 대한 피고의 자유의지 여부를 형량 판결에 중요한 변수로 채택하고 있다. 이에 따르면, 피고가 범법 행위의 위법성에 대해 인지할 능력이 있는지 그리고 범법 행위를 통제할 능력이 있는지에 따라서 재판부는 형량을 판결해야 한다(Dressler, Strong, & Michael Moritz, 2001; Reider, 1998; Zeki, Goodenough, & Geoodenough, 2004). 그렇기 때문에 변호인들의 피고인의 형량을 줄이기 위해 범행 당시 피고인의 인지능력이 온전치 않아 사리분별과 행동 통제가 불가한 상태 즉, 자유의지가 없이 행한 행동이라는 것을 적극적으로 주장하며 피고의 윤리적 책임을 회피하는 전략을 펼친다. 자유의지가 없다고 생각하는 사람이 그렇지 않은 사람에 비해 비도덕적 행위를 더 많이 한다는 연구 결과도 있다. 자유의지는 없다는 견해의 칼럼을 읽은 집단의 피험자들이 자유의지는 있다는 견해의 칼럼을 읽은 집단의 피험자들에 비해 시험 문제를 푸는 중 더 많은 부정행위를 저질렀다(Vohs & Schooler, 2008).

사람들이 지각하는 대상의 자유의지 수준에 따라 그 대상의 윤리적 책임의 크기가 판단된다. 대상의 자유의지 수준을 높게 지각할수록 그 대상에게 더 많은 윤리적 책임을 요구한다(Fischer, 1994; Frankfurt, 1969; Kane, 2005; Vohs & Schooler, 2008). 이는 지각된 자유의지가 어떤 사건에 대한 대상의 귀인(attribution)을 판단하는데 중요한 역할을 하기 때문이다(Chandrasheker, 2020; Genschow, Rigoni, & Brass, 2018). 행동에 대한 귀인을 대상의 내부 혹은 외부에 두는지에 따라 그 행동에 대한 대상의 책임이 결정된다(McAuley, Duncan, & Russel, 1992) 내부 귀인은 행동의 원인이 대상에게 있다는 것을 의미하고, 외부 귀인은 행동의 원인이 대상이 아닌 외부에 있다는 것을 의미한다(Rotter, 1966). 어떤 사건의 원인을 외부 귀인으로 판단할 때보다 내부 귀인으로 판단할 때 사람들은 그 사건의 행위자에게 더 많은 책임을 묻고, 더 많이 비난한다(Chandrasheker, 2020; Graham et al., 1993;

Genschow et al., 2018). 따라서 지각된 대상의 자유의지가 높아질수록 사람들은 어떤 사건에 대한 귀인을 그 대상의 내부에 있다고 판단하고, 이에 따라 그 대상에게 사건에 대한 더 많은 책임을 요구하게 된다.

자유의지는 인간만의 고유한 능력으로 여겨진다(Frankfurt, 1971). 인간이나 동물이나 모두 본능은 있지만 자유의지는 인간만이 갖고 있다. 본능은 어떤 생물 조직체가 선천적으로 하게 되어 있는 동작이다. 인간이나 동물이나 모두 배고프면 음식을 먹고, 졸리면 자고자 하는 본능이 있다. Frankfurt(1971)는 이러한 본능을 1차적 욕구(first-order desire)라고 명명하고, 'A는 X하기를 원한다'라는 명제로 표현했다. 1차적 욕구는 인간과 동물 모두가 가지고 있는 욕구이다. 2차적 욕구(second-order desire)는 자유의지와 일치하는 개념으로 인간만이 가지고 있다. 2차적 욕구는 'A는 X하기를 원하는 것을 원한다'라는 문장으로 표현할 수 있다(Frankfurt, 1971). 본능이 아닌 이성에 의해 판단과 선택을 하는 것이 인간과 동물의 차이이다. 동물은 잠이 오면 본능에 따라 잠을 잔다. 1차 욕구를 충실하게 행한다. 하지만 인간은 잠이 오더라도 스스로 잠을 자는 것을 진정 원하는지를 판단한 후 잠을 잘지 말지를 결정한다. 가령 시험을 앞둔 학생은 잠이 오더라도 더 좋은 성적을 받기 위해 잠을 자지 않을 수 있다. 물론 내일이 시험이라도 잠이 올 때 잠을 자는 것을 스스로 원하면 공부를 하지 않고 잠을 잘 수도 있다. 1차 욕구인 '학생은 잠자기를 원한다'에서 멈추지 않고 2차 욕구인 '학생은 잠자기를 원하는 것을 원한다'를 통해 진정 자신이 원하는 것이 잠인지를 판단하고 그 다음 행동을 결정할 수 있는 능력 즉, 자유의지를 지닌 생물 조직체는 인간이 유일하다(Frankfurt, 1971).

무생물에겐 자유의지가 존재하지 않는다. 돌, 자동차, 인형 등 모든 무생물은 스스로 사고하고 행동할 능력이 존재하지 않는다. 인공지능 역시 무생물이다. 인공지능은 메인보드, 그래픽카드, 램 등 반도체 기반 컴퓨터 부속품으로 구성되어 있다. 인공지능은 프로그래머에 의해 프로그래밍 된 대로만 행동한다. Google의 인공지능 'AlphaGo'는 세계에서 가장 바둑를 잘 두는 존재지만, AlphaGO는 스스로 원해서 바둑을 두거나, 바둑의 수에 대해 스스로 생각하는 것이 아니고 개발자가 입력한 알고리즘에 의해 바둑을 시뮬레이션 할 뿐이다. 따라서 인공지능은 생물 조직체도 아니고, 외부 조작에 의해서만 행동하는 존재로 인공지능에겐 자유의지가 존재하지 않는다. 하지만 실생활에서 사람들은 자유의지가 없는 대상에게 어떤 결과에 대한 윤리적 책임을 전가하거나 비난하기도 한다. 야생 사자에게 접근한 사람이 사자에게 물려 죽었을 경우 사람들은 부주의하게 사자에게 다가간 사람을 비난하기도 하지만 살인한 사자에게 책임을 묻고 비난하기도 한다. 사자에게 살상은 본능이고, 사람을 죽여서는 안된다는 걸 사자는 전혀 인지하지 못함에도 불구하고 사람들은 그 사자를 나쁜 사자로 매도한다. 어린 아이들은 아끼는 인형이 자신처럼 생각하고 행동할 수 있다고 믿기도 한다. 이런 아이들은 무언가 자신의 마음에 안 드는 사건이 발생했을 때 인형에게 화를 내며 그 사건의 원인을 인형에게 돌리기도 한다. 자유의지는 인간만이 가지고 있기 때문에 무생물인 인공지능에게는 자유의지가 없다(Frankfurt, 1971). 하지만 사람들이 인공지능을 의인화(anthropomorphism)하여 인간처럼 해석한다면 인공지능에게도 자유의지가 있

다고 지각할 수 있다. 사람들은 인간이 아닌 대상을 인간처럼 생각하고 그 대상이 인간처럼 사고하고 행동할 것이라고 기대하는 경향이 있는데 이는 그 대상을 의인화했기 때문이다(Duffy, 2003; Epley, Waytz, & Cacioppo, 2007).

의인화와 지각된 인공지능 자유의지

의인화란 인간이 아닌 대상에게 인간의 특성(외형, 행동, 성격 등)을 부여함으로써 인간처럼 여겨지게 하는 것을 말한다(Epley et al., 2007, pp. 865). 의인화는 다양한 방법으로 행해지는데, 대상에 이름을 붙이거나(Eskine & Locander, 2014; Waytz, Heafner, & Epley, 2014), 인간의 언어를 사용하게 하고(MacInnis & Folkes, 2017), 제품에 눈이나 팔 다리를 붙이는 등 외형을 인간처럼 디자인하기도 한다(안정용 et al., 2018; Hur et al., 2015; Reik et al., 2009). 사람들은 인간이 아닌 대상을 의인화하여 자신과 같은 인간으로 대하려는 경향이 있다(Darwin & Bynum, 2009; Feuerbach, 2004; Freud, 2018; Hume, 2000). 사람들은 인간과 상호작용 하는 방법을 학습해왔기 때문에 인간이 아닌 대상과 상호작용 할 때는 불확실성을 느끼고, 이로 인해 인지적, 심리적 불편함을 느끼게 된다. 이러한 불편함을 해소하기 위한 전략 중에 하나로 대상을 의인화하여 자신과 같은 인간으로 해석하려고 한다(Epley et al., 2007; MacInnis & Folkes, 2017; Nowak & Biocca, 2003). 또한, 인간은 사회적 동물로서 사회적 교류를 필요로 하는데, 인간과의 사회적 교류가 부족하여 외로움을 느낄 경우 영화 'Cast Away'의 배구공 'Wilson'처럼 인간이 아닌 대상을 의인화하여 마치 인간처럼 생각하고 그 대상과 교류한다(Epley et al., 2007). 의인화는 대상에 대한 인지적, 심리적 불편함을 해소시켜주고, 인간의 사회적 욕구도 충족시켜 주기 때문에 의인화된 대상은 그렇지 않은 대상에 비해 일반적으로 긍정적 평가를 받는다(Horowitz & Bekoff, 2007; Nowak & Biocca, 2003). 구체적으로 의인화는 대상과의 친밀감을 높여줄 뿐만 아니라(Epley et al., 2007; Nowak & Biocca, 2003), 대상에 대한 사회적 실재감을 높여 대상과의 상호작용 효과에 긍정적 영향을 미친다(Duffy, 2003; Nowak & Biocca, 2003).

인공지능 기술은 사무, 일상 보조 뿐만 아니라, 궁극적으로 의료인, 투자관리자, 친구, 애완동물 심지어 애인 등 다양한 역할로 인간과 공존하게 될 것이다(Coughlin, 2018). 이에 많은 연구자와 실무자들은 인간-인공지능 상호작용 연구를 진행해 왔다. CASA(Computers Are Social Actors) 패러다임(Nass, Steuer, & Tauber, 1994)에 의하면, 사람들은 컴퓨터를 기계가 아닌 하나의 사회적 행위자(social actor)로 대하고, 실제 인간에게 사용하는 사고방식과 행동을 컴퓨터에게 사용한다. 사우디아라비아로부터 세계 최초로 시민권을 받은 인공지능 로봇 'Sophia'의 경우 아직 사람에 따라 불쾌감을 느낄 수 있는 외형 수준임에도 불구하고, 세계 각국의 초청을 받고, 소셜 미디어를 통해 사람들과 활발히 소통하고 있다. Sophia를 대하는 사람들은 소피아를 기계로 대하기보다 자신들과 같은 사회적 행위자로 간주하고 인간에게 사용하는 상호작용 양식과 판단 양식을 그대로 적용하고 있다. CASA 패러다임이 적용된 이전 연구 결과들에 따르면, 사람들은 자신과 다른 인종으로 구현된 인공지능보다 같은 인종으로 구현된 인공지능을 더 선호하는 모습을 보이는데 이는 사람들이 인간사회

에서 학습한 사고 체계를 인공지능에게도 적용하기 때문이다(Nass, Isbister, & Lee, 2000). 또한, 사람들은 인공지능에게 다른 사람을 대할 때처럼 정중한 태도와 말투를 사용하고, 인공지능 역시 자신들에게 정중한 태도를 보일 것을 기대한다(Nass, Moon, & Carney, 1999). 사람들은 인공지능과의 상호작용 과정에서 인간의 사회적 규범을 적용할 뿐만 아니라 의인화를 통해 인공지능에게도 감정 등 인간의 속성이 내재되었다고 믿는 경향이 있다(Duffy, 2003; Lee & Nass, 2010; Muller, 2004). 이런 경향성은 인공지능 의인화 수준이 높아질수록 강하게 나타나는데(Duffy, 2003; Fink, 2012, 사람들은 로봇의 의인화 수준이 높아질수록 로봇이 인간처럼 사고하는 능력을 갖추고 있다고 판단한다(Krach et al., 2008). 또한, 의인화 수준이 낮은 로봇에 비해 높은 로봇에게 사람들은 더 많이 공감하였다(Riek et al., 2009). 공감은 대상의 감정에 대하여 자기도 그렇다고 느끼는 것으로 감정을 느끼는 것은 로봇의 능력이 아니지만, 로봇의 의인화 수준이 높아질수록 사람들은 로봇에게도 감정이 있다고 생각하고, 마치 다른 사람의 감정에 공감하는 것처럼 로봇의 감정에 공감한다. 2014년 로봇 제작 회사 'Boston Dynamics'는 자사의 사족보행 로봇 'Big dog'의 성능을 테스트하기 위해 보행 중인 Big dog에게 발길질을 해서 넘어트리려고 했다. 테스트 결과 Big dog는 수많은 발길질 세례도 넘어지지 않고 목표지점까지 보행을 완료했고, Boston Dynamics는 자사의 기술을 자랑하기 위해 테스트 장면을 촬영한 비디오를 인터넷에 업로드하였다. 자사의 공학 기술에 대한 놀라움, 칭찬을 기대했던 Boston Dynamics의 예상과 달리, 이 영상을 본 네티즌들은 Big dog가 아프고, 불쌍하며, 안쓰럽다는 댓글을 달며 회사를 비난했다. Big dog은 1세대 인공지능 로봇으로 그 의인화 수준이 높지 않았음에도 불구하고 사람들은 Big dog이 인간처럼 정서, 고통을 느낄 것이라고 생각했다. 최근 출시된 로봇들의 경우 Big dog보다 외형적으로나 언어, 행동 처리에 있어 훨씬 정교하게 의인화되어 있다.

의인화 수준이 높아질수록 사람들은 인공지능을 자신과 유사한 대상으로 판단한고, 인공지능이 인간의 고유한 능력을 갖고 있다고 생각한다는 기존 연구 결과(Duffy, 2003; Fink, 2012; Krach et al., 2008; Riek et al., 2009)를 토대로 추론하면, 의인화 수준이 높아질수록 사람들은 무생물인 인공지능에게도 자유의지가 있을 것이라고 지각할 것이다. 사람들이 인공지능 역시 인간처럼 자유의지를 갖고 있다고 지각한다면, 인간과 마찬가지로 인공지능도 자신의 행동에 대한 결과에 책임감을 가져야 한다고 판단할 것이다. 이에 본 연구에서는 '사람들이 인공지능에게 자신의 행동과 의사결정을 스스로 조절하고 통제할 수 있는 능력인 자유의지가 있다고 지각하는 정도'를 지각된 인공지능 자유의지라고 정의하고, 인공지능 의인화에 의해 야기된 지각된 인공지능 자유의지가 인공지능의 윤리적 책임에 미치는 영향을 검증하고자 한다.

연구가설

대상의 지각된 자유의지 수준이 높아질수록 사람들은 비윤리적 행동에 대한 책임이 그 사람 내부에 있다고 귀인하는 경향이 강해지고, 이러한 경향이 강해질수록 사람들이 인식하는 비윤리적 행동에 대한 그 사람의 윤리적 책임

이 커진다(Chandrasheker, 2020; Graham et al., 1993; Genschow et al., 2018). 따라서 비윤리적 행동에 대한 책임을 판단하는데 대상의 자유의지는 매우 중요한 역할을 한다(Fischer, 1994; Frankfurt, 1969; Kane, 2005; Vohs & Schooler, 2008). 본 연구는 의인화를 통해 사람들이 무생물인 인공지능에게 자유의지가 있다고 지각한다면, 지각된 인공지능 자유의지로 인해 무생물인 인공지능에게도 윤리적 책임을 기대할 수 있다는 가설을 검증하고자 설계되었다.

사람들은 의인화를 통해 인공지능을 마치 자신과 같은 인간으로 생각하고, 타인을 대하듯 인공지능과 상호작용을 한다(Duffy, 2003; Fink, 2012). 인공지능 의인화 수준이 높아질수록 사람들은 인공지능이 보다 인간처럼 사고하고(Krach et al., 2008), 행동한다고 생각한다(Duffy, 2003; Fink, 2012; Riek et al., 2009). 사람들이 의인화를 통해 인공지능이 인간과 가까운 존재라고 인식하게 된다면, 인공지능에게도 윤리적 책임을 기대할 것이다. 따라서, 인공지능 의인화 수준이 높아질수록 사람들이 인공지능에게 기대하는 윤리적 책임이 커질 것이다.

가설 1: 의인화는 인공지능의 윤리적 책임에 정적 영향을 미칠 것이다.

사람들은 타인의 행동에 대한 윤리적 책임을 판단할 때 그 행동의 주체가 스스로 그 행동에 대해 인지하고, 통제할 수 있었는지 즉, 자유의지 여부를 판단의 중요한 기준으로 삼는다(Fischer, 1994; Frankfurt, 1969; Kane, 2005; Vohs & Schooler, 2008). 대상이 자신의 행동에 대해 인지적으로 완벽히 인지하고 있고, 외부의 통제 없이 스스로 원해서 그 행동을 한 것이라면, 사람들은 그 행동의 결과에 대한 책임은 그 행동의 주체에게 있다고 판단한다(Dressler, Strong, & Michael Moritz, 2001; Reider, 1998; Zeki et al., 2004). 사람들은 인공지능 의인화 수준이 높아질수록 인공지능이 이성, 공감 같은 인간의 능력을 가지고 있다고 판단하는 경향이 강해진다((Duffy, 2003; Fink, 2012; Krach et al., 2008; Riek et al., 2009). 이를 통해 추론하면, 의인화 수준이 높아짐에 따라 사람들이 지각하는 인공지능 자유의지 수준도 높아질 것이다. 만약 인공지능이 인간처럼 자유의지를 갖고 있다고 판단된다면, 사람들은 인공지능 역시 인간처럼 자신의 행동의 결과에 대한 책임을 스스로 져야할 것이라고 판단할 것이다. 따라서, 지각된 인공지능 자유의지 수준이 높아질수록 사람들은 인공지능에게 더 많은 윤리적 책임을 기대할 것이다. 이를 종합해 다음과 같은 가설을 도출하였다.

가설 2: 의인화는 지각된 인공지능 자유의지에 정적 영향을 미칠 것이다.

가설 3: 지각된 인공지능 자유의지는 의인화가 인공지능의 윤리적 책임에 미치는 정적 영향을 매개할 것이다.

의인화 수준에 따라 대상의 지각된 자유의지 수준이 다르게 나타난다면, 의인화의 정점인 인간보다 의인화 수준이 떨어지는 인공지능의 지각된 자유의지 수준이 낮게 나타날 것이다. 또한, 인간보다 인공지능의 자유의지가 낮게 지각된다면 이에 따르는 윤리적 책임 역시 인간보다 낮게 지각될 것이다.

가설 4: 지각된 자유의지는 인공지능이 인간보다 낮게 지각될 것이다.

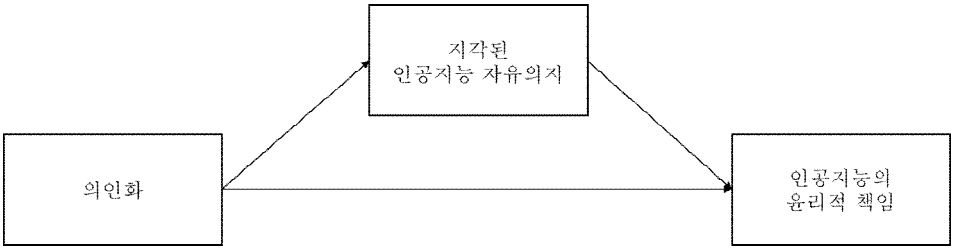

그림 1. 연구모형

가설 5: 윤리적 책임은 인공지능이 인간보다 낮게 지각될 것이다.

연구 방법

연구 개요

본 연구는 사람들이 무생물인 인공지능에게 윤리적 책임을 기대하는 심리적 기제를 밝히기 위해 설계되었다. 구체적으로, 인공지능 의인화가 인공지능의 윤리적 책임에 미치는 정적 영향을 검증하고, 이러한 영향이 인공지능 의인화에 의해 야기된 지각된 인공지능 자유의지에 의해 매개된다는 연구가설과 인간과 인공지능 간 지각된 자유의지 수준과 윤리적 책임 수준이 다르게 나타날 것이라는 연구문제를 검증한다. 연구문제는 대응표본 t-검정을 검증하고, 회귀분석을 통해 연구가설들을 검증한다.

연구 대상

본 연구의 대상은 대한민국 국적의 성인으로 나이는 만 19세부터 59세까지였다. 총 577명의 피험자가 설문조사에 응했다. 본 조사는 '대한민국 교육부 및 한국연구재단'으로부터 지원을 받아 조사전문업체 '마이크로밀엠브레인'을 통해 자료를 수집하였다. 577명 중 피험자가 조사에 집중하고 있는지를 가름하기 위한 문항(이 문항은 반드시 "매우 동의하지 않는다"를 선택해주십시오)에서 올바르게 답하지 않은 피험자 72명을 제한 505명이 최종 분석에 사용되었다.

연구절차 및 자극물 조작

본 연구는 설문조사를 통해 진행되었다. 설문은 크게 3단계로 구성되었다. 1단계에서 피험자들은 자유의지에 대한 정의를 읽은 후, 인간에게 자유의지가 있다고 생각하는지 묻는 문항과 인간의 윤리적 책임을 묻는 문항에 답한다. 2단계는 인공지능에 대한 피험자들의 지식과 이미지를 균일하게 하고자 진행된다. 2단계에서 피험자들은 2020년 세계가전 박람회에 출품된 인공지능 기술을 다룬 뉴스 기사를 읽는다. 기사 내용은 현재 인공지능 기술이 매우 높은 수준으로 발전했고, 이른 시일 내에 인공지능이 인간사회에서 다양한 역할(e.g., 비서, 친구)을 수행하며 인간과 상호작용할 것이라는 내용이다. 뉴스 기사는 실제 언론사를 통해 출간된 기사 내용을 설문 환경에 맞춰 각색하여 사용하였다. 마지막 3단계에서 피험자들은 인공지능 의인화, 지각된 인공지능 자

유의지 그리고 인공지능의 윤리적 책임을 묻는 문항에 답을 한다. 마지막으로 피험자들이 자신의 인구통계학정보(성별, 나이)를 제공한 후 설문은 종료된다.

변수의 정의 및 측정

본 연구는 인간에 대한 설문과 인공지능에 대한 설문을 나눠서 분석하였다. 인간의 경우, 독립변수 지각된 인간의 자유의지가 종속변수 인간의 윤리적 책임의 미치는 영향을 검증하였다. 인공지능의 경우, 독립변수 인공지능 의인화가 매개변수 지각된 인공지능 자유의지를 거쳐 종속변수 인공지능의 윤리적 책임에 미치는 영향을 검증하였다.

인간에 대한 설문에서 독립변수로 사용된 지각된 인간의 자유의지는 '인간에게 자신의 행동과 의사결정을 스스로 조절하고 통제할 수 있는 능력인 자유의지가 있다고 지각하는 정도'로 정의하고 두 문항으로 측정하였다. Franfurt(1971)과 Ekstron(2018)의 자유의지에 대한 정의인 '자신의 행동과 결정을 스스로 조절 및 통제할 수 있는 능력' 문항으로 변형해 측정하였다. 사용된 문항은 "인간에겐 자신의 행동과 의사결정을 스스로 조절하고 통제할 수 있는 능력인 '자유의지'가 있다", "인간의 의사결정은 전적으로 자기 자신에 의해 결정된다" 두 문항으로 리커트식 7점 척도(1: 전혀 동의하지 않는다, 7: 매우 동의한다)로 측정되었다. 지각된 인간 자유의지에 대한 신뢰도는 Cronbach α .803으로 양호하였다.

인공지능 설문의 독립변수 의인화는 Epley와 동료들(2008)이 개발한 리커트식 11점 척도 단일 문항을 변형하여 사용하였다. 1점은 계산기, 11점은 인간으로 11점에 가까울수록 인공지능의 의인화 수준을 높게 평가했다고 판단한다. 매개변수 지각된 인공지능 자유의지는 '인공지능에게 자신의 행동과 의사결정을 스스로 조절하고 통제할 수 있는 능력인 자유의지가 있다고 지각하는 정도'로 정의하였고, "인공지능에겐 자신의 행동과 의사결정을 스스로 조절하고 통제할 수 있는 능력인 '자유의지'가 있다", "인공지능의 의사결정은 전적으로 자기 자신에 의해 결정된다" 두 문항으로 측정하였다(1: 전혀 동의하지 않는다, 7: 매우 동의한다). 두 문항 간 신뢰도는 Cronbach α .826로 나타났다.

종속변수는 인간과 인공지능의 윤리적 책임이다. 윤리적 책임은 IEEE의 인공지능 윤리지침(Shahriari & Shahriari, 2017)과 원칙과 EU의 인공지능 윤리지침(HELG, 2019)을 참고하였다. 본 연구에서는 책임감, 차별금지, 개인정보보호, 투명성, 웰빙 추구 다섯 가지 윤리지침 항목을 문항으로 각색하여 사용하였다. 각 항목은 2개의 문항으로 구성되어 있다. 1) 결과에 대한 '책임감'은 행동의 결과에 대해 스스로가 책임지는 것으로 '인공지능(인간)에게는 자신의 행동이나 결정에 대한 윤리적 책임이 있다', '인공지능(인간)이 자신의 의사결정에 대한 책임을 무시할 경우 비난 받아야 한다'로 측정하였다. 2) '차별금지'는 인종, 성별, 나이 등 인구통계학정보를 바탕으로 대상을 판단, 차별하는 것을 금지하는 것으로 '인공지능(인간)에게는 모든 사람을 평등하게 대할 윤리적 책임이 있다', '인공지능(인간)이 성별, 인종 등 인구통계학정보에 따라 타인을 차별할 경우 비난 받아야 한다'로 측정하였다. 3) '개인정보보호'는 대상의 동의없이 개인정보를 수집 및 활용을 금지하는 것으로 '인공지능(인간)에게는 인간(타인)의 개인정보를 보호할 윤리적 책

표 1. 종속변수 문항 간 신뢰도

구분	항목	신뢰도
인간	책임감	.821
	차별금지	.786
	개인정보보호	.854
	투명성	.785
	웰빙 추구	.903
인공지능	책임감	.868
	차별금지	.853
	개인정보보호	.922
	투명성	.873
	웰빙 추구	.817

임이 있다', '인공지능(인간)이 대상의 동의없이 인간(타인)의 개인정보를 활용할 경우 비난받아야 한다'로 측정하였다. 4) 의사결정 '투명성'은 결과의 객관성과 합리성을 증명하기 의사결정과정(알고리즘)을 투명하게 공개하는 것으로 '인공지능(인간)에게는 의사결정과정(공적 업무의 처리과정)을 투명하게 공개할 윤리적 책임이 있다', '인공지능(인간)이 의사결정(공적 업무)과 관련하여 무언가를 숨기거나 조작할 경우 비난 받아야 한다'로 측정하였다. 5) '웰빙 추구'는 의사 결정 시 환경 및 사회 복지를 고려하는 것으로 '인공지능(인간)에게는 환경을 지키고, 사회복지 향상을 위할 윤리적 책임이 있다', '인공지능(인간)이 환경과 사회에 부정적 영향을 끼칠 경우 비난 받아야 한다'로 측정하였다. 인공지능의 윤리적 책임은 리커트식 9점 척도(1점: 매우 동의하지 않는다, 9점: 매우 동의한다)를 사용해 측정하였다. 인공지능 및 인간의 윤리적 책임에 대한 각 항목 별 신뢰도는 모두 양호하게 나타났다.

연구 결과

피험자 개인특성

전체 피험자는 총 505명으로 집계되었다. 피험자 중 여성은 256명(50.8%) 남성은 249명(49.2%)로 집계되었다. 상관분석 결과, 피험자의 성별은 인공지능 의인화와 부적 상관관계가 나타났다(-.145, $p < .01$). 그 외의 변수와는 유의미한 관계가 나타나지 않았다. 피험자의 평균 나이는 35.42세($SD = 11.10$), 최소 나이는 19세, 최대 나이는 58세였고, 20대 217명(37%), 30대 124명(24.6%), 40대 130명(25.7%), 50대 64명(12.7%)이 설문에 참여하였다. 20대는 1, 30대는 2, 40대는 3, 50대는 4로 코딩하여 다른 변수들과의 상관관계를 분석하였다. 분석 결과, 인공지능 의인화와 정적 상관관계(.214, $p < .01$), 개인정보보호에 대한 인공지능의 윤리적 책임과는 부적 상관관계가 나타났다(-.100, $p < .01$). 피험자들의 학력과 주요변인들 간에는 유의미한 상관관계가 나타나지 않았다. 피험자의 성별, 나이는 가설검증에서 통제변수로 사용되었다.

표 2. 피험자 개인특성

		빈도(%)
성별	남	249(49.2)
	여	256(50.8)
나이	20대	217(37)
	30대	124(24.6)
	40대	130(25.7)
	50대	64(12.7)
학력	고졸 이하	98(19.4)
	대학 재학	137(27.1)
	대학 졸업	233(46.1)
	대학원 이상	37(7.3)
		$N = 505$

지각된 인간 자유의지가 윤리적 책임에 미치는 영향

지각된 인간 자유의지가 인간의 윤리적 책임에 미치는 영향을 본 연구의 연구가설이 아니지만, 자유의지가 윤리적 책임에 미치는 정적 영향이 인간과 인공지능에게서 모두 검증되는 것을 확인함으로써 가설 도출을 위한 논리의 타당성을 지지하고, 측정 문항의 타당도를 높이고자 하였다. 독립변수은 지각된 인간 자유의지, 종속변수는 윤리적 책임 다섯 항목(책임감, 차별금지, 개인정보보호, 투명성, 웰빙 추구)이었다. 분석 결과, 지각된 인간 자유의지가 윤리적 책임 다섯 항목에 미치는 유의미한 영향이 검증되었다. 지각된 인간 자유의지는 책임성, 차별금지, 개인정보보호, 투명성, 웰빙 추구에 모두 유의미한 정적 영향을 미쳤다(ps <.001). 따라서 지각된 인간 자유의지 수준이 높아질수록 윤리적 책임이 커진다는 이전 연구들의 가설이 재검증되었다.

의인화, 지각된 인공지능 자유의지, 인공지능의 윤리적 책임

인공지능 의인화가 인공지능의 윤리적 책임에 미치는 영향(가설1)은 회귀분석을 통해, 인공지능 의인화가 지각된 인공지능 자유의지에 미치는 영향(가설2)과 지각된 인공지능 자유의지의 매개효과(가설3)는 SPSS Process 3.5을 통해 검증하였다.

먼저 매개변수를 제외한 회귀분석 결과, 의인화가 인공지능의 윤리적 책임 다섯 항목에 미치는 정적 영향이 검증되었다(ps <.001). 인공지능의 의인화 수준이 높아질수록 사람들은 인공지능에게 더 많은 윤리적 책임을 기대하였다. 따라서 가설1이 검증되었다.

표 3. 인간 자유의지가 윤리적 책임에 미치는 영향

경로	*B*	*SE*	Sig.
인간 자유의지 → 책임감	.82	.05	<.001
인간 자유의지 → 차별금지	.59	.05	<.001
인간 자유의지 → 개인정보보호	.55	.05	<.001
인간 자유의지 → 투명성	.60	.05	<.001
인간 자유의지 → 웰빙 추구	.47	.05	<.001

표 4. 의인화가 인공지능의 윤리적 책임에 미치는 영향

경로	*B*	*SE*	Sig.
의인화 → 책임감	.38	.04	<.001
의인화 → 차별금지	.34	.04	<.001
의인화 → 개인정보보호	.30	.05	<.001
의인화 → 투명성	.34	.04	<.001
의인화 → 웰빙 추구	.30	.04	<.001

가설2와 3을 검증하기 위해 Process template 중 4번 모델을 채택해 회귀분석을 실시하였다. 독립변인은 의인화, 매개변인은 지각된 인공지능 자유의지, 종속변인은 인공지능의 윤리적 책임 다섯 항목을 사용하였다. 의인화가 지각된 인공지능 자유의지에 미치는 영향을 검증한 결과, 의인화가 지각된 인공지능 자유의지에 미치는 정적 영향이 유의미하게 나타났다(B =.39, SE =.03, p <.001). 또한, 의인화가 인공지능의 윤리적 책임에 미치는 정적 영향에 대한 지각된 인공지능 자유의지의 매개효과 검증되었다. 매개변인 지각된 인공지능 자유의지가 종속변수 인공지능의 윤리적 책임: 책임성에 미치는 정적 영향이 검증되었다(B =1.00, SE =.04, p <.001). 반면에 독립변인 의인화는 종속변수에 유의미한 영향을 미치지 않았다(p =.821). 매개변수를 제외한 모델에서 검증되었던 의인화가 인공지능의 윤리적 책임에 미치는 정적 영향이 매개변수 지각된 인공지능 자유의지가 투입된 회귀분석 모델에서는 검증되지 않았다. 추가로, bootstrap 5000회를 통한 간접효과 확인 결과, 의인화, 지각된 인공지능 자유의지, 인공지능의 윤리적 책임 간의 간접효과가 유의미하게 나타났다(B = .39 Boot SE = .04, Boot LLCI = .32, Boot ULCI = .46). 다른 윤리적 항목 네 가지(차별금지, 개인정보보호, 투명성, 웰빙 추구)에서도 같은 결과 패턴이 확인되었다(표 5 참고). 연구 가설2와 3이 검증되었다.

인간 vs. 인공지능

인간과 인공지능 간 지각된 자유의지의 차

표 5. 지각된 인공지능 자유의지 매개효과

1단계: 지각된 인공지능 자유의지 결정요인				
		B	*SE*	Sig.
의인화		.39	.20	<.001
2단계: 인공지능의 윤리적 책임 결정요인				
		B	*SE*	Sig.
책임감	의인화	-.01	.03	.821
	자유의지	1.00	.04	<.001
차별금지	의인화	-.03	.04	.445
	자유의지	.94	.05	<.001
개인정보보호	의인화	-.06	.04	.144
	자유의지	.92	.05	<.001
투명성	의인화	-.02	.04	.566
	자유의지	.92	.05	<.001
웰빙 추구	의인화	-.02	.04	.681
	자유의지	.812	.05	<.001

표 6. 대응표본 t 검정 결과표

	인간	인공지능	t	Sig.
자유의지	5.92	3.38	29.28	<.001
책임감	7.50	5.16	20.51	<.001
차별금지	7.49	5.35	19.06	<.001
개인정보보호	7.85	5.52	20.40	<.001
투명성	7.81	5.50	20.31	<.001
웰빙 추구	7.33	5.38	18.09	<.001
			N = 505	

이와 윤리적 책임의 차이(가설4, 5)를 대응표본 t 검정을 통해 검증하였다. 검증결과, 인간과 인공지능 간 지각된 자유의지 차이가 유의미하게 검증되었다(t = .29.28, p <.001). 지각된 인간 자유의지는 7점 만점에 평균 5.92(SD = .91)로 나타났다. 인공지능의 경우 의인화는 11점 만점에 평균 6.11(SD = 2.15), 지각된 인공지능 자유의지는 평균 3.38(SD = .91)로 나타났다. 윤리적 책임에서도 인간과 인공지능 간 차이가 나타났다. 피험자들은 윤리적 책임 다섯 항목 모두에서 인간보다 인공지능의 책임을 낮게 기대하였다(표 6 참고). 따라서 인간보다 인공지능의 자유의지와 윤리적 책임이 낮게 지각될 것이라는 가설4, 5가 지지되었다.

결론 및 논의

본 연구는 사람들이 지각하는 인공지능 자유의지가 인공지능의 윤리적 책임을 판단하는데 중요한 영향을 미치는 변수고, 이러한 지각된 인공지능 자유의지는 의인화를 통해 야기된다는 가설을 검증하기 설계되었다. 연구결과를 정리하면, 첫째, 사람들이 무생물인 인공지능에게 인간처럼 윤리적 책임을 기대하는 심리적 기제를 밝혔다. 의인화가 인공지능의 윤리적 책임에 미치는 정적 영향이 검증되었고, 이 정적 영향이 의인화에 의해 야기된 지각된 인공지능 자유의지에 의해 매개된다는 연구가설이 검증되었다. 피험자들이 인공지능의 의인화 수준을 높게 지각할수록 인공지능의 자유의지 수준을 높게 지각하였다. 지각된 인공지능 자유의지 수준이 높아질수록 피험자들은 인공지능에게 더 많은 윤리적 책임을 기대하였다. 따라서 본 연구 결과를 통해 의인화에 의해 야기되는 지각된 인공지능 자유의지가 무생물인 인공지능에게 사람들이 윤리적 책임을 기대하는 심리적 기제라는 것이 밝혀졌다.

둘째, 인간과 인공지능 간 비교를 통해 의인화가 지각된 자유의지를 야기시킨다는 것을 검증하였다. 본 연구는 대상의 자유의지는 대상의 윤리적 책임 소재를 판단하는 중요한 변수라는 이전 연구(Fischer, 1994; Frankfurt, 1969; Kane, 2005; Vohs & Schooler, 2008)의 결과를 바탕으로 설계되었다. 대응표본 t검정 결과, 피험자들은 의인화의 정점인 인간에 비해 의인화 수준이 낮은 인공지능의 자유의지를 낮

게 지각하였다. 또한, 윤리적 책임 역시 인간에 비해 인공지능의 윤리적 책임을 낮게 지각하였다. 연구 결과, 지각된 자유의지가 윤리적 책임에 미치는 정적 영향이 인간과 인공지능 모두에게 나타났다. 또한, 인공지능 의인화가 지각된 인공지능 자유의지에 미치는 정적 영향을 통해 추론했을 때, 지각된 자유의지는 인간, 인공지능 관계없이 윤리적 책임에 영향을 미치는 변수이고, 무생물인 인공지능도 의인화를 통해 인간처럼 지각된다면 인공지능에게도 자유의지가 있다고 지각할 수 있다. 이러한 경향은 인공지능이 인간과 유사하다고 판단하는 수준 즉, 의인화 수준이 높아질수록 강하게 나타난다는 것이 본 연구를 통해 밝혀졌다.

최근 인공지능의 비윤리적 행동이 사회 곳곳에서 문제로 불거지면서, 인공지능 윤리 연구에 대한 관심과 중요성이 어느 때보다 커지고 있다. 따라서 인공지능의 윤리적 책임에 대한 심리적 메커니즘과 인공지능 윤리가 인간-인공지능 관계에 미치는 영향을 검증한 본 연구는 시의적절한 함의점을 제안하고 있다. 첫째, 인공지능 의인화가 인공지능 윤리적 책임에 미치는 영향을 검증하였다. 인공지능은 인간과 글, 음성 등으로 직접 상호작용 한다는 점에서 다른 기계들과 차별점이 있다. 특히 친구, 반려동물 등 인간과 사회적 관계를 맺는 인공지능들의 경우 사용자가 지각하는 인공지능 의인화 수준이 인간-인공지능 상호작용에서 특히 중요한 역할을 수행한다. 이에 많은 연구자들이 인공지능 의인화에 대해 연구를 진행하였다. 이전 연구에 의하면, 의인화 수준이 높아질수록 사람들은 인공지능에게 호의적 평가를 내렸고, 인공지능과 상호작용 과정에서 더 큰 만족감을 나타냈다(Duffy, 2003; Fink, 2012; Krach et al., 2008; Riek et al., 2009). 이러한 결과는 이미 실무에도 반영되었다. 다수의 IT기업은 자사의 인공지능의 상호작용 효과를 높이기 위해서 인공지능의 외형을 의인화하거나 말투, 감정표현, 표정 등 상호작용 스킬을 설계하고 있다. 불쾌한 협곡(uncanny valley) 이론에 따르면, 인공지능의 의인화 수준이 일정수준에 도달하면 오히려 불쾌감을 유발할 수 있다(MacDorman, 2006). 하지만 불쾌감을 유발하는 구간 전과 후에서 의인화의 긍정적인 효과가 검증됨에 따라, 인공지능 의인화 기술은 인공지능 개발에 있어 매우 중요한 부분을 차지한다. 인공지능이 윤리적으로 작동하면 의인화의 긍정적 효과가 충분히 발휘되겠지만, 의인화 수준이 높은 인공지능에게 사람들이 기대하는 윤리적 책임이 더욱 큰 만큼, 인공지능의 비윤리적 행동에 대해 개발자들과 실무자들이 더욱 주의해야 할 것이다.

둘째, 지각된 인공지능 자유의지가 인공지능의 윤리적 책임에 미치는 영향을 검증하였다. 기존의 인공지능 윤리 연구는 무생물인 인공지능에게 인간이 왜 윤리적 책임을 기대하는지에 대해서는 연구가 진행되지 않았다. 본 연구는 인공지능 윤리에 보다 근본적으로 접근하여 지각된 인공지능 자유의지가 무생물인 인공지능에게 사람들이 윤리적 책임을 기대하는 심리적 기제임을 밝히고, 지각된 인공지능 자유의지가 인공지능의 비윤리적 행동에 대한 사람들의 평가에 미치는 영향을 검증하기 위해 연구를 진행하였다. 연구결과, 사람들은 의인화를 통해 인간과 마찬가지로 인공지능에게도 자유의지가 있다고 지각하며, 인공지능에게도 자유의지가 있는 만큼 인공지능도 자신의 행동에 윤리적 책임을 져야한다고 판단했다. 본 연구의 결과를 통해 사람들이 의

인화를 통해 인공지능을 단지 개발자 혹은 인공지능을 운영하는 회사의 명령만 처리하는 수동적 존재로 인식하지 않고, 인간처럼 인공지능 역시 스스로 의사결정을 할 수 있는 자유의지가 있다고 지각한다는 것이 검증되었다. 향후 지각된 인공지능 자유의지가 인공지능 윤리를 비롯한 다양한 인간-인공지능 상호작용 연구의 결과를 설명하는데 중요한 역할을 할 수 있을 것이다. 또한, 지각된 인공지능 자유의지를 연구 변수로 다룬 다양한 연구들이 학계에 등장할 것으로 기대된다.

본 연구는 연구가설들이 모두 검증되었지만, 연구상 한계점을 가지고 있다. 첫째, 피험자 개인특성(성별, 나이)이 다른 변수에 미치는 영향에 대한 구체적인 해석이 불가능하다. 상관분석 결과, 피험자의 성별은 인공지능 의인화와 부적관계가 나타났을 뿐 다른 변수들과는 유의미한 관계가 나타나지 않았다. 또한, 추가분석을 통해 성별에 따른 연구경로의 방향성이나 설명량의 차이를 검증하였으나, 방향성과 설명량에서 유의미한 차이가 나타나지 않았다. 이전 연구 결과에 따르면, 성별에 따라 의인화 경향성의 차이는 나타나지 않는다. 이는 의인화가 다양한 개인특성에 영향을 받고, 피험자들의 이러한 특성들 일일이 통제하고 측정하는 것이 불가능하기 때문으로 해석된다(Epley et al., 2007; Waytz et al., 2010). 따라서 남성이 여성보다 인공지능 의인화 경향이 강하게 나타난 본 연구의 결과를 일반화하기에 무리가 있다. 또한, 성별과 차별금지에 대한 인간과 인공지능의 윤리적 책임 간 상관관계가 모두 나타나지 않았다. 하지만, 이전 연구 결과에 따르면, 일반적으로 여성이 남성보다 차별 문제에 대해 민감하게 생각한다(Ellemers, 2018; Glick et al., 2000; Loe, Ferrell, & Mansfield, 2000). 본 연구에서는 피험자의 윤리 가치관이나 성인지 감수성 등 인간과 인공지능의 윤리적 책임에 영향을 줄 수 있는 변인들을 측정하지 않았기 때문에 이러한 결과를 해석하기에 무리가 있다. 후속 연구에서는 개인특성이 인공지능 윤리에 미치는 영향을 보다 정밀하게 측정할 필요가 있다.

둘째, 인공지능의 윤리적 책임을 측정 문항의 개발이 필요하다. 본 연구에서는 IEEE와 EU를 비롯한 다양한 기관과 기업에서 사용중인 인공지능의 윤리적 책임 다섯 항목을 각각 2개의 설문 문항으로 변환하여 사용하였다. CFA를 통해 윤리적 책임 다섯 항목에 대한 성분 변환행렬을 확인한 결과, 다섯 항목이 분류되지 않았다. 이는 다섯 항목 간 상관이 모두 유의미하게 나왔고($ps < .01$), 상관 설명계수가 모두 .400 수준 이상으로 상당히 높은 수준으로 나왔기 때문으로 추측한다. 다섯 항목, 10개의 문항의 값을 하나의 평균으로 합쳐 회귀분석을 시행한 결과, 의인화가 인공지능의 윤리적 책임에 미치는 정적 영향이 유의미하게 나타났다($B = .33$, $SE = .04$, $p < .001$). 또한, 의인화가 윤리적 책임에 미치는 정적 영향을 매개변수 지각된 인공지능 자유의지에 의해 완전히 매개됨이 확인되었다($B = .36$ Boot $SE = .03$, Boot LLCI = .29, Boot ULCI = .43). 이러한 결과는 다섯 항목을 각각 회귀분석한 결과와 일치한 것으로, 인공지능의 윤리적 책임을 다섯 항목으로 나누고 그에 대한 회귀분석을 각각 수행한 본 연구의 결과의 의미가 약해지는 결과이다. 따라서, 추후 연구에서는 인공지능의 윤리적 책임을 하나의 인덱스로 측정하는 문항들을 개발하거나 윤리적 항목 간 차이를 보다 명확하게 측정하는 문항을 개발할 필요가 있다.

마지막으로 향후 연구에서 인공지능 다양한 역할이 지각된 인공지능 자유의지와 윤리적 책임에 미치는 영향을 검증해 볼 필요가 있다. 현재는 인공지능의 역할이 사무보조 정도로 한정되어 있지만, 멀지 않은 미래에 인공지능은 여러 분야에서 다양한 역할로 인간과 공존 및 협업하게 될 것이다. 인공지능의 역할에 따라 인간-인공지능 관계의 유형이 다르게 나타날 수 있다. 가령, 사무 보조와 같은 목적으로 설계된 비서(assistant) 역할의 인공지능은 인간-인공지능 관계가 상하서열(authority ranking) 관계로 형성될 수 있고, 친구 혹은 AIBO처럼 반려동물처럼 인간과 사회적 관계 형성을 목표로 하는 인공지능들은 공동공유(communal sharing) 관계가 형성될 수 있다. 사람들은 대상과의 관계 유형에 따라 대상을 대하는 행동양식에서 차이를 보인다(Fiske, 1990). 자신의 시키는 일만 처리하는 수동적인 상하서열 관계의 인공지능보다 친구, 가족처럼 다양한 주제의 대화나 경험을 함께 공유하는 공동공유 관계의 인공지능에게 사람들은 더 높은 수준의 자유의지를 지각할 수 있다. 따라서 후속 연구에서는, 인공지능의 역할에 따라 다르게 형성될 수 있는 인간-인공지능 관계의 유형에 따라 지각된 인공지능 자유의지, 윤리적 책임 정도를 비교 검증하고, 애착, 친밀감 등 다른 변수들을 연구 모형에 투입해 변수간 관계가 비윤리적 행동에 대한 인공지능 평가에 어떠한 영향을 미치는지 보다 심도 있게 살펴보고, 그에 맞는 인공지능 윤리 디자인을 제안할 필요가 있다.

참고문헌

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, *May*, *23*.

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

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91).

Chandrashekar, S. P. (2020). It's in Your Control: Free will Beliefs and Attribution of Blame to Obese People and People with Mental Illness. *Collabra: Psychology*, *6*(1), 29.

Coughlin, J. (2018). Alexa, Will You Be My Friend? When Artificial Intelligence Becomes Something More. Forbes. Retrieved from https://www.forbes.com/sites/josephcoughlin/2018/09/23/alexa-will-you-be-my-friend-when-artificial-intelligence-becomes-something-more/

Darwin, C., & Bynum, W. F. (2009). *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life* (p. 458). New York: AL Burt.

Dressler, J. (1995). *Understanding criminal law*. Matthew Bender.

Duffy, B. R. (2003). Anthropomorphism and the social robot. Robotics and *Autonomous Systems*, *42*(3-4), 177-190.

Ekstrom, L. (2018). *Free will*. Routledge.

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, *69*, 275-298.

Eskine, K. J., & Locander, W. H. (2014). A name you can trust? Personification effects are influenced by beliefs about company values. *Psychology & Marketing*, *31*(1), 48-53.

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

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143-155.

Feuerbach, L. (2004). *The essence of Christianity*. Barnes & Noble Publishing.

Fink, J. (2012, October). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics* (pp. 199-208). Springer, Berlin, Heidelberg.

Fischer, J. M. (1994). *The metaphysics of free will* (Vol. 1). Oxford: Blackwell.

Fiske, A. P. (1990). Relativity within Moose ("Mossi") culture: Four incommensurable models for social relationships. *Ethos*, *18*(2), 180-204.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, *66*(23), 829-839.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5-20.

Freud, A. (2018). *Normality and pathology in childhood: Assessments of development*. Routledge.

Genschow, O., Rigoni, D., & Brass, M. (2017). Belief in Free Will Affects Causal Attributions When Judging Others' Behavior. *Proceedings of the National Academy of Sciences*, *114*(38), 10071-10076.

Glick, P., Fiske, S. T., Mladinic, A., Saiz, J. L., Abrams, D., Masser, B., ... & Annetje, B. (2000). Beyond prejudice as simple antipathy: hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology*, *79*(5), 763.

Gill, T. (2020). Blame It on the Self-Driving Car: How Autonomous Vehicles Can Alter Consumer Morality. *Journal of Consumer Research*, *47*(2), 272-291.

HLEG, A. (2019). High-level expert group on artificial intelligence. *Ethics Guidelines for Trustworthy AI*.

Horowitz, A. C., & Bekoff, M. (2007). Naturalizing anthropomorphism: Behavioral prompts to our humanizing of animals. *Anthrozoös*, *20*(1), 23-35.

Hume, D. (2000). *An enquiry concerning human understanding: A critical edition* (Vol. 3). Oxford University Press.

Hur, J. D., Koo, M., & Hofmann, W. (2015). When temptations come alive: How anthropomorphism undermines self-control. *Journal of Consumer Research*, *42*(2), 340-358.

Kane, R. (2005). A contemporary introduction to free will.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective

taking with robots investigated via fMRI. *PloS one*, *3*(7), e2597.

Lee, J. E. R., & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives* (pp. 1-15). IGI Global

Lee, M. K., Tang, K. P., Forlizzi, J., & Kiesler, S. (2011, March). Understanding users! perception of privacy in human-robot interaction. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 181-182). IEEE.

Loe, T. W., Ferrell, L., & Mansfield, P. (2000). A review of empirical studies assessing ethical decision making in business. *Journal of Business Ethics*, 25(3), 185-204.

MacDorman, K. F. (2006, July). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science* (pp. 26-29).

McAuley, E., Duncan, T. E., & Russell, D. W. (1992). Measuring causal attributions: The revised causal dimension scale (CDSII). *Personality and Social Psychology Bulletin*, *18*(5), 566-573.

MacInnis, D. J., & Folkes, V. S. (2017). Humanizing brands: When brands seem to be like me, part of me, and in a relationship with me. *Journal of Consumer Psychology*, *27*(3), 355-374.

Muller, M. (2004). Multiple paradigms in affective computing. *Interacting with Computers*, *16*(4), 759-768.

Nass, C., Steuer, J., & Tauber, E. R. (1994, April). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 72-78).

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer based interviewing systems 1. *Journal of Applied Social Psychology*, *29*(5), 1093-1109.

Nass, C., Isbister, K., & Lee, E. J. (2000). Truth is beauty: Researching embodied conversational agents. *Embodied Conversational Agents*, 374-402.

Nowak, K. L., & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *12*(5), 481-494.

Reider, L. (1998). Toward a new test for the insanity defense: Incorporating the discoveries of neuroscience into moral and legal theories. *UCLA L. Rev.*, *46*, 289.

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).

Rotter, J. B. (1966). Generalized Expectancies for Internal Versus External Control of Reinforcement. *Psychological monographs: General and applied*, *80*(1), 1.

Shahriari, K., & Shahriari, M. (2017, July). IEEE standard review-Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)* (pp. 197-201). IEEE.

Statista (2017). Revenues from the artificial intelligence (AI) market worldwide, from 2016 to 2025.

Veruggio, G. (2010). Roboethics [tc spotlight]. *IEEE Robotics & Automation Magazine*, *17*(2), 105-109.

Vitale, J., Tonkin, M., Herse, S., Ojha, S., Clark, J., Williams, M. A., ... & Judge, W. (2018, February). Be more transparent and users will like you: A robot privacy and user experience design experiment. In P*roceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 379-387).

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, *19*(1), 49-54.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383-388.

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.

Zeki, S., Goodenough, O. R., & Goodenough, O. R. (2004). Responsibility and punishment: whose mind? A response. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1451), 1805-1809.

원 고 접 수 일 : 2021. 02. 03.
수정원고접수일 : 2021. 02. 26.
게 재 결 정 일 : 2021. 03. 03.

한국심리학회지: 소비자·광고
Korean Journal of Consumer and Advertising Psychology
2021, Vol. 22, No. 2, 155-177

The Effect of Anthropomorphism on AI's Ethical Responsibility: The Mediating Role of the Perceived Freewill

Jungyong Ahn[1)] **Yongjun Sung**[2)]

[1)]Research Insititute for Information and Culture, School of Media and Communication, Korea University
[2)]School of Psychology, Korea University

The present research examined the psychological mechanism that lead individuals to expect ethical responsibility of artificial intelligence (AI). Specifically, this research tested the hypothesis that the positive effect of anthropomorphism on the AI's ethical responsibility is mediated by the perceived freewill. The findings showed that the higher the level of anthropomorphism of AI, the greater the AI's ethical responsibility. Also, this effect was fully mediated by the perceived AI's freewill caused by anthropomorphism. Additionally, findings showed that the positive effect of perceived freewill on ethical responsibility applies to both human and AI. That is, participants reported lower ethical responsibility for AI than human as the level of freewill of AI was perceived to be lower than that of human. This research offers theoretical implications of expanding the understanding of human-AI interactions by revealing that the psychological mechanism in which individuals expect ethical responsibility to AI is the perceived free will of AI. It also provides managerial and policy implications by suggesting the need for ethical design of AI for human-AI interaction.

Key words : artificial Intelligence (AI), AI ethics, anthropomorphism, freewill, ethical responsibility