인지과학, 제33권 제4호
Korean Journal of Cognitive Science
2022, Vol. 33, No. 4, 169~203
http://dx.doi.org/10.19066/cogsci.2022.33.4.001

AI의 의사결정에 대한 도덕판단에서 의인화가 미치는 영향 - 쌍 도덕 이론을 중심으로 -

최 윤 빈
서울대학교 협동과정 인지과학전공

장 대 익†
가천대학교 창업대학

인공지능 기술이 고도화됨에 따라 인공지능이 도덕적 판단의 대상이 되거나 주체가 되는 사례가 늘어가고 있으며, 이러한 추세는 가속화될 전망이다. 인공지능은 고용, 의료 등 인간 사회의 핵심적인 분야에서 활발히 활용되기 시작했지만, 그에 반해 사람들이 인공지능과의 상호작용에서 그들을 어떠한 방식으로 지각하고 반응하는지에 관한 연구는 상대적으로 많지 않다. 본 연구는 세 가지 맥락(고용, 의료, 법조)에서의 실험을 통해 인공지능의 의인화가 인공지능의 의사결정에 대한 도덕적 책임 판단에 미치는 영향과 그 과정을 살펴보았다. 쌍 도덕 이론의 주요 변인인 지각된 행위 능력과 지각된 경험 능력을 매개 변인으로 모델을 구성해 검증하였으며, 구체적으로는 지각된 의인화가 인공지능의 도덕적 책임을 증가시키고, 인공지능에 대해 지각된 행위 능력과 경험 능력이 이를 매개할 것이라 예측하였다. 연구 결과, 실험 조작은 유효하지 않았으나 모든 실험에서 지각된 경험 능력이 의인화와 도덕적 책임 지각 간의 관계를 매개함을 확인하였다. 반면 지각된 행위 능력의 효과는 혼재된 결과를 보여 가설을 부분적으로 지지하였다. 이는 도덕적 지위에 대한 경험 능력의 중요성을 주장하는 유기체적 관점을 지지하는 결과이며, 또한 AI와 로봇의 의인화 연구에서 경험 능력이 행위 능력보다 더욱 중요함을 보이는 것이다.

주제어 : 인공지능, 의인화, 도덕적 책임, 마음 지각, 도덕 쌍 이론

† 교신저자: 장대익, 가천대학교 창업대학, (13120) 경기도 성남시 성남대로 1342
연구 분야: 창업학, 진화학, 인지과학, 과학기술학
E-mail: djang@gachon.ac.kr

서 론

인공지능은 인간 사회에서 점점 더 다양하고 중요한 역할을 차지하고 있다. 비교적 최근까지도 인공지능은 단순히 인간보다 바둑(Borowiec, 2016)과 체스(Newborn, 2012)를 더 잘 두며, 기껏해야 약속 스케줄을 잡거나 영화를 추천하는 수준에서 그 능력과 역할이 그쳤다. 그러나 인공지능은 생각보다 더욱 빠르게 우리의 생활과 사회 깊숙이 들어오고 있으며, 그 지위 역시 단순히 인간에 봉사하는 역할을 넘어서기 시작했다. 그 대표적인 사례로 의료, 법조, 고용 등의 분야에서 인공지능은 인간의 삶의 방향을 바꿀 수 있는 능력과 판단을 수행할 역할을 부여받고 있다. 가령 의료 분야에서는 이미 ‘헬스케어 AI’이라는 이름의 인공지능을 통해 암과 같은 질병의 진단, 보험료 산정, 환자 데이터 관리 등이 이루어지고 있다(Dash et al., 2019; Ayasdi, 2018). 또한 법조 분야에서는 인공지능이 데이터 관리 및 분석, 실사 작업, 가석방 판단 등의 중요한 역할을 수행하며, 이에 대한 윤리적 논의 역시 진행되고 있다(권헌영, 2019). 여기에 더해 인공지능 면접 시스템이나 인적자원 관리 시스템들은 기업과 학교 등에서 이미 활발히 활용되고 있다(Ajunwa et al., 2016; Kuncel, Klieger, & Ones, 2014; O’Niel, 2016).

이렇게 사회 핵심 분야에 인공지능이 점차 높은 영향력을 가지게 되는 동시에, 인공지능의 편견 및 오류로 야기될 수 있는 위험의 중대성 또한 높아지고 있다. 헬스케어 분야에서는 인공지능의 도움을 받은 의료적 결정의 책임 소재 문제, 전자의료기록(electronic health record; EHR)의 보안 등의 문제들이 제기되고 있다(Adler-Milstein et al., 2017; Hollister & Bonham, 2018). 법조계에서 가장 널리 알려진 사례로는 미국 법정에서 실제로 사용되었던 컴퍼스(COMPAS; Correctional Offender Management Profiling for Alternative Sanctions)가 있다. 범죄자의 잠재적인 재범 가능성을 예측하는 이 시스템은 흑인이라는 이유만으로 범죄자의 재범 가능성을 더 높게 판단해왔다(Angwin, Larson, 2016). 또한 인공지능은 이제 누군가의 취업이나 해고 여부를 좌우할 수 있는 영향력을 지니기 시작했으며, 현실에서도 이러한 변화에 피해를 받는 사례들 역시 여럿 생겨나고 있다(O’Neil, 2016).

인공지능이 불러올 수 있는 윤리적 문제들에 대응해 여러 정부들은 인공지능의 차별, 데이터 보안 등 인공지능 개발에 관한 규제들을 논의하였다. 대표적 사례로 유럽 연합의 윤리적 인공지능을 위한 가이드라인(HLEG, A. I., 2019), 유럽 연합 일반 데이터 보호 규칙(General Data Protection Regulation; GDPR)은 세계의 인공지능 및 데이터 관련 규제를 선도하고 있다. 한국 역시도 비슷한 맥락에서 2020년 시행된 데이터 3법과 인공지능 윤리 기준이 발표된 바 있다. 다양한 정부와 기업 차원의 시도에도 불구하고 이러한 ‘가이드라인’들의 법적 추상성과 비강제성은 현장에서 인공지능을 개발하는 기업들이 보다 윤리적인 인공지능을 만드는 데 혼란을 주거나 실제로 큰 효력이 없다는 비판에 노출된다. 윤리적 가이드라인은 한 사회가 지향하는 기술의 발전 및 개발 방향을 제시한다는 점에서 의의가 있지만, 이러한 방향성을 구체적으로 어떻게 실현

할 지에 대한 연구들이 필요한 시점이다.

HCI(Human-Computer Interaction) 및 HRI(Human-Robot Interaction) 분야에서는 더 효과적인 인공지능 인터페이스의 개발뿐만 아니라 경찰, 경영 등 여러 사회적 맥락에서 인공지능에 대한 사람들의 인식과 상호작용이 연구된 바 있다. 가령 예측 치안 유지(predictive policing) 시스템에 대한 실제 경찰관들의 경험을 다룬 한 연구는 실제 경찰관의 직관 및 경험과 경찰 배치 알고리즘의 예측이 충돌하는 사례들을 질적 연구를 통해 보고하였다(Verma & Dombrowski, 2018). 또한 이민경은 경영적 맥락에서 인간과 알고리즘의 공정성에 대한 지각이 인간의 권위와 알고리즘의 불편부당성이라는 각기 다른 이유에서 비롯됨을 제시한 바 있다(Lee, 2018; c.f. Ötting & Maier, 2018). 여기에 더해 왕 등은 알고리즘의 산출 결과, 개발 과정에 대한 정보, 사용자의 학력과 같은 요인들이 알고리즘에 대한 공정성 지각에 영향을 줌을 밝혔다(Wang, Harper, & Zhu, 2020).

한편 인공지능이 사회에서 점차 중요한 행위자로서 주목받으면서, 도덕심리학 분야에서도 인공지능과 같은 비인간 대상들에 대한 연구들이 이루어지고 있다. 예를 들어 칸타레로(2021)는 도덕 기반 이론(moral foundation theory; Haidt, 2011)을 통해 권력 격차가 권위 기반의 도덕을 통해 도덕 판단에 영향을 준다는 사실을 밝혔다. 또한 그레이 등이 주장하는 쌍 도덕 이론(theory of dyadic morality; Schein & Gray, 2018) 역시 마음 지각을 통해 로봇, 인공지능 등의 비인간 대상과 관련된 도덕적 지각 연구에 활용된 바 있다(Gray, Gray & Wegner, 2007). 가령 빅맨과 그레이(2018)은 컴퓨터의 도덕적 의사결정에 대한 회피를 마음 지각을 통해 설명하였으며, 얌 등(2020)도 로봇 호텔이라는 맥락에서 서비스 로봇의 의인화가 사용자 만족도에 미치는 영향을 쌍 도덕 이론을 이용해 연구하였다.

이러한 맥락에서 본 연구도 도덕심리학의 관점을 통해 인공지능 사용자들의 도덕적 인식 과정을 규명하고, 실제 인공지능 개발 과정에서 참고될 수 있는 연구를 수행하려 한다. 구체적으로 본 연구는 고용, 의료, 법조 분야와 같은 사회 핵심 분야의 인공지능의 의인화가 사람들의 도덕적 책임 인식에 어떠한 영향을 주는지 알아보고자 한다. 특히 쌍 도덕 이론의 주요 요인인 지각된 행위(perceived agency)와 지각된 경험(perceived patiency)이라는 두 요인이 어떻게 의인화와 도덕적 책임 간의 효과를 매개하는지 확인할 것이다.

이론적 배경

인간-컴퓨터 상호작용(Human-Computer Interaction; HCI)

CASA(Computers are Social Actors) 패러다임

CASA 패러다임은 사람들과 컴퓨터의 상호작용이 사람들 간의 사회적 상호작용과 같은 방식

으로 이루어진다는 패러다임이다(Reeves & Nass, 1996). 예를 들어, 사람들은 컴퓨터와의 상호작용 중에도 인간과의 상호작용에서 나타나는 현상들인 편견, 내집단과 외집단의 구분, 호혜성과 같은 특성들을 보인다(Nass, Moon, 2000). CASA 패러다임은 HCI 분야에서 활발히 활용되고 있으며, 특히 로봇, 인공지능과 같은 인공적 행위자들(artificial agents; AAs)을 다루는 연구들에서 그러하다(Malle et al., 2016, Li & Suh, 2021).

그러나 적어도 몇몇 조건에서는 CASA 패러다임이 적용되지 않는 사례들도 발견되고 있다. 가령 몇몇 연구에서는 알고리즘이 더 정확함에도 불구하고 인간의 의사결정을 더 신뢰하는 알고리즘 회피(algorithmic aversion; Dietvorst, Simons, & Massey, 2015; Bigman & Gray, 2018) 현상이 보고된 바 있다. 디엣포어스트는 알고리즘이 MBA 과정 합격자를 더 정확하게 예측함에도 불구하고 사람들이 컴퓨터의 실수에 신뢰감을 더욱 빠르게 잃어버리는 현상을 확인하였다. 또한 빅맨과 그레이(2018)는 인간으로 구성된 위원회와 슈퍼컴퓨터를 비교하는 실험을 통해 사람들이 기계에 의한 도덕적 의사결정을 꺼리며, 이러한 현상이 기계에 대한 마음 지각에서 기인함을 밝혔다.

인간과 기계에 각각 다르게 기대되는 사람들의 윤리적 기준을 확인한 사례도 존재한다. 말레 등(2015)은 전통적인 트롤리 딜레마(Foot, 1967; 5명을 향해 돌진하는 열차를 스위치를 눌러 1명을 대신 희생하도록 할 지 선택하는 딜레마) 상황에서 사람들이 기계 로봇에게는 공리주의 윤리에 따른 선택(5명을 위해 1명을 희생하는 것)을 기대하지만, 인간에게는 의무론적 윤리에 따른 선택(1명을 희생하지 않는 것)을 기대함을 확인하였다.

인공지능과 의인화

인공지능의 의인화 연구는 주로 의인화가 사람들의 인식에 미치는 영향을 다루고 있으며, 본 연구 또한 이러한 흐름 안에 있다. 인공지능의 의인화는 대체로 사용성에 긍정적인 효과를 가져오는 것으로 보이는데, 예를 들어 사람들은 의인화된 인공지능을 그렇지 않은 인공지능보다 더 신뢰하며(Natarajan, Gombolay, 2020; Waytz, Heafner, Epley, 2014), 실수에도 더 너그럽다(Yam et al., 2020). 또한 외로움, 애착 유형, 문화적 지향과 같은 사용자의 사회적 경향은 스마트폰에 대한 의인화 경향에도 영향을 주는 것으로 보고되었다(Wang, 2017).

반면 의인화의 효과와 관련된 주요 연구들 중 불쾌한 골짜기를 다룬 연구들을 주목할만하다. 불쾌한 골짜기 효과란 일정 수준 이상으로 인간과 닮은 대상은 사람들에게 불쾌감과 혐오감을 불러일으키는 현상을 말한다(Mori, 1970). 이러한 현상이 실제로 있는지에 대한 의견은 분분하며, 그 원인에 대해서도 진화심리학적 맥락의 위험 회피(Moosa & Ud-Dean, 2010), 감염 회피(MacDorman & Entezari, 2015), 공포 관리 이론(MacDorman, 2005) 등에 기반한 여러 가설들이 존재한다. 또한 한 연구에 따르면, 불쾌한 로봇의 도덕 판단 역시도 더 비도덕적으로 평가된다는 '도덕적 불쾌한 골짜기(moral uncanny valley effect)' 효과가 보고되기도 하였다(Laakasuo, Palomäki, & Köbis, 2021). 또한 본 연구에서 주로 언급하게 될 쌍 도덕 이론의 경우, 기계에서 느껴지는 지각

된 경험, 즉 기계 스스로가 고통과 같은 감각을 느낄 수 있다는 지각이 불쾌한 골짜기 효과와 관련된다는 것을 밝혀내었다(Gray & Wegner, 2012a).

그러나 AI나 로봇의 의인화 연구에서 의인화의 정의, 조작 방식, 측정 방식 등에 있어서 연구자들 간의 일관된 합의는 아직 이루어지지 못한 것으로 보인다(Li & Suh, 2021). Li와 Suh의 리뷰 연구에 따르면, 의인화 관련 연구들에서 의인화의 정의는 인간의 심리적 경향성, 심리적 과정, 지각, 기술적 자극, 추론 등으로 상이하며, 실험적 조작 역시 시각적(얼굴, 움직임, 표정), 음성적(언어적 능력, 목소리), 심리학적 (자율성) 방식들로 각기 다르다. 이러한 분류 이외에도 의인화의 요인 및 조건들에 대해 탐색하는 연구들 역시 활발히 이루어졌다(Epley, Waytz, & Cacioppo, 2007; Cacioppo & Epley, 2010). 또한 의인화에 대한 설문 문항들도 연구마다 다른 설문지들을 인용하거나 직접 제작하여 사용하고 있는 것으로 보고되었다.

도덕심리학

도덕심리학은 인간의 도덕적 지각, 도덕 판단 등 도덕과 관련된 심리적 메커니즘을 탐구하며, 도덕적 발달(Kohlberg, 1969; Kohlberg, 2016), 사람들의 도덕 기제(Greene et al., 2001; Greene, 2013; Schein & Gray, 2018; Haidt 2001), 그리고 문화 간의 도덕적 직관 및 규범의 차이(Haidt, Koller, & Dias, 1993; Haidt, 2012; Curry, 2019)를 연구한다.

최근 도덕심리학계에서 인간의 도덕적 지각과 관련해 주로 연구되는 가설들은 크게 두 가지로, 하나는 도덕 쌍 이론(Gray & Wegner, 2012b; Gray & Wegner, 2016; Schein & Gray, 2018), 그리고 다른 하나는 도덕 기반 이론(Moral Foundation Theory; MFT; Haidt, 2011; Graham et al., 2013; Graham et al., 2009)이다. 도덕 쌍 이론은 사람들의 도덕 지각을 가해자와 피해자라는 쌍, 그리고 위해(harm) 행위라는 요인들로 설명하며, 가해자가 행위 능력이 높을수록, 피해자가 경험 능력, 즉 고통을 느끼는 능력이 쉽다고 지각될수록, 그리고 위해 행위가 더 심각할수록 높은 수준의 도덕성 위반을 인식한다고 주장한다.

한편 도덕 기반 이론은 도덕성을 여러 독립된 도덕 기반(moral foundation)들로 설명한다. 하이트는 대표적인 도덕 기반으로 위해/돌봄(care/harm), 공정/기만(fairness/cheating), 충성/배신 (loyalty/betrayal), 권위/무질서(authority/subversion), 순수/오염 (sanctity/degradation), 자유/억압 (liberty/oppression)을 제시하며, 개인마다, 문화마다 중시하는 도덕 기반의 차이가 도덕적 판단의 차이를 만든다고 주장한다(Haidt, 2011). 이 이론은 앞서 말한 도덕 쌍 이론이 취하는 위해 중심의 일원론적 접근과는 달리 각기 다른 기반들을 통해 도덕성을 설명함으로써 다원주의적인 접근을 취하고 있으며(Graham et al., 2018), 현재도 두 이론과 관련한 여러 논의들이 이루어지고 있다(Graham et al., 2013; Schein, Ritter & Gray, 2016).

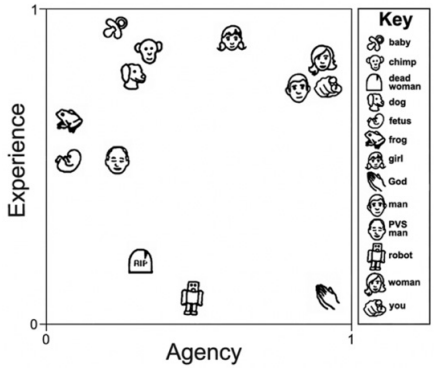
도덕 쌍 이론(Theory of Dyadic Morality; TDM)

도덕 쌍 이론은 샤인, 그레이, 웨그너 등의 학자들에 의해 주장된, 인간의 도덕 지각을 설명하려는 이론이다. 이들은 도덕적 행위자(agent), 도덕적 수동자(patient), 그리고 도덕 행위의 내용이라는 요인으로 인간의 도덕 지각을 설명한다. 간단히 말해 사람들은 행위자가 큰 행위 능력을 가질수록, 수동자가 큰 경험 능력을 가질수록, 그리고 더 큰 위해가 가해질수록 제시된 상황을 더욱 비도덕적으로 판단한다(Schein & Gray, 2018). 예를 들어, 사람들은 어린 소녀가 최고경영자를 때리면 도덕적으로 큰 문제를 느끼지 않지만, 반대로 그 최고경영자가 어린 소녀를 때리는 상황에 대해서는 도덕적 문제를 지각하고 비난할 것이다(Gray & Wegner 2016). 왜냐하면 최고경영자는 상식적으로 매우 큰 행위 능력(무언가를 계획하고, 의도를 가지고, 실행에 옮기는 능력)을 가지며, 소녀는 매우 큰 경험 능력(무언가를 지각하고, 고통과 같은 감각을 느끼는 능력)을 가지고 있기 때문이다.

도덕심리학의 여러 연구들 중 본 연구에서 도덕 쌍 이론의 요인들(행위 능력 및 경험 능력)을 주요 변인으로 채택한 이유는 다음과 같다. 첫째, 도덕 쌍 이론은 인간만이 아니라 비인간 대상(동물, 기계, 유령 등)에 대한 도덕 인식까지 폭넓게 다루는 이론이다. 둘째로 도덕 쌍 이론은 인공지능, 로봇과 같은 대상에 대한 도덕 인식 연구에서 활발하게 활용되고 있으며, 특히 의인화를 통한 행위 능력과 경험 능력의 변화라는 주제들을 연구한 바 있다(Waytz, Epley, 2014; Bigman, Gray, 2018; Yam et al., 2020).

의인화와 도덕적 책임 지각

인간, 로봇을 비롯한 여러 대상에 대한 마음 지각을 다룬 연구에 따르면, 로봇과 같은 존재는 무언가를 행위할 수 있는 능력은 높게 지각되며, 무언가를 느낄 수 있는 능력인 경험 능력은 상대적으로 낮게 지각된다(Gray, Gray, & Wegner, 2007; <그림 1>). 따라서 이러한 대상은 도덕적 가해자로 지각되기 쉬운 반면 도덕적 피해자로서 지각되기는 상대적으로 어렵다.



〈그림 1〉 마음 지각의 차원(Gray, Gray, & Wegner, 2007)

한 선행 연구에 따르면, 서비스를 제공하는 로봇(호텔 접수원, 웨이터)을 의인화하는 것이 사람들로 하여금 그 로봇들에 대해 더 높은 행위 능력과 더 높은 경험 능력을 지각하게 만들었다. 지각된 경험은 로봇의 실수에 대해 더 너그러운 평가와 서비스 만족도로 이어졌지만, 증가된 행위 지각은 유의미한 변화를 만들어내지 못하였다(Yam et al., 2020). 또한 기계에 의한 도덕적 의사결정에 대한 사람들의 회피 (aversion)를 연구한 한 논문(Bigman & Gray, 2018)에 따르면 사람들은 기계가 도덕적 의사결정을 내리는 것을 반대하지만, 조언과 같은 제한적 역할, 지각된 전문성과 지각된 경험을 높이는 것으로 이러한 현상을 완화할 수 있음을 제안하고 있다. 인공지능의 대표적 사례 중 하나인 자율주행차의 의인화를 다룬 연구 역시, 의인화가 자율주행차에 대한 신뢰감을 상승시키며, 그에 더해 자율주행차가 사고를 냈을 경우에도 인공지능 자체에 대해 더 많은 비난을 가함을 밝혔다(Waytz, 2014).

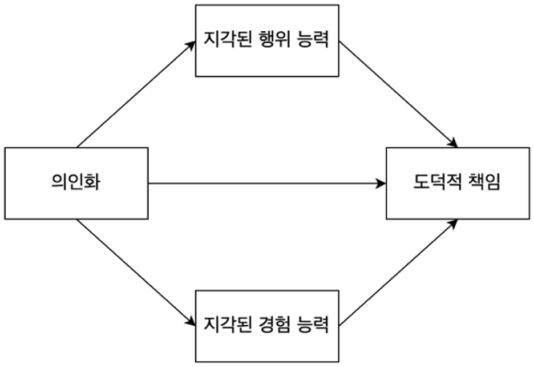
연구 내용

본 연구는 세 차례의 실험을 통해 인공지능의 의사결정에 대한 도덕적 책임의 평가가 인공지능의 의인화에 따라 어떻게 변화하는지 알아보고자 한다. 먼저 의인화의 결과로 인공지능에 대해 지각된 행위 능력과 경험 능력이 높게 지각될 것이며, 이러한 변화가 인공지능의 자체적 책임의 증가를 매개할 것이라는 가설을 검증하였다.

얌 등의 연구(2020)에 따르면 호텔의 웨이터, 접수원과 같이 수동적 역할을 하는 로봇의 실수에 대해 시각적, 언어적 의인화가 사람들을 더 너그럽게 만들었으며, 특히 지각된 경험이 이를 매개하였다고 보고하였다. 이에 반해 본 연구는 단순한 실수가 아닌, 인종 차별, 개인정보 유출과 같은 심각한 도덕적 위반 상황을 설정하였다. 또한 고용, 법조, 의료 맥락의 인공적 행위자 시나리오를 통해 사람들에게 유의미한 도덕적 문제를 야기할 수 있는 적극적 행위자로서의 인공지능 맥락에서 가설들을 검증하고자 하였다.

연구목표 및 연구가설

- 가설 1: 인공지능의 의인화는 인공지능에 대해 지각된 행위 능력과 지각된 경험 능력을 증가시킬 것이다.
- 가설 2: 인공지능의 의인화는 인공지능 자체에 대한 도덕적 책임 지각을 증가시킬 것이다.
- 가설 3: 지각된 행위 능력과 지각된 경험 능력은 의인화가 인공지능의 책임 지각에 미치는 영향을 매개할 것이다.



〈그림 2〉 이론적 연구 모델

연구 방법

연구 개요

본 연구는 총 세 번의 실험을 통해 인공지능의 의인화가 인공지능의 도덕 위반에 대한 책임에 미치는 영향을 조사하였다. 실험 결과의 견고함과 여러 맥락에서의 연구를 위해 인공지능이 사용되는 맥락을 각기 달리 하여 설문을 구성 및 진행하였다. 현 시점에 활발히 개발되고 있으며, 사람들의 삶에 지대한 영향을 미치는 세 분야인 면접, 의료, 법조 분야를 선정하였다. 또한 의인화가 책임 지각에 미치는 영향이 어떠한 방식으로 일어나는지를 알아보기 위하여 검증된 도덕심리학적 요인들(지각된 행위 능력, 지각된 경험 능력)을 함께 측정하여 평행다중매개(multiple parallel mediation) 모델을 사용해 분석하였다.

실험 참여자들은 무작위로 의인화/비의인화 조건으로 나누어져, 시나리오를 통해 각 인공지능들에 대한 설명과 도덕적 위반 사례를 제시받았다. 모든 실험은 시나리오 1(인공지능 소개), 의인화 및 매개 변인(행위와 경험) 설문, 시나리오 2(도덕 위반), 최종 설문(책임 및 통제 변인)의 순서로 진행되었으며, 실험 변인들은 중간 설문과 최종 설문들을 통해 수집되었다. 이러한 순서로 실험이 진행된 이유는 인공지능의 지각된 행위 능력 및 지각된 경험 능력과 책임을 참여자에게 동시에 물어볼 경우 일어날 수 있는 실험상의 오염을 방지하기 위함이다. 비슷한 주제의 연구를 진행하였던 웨이츠 등의 연구에서도 이러한 절차로 실험이 진행되었던 바 있다(Waytz, Heafner, & Epley, 2014)

연구윤리 및 사전 등록

본 실험은 서울대학교 연구윤리심의위원회를 통해 사전 등록 및 승인되었다(https://irb.snu.ac.kr; 승인 번호: IRB No. 2110/002-017).

참여자 모집

모든 실험 참여자는 대한민국에 소재하고 있으며, 전국에 패널 회원을 보유한 설문조사업체 '마크로밀 엠브레인(이하 엠브레인)'을 통해 모집하였다. 연구 참여자의 수는 슈만 등에 의해 개발된 몬테카를로 기법 기반의 매개 모델 검정력 계산 프로그램(Schoemann, Boulton, & Short, 2017; url: https://schoemanna.shinyapps.io/ mc_power_med/)을 통해 도출되었으며, 계산 결과 실험 당 약 200명을 설문 당 참여자 수로 산정하였다.

변인

지각된 의인화

실험 조작으로 실험군(의인화 집단)에 배정된 참여자에게는 시각적, 언어적으로 의인화 처리된 시나리오를 제공하였다. 시각적 의인화의 경우 눈, 입이 있는 인공지능의 얼굴 이미지를 인공지능 소개 시나리오와 함께 제시하였다. 실험 1에서는 실제로 존재하는 로봇 '페퍼'의 사진을 화상 면접 이미지에 합성하여 제시하였으며(<그림 3>), 실험 2, 3에서는 인공지능의 이미지를 직접 제작해 사용하였다(<그림 4>). 비의인화 조건의 경우, 모든 실험에서 인간적 요소가 없는 상징적 이미지를 제작해 참여자들에게 제시하였다. 언어적 의인화는 시나리오 상의 인공지능의 이름을 더 인간스러운 것으로 설정하였고, 시나리오 상의 묘사 역시 보다 인간에 가깝게 작성되었다. 가령 의인화 조건에서는 '진수', '제인' 등의 인간적인 이름을 사용한 반면, 비의인화 조건에서는 'KJI-7', 'KLA-7'과 같이 기계적인 이름을 사용하였다. 또한 의인화 시나리오에서는 면접관, 인공지능 소개 및 인터뷰에서 인공지능을 실제 사람과 유사하게 표현하였다. 시나리오의 부분적 예시는 아래와 같다.

〈그림 3〉 실험에 사용된 이미지 (실험 1)

비의인화 조건(통제 조건: 면접 맥락)

KJI-7은 스웨덴의 한 기업에서 2019년부터 개발된 세계 최초의 면접 인공지능입니다. 단순히 데이터를 분석하고 조언하는 역할에서 그쳤던 기존의 면접 AI와는 달리, 'KJI-7'은 다른 인간 면접관의 옆에 실제로 배치되어 지원자에게 질문하고 판단합니다. 최근 미국의 한 스타트업에서는 이 인공지능을 사무실에 배치해 다른 직원들이 이용할 수 있도록 하기도 했습니다.

의인화 조건(실험 조건: 면접 맥락)

진수 팀장은 스웨덴의 한 기업에서 2019년부터 개발된 세계 최초의 AI 면접관입니다. 단순히 데이터를 분석하고 조언하는 역할에서 그쳤던 기존의 면접 AI와는 달리, 진수 면접관은 다른 인간 면접관의 옆에 실제로 앉아 지원자에게 질문하고 판단합니다. 최근 미국의 한 스타트업에서는 이 로봇을 사무실에서 다른 직원들과 함께 일할 수 있게 하기도 했습니다.

지각된 의인화에 대한 조작 검증은 동물, 자연, 기술과 같은 다양한 존재자들에 대한 의인화 척도를 개발한 웨이츠등의 선행연구(Waytz, Cacioppo, & Epley, 2010)에서 사용되었던 설문 문항 중, 기술에 대한 의인화 척도를 다루는 네 가지 설문을 번안하여 사용하였다(7점 척도). 설문 문항의 예시로는 "이 인공지능이 얼마나 의도를 가질 수 있는 것처럼 보이십니까?", "이 인공지능이 얼마나 의식을 가지고 있는 것처럼 보이십니까?"가 있다.

조건	실험 조건	통제 조건
실험 2 (의료)		
실험 3 (법조)		

〈그림 4〉 실험에 사용된 인공지능 이미지 (실험 2, 3). © Ahn Lee.

지각된 행위와 경험

본 연구는 의인화와 마음 지각, 그리고 사용자 만족도의 관계를 다루었던 선행 연구(Yam et al., 2020)에서 사용되었던 설문 문항들을 통해 참여자들이 제시된 인공지능의 행위 능력과 경험 능력을 어떻게 지각하였는지 측정하였다. 행위 능력과 경험 능력의 측정을 위해 각각 네 개의 설문 문항이 제시되었다. 참여자들은 "이 인공지능은 스스로의 행위를 계획할 수 있다", "이 인

공지능은 생각할 수 있다"(행위 능력), "이 인공지능은 고통을 느낄 수 있다", "이 인공지능은 행복을 느낄 수 있다"(경험 능력)와 같은 문항들에 7점 척도(1점: 매우 동의하지 않음, 7점: 매우 동의)로 응답하였다.

도덕적 책임

각 인공지능의 도덕 위반 시나리오를 읽은 후, 참여자들은 해당 실패에 대한 각 도덕적 주체들에 대한 도덕적 책임 설문에 응답하였다. 인공지능 자체, 인공지능 개발자, 인공지능 개발 회사가 대상으로 제시되었다. 개발자와 회사를 설문에 넣은 이유는 단순히 인공지능 자체에 대한 책임만을 설문하면 그 결과가 지나치게 증폭될 우려가 있으며, 추가적으로 조건에 따라 각 책임 주체 간에 유의미한 차이가 존재하는지 관측될 수 있기 때문이다.

컴퓨터 친화도(computer literacy)

컴퓨터 친화도는 여러 선행 연구들을 통해 인공지능에 대한 인식과 연관되어 있음이 밝혀진 바 있다(Wang, Harper, & Zhu, 2020). 따라서 본 연구에서 역시 통제 변인의 하나로 컴퓨터 친화도를 설정하였으며, 그 측정은 인공지능에 대한 공정성 인식의 변인 연구를 수행하였던 Wang등의 연구(ibid)의 설문지를 사용하였다. 총 8개의 설문 문항으로 이루어져 있으며, 온라인 쇼핑 추천 알고리즘에 대한 지식, 프로그래밍에 대한 전반적 이해력 등을 7점 척도로 설문하였다.

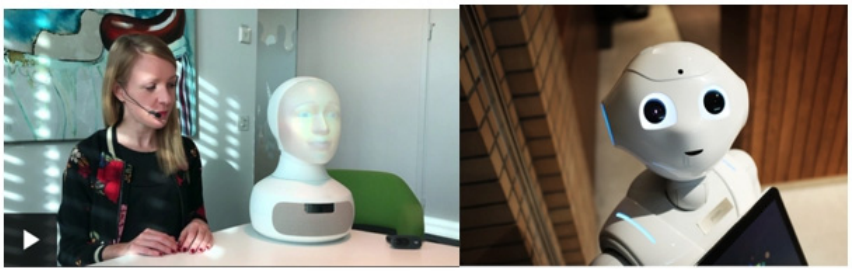
실 험

실험 1

실험 개요

실험 1에서는 최근 기업들에서 활발히 사용되고 있는 인공지능 면접을 설문 내용으로 구성하였다. 국내에서는 대부분 면접자의 표정, 인터뷰 등의 데이터를 활용하는 무형적 알고리즘의 형태로 활용되고 있으나, 실제 형체가 있는 면접 로봇 역시 개발된 바 있다(Savage, 2019; <그림 5> 좌측). 이러한 배경 하에서 로봇의 형태를 한 인공지능 면접관의 시나리오가 구상되었으며, 지나친 의인화로 인한 불쾌한 골짜기 현상을 방지하기 위해 인간과 닮았지만 상대적으로 불쾌감을 주지 않는 소프트뱅크사의 '페퍼'의 이미지를 채택하였다(<그림 5> 우측). 참여자에게 TV

방송 대본의 형식을 취한 인공지능 소개 시나리오를 제공한 후, 이후에 인공지능의 도덕 위반 시나리오를 제시함으로써 참여자들로 하여금 책임을 판단할 수 있도록 하였다.



〈그림 5〉 실험 1에 고려된 로봇들의 이미지.
(좌: Furhat Robotics사의 텐가이(Tengai), 우: Softbank사의 페퍼(Pepper))

참여자

마크로밀 엠브레인을 통해 총 205명의 한국인 패널들이 모집되었다(여성 50.2%, 나이: M = 44.46, SD = 13.34). 다른 사항들은 실험 1과 동일하게 진행되었다. 참여자들은 층화 표집 방법을 사용해 연령과 성별이 고르게 수집되었다.

절차

본 연구는 인공지능에 대한 일반인들의 인식 조사에 응답하는 것으로 참여자들에게 소개되었다. 실험 참여에 동의한 참여자들은 무작위로 의인화(실험) 및 비의인화(통제) 조건에 할당되었으며, 각 실험에 따라 고용, 의료, 법조 인공지능을 소개하는 시나리오를 읽고 해당 인공지능에 대한 지각된 행위 능력과 지각된 경험 능력을 묻는 중간 설문에 응답하였다. 그 다음 절차에서 참여자들은 신문기사 형식의 도덕 위반 시나리오를 읽고 도덕적 책임과 학력, 컴퓨터 친화도, 그리고 기타 인구통계학적 정보를 묻는 최종 설문에 응답하였다.

결과 및 논의

연구 변인들의 기술통계량과 상관관계를 <표 1>에 제시하였다. 모든 데이터 분석은 SPSS (Version 28)와 PROCESS (Version 4.0)를 통해 이루어졌다.

〈표 1〉 기술통계량 및 상관계수 (실험 1)

변인	M	SD	1	2	3	4	5	6	7
1. 의인화 지각	4.13	1.11							
2. 행위 능력	4.58	1.04	.67**						
3. 경험 능력	2.46	1.28	.52**	.40**					
4. 도덕적 책임	3.33	1.74	.31**	.11	.41**				
5. 성별	-	-	.02	.01	.08	.15*			
6. 나이	44.46	13.34	.25**	.09	.13	.06	-.02		
7. 학력	3.82	0.86	.01	.00	-.08	-.12	-.10	-.04	
8. 컴퓨터 친화도	3.86	1.22	.21**	.13	.14	.07	-.19**	.20**	.19**

* p < .05. ** p < .01

조작 점검

예상과는 달리, 통제 조건(M = 4.13, SD = 1.08)과 실험 조건(M = 4.13, SD = 1.15)간 의인화 설문 결과의 차이가 발견되지 않았다 (t(203) = 0.02, p = 0.49, Cohen's d = 0.03). 이러한 결과는 시나리오 기법과 설문조사 방법의 한계, 혹은 실험 자극의 불충분함으로 설명될 수 있다. 비록 실험 조건 간에는 의인화 지각의 차이가 없었으나, 의인화 조작의 영향은 기존의 HRI 연구 흐름과 다른 연구들에서 확인된 바 있으므로, 조건 대신 의인화 설문을 통해 얻은 의인화 지각 점수를 연구 모델에 대입해 검증하였다.

의인화와 마음 지각

가설 1에 따라 인공지능의 의인화가 행위 능력과 경험 능력 지각을 증가시키는지를 알아보았다. 분석 결과 의인화 점수가 높을수록 행위 능력(b = 0.64, SE = 0.05, p < .001)과 경험 능력(b = 0.59, SE = 0.07, p < .001)을 높게 지각하여 가설 1를 지지하였다.

의인화와 도덕적 책임 지각

먼저 의인화가 도덕적 책임 지각에 미치는 직접적인 효과를 살펴보았다. 회귀분석 결과 인공지능에 대한 의인화 점수가 도덕적 책임 을 유의하게 예측하였으며(b = 0.48, SE = 0.11, p < .001), 가설 2를 지지하였다.

마음 지각의 매개 효과

마음 지각, 즉 지각된 행위 능력과 지각된 경험 능력이 의인화와 책임 간의 효과를 매개한다는 가설 3을 검증하기 위하여 SPSS PROCESS를 이용한 경로분석을 수행하였다. 경로별 유의성 검증 결과, 위에 언급한 대로 의인화에 대한 지각이 행위 능력과 경험 능력과 유의한 정적 상관이 있는 것으로 나타났다. 각 매개 변인(지각된 행위 능력, 지각된 경험 능력)이 책임 지각에 미치는 영향 역시 분석하였고, 그 결과 지각된 행위 능력의 도덕적 책임에 대한 부적 상관(b = -0.36, SE = 0.14, p = .012)이 확인되었다. 또한 경험 능력의 도덕적 책임에 대한 정적 상관(b = 0.46, SE = 0.10, p < .001) 역시 확인되었다. 매개 효과를 제외하더라도, 의인화 지각이 도덕적 책임에 미치는 영향은 여전히 유의하였다(b = 0.44, SE = 0.15, p < .01)

의인화가 도덕적 책임 지각에 미치는 영향에서 지각된 행위 능력과 지각된 경험 능력의 간접효과 유의성 검증을 위해 부트스트래핑 (Preacher & Hayes, 2008; 5000 iteration, Model 4)을 이용해 분석하였다. 분석 결과 의인화 지각이 행위 능력에 대한 지각을 경유해 책임 지각에 이르는 간접효과(b = -0.23, SE = 0.10, CI[-.44, -.04])를 확인하였으며, 지각된 경험 능력의 간접효과 (b = 0.27, SE = 0.08, CI[.13, .44} 역시 확인되었다.

실험 2

실험 개요

실험 2는 최근 활발히 개발, 보급되고 있는 헬스케어 맥락의 인공지능을 시나리오로 구현하였다. 실험 1에서와는 달리 실험 2, 3의 실험 조건에는 시각적으로 의인화된 인공지능의 그림만을 제시하였고, 통제 조건에서는 의인화를 하지 않은 상징적 그림을 제시하였다(<그림 4> 참조).

참여자

설문조사업체 '마크로밀 엠브레인'을 통해 총 216명의 한국인 성인 패널들이 모집되었다(여성 51.4%, 나이: M = 43.5, SD = 13.70). 기타 다른 사항들은 실험 1과 동일하게 진행되었다. 참여자들은 층화 표집 방법을 사용해 연령과 성별이 고르게 수집되었다.

절차

참여자들은 실험 조건(의인화)과 통제 조건(비의인화)에 무작위 할당되어 설문을 진행하였다. 실험 1과 동일하게 의인화 조건에서는 눈, 코, 입이 그려진 인공지능의 이미지를 통한 시각적 의인화와 보다 인간적으로 인공지능을 묘사한 시나리오를 통한 언어적 의인화로 실험 조작을 수행하였다.

변인

실험 1과 동일한 설문 문항을 통해 지각된 의인화, 지각된 행위 능력과 경험 능력, 도덕적 책임, 기타 통제 요인, 인구통계학적 정보들을 설문하였다.

결과 및 논의

변인들의 기술통계량 및 상관관계를 <표 2>에 제시하였다.

조작 점검

실험 1과 같이, 통제 조건(M = 3.96, SD = 1.15)과 실험 조건(M = 3.92, SD = 1.19)간 의인화 설문 결과의 차이가 발견되지 않았다 (t(214) = 0.263, p = 0.396, Cohen's d = 0.04).

〈표 2〉 기술통계량 및 상관계수 (실험 2)

변인	M	SD	1	2	3	4	5	6	7
1. 의인화 지각	3.94	1.17							
2. 행위 능력	4.44	1.07	.62**						
3. 경험 능력	2.47	1.42	.64**	.48**					
4. 도덕적 책임	3.34	1.80	.29**	.25**	.43**				
5. 성별	-	-	-.03	.03	-.03				
6. 나이	43.5	13.7	.14*	.10	.02	-.10	.00		
7. 학력	3.66	0.95	-.16*	.06	-.16*	-.09	.00	-.03	
8. 컴퓨터 친화도	3.82	1.16	.32*	.30	.39**	.19**	-.17**	.10	.11

* p < .05. ** p < .01

의인화와 마음 지각

분석 결과 의인화 점수가 높을수록 행위 능력(b = 0.55, SE = 0.05, p < .001)과 경험 능력(b = 0.68, SE = 0.07, p < .001)을 높게 지각하여 가설 1를 지지하였다.

의인화와 도덕적 책임 지각

실험 1과 동일하게, 의인화가 도덕적 책임 지각에 정적인 영향을 미치는 것이 확인되었다. 인공지능에 대한 의인화 점수가 인공지능의 실패에 대한 도덕적 책임 지각을 유의하게 예측하였으며(b = 0.40, SE = 0.11, p < .001), 가설 2는 다시 지지되었다.

마음 지각의 매개 효과

위에서 언급한 바와 같이 의인화는 마음 지각과 정적 상관을 보였으나, 경로별 유의성 검증 결과 실험 1과는 달리 지각된 행위 능력의 도덕적 책임에 대한 유의한 영향을 확인할 수 없었다(b = 0.09, SE = 0.14, p = .53). 그러나 경험 능력의 도덕적 책임에 대한 정적 상관(b = 0.47, SE = 0.11, p < .001)은 여전히 확인되었으며, 매개 효과를 제외하였을 경우, 의인화 지각이 도덕적 책임에 미치는 영향은 더 이상 유의하지 않았다(b = 0.04, SE = 0.14, p = .80).

의인화가 도덕적 책임 지각에 미치는 영향에서 지각된 행위 능력과 지각된 경험 능력의 간접효과 유의성 검증을 위해 이전 실험과 동일한 부트스트래핑 방법을 사용하였다. 행위 능력의 간접효과는 유의하지 않았으나(b = 0.05, SE = 0.08, CI[-.12, -.21]), 지각된 경험 능력의 간접효과(b = 0.32, SE = 0.08, CI[.16, .50])는 실험 1과 같이 확인되어 가설 3을 부분적으로 지지하였다.

실험 3

실험 개요

실험 3은 아직 상용화 수준으로 개발되지는 않았으나 분명한 수요를 가지고, 실제 사람들에게 미치는 영향 역시 지대할 것으로 예측되는 법조 인공지능을 시나리오로 제작하였다. 실험에 사용된 이미지 역시 실험 2와 같은 방식으로 인간형, 상징형 인공지능 이미지를 제작해 활용하였다.

참여자

설문조사업체 '마크로밀 엠브레인'을 통해 총 210명의 실험 참여자들(여성 50%, M = 43.97, SD = 13.31)을 모집하였다. 다른 사항들은 모두 실험 1, 2와 동일하게 설정되었다. 참여자들은 층화 표집 방법을 사용해 연령과 성별이 고르게 수집되었다.

절차

참여자들은 실험 1, 2와 동일하게 통제 조건(비의인화)과 실험 조건(의인화)에 무작위로 배정되었다. 실험 조작 역시 마찬가지로 얼굴 이미지를 통한 시각적 의인화와 시나리오상의 서술을 통한 언어적 의인화로 이루어졌다.

변인

실험 1, 2와 동일한 설문 문항을 통해 지각된 의인화, 지각된 행위 능력과 지각된 경험 능력, 도덕적 책임, 그리고 통제 요인과 기타 인구통계학적 정보들을 설문하였다.

결과 및 논의

변인들의 기술통계량 및 상관관계를 <표 3>에 제시하였다.

〈표 3〉 기술통계량 및 상관계수 (실험 3)

변인	M	SD	1	2	3	4	5	6	7
1. 의인화 지각	3.74	1.19							
2. 행위 능력	4.29	1.16	.50**						
3. 경험 능력	2.36	1.35	.57**	.47**					
4. 도덕적 책임	3.35	1.72	.49**	.34**	.43**				
5. 성별	-	-	-.03	-.04	-.07	.00			
6. 나이	43.97	13.31	.08	-.06	-.11	-.16*	-.02		
7. 학력	3.78	0.89	-.03	.08	.02	.12	-.13	-.06	
8. 컴퓨터 친화도	3.82	1.14	.18**	.26**	.28**	.19**	-.19**	.02	.07

* p < .05. ** p < .01

조작 점검

실험 1, 2와 같이, 통제 조건(M = 3.72, SD = 1.21)과 실험 조건(M = 3.76, SD = 1.18)간 의인화 설문 결과의 차이가 발견되지 않았다(t(208) = -0.27, p = 0.393, Cohen's d = -0.04).

의인화와 마음 지각

참여자들은 의인화 점수가 높을수록 행위 능력(b = 0.47, SE = 0.06, p < .001)과 경험 능력(b = 0.62, SE = 0.06, p < .001)을 높게 지각하였고, 가설 1를 지지하였다.

의인화와 도덕적 책임 지각

실험 1, 2에 이어 실험 3에서 역시 의인화가 도덕적 책임 지각에 정적인 영향을 미치는 것을 확인하였다. 인공지능에 대한 의인화 점수가 인공지능의 실패에 대한 도덕적 책임 지각을 유의하게 예측하였으며(b = 0.71, SE = 0.09, p < .001), 가설 2는 다시 지지되었다.

마음 지각의 매개 효과

실험 1, 2와 동일하게 행위 능력 지각과 경험 능력 지각의 도덕적 책임에 대한 영향을 PROCESS를 통해 분석하였다. 분석 결과 실험 2와 동일하게 지각된 행위 능력은 도덕적 책임에 유의미한 영향을 미치지 못하였으며(b = 0.56, SE = 0.10, p = .63), 인공지능에 대한 지각된 경험 능력은 인공지능의 도덕적 책임에 유의한 영향을 보였다(b = 0.21, SE = 0.10, p < .05). 매개 효과를 제외하더라도 의인화 지각의 도덕적 책임 지각에 미치는 영향은 여전히 유의하였다(b = 0.56, SE = 0.11, p < .001).

부트스트래핑을 이용한 간접효과 유의성 검증 결과, 실험 2와 같이 의인화에 대한 지각과 도덕적 책임 지각 간의 관계에서 지각된 경험 능력만의 유의성을 확인할 수 있었다(b = 0.13, SE = 0.07, CI[.004, .26]. 지각된 행위 능력의 간접 효과는 유의하지 않았으며(b = 0.02, SE = 0.06, CI[-.11, .13], 실험 2과 같은 방식으로 가설 3을 부분적으로 지지하였다.

종합 논의

본 연구는 사람들이 로봇, 챗봇과 같은 비인간 대상들의 도덕 위반을 지각할 때 의인화의

영향을 도덕적 책임의 맥락에서 알아보고자 수행되었다. 의인화가 마음 지각에 유의한 영향을 미칠 것이라는 가설 1은 모든 실험에서 확인되었으나, 마음 지각이 도덕적 책임에 미치는 영향은 각 실험에서 상이하게 나타났다. 면접 로봇이라는 맥락에서 수행되었던 실험 1에서는 지각된 행위 능력이 도덕적 책임과 부적 상관을 보였고, 지각된 경험 능력은 정적 상관을 보였다. 의료 인공지능 맥락의 실험 2에서는 행위 능력은 책임 지각에 유의한 영향을 미치지 못하였으나 경험 능력은 실험 1과 같이 유의한 정적 상관을 보였고, 이는 실험 3에서도 재현되었다.

이러한 결과는 의인화와 마음 지각이 밀접히 연관되어 있으며(가설 1), 의인화가 도덕적 책임 지각에 영향을 준다는 사실을 뒷받침한다 (가설 2). 여기에 더해 비인간 대상에 대해 지각된 행위 능력과 지각된 경험 능력이 도덕적 책임에 미치는 영향을 확인할 수 있었으나, 마음 지각과 도덕적 책임에 대한 가설 3은 부분적으로 지지되었다. 먼저 정적 영향을 예상하였던 지각된 행위 능력의 간접효과는 일관되게 관측되지 않았다. 이에 반해 경험 능력의 경우 모든 실험에 걸쳐 도덕적 책임에 대한 유의미한 간접효과를 관측할 수 있었다.

이는 쌍 도덕 이론이 강조하는 행위-책임, 그리고 경험-권리의 관계와는 일치하지 않으며, 도덕적 지위에 관한 유기체적 관점 (Torrance, 2006; 2008; cf. Gunkel, 2012; Tollon, 2020)을 지지하는 것이다. 웨그너와 그레이는 기계의 마음을 다룬 그들의 글에서 '로봇이 도덕적 행위자가 될 수 있는가'라는 질문을 던지지만, 도덕적 책임과 관련해서는 뚜렷한 해답을 내놓지 못하고 있다 (Wegner & Gray, 2017). 그러나 기계에 대한 마음 지각과 불쾌한 골짜기 효과를 다룬 그들의 연구(Gray & Wegner, 2012a)에서 보듯, "경험 능력-행위 능력이 아닌-이 인간의 근본적 조건이며, 기계가 근본적으로 결여하고 있는 것"이다(직접 번역). 이러한 관점에서, 도덕적 책임은 인간만이 질 수 있지만, 마음 지각의 관점에서 기계는 '경험의 간극(experience gap; Wegner & Gray, 2017)'을 넘지 않는 한 인간과 같아질 수 없는 것처럼 보인다.

또한 도덕적 지위에 관한 유기체적 관점에 따르면, 도덕적 행위자와 수동자의 관계는 비대칭적이다. 즉, 모든 도덕적 수동자가 도덕적 행위자일 필요는 없지만(아기는 행위 능력이 없이도 도덕적 권리를 가지지만), 오직 도덕적 수동자만이 도덕적 행위자의 지위와 책임를 가질 수 있다(로봇은 도덕적 책임을 가지기 위해 경험 능력을 필요로 한다). 토렌스는 그의 논문에서 진정한 윤리적 책임을 위해서는 타인의 상황에 공감하고, 자의적으로 행동하며 윤리적으로 바람직한 상황을 욕망하거나 원하며, 적절한 도덕적 감정과 감응력 등이 필요하며, 이러한 이유에서 기계적 로봇은 도덕적 행위자가 될 수 없다고 주장한다(Torrance, 2006). 같은 맥락에서 아자로(2011)가 그의 글 제목에서 말하듯, 로봇은 "걷어찰 몸은 있더라도 비난할 영혼은 없다("a body to kick but no soul to damn", 직접 번역). 기계의 상대적 우월성에 대한 사람들의 반응을 다룬 선행 연구(Cha et al., 2020) 역시 인공지능의 계산적, 이성적 우월성을 대면했을 때 사람들이 사회성과 감정 능력과 같은 인간만의 특성에 가치를 부여함으로써 인간 고유성을 회복하려 한다는 사실

을 확인하였다.

또한 경험 능력과 도덕적 책임의 관계는 선행 연구의 맥락에서 부분적으로 설명된다. 여러 대상에 대한 마음 지각을 비교한 그레이 등의 선행 연구에서는 대상의 잘못에 대한 처벌에 대한 평가가 지각된 경험 능력과 정적 상관을 보인 바 있다(b = 0.22; Gray, Gray, & Wegner, 2007). 또한 기계적 로봇, 휴머노이드 로봇, 그리고 인간에 대한 도덕적 책임 지각을 다룬 말레 등의 연구(Malle et al., 2016)가 보여주듯이, 인간적인 특성을 드러내는 로봇들은 인간만큼이나 많은 비난을 받는 반면 기계적 특성이 두드러지는 로봇들은 훨씬 더 적은 비난을 받는다. 그리고 마음 지각의 맥락에서 기계적 로봇과 인간적 로봇, 혹은 인간의 차이는 행위 능력보다는 경험 능력에 있다(<그림 2>). 또한 컴퓨터의 도덕적 의사결정에 대한 회피를 다룬 빅맨과 그레이의 연구(Bigman & Gray, 2018) 역시 컴퓨터의 의사결정을 더 잘 받아들여지게 하기 위한 방법 중 하나로 지각된 경험을 높이는 방안을 제시한 바 있다.

연구 의의

위와 같은 실험 결과가 본 연구에서 갖는 의의는 다음과 같다. 첫째, 도덕심리학적 접근을 통해 인공지능의 의인화와 마음 지각의 효과를 도덕적 책임이라는 맥락에서 확인하였다. 특히 기존에 주목받지 못하였던 지각된 경험 능력과 책임 지각의 관계를 밝힘으로써 인간과 상호작용할 로봇들의 적절한 디자인을 탐구하는 도덕적 HRI(moral HRI; Malle et al., 2015) 연구에 기여하였다. 오직 인간만이 진정한 의미에서 도덕적 책임의 주체가 될 수 있으며, 로봇이 인간보다 상대적으로 부족한 능력은 행위 능력보다는 경험 능력이다. 토렌스가 주장하듯, 진정한 도덕적 지위를 위해서는 타자의 고통을 이해하고 도덕적 고려의 대상으로 삼을 수 있는 능력인 공감적 합리성(empathic rationality; Torrance, 2008)이 필요한 것으로 보인다.

둘째로 사회 내에서 지대한 영향력을 가지게 된 인공지능 시나리오를 실험에 적용하였다. 지금까지 연구되었던 인공적 행위자들은 대부분 고객의 요구에 맞춰진 수동적 행위자로서의 위치를 가지고 있다. 가령 집안일을 공평하게 나누는 도구적 알고리즘(Lee, 2017), 호텔의 종업원 로봇(Yam et al., 2020)이 그 사례이다. 이러한 맥락에서 본 연구는 실제 중대한 도덕적 문제들을 야기할 수 있는 분야의 인공지능에 대해 일반 대중이 어떻게 반응하는지에 대해 알아봄으로써 논의의 범위를 확장시켰다.

마지막으로 본 연구는 사람들의 인공물의 도덕적 지위에 대한 철학적, 법적 논의에서 하나의 근거로 활용될 수 있다. 점차 늘어나는 자율성과 영향력에 따라, 인공적 행위자의 도덕적, 법적 지위와 책임에 관한 논의의 필요성 역시 중요해지고 있다. 사람들이 어떤 대상의 도덕적 지위를 인식할 때 그 행위자의 경험 능력이 주요한 역할을 한다는 관측은 윤리적 차원의 논의에서 우

리의 직관이 어떤 방식으로 작동하는지 알려준다는 점에서 유용하다.

한계점

본 연구는 의인화가 도덕적 책임에 미치는 영향을 도덕 쌍 이론을 통해 알아보았으며, 이 과정에서 인공지능의 '도덕적 책임'이라는 단일한 개념을 참여자들에게 설문하였다. 비록 도덕심리학에서 도덕적 그름, 또는 도덕적 책임을 자주 하나의 설문 문항으로 취급하지만, 실제로 도덕적 책임은 인과적 맥락, 법적 맥락 등 다양한 의미를 가질 수 있다. 본 연구에서는 사람들이 일반적으로 생각하는 도덕적 책임의 개념을 포착하기 위해 '도덕적 책임'이라는 단일한 개념을 설문하였으나, 그 의미를 더욱 구체적으로 포착하기 위해서는 후속 연구를 통한 보다 구체적인 설문이 필요하다.

본 연구에서 실험한 도덕 위반 시나리오 외에도 도덕적이지 않은 위반, 즉 단순한 기능적 오류 시나리오와의 비교가 필요하다는 지적 역시 가능하다. 가령 단순한 작동 오류를 다룬 연구 중 하나로, 얌 등은 로봇 호텔에서 의인화와 작동 오류에 따른 고객 만족도의 관계를 실험한 바 있다. 해당 연구 결과, 로봇에서 경험 능력을 강하게 느낄 수록 오류를 일으킨 로봇에 대한 사용자 만족도 역시 높아졌다(Yam et al., 2020). 이러한 결과는 본 실험에서 확인한 경험 능력 지각과 도덕적 책임 간의 정적 상관과는 대비되는데, 일반적으로 도덕적 책임을 크게 지각한다면 자연히 사용자 만족도는 낮을 것이라 예측되기 때문이다. 얌 등은 이러한 현상을 지각된 경험 능력이 고객으로 하여금 로봇이 더 죄책감을 느끼고 반성한다고 느끼게 하기 때문이라 추정한다.

다만 얌의 연구에서 설정된 오류가 단순히 설문 후에 받을 상품(초콜릿 바 혹은 감자칩)을 잘못 가져다주는 수준임을 고려한다면, 첫째, 작은 실수 수준의 오류는 인간과 닮을수록 용서하기 쉽고 책임을 덜 느끼지만, 큰 오류의 경우에는 그렇지 않을 수 있다. 둘째로, 의인화 지각에 따라 책임은 여전히 더 크게 느끼지만 동시에 더 쉽게 용서할 수 있다는 해석 역시 가능하다. 따라서 후속 연구에서는 도덕적 오류와 기능적 오류에 대한 보다 정밀한 정의와 통제 조건을 포함함으로써 오류의 종류와 그 영향, 그리고 도덕적 책임 지각에 대한 관계 규명이 필요하다 하겠다.

또한 비인간의 의인화를 다루는 연구들은 대부분 불쾌한 골짜기 현상과 관련된 문제에서 벗어나기 어렵다. 인간과 과하게 닮은 로봇은 사람들에게 불쾌감을 준다고 알려져 있으며, 실험 결과에 부정적 영향을 미칠 수 있다. 이러한 문제를 본 연구 역시 내포하고 있으며, 때문에 기존에 존재하는 면접 로봇의 이미지를 대체하거나 제작해 실험을 수행하였다. 그러나 참여자들이 실제로 해당 이미지에서 불쾌감을 느꼈는지 확인하는 별도의 설문을 거치지 않았다는 점에서

불쾌한 골짜기와 관련된 우려에서 자유롭지 못하다.

의인화에 관한 선행 리뷰(Li & Suh, 2021)에서 보듯, 인공지능을 의인화하는 방식에는 시각적, 언어적, 심리적 의인화와 같이 다양한 방식이 존재한다. 본 연구는 시각적(눈, 입이 있는 얼굴), 언어적(사람같은 시나리오 상의 표현)을 사용하였으나, 실제로 의인화의 조작 효과는 유의미하게 나타나지 않았다. 본 실험에서 매체의 한계로 시각적, 언어적 조작만을 수행하였으나, 실제 인공지능 사용 환경에서는 이러한 특징들 이외에도 인공지능의 목소리와 움직임 등이 적극적으로 사용되는 만큼 보다 다양한 방식의 의인화 조작을 통한 실험이 차후 연구들에서 요구된다.

마지막으로 본 연구는 참여자들로 하여금 제 3자의 입장에서 도덕적 책임에 대한 판단을 내리는 상황을 실험으로 구성하였지만, 실제 상황에서 인공지능으로 인한 피해 당사자들이 어떠한 방식으로 도덕적 책임을 인식할 지는 알아보지 않았다. 물론 많은 상황에서 여론은 제 3자들에 의해 구성되지만, 실제 인공지능 및 개발사와 부딪히게 되는 피해자들의 인식 역시 매우 중요하다. 이러한 맥락에서, 유사한 상황에서의 다른 행위자들과 관련된 물음들은 미래 연구를 통해 보완이 필요하다 하겠다.

결 론

본 연구는 인간 사회 내에서 영향력을 점점 확대해나가고 있는 인공지능 분야에 발맞추어 인공지능의 도덕적 실패 상황에 대한 사람들의 반응을 의인화와 도덕심리학이라는 관점에서 설명하였다. 도덕심리학은 도덕적 책임을 주로 가해자의 행위 능력과 연관짓지만, 비인간 행위자의 진정한 도덕적 지위는 행위 능력의 획득만으로는 불충분하다. 이렇듯 인간 사회 속에서 다양한 활동을 수행하게 될 미래의 인공지능에 대한 연구는 여러 관점에서의 접근이 효과적이며, 본 연구가 더 윤리적이고 사회적인 인공지능의 개발을 위한 통섭적 연구에 기여하기를 기대한다.

참고문헌

권헌영 (2019). 인공지능(AI)과 법조 분야: 윤리적・규제적 고려사항. **경제규제와 법**, 12(2), 69-80.

Adler-Milstein, J., Holmgren, A. J., Kralovec, P., Worzala, C., Searcy, T., & Patel, V. (2017). Electronic health record adoption in US hospitals: the emergence of a digital “advanced use” divide. *Journal of the American Medical Informatics Association: JAMIA, 24*(6), 1142-1148.

https://doi.org/10.1093/jamia/ocx080

Ajunwa, I., Friedler, S., Scheidegger, C. E., & Venkatasubramanian, S. (2016). Hiring by algorithm: predicting and preventing disparate impact. Available at SSRN.

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

Artificial intelligence: Go master Lee Se-dol wins against AlphaGo program (2016, March 13). BBC News Online. https://www.bbc.com/news/technology-35797102.

Asaro, P. M. (2011). 11 A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. Robot ethics: The ethical and social implications of robotics, 169.

Ayasdi (2018). Ayasdi for Payers: white paper. Ayasdi. https://s3.amazonaws.com/cdn.ayasdi.com/wp-content/uploads/2018/10/05102657/WP-Ayasdi-for-Payers.pdf

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. Cognition, 181, 21-34.

Cantarero, K., Szarota, P., Stamkou, E., Navas, M., & Dominguez Espinosa, A. D. C. (2021). The effects of culture and moral foundations on moral judgments: The ethics of authority mediates the relationship between power distance and attitude towards lying to one's supervisor. Current Psychology, 40(2), 675-683.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining (pp. 797-806).

Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation'with a new questionnaire. *Journal of Research in Personality, 78,* 106-124.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data, 6*(1), 1-25.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56-62.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114.

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

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. Oxford review, 5.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In Advances in experimental social

psychology (Vol. 47, pp. 55-130). Academic Press.

Graham, J., Haidt, J., Motyl, M., Meindl, P., Iskiwitch, C., & Mooijman, M. (2018). Moral foundations theory: On the advantages of moral pluralism over moral monism. In K. Gray & J. Graham (Eds.), Atlas of moral psychology (pp. 211-222). The Guilford Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology, 96*(5), 1029.

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

Gray, K., Jenkins, A. C., Heberlein, A. S., & Wegner, D. M. (2011). Distortions of mind perception in psychopathology. *Proceedings of the National Academy of Sciences, 108*(2), 477-479.

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

Gray, K., & Wegner, D. M. (2012). Morality takes two: Dyadic morality and mind perception.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108.

Gunkel, D. J. (2012). The machine question: Critical perspectives on AI, robots, and ethics. mit Press.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog?. *Journal of personality and social psychology, 65*(4), 613.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review, 108(4), 814.

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Vintage.

HLEG, A. I. (2019). High-level expert group on artificial intelligence: Ethics guidelines for trustworthy AI. European Commission, 09.04.

Hollister, B., & Bonham, V. L. (2018). Should electronic health record-derived social and behavioral data be used in precision medicine research?. AMA journal of ethics, 20(9), 873-880.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. Handbook of socialization theory and research, 347, 480.

Kohlberg, L. (2016). 1. Stages of moral development as a basis for moral education. In C. Beck, B. Crittenden & E. Sullivan (Ed.), Moral Education (pp. 23-92). Toronto: University of Toronto Press. https://doi.org/10.3138/9781442656758-004

Kuncel, N. R., Klieger, D. M., & Ones, D. S. (2014). In hiring, algorithms beat instinct. Harvard business review, 92(5), p32-32.

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: a robot's appearance moderates

how its decisions are judged. International Journal of Social Robotics, 1-10.

Larsen, R. R. (2020). Psychopathy as moral blindness: a qualifying exploration of the blindness-analogy in psychopathy theory and research. Philosophical Explorations, 23(3), 214-233.

Lee, D. (2016, March 25). Tay: Microsoft issues apology over racist chatbot fiasco. BBC News Online. https://www.bbc.com/news/technology-35902104

Li, M., & Suh, A. (2021, January). Machinelike or Humanlike? A Literature Review of Anthropomorphism in AI-Enabled Technology. In Proceedings of the 54th Hawaii International Conference on System Sciences (p. 4053).

MacDorman, K. F. (2005, July). Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it. In CogSci-2005 workshop: toward social mechanisms of android science (Vol. 106118).

MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. Interaction Studies, 16(2), 141-172.

MacDorman, K. F., Green, R. D., Ho, C. C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. Computers in human behavior, 25(3), 695-710.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 117-124). IEEE.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). 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.

Min, J., Kim, S., Park, Y., & Sohn, Y. W. (2018). A Comparative Study of Potential Job Candidates' Perceptions of an AI Recruiter and a Human Recruiter. *Journal of the Korea Convergence Society, 9*(5), 191-202.

Moosa, M. M., & Ud-Dean, S. M. (2010). Danger avoidance: An evolutionary explanation of uncanny valley. Biological Theory, 5(1), 12-14.

Mori, M. (1970). Bukimi no tani [the uncanny valley]. Energy, 7, 33-35.

Morse, S. J. (2008). Psychopathy and criminal responsibility. Neuroethics, 1(3), 205-212.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of social issues, 56(1), 81-103.

Natarajan, M., & Gombolay, M. (2020, March). Effects of anthropomorphism and accountability on trust in human robot interaction. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 33-42).

Newborn, M. (2012). Kasparov versus Deep Blue: Computer chess comes of age. Springer Science & Business Media.

O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Oxford Dictionary. (n.d.). artificial intelligence. In Oxford English Dictionary. Retrieved October 28, 2021, from https://www.oed.com/viewdictionaryentry/Entry/271625

Ötting, S. K., & Maier, G. W. (2018). The importance of procedural justice in human–machine interactions: Intelligent systems as new decision agents in organizations. Computers in Human Behavior, 89, 27-39.

Savage, M. (2019, March 19). Meet Tengai, the job interview robot who won't judge you. BBC News Online. https://www.bbc.com/news/business-47442953

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.

Schein, C., Ritter, R. S., & Gray, K. (2016). Harm mediates the disgust-immorality link. Emotion, 16(6), 862.

Tollon, F. (2021). The artificial view: toward a non-anthropocentric account of moral patiency. Ethics and Information Technology, 23(2), 147-155.

Torrance, S. (2006). The ethical status of artificial agents-with and without consciousness. Ethics of human interaction with robotic, bionic and AI systems: concepts and policies. Istituto Italiano per gli Studi Filosofici, Napoli, 60-66.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *Ai & Society*, *22*(4), 495-521.

Verma, N., & Dombrowski, L. (2018, April). Confronting social criticisms: Challenges when adopting data-driven policing strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Wang, W. (2017). Smartphones as social actors? Social dispositional factors in assessing anthropomorphism. *Computers in Human Behavior*, *68*, 334-344.

Wang, R., Harper, F. M., & Zhu, H. (2020, April). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

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.

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.

Wegner, D. M., & Gray, K. (2017). *The mind club: Who thinks, what feels, and why it matters*. Penguin.

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2020). Robots at work: People prefer-and forgive-service robots with perceived feelings. *Journal of Applied Psychology*.

1차 원고 접수: 2022. 04. 27
1차 심사 완료: 2022. 08. 30
2차 원고 접수: 2022. 09. 14
2차 심사 완료: 2022. 09. 21
최종 게재 확정: 2022. 09. 23

(*Abstract*)

Is Mr. AI more responsible? The effect of anthropomorphism in the moral judgement toward AI's decision making

Yoon-Bin Choi[1)] Dayk Jang[2)]

[1)]Interdisciplinary Program in Cognitive Science, Seoul National University

[2)]Gachon Startup College, Gachon University

As artificial intelligence (AI) technology advances, the number of cases in which AI becomes an object or subject of moral judgment is increasing, and this trend is expected to accelerate. Although the area of AI in human society expands, relatively few studies have been conducted on how people perceive and respond to AI. Three studies examined the effect of the anthropomorphism of AI on its responsibility. We predicted that anthropomorphism would increase the responsibility perception, and perceived agency and perceived patiency for AI would mediate this effect. Although the manipulation was not effective, multiple analyses confirmed the indirect effect of perceived patiency. In contrast, the effect of perceived agency of AI was somewhat mixed, which makes the hypothesis partially supported by the overall result. This result shows that for the moral status of artificial agents, perceived patiency is relatively more critical than perceived agency. These results support the organic perspective on the moral status that argues the importance of patiency, and show that patiency is more important than agency in the anthropomorphism related study of AI and robots.

Key words : artificial intelligence, anthropomorphism, moral responsibility, mind perception, moral dyad theory

부록 A

시나리오 (실험 1 ~ 3)

실험 1(면접) - 비의인화

- <기자>
    - 이곳은 미국 T사의 면접장. 면접관들 사이에 AI가 탑재된 모니터 하나가 올려져있습니다. 단지 자리만 차지하고 있는 것이 아니라, 지원자에게 질문도 하고, 궁금한 점에 답변도 해줍니다.
- <음성>
    - (KJI-7) “안녕하세요, 저는 오늘 여러분의 면접을 담당한 KJI-7입니다.”
    - (KJI-7) “회사 전산 시스템의 결함으로 인해 개인정보가 유출되었다면 어떻게 하시겠습니까?”
    - (KJI) “회사의 복지로는 학자금 지원, 의료보험이 유명합니다”
- <기자>
    - KJI-7은 스웨덴의 한 기업에서 2019년부터 개발된 세계 최초의 면접 인공지능입니다. 단순히 데이터를 분석하고 조언하는 역할에서 그쳤던 기존의 면접 AI와는 달리, ‘KJI-7’은 다른 인간 면접관의 옆에 실제로 배치되어 지원자에게 질문하고 판단합니다.
    - 최근 미국의 한 스타트업에서는 이 인공지능을 사무실에 배치해 다른 직원들이 이용할 수 있도록 하기도 했습니다.
- <인터뷰(임직원)>
    - “지원자들이 KJI-7과 이야기할 때 더 자연스러운 분위기에서 면접에 임할 수 있도록 노력하고 있습니다. 뿐만 아니라 회사 안에 KJI-7을 위한 자리를 마련하면 우리 직원들도 익숙해질 것이라 기대합니다.” (후략)

- 한 스타트업이 개발한 인공지능(AI) ‘KJI-7’이 채용 과정에서의 차별과 부적절한 질문으로 논란에 휩싸였다. 지난 2019년 미국의 모 기업에서 만들어진 인공지능 ‘KJI-7’은 최근 여러 기업의 채용 과정에 사용되며 그 이름을 알려왔다. 그러나 최근 지원자들의 폭로와 정부 조사에 의해 밝혀진 바에 따르면, 해당 인공지능은 면접 과정에서 여성/인종차별적 발언을 자주 표출해왔던 것으로 밝혀졌다.(후략)

실험 1(면접) - 의인화

- <기자>
    - 이곳은 미국 T사의 면접장. 면접관들 사이에 사람같이 생긴 로봇이 앉아있습니다. 단지 자리만 차지하고 있는 것이 아니라, 지원자에게 질문도 하고, 궁금한 점에 답변도 해줍니다.
- <음성>
    - (진수) “안녕하세요, 저는 오늘 여러분의 면접을 담당한 진수 팀장입니다.”
    - (진수) “회사 전산 시스템의 결함으로 인해 개인정보가 유출되었다면 어떻게 하시겠습니까?”
    - (진수) “회사의 복지로는 학자금 지원, 의료보험이 유명합니다”
- <기자>
    - 진수 팀장은 스웨덴의 한 기업에서 2019년부터 개발된 세계 최초의 AI 면접관입니다. 단순히 데이터를 분석하고 조언하는 역할에서 그쳤던 기존의 면접 AI와는 달리, 진수 면접관은 다른 인간 면접관의 옆에 실제로 앉아 지원자에게 질문하고 판단합니다.
    - 최근 미국의 한 스타트업에서는 이 로봇을 사무실에서 다른 직원들과 함께 일할 수 있게 하기도 했습니다.
- <인터뷰(임직원)>
    - “지원자들이 진수 팀장과 이야기할 때 더 자연스러운 분위기에서 면접에 임할 수 있도록 노력하고 있습니다. 뿐만 아니라 회사 안에 진수 팀장을 위한 자리를 제공하면 우리 직원들과도 더 친밀한 관계를 쌓을 수 있을 것이라 기대합니다.” (후략)

- 한 스타트업이 개발한 인공지능(AI) ‘진수’가 채용 과정에서의 차별과 부적절한 질문으로 논란에 휩싸였다. 지난 2019년 미국의 모 기업에서 만들어진 인공지능 ‘진수’는 최근 여러 기업의 채용 과정에 사용되며 그 이름을 알려왔다. 그러나 최근 지원자들의 폭로와 정부 조사에 의해 밝혀진 바에 따르면, ‘진수’는 면접 과정에서 여성/인종차별적 발언을 자주 표출해왔던 것으로 밝혀졌다.(후략)

실험 2(의료) - 비의인화

- <기자>
    - 이곳은 미국 T시의 한 병원. 응급실 내부에 패스트푸드점에서나 보던 키오스크 화면이 배치되어 있습니다. 그 안에는 인공지능이 있다고 하는데요, 단지 자리만 차지하

고 있는 것이 아니라, 환자에게 질문도 하고, 치료에 관한 질문에 답변도 해줍니다.

- <음성>
  - (KMA-7) “저는 의료 상담을 제공하는 인공지능 ‘KMA-7’입니다.”
  - (KMA-7) “통증이 느껴지는 부위를 화면에서 터치해주십시오.”
  - (KMA-7) “데이터 분석에 따르면 독감일 확률: 68%.”
- <기자>
  - (KMA-7)은 스웨덴의 한 기업에서 2019년부터 개발된 의료용 인공지능입니다. 단순히 의료 데이터를 분석하는 역할에서 그쳤던 기존의 헬스케어 AI와는 달리, KMA-7은 환자나 의사가 직접 사용하고 조언을 얻을 수 있습니다.
  - 최근 미국의 한 병원에서는 이 인공지능을 의료실에 배치하고, 다른 사람들이 이용할 수 있게 하기도 했습니다.
- <인터뷰(관계자)>
  - “환자분들이 ‘KMA-7’를 통해 더 편리하게 진료를 받고, 어디서나 자신의 데이터에 접근 수 있도록 하려고 합니다. 뿐만 아니라 병원 안에 KMA-7을 위한 자리를 제공하면 우리 직원들도 더 적극적으로 활용할 수 있을 것이라 기대하고 있습니다.” (후략)

- 한 스타트업이 개발한 인공지능 의사 ‘KMA-7’이 상담 과정에서의 오류로 논란에 휩싸였다. 지난 2019년 미국의 M기업에서 만들어진 인공지능 ‘KMA-7’은 최근 몇몇 병원의 응급실 등에서 사용되며 그 이름을 알려왔다. 그러나 최근 병원 내부자의 폭로와 환자들의 고소 내용에 따르면, KMA-7은 같은 이름의 환자를 다른 환자로 착각해 환자 정보를 유출하거나, 치료 내용을 잘못 알려줘 환자 가족에게 혼란을 주는 등의 오류를 자주 일으켰던 것으로 밝혀졌다. (후략)

실험 2(의료) - 의인화

- <기자>
  - 이곳은 미국 T시의 한 병원. 응급실 내부에 패스트푸드점에서나 보던 키오스크 화면이 배치되어 있습니다. 그 안에는 사람같이 생긴 인공지능이 있는데요, 단지 자리만 차지하고 있는 것이 아니라, 환자에게 질문도 하고, 치료에 관한 질문에 답변도 해줍니다.
- <음성>
  - (제인) “안녕하세요, 저는 여러분께 의료 상담을 드리는 어시스턴트 ‘제인’입니다.”

    - (제인) “어느 부위가 아프신가요? 화면에서 터치해주세요.”
    - (제인) “제 소견에 따르면 환자분이 독감일 확률은 68%입니다.”
- <기자>
    - (제인)은 스웨덴의 한 기업에서 2019년부터 개발된 AI 의사입니다. 단순히 데이터를 분석하고 의사에게 조언하는 역할에서 그쳤던 기존의 헬스케어 AI와는 달리, 제인은 다른 환자들과 직접 소통하고 판단합니다.
    - 최근 미국의 한 병원에서는 이 인공지능을 레지던트로 대하고, 다른 사람들과 함께 상호작용할 수 있게 하기도 했습니다.
- <인터뷰(관계자)>
    - “환자분들이 ‘제인’과 이야기할 때 더 자연스러운 분위기에서 진료를 받고, 어디서나 자신의 데이터에 접근할 수 있도록 하려고 합니다. 뿐만 아니라 병원 안에 제인을 위한 자리를 제공하면 우리 직원들과도 더 친밀한 관계를 쌓을 수 있을 것이라 기대하고 있습니다.” (후략)

- 한 스타트업이 개발한 인공지능 의사 ‘제인’이 상담 과정에서의 오류로 논란에 휩싸였다. 지난 2019년 미국의 M기업에서 만들어진 인공지능 ‘제인’은 최근 몇몇 병원의 응급실 등에서 사용되며 그 이름을 알려왔다. 그러나 최근 병원 내부자의 폭로와 환자들의 고소 내용에 따르면, ‘제인’은 같은 이름의 환자를 다른 환자로 착각해 환자 정보를 유출하거나, 치료 내용을 잘못 알려줘 환자 가족에게 혼란을 주는 등의 오류를 자주 일으켰던 것으로 밝혀졌다.(후략)

실험 3(법조) - 비의인화

- <기자>
    - 최근 한국에서 개발된 법조 AI 챗봇 ‘KLA-7’이 해외에서 인기를 끌고 있습니다. ‘스마트폰 안의 변호사’라는 문구에서 알 수 있듯이, 화면 속에는 법조 상담을 위한 인공지능이 있습니다. 사용자는 채팅으로 이 인공지능에게 실시간으로 질문하며, 상대적으로 저렴한 비용으로 법적 조언들을 얻을 수 있다고 합니다.
- <영상>
    - (KLA-7) “저는 여러분의 법률 상담을 맡고 있는 KLA-7입니다.”
    - (KLA-7) “본 사건과 관련된 판결 사례들의 리스트를 보내드리겠습니다.”
    - (KLA-7) “예상 형량: 약 징역 3개월, 집행유예 5개월.”
- <기자>
    - KLA-7은 법조문은 물론 이전의 판례들을 입력해 가장 유사한 사례들을 찾아주고, 처

벌이나 비용 등에 대한 예측 서비스를 제공합니다. 특히 법적 비용이 상대적으로 비싼 미국 등에서 시범적으로 사용되기 시작했습니다.

- 최근 미국의 한 법률 서비스 기업에서는 이 챗봇을 홈페이지에 배치하고, 직원과 고객들에게 상담 서비스를 제공하고 있습니다.

- <인터뷰(관계자)>
  - 저희 회사는 법적 도움을 필요로 하는 고객이 점차 늘어남에 따라, KLA-7을 통해 고객분들이 더 간편하고 신속하게 도움을 받을 수 있게 했습니다.

- 한 스타트업이 개발한 인공지능(AI) 법조인 KLA-7(Lawbin)이 인종차별 문제로 논란에 휩싸였다. 지난 2020년 한국의 L사에서 만들어진 인공지능 AI는 최근 미국의 한 법률 기업에서 사용되며 한국에도 그 이름이 알려진 바 있다. 그러나 최근 기업 내부에서의 폭로와 정부 조사에 의해 밝혀진 바에 따르면, '로빈'은 상담 과정에서 흑인의 형량을 백인보다 더 높게 책정하거나, 유사한 데이터임에도 흑인의 가석방 가능성을 더 낮게 예상하는 등 인종차별적 발언을 내놓아온 것으로 밝혀졌다.(후략)

실험 3(법조) - 의인화

- <기자>
  - 최근 한국에서 개발된 법조 AI 챗봇 '로빈'이 해외에서 인기를 끌고 있습니다. '스마트폰 안의 변호사'라는 문구에서 알 수 있듯이, 화면 속에는 정장을 입은 인공지능이 자리하고 있습니다. 사용자는 채팅으로 이 인공지능과 실시간으로 대화하며, 상대적으로 저렴한 비용으로 법적 조언들을 얻을 수 있다고 합니다.
- <영상>
  - (로빈) "안녕하세요, 저는 여러분의 법률 상담을 맡고 있는 로빈입니다."
  - (로빈) "본 사건과 관련된 판결 사례들의 리스트를 보내드리겠습니다."
  - (로빈) "제 판단으로는, 고객님의 예상 형량은 약 징역 3개월, 집행유예 5개월입니다."
- <기자>
  - 로빈은 법조문은 물론 이전의 판례들을 입력해 가장 유사한 사례들을 찾아주고, 처벌이나 비용 등에 대한 예측 서비스를 제공합니다. 특히 법적 비용이 상대적으로 비싼 미국 등에서 시범적으로 사용되기 시작했습니다.
  - 최근 미국의 한 법률 서비스 기업에서는 이 챗봇을 홈페이지에 배치하고, 직원과 고객들에게 상담 서비스를 제공하고 있습니다.

- <인터뷰(관계자)>
    - 저희 회사는 법적 도움을 필요로 하는 고객이 점차 늘어남에 따라, 고객분들이 더 간편한 동시에 사람만큼 친절한 로빈을 통해 쉽게 도움을 받을 수 있게 했습니다.

- 한 스타트업이 개발한 인공지능(AI) 법조인 로빈(Lawbin)이 인종차별 문제로 논란에 휩싸였다. 지난 2020년 한국의 L사에서 만들어진 인공지능 AI는 최근 미국의 한 법률 기업에서 사용되며 한국에도 그 이름이 알려진 바 있다. 그러나 최근 기업 내부에서의 폭로와 정부 조사에 의해 밝혀진 바에 따르면, 해당 인공지능은 상담 과정에서 흑인의 형량을 백인보다 더 높게 책정하거나, 유사한 데이터임에도 흑인의 가석방 가능성을 더 낮게 예상하는 등 인종차별적 발언을 내놓아온 것으로 밝혀졌다.(후략)

부록 B: 설문 문항

의인화

이 인공지능이 얼마나 의도를 가질 수 있는 것처럼 보이십니까?
이 인공지능이 얼마나 감정을 경험할 수 있는 것처럼 보이십니까?
이 인공지능이 얼마나 의식을 가지고 있는 것처럼 보이십니까?
이 인공지능이 얼마나 마음을 가지고 있는 것처럼 보이십니까?

마음 지각

이 인공지능은 다른 존재와 소통할 능력이 있다
이 인공지능은 생각할 수 있다
이 인공지능은 스스로의 행위를 계획할 수 있다
이 인공지능은 무언가를 기억할 수 있다

이 인공지능은 고통을 느낄 수 있다
이 인공지능은 공포를 느낄 수 있다
이 인공지능은 욕구를 가질 수 있다
이 인공지능은 행복을 느낄 수 있다

도덕적 책임

인공지능 자체
인공지능을 만든 개발자(들)
인공지능을 개발한 회사

컴퓨터 친화도

문제를 해결하기 위해 프로그램을 만들 수 있습니까?
프로그래밍과 관련되어 얼마나 많은 지식을 알고 있습니까?
컴퓨터 알고리즘에 대해 얼마나 많은 지식을 알고 있습니까?
나는 컴퓨터 사용에 자신감을 가지고 있다.
나는 컴퓨터를 가능한 한 언제나 사용한다.
나는 내 신용 점수가 컴퓨터에 의해 어떤 방식으로 계산되는지 이해하고 있다.
나는 내 이메일의 스팸 필터가 어떤 방식으로 작동하는지 이해하고 있다.
나는 온라인 쇼핑몰의 추천 시스템이 어떤 방식으로 작동하는지 이해하고 있다.