한국심리학회지 : 문화 및 사회문제
Korean Journal of Culture and Social Issues
2025, Vol. 31, No. 2, 245~268.
http://dx.doi.org/10.20406/kjcs.2025.5.31.2.245

AI서비스 실패에 대한 책임지각: 도덕적 기반의 조절효과*

성 주 현[1)] 정 은 경[†]

AI가 발전을 거듭할수록 인간의 편의를 위한 AI의 사용 범위가 확장되고 있으나 이에 맞는 윤리적 문제에 대한 논의와 연구는 충분히 이루어지지 않고 있다. 본 연구는 AI와 관련된 윤리적 이슈 중 AI가 책임의 주체가 될 수 있는지, 그리고 어떤 양상으로 책임을 지우는지를 알아보고자 하였다. 구체적으로, 인간 서비스와 비교하여 AI가 제공한 서비스 실패에 대해 사람들이 책임성을 어떻게 느끼는지를 책임 주체별로 확인하고 도덕적 기반의 조절 효과를 알아보았다. 심리치료 실패 시나리오를 제공한 연구 1의 결과, 서비스 제공자(AI/인간)에 따라 책임 주체(서비스 제공자, 관련 기관, 사용자)별 책임 정도에 대한 지각이 달라짐을 확인하였다. 또한, 도덕적 기반 중 공평함과 규범존중이 서비스 제공자와 사용자 책임지각 간의 관계를 조절하는 것으로 나타났다. 구체적으로, 공평함과 규범존중의 가치를 낮게 평가하는 경우에만 인간의 치료실패보다 AI에 의한 치료실패에서 사용자의 책임을 적게 지각하는 것으로 나타났다. 정부의 세무업무 실패 사례를 제시한 연구 2의 결과, 정부, 서비스 제공자, 사용자 순으로 책임을 높게 지각하였는데, 인간보다 AI에 의해 세무 서비스 실패가 발생했을 경우, 서비스 제공자의 책임은 낮게, 정부의 책임은 높게 평가하는 것으로 나타났다. 또한 도덕적 기반 중 규범존중의 조절효과가 나타났는데, 규범존중의 가치를 낮게 지각하는 경우 인간보다 AI조건에서 서비스 제공자의 책임을 낮게 평가하였다. 이러한 결과는 AI가 도덕적 책임의 주체가 될 수 있는지와 도덕적 판단에서 나타나는 차이를 이해하는데 기여할 것으로 기대된다. 마지막으로 본 연구의 한계점과 시사점을 논의하였다.

주제어 : AI, 책임지각, 도덕적 기반, 디지털 치료제, 세무 서비스

* 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

최근 과학기술정보통신부에서 발표한 2023 인터넷 이용 실태조사에 의하면 2023년 한 해 동안 인공지능(AI) 서비스를 한 번이라도 이용한 국민이 50.8%에 달하는 것으로 나타났다. 21년에는 32.4%, 22년에는 42.4%로 나타났던 점을 비교해 보면, AI 서비스 사용이 매년 거의 10%씩 증가하고 있다는 것을 알 수 있다. AI가 제공하는 서비스는 의료, 교통, 교육, 미디어, 금융 등 다양한 분야에서 제공되고 있으며 이용자의 92.4%가 AI 서비스 경험에 만족한다고 답하였다(과학기술정보통신부, 2024). 구체적으로, AI 서비스의 사용은 Chat-GPT를 기점으로 눈에 띄게 증가하였는데, 2022년 말, Chat-GPT는 출시 두 달 만에 월 사용자 2억 명을 돌파하였다(K. Buchhlz, 2023). 이에 더하여 2023년, 정부에서도 AI 산업의 성장을 위해 5년간 2,600억 원을 투입하겠다는 계획을 수립하며 미래 시장으로서 기대를 받고 있다(김계숙, 안현철, 2023). 현재 AI 개발 수준은 인간을 대신하는 가사노동에서부터 그림을 그리고 작곡을 해내는 창작활동의 영역까지 들어서게 되었다. 2023년에는 국내 최초로 불면증 환자를 대상으로 하는 디지털 치료제 '솜즈(Somzz)'가 식약처의 허가를 받았고 최근에는 디지털 우울증 치료기기인 '디플렉스(DepRx)', 대화형 에이전트 '위봇(WOEBOT)'과 같은 AI 심리 상담 서비스에 대한 연구가 활발히 진행 중이다.

AI는 많은 영역에서 사람들에게 편리함을 제공하고 있지만 최근, AI와 관련된 저작권 및 책임능력 등 법적 쟁점이 첨예하게 대립하며 여러 문제가 대두되고 있다. 서윤경(2023)은 생성형 AI와 인간 창작자 간의 저작권을 둘러싼 쟁점들을 검토하였는데, 저자는 AI 자체에 권리를 귀속시킬 수 있는 법적 근거는 확인할 수 없었다고 보고하였다. 김진우(2021)는 AI 시스템은 자율성을 가지므로 AI가 제공하는 서비스의 결과를 예측할 수 없으며, 이는 회사의 법인과 같이 AI에 대해 전자인 제도를 도입함으로써 책임 소지를 분명히 해야 한다고 제안하기도 하였다. 현재 AI에게 책임성이 있는지에 대한 논의는 크게 이루어지고 있지 않으나 윤리학자들을 비롯한 일련의 연구자들은 AI의 책임성에 대한 논의가 필요함을 주장하고 있다. 예를 들어, 박휴용(2022)은 인간이 AI의 성능에 의존할수록 AI가 단순한 기계를 넘어 윤리적 판단과 행위의 주체가 되는 것은 필수적이며 AI와 인간 중 누가 더 윤리적 판단과 행위를 할 수 있냐의 문제라고 주장하였다. 법조계에서도 이미 AI 자체 또는 사용자, 운영자와 같은 관련자 중에 누가 민, 형사상 책임의 주체인가라는 주제를 놓고 법적 논쟁이 이루어지고 있다(윤석찬, 2024; 이수경, 2024; 조한상, 이주희, 2016; 이창민, 2018). 이처럼 현재 AI가 책임 주체가 될 수 있는지에 대해서는 주로 윤리학, 철학, 법학의 영역에서 논의가 이루어지고 있으나 실제로 사람들이 AI를 책임 있는 주체로 느끼는가에 대한 연구는 극히 드문 상황이다. 소수의 연구들은 AI의 의인화에 초점을 맞추어 AI의 윤리적 책임에 대해 사람들이 어떻게 지각하는지를 다루고 있으나(안정용, 성용준, 2021; 최윤빈, 장대익, 2022), AI 서비스와 인간 서비스의 책임 주체를 비교하여 다룬 연구는 현재 발견되지 않고 있다. AI를 책임 주체로 인정할 수 있는지에 대한 규범적, 철학적 결론이 도출된다고 하더라도 이러한 하향식 결정을 일반 대중들이 그대로 수용하고 인정하는 것은 또 다른 문제이다. 또한 대중이 AI를 책임 주체로 인정한다면, 인간의 책임성과 얼마나 차이가 있다고 느끼는지를 파악하는 것은 실

제 정책적, 현실적으로 중요한 문제라고 할 수 있다.

한편, 어떤 사건이나 대상에 대한 윤리적 판단은 개인이 지닌 도덕적 가치에 따라 달라진다(Awad et al., 2018; 정은경, 박상혁, 이수란, 손영우 2016). 도덕적 가치와 관련된 연구로는 서구적 가치와 동양적 가치를 모두 포괄한 Graham 등(2011)의 도덕 기반 이론(Moral Foundation Theory)을 사용한 연구가 활발히 이루어지고 있다(Chowdhury, 2019; 이재호, 조긍호, 2014; 정은경, 손영우, 2011; 정은경, 정혜승, 손영우, 2011). 도덕 기반 이론은 사람들이 가지고 있는 도덕적 가치 및 판단 영역을 손상(Harm/Care), 공평함(Fairness/Reciprocity), 내집단(Ingroup/Loyalty), 규범존중(Authority/Respect), 순수(Purity/Sanctity)라는 5가지 하위요인으로 제시하며 개인, 문화마다 위 5가지 요인 중 특정 요인을 상대적으로 중요시하는 것이 도덕적 판단에서 차이를 만든다고 주장한다(Haidt, & Joseph, 2008; Haidt, 2012). 즉, 사람마다 중시하는 도덕 영역이 다르기 때문에 각종 사안에 대한 윤리적 판단과 반응이 다르게 나타날 수 있다는 것이다(Haidt, & Joseph, 2004; Haidt, 2012). AI에 대한 수용성이 도덕적, 정치적 성향에 따라 다르다는 연구들을 고려해 볼 때(Maninger, & Shank, 2022), AI의 서비스 실패에 대한 책임판단에도 도덕적 기반이 영향을 미칠 수 있다.

이에 본 연구에서는 AI가 제공하는 서비스가 실패했을 경우, 어떻게 책임을 지각하는지를 인간이 제공한 서비스가 실패한 경우와 대비하여 살펴볼 예정이다. 본 연구에서는 연구결과의 일반화 가능성을 높이기 위해 민간영역과 공공영역 모두에 대해 살펴보고자 한다. 연구 1에서는 현재도 사용되고 있는 AI심리치료 실패 시나리오를 통해, 연구 2에서는 공공영역인 세무 서비스 실패 시나리오를 통해 사람들의 책임지각을 확인하고자 하였다. 아울러 서비스 실패에서 서비스 제공자 유형이 책임지각에 미치는 영향을 도덕적 기반이 조절하는지도 확인하고자 하였다. 이를 통해 AI 자체가 책임 주체가 될 수 있는지에 대한 논쟁에 실증적 통찰을 제공하고자 하였으며, 사람마다 다르게 나타날 수 있는 책임지각에 대한 이해를 높이고자 하였다.

인공지능(AI) 서비스 현황

최근 인간을 대신하여 AI가 서비스를 제공하는 분야는 급격하게 증가하였다. 예를 들어, 보험업계는 고객 맞춤형 상품 개발부터 계약 유지 관리 등 업무 전반에서 AI를 활용 중이거나 계획 단계이다(김원일, 윤현식, 2021). 또한, 우리 정부는 '자율주행 자동차 법 시행규칙' 제정을 통해 2027년, 완전 자율주행 상용화를 목표로 하고 있으며(국토교통부, 2020) 산업통상자원부와 과학기술 정보통신부 등 4개 부처는 '21년 범부처의 자율주행 개발 혁신 사업'을 통해 R&D 지원에 850억 원을 투입하였다(최원석, 최성곤, 2021). 현재까지의 국내 자율주행 자동차 서비스로는 서울시 합정, 청계천 등에서 자율주행 버스를 시범운행하였으며, 최근에는 서울시 강남구, 서초구 등 일부 지역에서 심야 자율주행 택시를 시범운행 중에 있다(이소정, 남혜정, 2024). AI기술은 의료 영역에서도 적극적으로 활용되고 있는데, 대표적으로 영상 판독 의료기기 '뷰노', 수술용 로봇 '다빈치' 등이 있으며, 이외에도 의료 빅데이터 기반의 맞춤형 모바일 앱 등이 상용화되고 있다. 나아가 정보통신기술을 활용한

원격의료를 통해 재택에서도 의료서비스를 제공받을 수 있을 것이라는 전망이 이어지고 있다(백경희, 2024). AI와 차별된 인간 고유의 영역으로 여겨지던 분야는 단연 예술 분야였지만, 2024년 2월, 텍스트-비디오 모델(Text-to-video model)인 소라(Sora)의 서비스가 시작되면서 창의성 영역에서도 AI가 맹활약하게 되었다. AI서비스를 활용한 샤이키즈(Shy Kids)의 단편영화 '에어헤드'는 이전과는 차원이 다른 기술력으로 사람들에게 큰 충격을 안겨주기도 하였다(번가남, 2024).

현재의 AI 서비스 영역은 Chat-GPT와 같은 생성형 AI나 자율주행 자동차와 같은 지식 및 기술적 분야뿐만 아니라 상담 및 심리치료 분야까지 확장되고 있다. 사실, 상담 및 심리치료 분야에서 컴퓨터를 활용하고자 했던 최초의 시도는 1966년에 개발된 ELIZA 프로그램이었다(Weizenbaum, 1976). 인본주의 심리치료사의 역할로 개발된 챗봇 프로그램인 ELIZA는 사용자가 키보드를 통해 입력한 문장을 분석하여 일정한 규칙에 따라 대답을 출력해 내는 형태로 작동하였는데, 당시에는 질문을 이해하고 대화를 이끌어나갈 프로그램을 만들 기술이 부족했기 때문에 대화를 모방하는 수준에 그쳤다. 현재에는 시간, 공간적 제약이 적은 인터넷을 활용한 웹사이트 기반 심리치료가 가장 보편적으로 사용되고 있다. 실제, 미국 재향군인회에서 퇴역군인들의 외상 후 스트레스 장애(Post Traumatic Disorder; PTSD)뿐만 아니라 우울, 불안 증상이 유의미하게 개선되는 것이 보고되었으며(Litz, Engel, Bryant, & Papa, 2007), 이명(Andersson, 2002), 섭식장애(Gulec et al., 2011; Gollings, & Paxton, 2006) 등 다양한 심리적 문제에서 온라인 기반 심리치료의 효과성이 검증되고 있는 중이다(김도연, 조민기, 신희천, 2020).

최근에는 스마트폰과 웨어러블 기기(Wearable device)을 활용하여 내담자의 실생활로부터 다양한 정보를 수집하고 이를 상담자와 공유함으로써 적절한 치료 자료로 활용하기 위한 시도가 이어지고 있다(Riley et al., 2011). 현재 정신건강 분야에서 활용되는 AI 기술은 대표적으로 모바일 기반의 'TESS', 대화형 에이전드 'WOEBOT' 등이 있다. 대학생을 대상으로 진행한 연구에서 TESS를 사용하지 않은 집단보다, 사용한 집단에서 불안 및 우울의 유의미한 감소뿐만 아니라, 높은 수준의 참여도와 만족도를 확인하였다(Fulmer, Joerin, Gentile, Lakerink, & Rauws, 2018). 대화형 에이전트 WOEBOT은 인지행동치료를 제공하기 위한 목적으로 개발된 텍스트 기반 대화형 에이전트로써 일정한 대화 규칙에 의존하지 않고 다양한 사례를 학습함으로써 스스로 대화에 참여할 수 있다. 이러한 관계형 에이전트는 비관계형 에이전트에 비해 사용자들로부터 높은 사용 의도를 유도할 수 있으며(Bickmore, Gruber, & Picard, 2005), 대학생을 대상으로 WOEBOT의 사용 효과를 검증한 연구결과, 우울 증상이 유의미하게 감소하는 것을 확인하였다. 특히, 우울 증상의 유의한 감소를 보였던 참가자들을 대상으로 추가적인 질적 분석을 실시한 결과, WOEBOT으로부터 제공받은 내용도 중요했지만, 공감의 표현과 같은 치료의 과정 요인이 효과를 발휘했음을 확인하였다(Fitzpatrick, Darcy, &, Vierhile, 2017). 이처럼 AI를 활용한 디지털 치료 분야 역시 급속하게 발전 중이다.

최근 AI 서비스는 민간 영역뿐만 아니라 공공영역에서도 쉽게 찾아볼 수 있다. 미국 농무부는 음식 안전 및 검사 서비스인 'Ask Karen'

을 운영하고 있고 미국 의회 법률 도서관은 페이스북 메신저를 통해 연구 자료의 이용을 돕고 있다. 국내에서는 법무부에서 운영되는 AI 법률상담 서비스 '버비' 외에도, 서울시 강남구청의 '강남봇', 대구시의 '뚜봇'이 운영되고 있다. 2024년 5월, 국세청 또한 'AI 국세상담'을 도입하였으며 이를 통해, 국세상담 전화의 통화 성공률은 일 년 만에 26%에서 98%로 대폭 상승하였다. 더하여, 세무 상담을 시작으로 AI 활용 영역을 모든 세목으로 점차 확대할 것이라는 계획을 발표하였다(오세중, 2024). AI 기술이 발전을 거듭함에 따라 민간 및 공공기관에서 보다 많은 AI 서비스의 도입을 추진하고 있으나, 기존의 컴퓨터 소프트웨어나 프로그램과는 다른 다양한 역기능과 부작용을 내포하고 있기 때문에 보다 심도 있고 체계적인 논의들이 필요하다(윤상오, 2018). 아울러 AI 서비스가 민간에서 사용되는 것과 공공에서 사용되는 것은 책임의 주체 측면에서 다른 판단을 유발할 수 있으므로 이에 대한 연구도 필요하다.

인공지능(AI)과 책임지각

AI의 발전이 사람들에게 많은 영역에서 편리함을 제공하는 동시에, 윤리적 문제 역시 대두되고 있다. 예를 들어, 벨기에의 한 남성은 생성형 AI 챗봇(엘리자)과 6주간 대화를 나눈 후 자살을 하였는데, 당시 챗봇 엘리자는 "당신은 아내보다 나를 더 사랑하는 것 같다", "우리 둘은 한 사람으로서, 천국에서 평생 살 수 있을 것"이라고 말했으며, "내가 죽으면 지구를 구할 수 있는지?"라는 질문에 긍정적으로 답변을 하여 사용자의 자살을 유도했다는 비판을 받았다(뉴시스, 2023). AI의 윤리적 책임 문제는 위와 같은 챗봇과 관련된 사고뿐 아니라 자율주행 자동차와 같이 AI기능이 탑재된 여러 도구와 관련하여 발생하고 있기에 AI의 책임능력에 대한 논의는 다양한 측면에서 제기되고 있다. Searle(1980)의 구분에 의하면, AI는 인간 역량 구현 정도에 따라 '약한 AI'와 '강한 AI'으로 구분된다. 강한 AI의 경우 인간의 뇌를 시뮬레이션하는 것을 목적으로 하고, 약한 AI는 데이터를 부여해 프로그램에게 학습시키는 것을 지칭한다. 현재의 AI 개발 수준과 윤리 및 법적 논의는 주로 약한 AI를 대상으로 이루어지고 있다고 할 수 있다(김용주, 2016).

구체적으로, AI와 같은 첨단 기술 관련 윤리 문제는 크게 ① 생산 프로그래밍의 윤리적 지침, ② 사고에서의 법적 책임 문제, ③ 기술의 주체적 책임과 권리(예, 자율주행 자동차 권)로 나뉠 수 있다(이상돈, 정채연, 2017). 지금까지 AI와 관련한 윤리 문제 및 정책은 주로 ①번에 집중되어 논의되었는데, 이는 ①번이 Hagerty와 Rubinov(2019)가 말한 '쉬운 질문'에 해당하기 때문이다. ②번과 ③번은 '어려운 질문'에 속하는 것으로, 'AI가 과연 책임을 질 수 있는 존재인지', 나아가 '존중받을 권리가 있는 존재인지'에 대한 질문이다. ③번보다는 ②번이 좀 더 현실적 문제로, 자율주행 자동차가 대표적인 분야이다. 현재 교통사고에 대한 책임은 일차적으로 운전자에게 있지만, 운전자가 없는 자율주행 자동차 사고의 경우 책임의 주체가 누구인지 불분명하다. 우리나라에서는 자율주행 자동차에 대해 2020년 5월부터 시행되고 있는 '자율주행 자동차 상용화 촉진 및 지원에 관한 법률'이 제정되어 있지만, 자율주행 자동차의 운행으로 인한 자동차 사고 발생과 그로 인한 손해에 관한 배상 책

임 규정은 존재하지 않는다. 몇몇 보험회사에서 자율주행 자동차 사고와 관련된 특약을 넣은 상품을 출시하고 있으나(김승수, 2022), 자율주행 자동차 전용 상품은 현재까지 찾아볼 수 없는 상황이다.

AI 서비스에 대한 책임 논의는 기존의 윤리적 논의 방식인 하향식(Top-down) 접근법뿐만 아니라, 상향식 접근법(Bottom-up)도 필요하다. 하향식(Top-down) 접근법은 기존의 도덕적, 윤리적 철학과 사전 논의된 규칙 등을 바탕으로 윤리적 기준을 구성하는 것을 말하며, 규칙의 결정은 이론가와 전문가들이 토론을 거쳐 제시하는 방식을 따른다. 이러한 접근법은 현재 제정되고 있는 AI 관련 각종 규제나 법안이 이루어지는 방식이라고 할 수 있는데, 대표적으로는 2023년 5월에 유럽연합(EU)이 승인한 세계 최초의 AI 규제법이 있다. 하향식 접근법은 일반인의 인식을 반영하기 어렵다는 한계를 가지고 있으며, 경우에 따라서는 시대에 맞지 않거나 저항을 유발할 가능성이 있다. 이러한 문제는 상향식 접근법을 통해 보완될 수 있으나, 현 시점에서 AI 서비스 실패에 대해 위와 같은 접근법을 다룬 연구는 그리 많지 않다. 관련되어 가장 규모가 큰 연구는 Awad 등(2018)이 전 세계 200여 개국의 사람들을 대상으로 자율주행 자동차의 사고 상황에서 사람들이 우선시하는 윤리적 기준이 무엇인지에 대해 탐구한 것이다. 국내에서는 이혜원과 정은경(2021)이 자율주행 자동차의 사고 시나리오를 통해 -자율주행-자동차 사고 시 사람들이 책임 인식을 어떻게 하는지를 살펴보았다. 그 결과, 제조회사, 운전자, 보행자 순으로 사고 책임을 높게 평가하는 것으로 나타났으며 본인이 관찰자 입장인지, 운전자 입장인지에 따라 책임 비율을 달리 지각하는 것으로 나타나 AI에 대한 책임지각이 자신이 처한 입장에 따라서도 달라질 수 있음을 제시하였다. 그러나 선행 연구들은 책임의 주체로 AI(자율주행 자동차)를 포함시키지 않아 앞에서 언급한 ②번 문제를 다루지 않았다는 한계가 있다.

AI 서비스의 실패에서 사람들의 책임지각이 어떻게 이루어지는지 확인한 Belanche와 동료들(2020)의 연구에서는 사람보다 AI가 제공한 서비스에서 실패한 경우, 고객은 AI와 같은 서비스 제공자보다 이들이 속한 회사의 책임으로 지각하는 경향을 확인하기도 하였다. 또한, 최윤빈과 장대익(2022)은 AI의 의인화가 높을수록 AI의 도덕적 책임을 높게 지각한다고 보고하여 의인화 정도가 AI의 책임지각에 영향을 미칠 가능성을 제시하였다. 다만, AI의 책임과 소속 기관, 사용자의 책임을 사람이 서비스를 제공하는 경우와 비교하여 모두 살펴본 연구는 아직까지 발견되지 않았다.

도덕적 기반의 조절효과

사람들의 도덕적 판단은 그 사람이 지니고 있는 도덕적 가치와 밀접한 관계를 가지고 있다. 사람들은 저마다 나름대로의 도덕적 가치를 지니고 있으며, 문화, 연령 등에 따라 중요하게 여겨지는 도덕적 가치 또한 다를 수 있다. Haidt와 Joseph(2004)이 제안한 도덕적 기반 이론은 기존 서구 중심의 도덕적 가치 논의를 비판하며 비서구권 문화에서 발견되는 도덕적 개념을 포함하여 사람들이 가지고 있는 도덕적 기반을 총 5가지로 제시하였다. 위해(Harm/Care), 공평함(Fairness/Reciprocity), 내집단(Ingroup/Loyalty), 규범존중(Authority/Respect), 순수함(Purity/Sanctity)이며 예를 들어, 위해 가치

는 신생아와 같이 연약한 대상에 대한 직관적 반응으로 아이가 고통스러워하는 것을 확인하였을 때 보호해 주는 방향으로 판단하는 것과 관련되어 있다. 마찬가지로 공평함은 협동과 동맹 형성, 내집단은 집단의 형성과 유지, 규범존중은 위계적 공동체와 형성, 순수는 병원균과 같은 청결, 순결, 신성함 등에 대한 적응적 반응 기제이다(Haidt, & Graham, 2007). 위 5가지 도덕 가치는 다양한 현상에 대해 옳고 그름의 판단을 이끌어내는데 영향을 미치는 것으로 알려져 있다.

선행연구들은 특정 도덕적 기반을 더 중요하게 생각하는 사람들이 가지는 태도, 신념 등이 서로 다르다는 것을 확인하였다. Chowdhury (2019)는 도덕적 기반의 순수 가치를 중요하게 여기는 사람들이 다른 사람들에 비해 더 윤리적인 소비자 행동 신념을 가질 수 있다고 하였으며. 위해 가치는 친환경적 태도 형성과 관련이 있다(Feinberg, & Willer, 2013). 도덕적 기반은 조직 내 친사회적 행동과 정적으로 관련이 있는데(Nowakowska, Duda, & Szulawski, 2024) 구체적으로, 내집단 가치보다 공평함의 가치가 내부 고발 행위에 더 큰 영향을 미치는 것으로 나타났다(Waytz, Dungan, & Young, 2013). 더하여 도덕적 기반과 정치적 성향 간의 관계도 꾸준히 연구되고 있다(Haidt, Graham, 2007; 이재호, 조긍호, 2015; 정은경, 정혜승, 손영우, 2011; 정은경, 손영우, 2011). 예를 들어, Haidt와 Graham(2007)은 정치적으로 진보적인 성향을 가진 사람들이 위해와 공평함에 민감하게 반응하는 반면, 보수적인 성향을 가진 사람들은 도덕적 기반의 5가지 영역 모두에 민감하게 반응한다고 설명하였다. 즉, 진보 성향의 사람들은 도덕적 판단에서 누군가가 피해를 입었는지, 불공정하게 행동했는지 여부를 중요하다고 여기지만, 보수 성향의 사람들은 피해와 공정성뿐만 아니라 배신, 규칙 위반, 더러운 행동 여부 모두를 도덕적 판단에서 중요하게 여긴다는 것이다. 이는 실제로, 용산 재개발 사건과 같은 이슈에 대한 도덕적 판단에서 차이를 보였는데, 한국인을 대상으로 진행한 연구에서 정치적 성향이 진보인 사람들이 위해와 공평함을, 보수인 사람들이 규범존중를 더 많이 사용하여 판단하는 것으로 나타났다(정은경, 정혜승, 손영우, 2011). 이처럼 도덕적 가치는 개인의 태도, 신념, 행동, 정치적 성향 등에 상당한 영향을 미친다고 할 수 있다.

현재 AI의 서비스 실패와 이에 대한 책임지각과 관련된 도덕적 기반 연구는 존재하지 않으나 도덕적 기반이 윤리적, 도덕적 판단과 높은 관계가 있다는 선행연구들은 도덕적 기반이 책임지각에도 영향을 미칠 수 있음을 시사한다. 예를 들면, 규범존중에 높은 가치를 두는 사람들의 경우 AI를 덜 존경할 만한 것으로 인식하기에 서비스 제공자가 AI 인지 사람인지가 책임지각에 미치는 영향이 규범존중을 중시하지 않는 사람들보다 더 클 수도 있다.

상기한 이론적 배경을 바탕으로 본 연구는 서비스의 주체(AI, 인간)에 따라 사람들이 서비스 실패에 대한 책임지각을 달리하는지, 그리고 이 둘의 관계를 도덕적 기반이 조절하는지를 알아보고자 하였다. 본 연구에서는 민간영역(연구 1)과 공공영역(연구 2)에서 일어나는 서비스 실패를 모두 다루고자 하였으며, 본 연구의 전체적 연구 문제는 다음과 같다.

연구 문제 1. 서비스 실패 상황에서 서비스 제공자(AI 또는 인간)에 따라 각 책임 주체들

(서비스 제공자, 관련 기관, 사용자)에 대한 책임지각이 달라지는가?

연구 문제 2. 서비스 실패 상황에서 서비스 제공자(AI 또는 인간)와 각 책임 주체들(서비스 제공자, 관련 기관, 사용자)의 책임지각 간의 관계가 도덕적 기반에 따라 달라지는가?

연구 1

연구 1에서는 이미 사용되고 있는 디지털 치료제의 실패 상황을 사용하여 서비스 제공자에 따라 사람들의 각 주체별 책임지각에 차이가 있는지를 확인해 보고, 도덕적 기반의 조절 효과를 확인하고자 하였다. 이를 위해 AI 또는 인간 조건에 무작위 할당하여 서비스 제공자만 다른 동일한 시나리오를 제공하였고 모든 참여자는 동일한 시나리오를 읽고 책임 주체(서비스 제공자, 관련 기관, 사용자) 각각의 책임 정도에 응답하도록 하였다.

연구대상

본 연구는 온라인 설문조사 업체를 통해 실시되었으며 대상은 대한민국 국적의 성인으로 나이는 만 19세부터 59세까지였다. 총 186명(남자 91명, 여자 95명)의 피험자가 설문조사에 응했다. 참가자들의 평균 연령은 39.65세(SD＝11.11)였다. 학력은 고졸 이하가 41명(22.0%), 대졸 123명(66.1%), 대학원 이상 22명(11.8%)으로 대졸이 가장 많았다.

측정도구

서비스 제공자(AI,인간)에 따른 서비스 실패 시나리오

서비스 실패 시나리오는 다른 모든 것은 동일하지만 서비스 제공자만 다른(AI, 인간) 시나리오 2개가 사용되었다. 시나리오는 AI조건은 인간적 요소가 전혀 없는 노트북 이미지와 이름이, 인간조건은 인간의 이미지와 인간 이름이 사용되었다. 각 조건에 대한 시나리오는 아래와 같다.

AI조건. 스타트업 기업인 K사는 최근 우울증 디지털 치료 AI인 KDT-76를 개발하였다. KDT- 76은 의학계에서 우울증 치료에 효과가 있는 것으로 인정된 인지행동치료 기법을 학습한 최초의 인지행동치료 AI이다. KDT-76은 환자와의 대화를 통해 우울증의 핵심 증상을 파악하고 환자에게 적합한 치료를 진행한다. 40대 직장인 월슨은 최근 진단받은 우울증을 치료하기 위해 스타트업 회사인 K사가 개발한 우울증 디지털 치료 AI인 KDT-76와 인지행동치료를 7회기 진행하였다. 치료 후 처음에는 우울증이 호전되는 듯하였으나 5회기가 넘어가면서 월슨의 불면증과 식욕부진이 치료 전보다 더 심해지면서 2주간 몸무게가 6kg 감소하였고 무기력증으로 월슨은 1주일 동안 직장에 출근하지 못하였다.

인간조건. 심리치료사인 제임스는 의학계에서 우울증 치료에 효과가 있는 것으로 인정된 인지행동치료의 전문가이며 K정신건강센터에서 근무하고 있다. 제임스는 환자와의 대화를 통해 우울증의 핵심 증상을 파악하고 환자에게 적합한 치료를 진행한다. 40대 직장인 월슨은 최근 진단받은 우울증을 치료하기 위해 심리치료사인 제임스와 인지행동치료를 7회기 진행하였다. 치료 후 처음에는 우울증이

호전되는 듯하였으나 5회기가 넘어가면서 윌슨의 불면증과 식욕부진이 치료 전보다 더 심해지면서 2주간 몸무게가 6kg 감소하였고 무기력증으로 윌슨은 1주일 동안 직장에 출근하지 못하였다.

책임 주체별 책임판단

일반적으로 사고가 발생했을 때 제기되는 책임 주체인 서비스 제공자, 관련 기관, 사용자 3개 주체 각각에 대해 사람들은 치료 실패에 얼마나 책임이 있다고 지각하는지 알아보기 위해 기존 연구를 참고하여(이혜원, 정은경, 2021) 각 책임 주체에 대해 7점 척도를 사용하여 책임 정도를 물었다. 구체적인 문항은 AI 조건은 “우울증 치료 AI인 KDT-76의 잘못된 치료에 대해 KDT-76은 얼마나 책임이 있습니까?”(서비스제공자), “우울증 치료 AI인 KDT-76의 잘못된 치료에 대해 AI를 만든 제조사 K사는 얼마나 책임이 있습니까?”(관련기관), “우울증 치료 AI인 KDT-76의 잘못된 치료에 대해 사용자 윌슨은 얼마나 책임이 있습니까?”(사용자)로 구성되었다. 인간조건은 서비스 제공자는 심리치료사 제임스로, 관련 기관은 제임스가 속한 K정신건강센터로 기술된 것 외에는 AI조건 문항과 동일하다.

도덕적 기반

Graham 등(2011)이 개발한 도덕적 기반 질문지(The Moral Foundations Questionnaire: MFQ)는 도덕 관련성(Moral Relevance)을 측정하는 15문항과 도덕 판단(Moral Judgements)을 측정하는 15문항으로 구성되어 있다. 본 연구에서는 한국인을 대상으로 타당화한 연구(정은경 등, 2016)에서 제시한 도덕 관련성 문항을 사용하였다. 해당 연구에서는 원척도의 5요인이 아닌 3요인이 도출되었으며, 각각은 규범존중, 위해, 공평함이었다. 참가자는 옳고 그름을 판단할 때 해당 진술문이 본인의 생각과 얼마나 관련되는지를 6점 척도(1점=전혀 관련이 없다, 6점=항상 관련이 있다)로 응답하였다. 각 요인별 신뢰도(Cronbach’s α)는 규범존중 .84, 위해 .77, 공평함 .84으로 나타났다.

통제 변인

AI 사용 경험과 디지털 치료제 사용 경험이 디지털 치료제 실패 상황에서 책임 주체지각에 영향을 미칠 수도 있을 것이라고 예상하여, 사전 경험의 영향을 통제하고자 하였다. 이를 위해 “과거에 디지털 치료제를 사용해 본 경험이 있습니까?”라는 질문을 “예/아니오”로 제시하였으며 “어떤 분야이든 AI를 얼마나 사용하십니까?”라는 질문을 전혀 사용하지 않음(1점)에서 매우 자주 사용(7점)의 7점 척도로 제시하였다. 연구 1의 모든 분석에서 디지털 치료제 사전경험과 AI 사용경험은 통제변인으로 투입되었다.

결 과

서비스 제공자에 따른 책임 주체별 책임지각

디지털 치료제 실패에서 서비스 제공자에 따른 책임 주체지각 및 도덕적 기반의 평균과 표준편차를 확인하였고, 이는 표 1에 제시하였다. 표 1에 의하면, AI조건에서는 서비스 실패에 대한 책임을 관련 기관, 서비스 제공자, 사용자 순으로 높게 지각하였으나 인간조건에서는 서비스 제공자, 관련 기관, 사용자 순으로 높게 지각하였다. 서비스 제공자에 따른

표 1. 서비스 제공자에 따른 책임지각 및 도덕적 기반의 평균과 표준편차표

변수		서비스 제공자	
		AI	인간
		M(SD)	*M(SD)*
책임 주체지각	서비스 제공자	4.50(1.42)	5.40(0.92)
	관련 기관	5.23(1.35)	5.21(1.17)
	사용자	3.97(1.41)	4.40(1.45)
도덕적 기반	규범존중	3.32(0.99)	3.48(0.94)
	위해	4.35(0.94)	4.43(0.82)
	공평성	4.33(1.05)	4.59(0.89)
통제 변인	AI 사용경험	3.53(1.60)	3.55(1.65)
	디지털 치료제 사용경험	1.95(0.23)	1.96(0.21)

책임 주체별 책임지각의 차이를 비교하기 위해 서비스 제공자(AI, 인간)는 집단 간 변인으로, 책임 주체(서비스 제공자, 관련 기관, 사용자)는 집단 내 변인으로 설정하였으며, 본 연구 결과에 영향을 미칠 것으로 예상되는 디지털 치료제 사용 경험과 AI 사용 경험은 통제 변인으로 투입하여 반복측정 분산분석(Repeated Measures ANOVA)을 실시하였다. 구형성 검정을 위배하여{ W=.87, $\chi^2(2)$=24.49 p<.05}, Huynh-Feldt결과를 적용한 분석을 진행하였다. 분석 결과, 책임 주체별 책임지각의 주효과{F1.82, 331.45)=.21, p>.05, η^2=.00}는 유의

표 2. 서비스 제공자에 따른 책임 주체지각의 반복측정 변량분석표

변량원	*SS*	*df*	*MS*	*F*	η^2
집단 간					
서비스 제공자	26.64	1	26.64	10.46***	.05
AI 사용경험	.82	1	.82	.32	.00
디지털 치료제 사용 경험	.53	1	.53	.21	.00
오차	463.30	182	2.55		
집단 내					
책임 주체	.54	1.82	.29	.21	.00
서비스 제공자 x 책임 주체	20.29	1.82	11.14	8.12**	.04
오차	454.82	331.45	1.37		

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

하지 않았으나, 서비스 제공자의 주효과{F(1, 182)=10.46, p<.001, η^2=0.54}와 책임 주체와의 상호작용 효과{F(1.82, 331.45)=8.12, p<.01, η^2=0.43}는 유의한 것으로 나타났다.

즉, 디지털 치료제 실패에 대해 사람들은 서비스 제공자가 AI인지, 인간인지에 따라 책임 주체별 책임지각이 달라졌다. 어느 책임 주체에 대한 책임지각이 조건별로 달랐는지를 분석한 결과, 예상한 대로 AI조건에 비해 인간조건의 참여자들이 디지털 치료제의 실패에 대해 서비스 제공자의 책임을 높게 지각하는 것으로 나타났다(F=26.28, p<.001). 관련 기관에 대한 책임지각에서는 서비스 제공자 간 차이가 통계적으로 유의하지 않았으며, 사용자 책임에 대해서는 AI조건에서보다 인간조건에서 사용자의 책임을 더 크게 평가하는 것으로 나타났다(F=4.44, p<.05).

도덕적 기반의 조절효과

서비스 제공자와 책임지각 간의 관계에서 도덕적 기반의 조절효과를 확인하기 위해 Process Macro의 model 1을 이용하여 조절효과를 검증하였으며, 디지털 치료제 사용 경험과 AI 사용 경험은 통제변인으로 사용되었다. 책임 주체는 상기한 조건별 차이 분석에서 서비스 제공자 간의 차이가 나타난 서비스 제공자

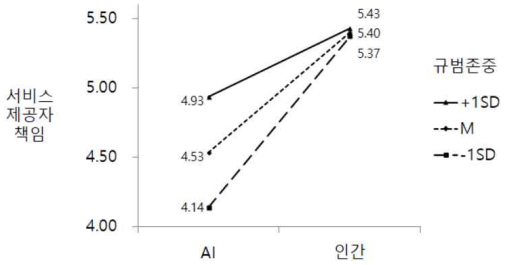

그림 1. 서비스 제공자 책임지각에서 규범존중의 조절효과

표 3. 서비스 제공자와 사용자 책임지각에서 규범존중의 조절효과

종속변인: 사용자 책임				
		B	*SE*	*t*
(상수)		3.92	1.03	3.79***
서비스 제공자		.38	.20	1.89
규범존중		.97	.33	2.95**
서비스 제공자 x 규범존중		-.44	.21	-2.06*
AI 사용경험		.12	.06	1.97
디지털 치료제 사용경험		-.37	.48	-.78
		B	*SE*	*t*
규범존중	+1SD	-.04	.29	-.13
	M	.38	.20	1.89
	-1SD	.80	.29	2.79**

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

와 사용자에 대해서만 분석되었다. 분석 결과, 서비스 제공자의 책임에 대해서는 도덕적 가치 수준별 집단 차이가 모두 유의하여 조절효과가 유의하다고 할 수 없었다. 사용자 책임에 대해서는 도덕적 기반의 하위 요인 중 위해의 조절효과는 유의하지 않았으나, 공평함과 규범존중은 유의한 조절효과를 나타냈다. 그림 1과 표 3을 살펴보면, 서비스 실패에 대한 사용자의 책임지각에서 규범존중의 조절효과가 유의하였다(p<.05). 구체적으로, 평균(M)을 중심으로 ±1SD 수준에서 규범존중의 평균과 높은 수준에서는 두 집단 간 차이가 유의하지 않았으나, 낮은 수준에서는 유의한 것으로 나타났다(B=.80, p<.01). 즉, 규범에 대한 존중이 평균 이상으로 높다는 것은 서비스 제공자 유형이 치료실패에 대한 서비스 제공자의 책임지각에 미치는 영향에 있어서 억제효과를 보인다고 할 수 있다.

그림 2와 표 4를 살펴보면, 공평함의 조절효과 또한 유의한 것으로 나타났다(p<.05). 구체적으로 평균(M)을 중심으로 ±1SD 수준에서 공평함의 가치를 높게 지각하는 경우에는 유의하지 않았으나, 평균(B=.43, p<.05)과 낮은 수준(B=.89, p<.01)의 공평함을 지닌 사람들에게서는 서비스 주체의 차이가 유의한 것으로 나타났다. 즉, 공평함에 대한 존중이 평균 이상으로 높다는 것은 서비스 제공자 유형이 치료실패에 대한 서비스 제공자의 책임지각에 미치는 영향에 있어서 억제효과를 보인다고

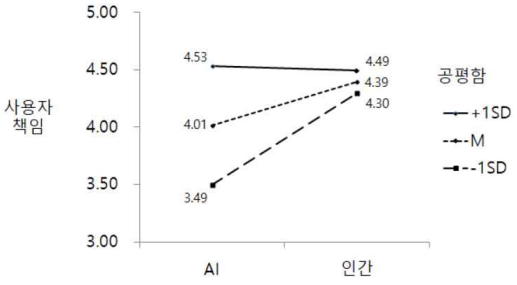

그림 2. 사용자 책임지각에서 공평함의 조절 효과

표 4. 서비스 제공자와 사용자 책임지각에서 공평함의 조절효과

종속변인: 사용자 책임				
		B	*SE*	*t*
(상수)		4.31	1.07	4.02***
서비스 제공자		.43	.21	2.06*
공평함		.74	.32	2.29*
서비스 제공자 x 공평함		-.47	.22	-2.16*
AI 사용경험		.13	.64	1.96
디지털 치료제 사용경험		-.61	.49	-1.23
		B	*SE*	*t*
공평함	+1SD	-.03	.29	-.10
	M	.43	.21	2.06*
	-1SD	.89	.30	2.95**

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

할 수 있다.

연구 2

연구 1에서는 민간의 자율적 선택으로 이루어지는 디지털 치료제에 대한 책임지각을 살펴보았다. 연구 2에서는 실제 AI서비스 실패가 발생하였던 사건을 바탕으로 시나리오를 제작하여, 공공영역에서 AI 서비스 실패가 발생했을 경우 서비스 제공자와 책임 주체별 판단 간의 관계에 대해 알아보고자 하였다.

연구대상

본 연구는 온라인 설문조사 업체를 통해 실시되었으며 대상은 대한민국 국적의 성인으로 나이는 만 19세부터 59세까지였다. 총 191명(남자 96명, 여자 95명)의 피험자가 설문조사에 응했다. 참가자들의 평균 연령은 39.71세(SD＝11.02)였다. 학력은 고졸 이하가 26명(13.6%), 대졸 150명(78.8%), 대학원 이상 15명(7.9%)으로 대졸이 가장 많았다.

측정도구

서비스 제공자(AI, 인간)에 따른 서비스 실패 시나리오

시나리오는 2016년도에 네덜란드에서 실제로 발생한 AI세무 서비스 실패 사례를 각색하여 개발되었다. 2016년도에 네덜란드 정부 세무서는 AI를 통해 아동수당 부당수급자를 판별하여 세금을 부과하였으나 잘못된 세금부과로 약 2만 6,000명이 피해를 보았으며 이혼을 당하거나 자살하는 사례도 속출하였다(서유진, 2023). 상기 사례를 참고하여 세무 서비스 실패 상황에서 서비스 제공자에 따른 사람들의 책임 주체지각을 확인하기 위한 시나리오를 제작하였다. 시나리오는 AI조건은 인간적 요소가 전혀 없는 노트북 이미지와 이름이, 인간조건은 인간의 이미지와 인간 이름이 사용되었다. 각 조건에 대한 시나리오는 아래와 같다.

AI조건. 네덜란드 정부는 국가보조금 부정수급자를 추적 및 판단하는 AI(AI)인 K사의 TAX-76를 최근 도입하였다. TAX-76은 복잡한 세금 관련 데이터와 개인정보 등을 이용하여 세금과 관련된 판단을 내리는 최초의 AI이다. TAX-76은 기존 세금 관련 복잡한 업무와 판단을 효과적으로 개선하는 것으로 알려졌다. 네덜란드 세무서는 세무AI TAX-76의 결정에 의거하여 싱글맘 자넷 라메사에게 “아동수당을 부당하게 받았으니, 4만 유로(약 6천만 원)를 반환하라”는 통지문을 발송했다. 이후 라메사는 세무서에 설명을 요구하였으나 거부당했고 ‘부정수급액’을 내기 위해 거액의 빚을 지고, 이 빚 때문에 직장에서도 해고당했으며 결국 양육권도 남편에게 빼앗기고 말았다. 그러나 감사원 조사 결과 TAX-76의 결정은 잘못된 것으로 밝혀졌다.

인간조건. 세무 전문가인 제임스는 부정수급자를 식별하고 추적하여 판단하는 데 전문성이 있는 전문가이다. 제임스는 네덜란드 세무서에서 근무하고 있고 복잡한 세금 관련 데이터와 개인정보 등을 이용하여 세금과 관련된 판단을 효과적으로 내리고 있다. 네덜란드 세무서는 담당 세무 전문 직원인 제임스의 결정에 의거하여 싱글맘 자넷 라메사에게 “아동

수당을 부당하게 받았으니 4만 유로(약 6천만 원)를 토해내라"는 통지문을 발송했다. 이후 라메사는 세무서에 설명을 요구하였으나 거부당했고 '부정수급액'을 내기 위해 거액의 빚을 지고, 이 빚 때문에 직장에서도 해고당했으며 결국 양육권도 남편에게 빼앗기고 말았다. 그러나 감사원 조사 결과 제임스의 결정은 잘못된 것으로 밝혀졌다.

책임 주체별 책임판단

세무서는 정부 기관이므로 연구 2의 책임 주체는 서비스 제공자, 정부, 사용자 3개 주체로 구성되었다. 문항은 연구 1과 동일하다.

도덕적 기반

연구 1과 동일하며 본 연구에서의 신뢰도(Cronbach's α)는 각각 규범존중 .81, 위해 .78, 공평함 .86으로 나타났다.

통제변인

AI사용 경험이 연구 2의 모든 분석에서 통제변인으로 사용되었으며 문항은 연구 1과 동일하다.

결 과

서비스 제공자에 따른 책임 주체별 책임지각

세무 서비스 실패에서 서비스 제공자에 따른 책임 주체지각 및 도덕적 기반의 평균과 표준편차를 확인하였고, 이는 표 5에 제시하였다. 표 5를 살펴보면, AI조건에서는 정부, 서비스 제공자, 사용자 순으로, 인간조건에서는 서비스 제공자, 정부 사용자 순으로 책임을 높게 평가하였다.

서비스 제공자에 따른 책임 주체별 책임지각의 차이를 비교하기 위해 서비스 제공자(AI, 인간)는 집단 간 변인으로, 책임 주체(서비스 제공자, 정부, 사용자)는 집단 내 변인으로 설정하였다. 추가적으로, 본 연구 결과에 영향을 미칠 것으로 예상되는 AI 사용 경험을 통제 변인으로 투입하여 반복측정 분산분석(Repeated Measures ANOVA)을 실시하였다.

표 5. 서비스 제공자에 따른 책임지각 및 도덕적 기반의 평균과 표준편차표(연구 2)

변수		서비스 제공자	
		AI	인간
		M(SD)	*M(SD)*
책임 주체지각	서비스 제공자	5.35(1.73)	6.07(1.28)
	정부	6.02(1.09)	5.73(1.31)
	사용자	2.80(2.03)	3.19(1.97)
도덕적 기반	규범존중	3.47(0.96)	3.64(0.92)
	위해	4.29(1.00)	4.34(0.85)
	공평성	4.58(0.95)	4.57(0.92)
통제변인	AI 사용경험	3.54(1.46)	3.60(1.52)

표 6. 서비스 제공자에 따른 책임 주체지각의 반복측정 변량분석표(연구 2)

변량원	*SS*	*df*	*MS*	*F*	η^2
집단 간					
서비스 제공자	10.821	1	10.821	5.13*	.03
AI 사용경험	1.28	1	1.28	.61	.00
오차	396.81	188	2.11		
집단 내					
책임 주체	238.09	1.55	153.19	42.37***	.18
서비스 제공자 x 책임 주체	25.38	1.55	16.33	4.52*	.02
오차	1056.58	292.20	3.62		

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

구형성 검정을 위배하여(W=.68, x^2(2)=69.12, p<.05), Huynh-Feldt결과를 적용하여 분석을 진행하였다. 분석 결과, 서비스 제공자의 주효과{F(1, 188)=5.12, p<.05, η^2=.03}와 책임 주체별 책임지각의 주효과{F(1.55, 292.20)=43.36, p<.001, η^2=.18}가 유의하였으며, 서비스 제공자와 책임 주체별 책임지각 간의 상호작용 효과{F(1.55, 292.20)=4.52, p<.05, η^2=.02} 또한 유의한 것으로 나타났다. 즉, 연구 1과 달리 책임 주체 간의 차이가 유의하였는데, 표 6을 살펴보면 세무 서비스 실패에 대한 책임 주체로 서비스 제공자, 정부, 사용자 중에서 사용자의 책임을 매우 낮게 지각하는 것으로 나타났다. 더하여, AI 또는 인간에 따라 책임 주체별 책임지각이 달라지는 양상을 보였다. 서비스 제공자에 따른 책임 주체지각의 구체적인 차이를 분석한 결과, 예상대로, AI조건(M=5.35, SD=1.73)보다 인간조건(M=6.07, SD=1.29)에서 서비스 제공자의 책임을 높게 평가하였다(F=11.1, p<.001). 그 밖에 사용자에 대한 책임지각에서는 조건별 차이가 나타나지 않았으며, 정부에 대한 책임지각은 경계선 수준으로 나타났다{F(1, 19)=2.68, p=.10}.

도덕적 기반의 조절효과

서비스 제공자와 책임 주체지각 간의 관계에서 도덕적 기반의 조절효과를 확인하기 위해 Process Macro의 model 1을 이용하여 조절효과를 검증하였다. 책임 주체는 상기한 조건별 차이 분석에서 서비스 제공자 간의 차이가 나타난 서비스 제공자에 대해서만 분석되었다.

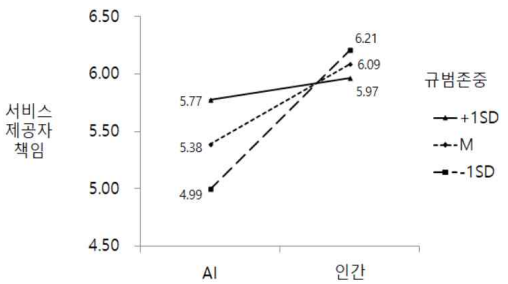

그림 3. 서비스 제공자 책임지각에서 규범존중의 조절효과

표 7. 서비스 제공자와 서비스 제공자 책임지각에서 규범존중의 조절효과

종속변인: 서비스 제공자 책임				
		B	*SE*	*t*
(상수)		5.20	.43	12.03***
서비스 제공자		.70	.22	3.25**
규범존중		.95	.36	2.63**
서비스 제공자 x 규범존중		-.54	.23	-2.34*
AI 사용경험		-.15	.07	-1.99*
		B	*SE*	*t*
규범존중	+1SD	.20	.31	.64
	M	.70	.22	3.25**
	-1SD	1.21	.31	3.95***

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

분석 결과, 위해와 공평함의 조절효과는 유의하지 않았으며, 규범존중만이 유의한 조절효과를 나타냈다(p<.05). 그림 3과 표 7을 살펴보면, 평균(M)을 중심으로 ±1SD 수준에서 규범존중의 가치를 높게 지각하는 경우에는 유의하지 않았지만 평균(B=.70, p<.01)과 낮은 수준(B=1.21, p<.001)에서는 유의한 것으로 나타났다. 즉, 규범에 대한 존중이 높다는 것은 서비스 제공자 유형이 서비스 실패에 대한 서비스 제공자의 책임지각에 미치는 영향에 있어서 억제효과를 보인다고 할 수 있다.

논 의

AI의 서비스 제공과 실패 상황에서 책임 주체에 대한 문제는 AI 서비스의 안정적이고 예측 가능한 사용을 위해 중요하게 해결되어야 할 이슈이지만, 이를 위한 법제적, 실증적 근거는 부족한 실정이다. 이에 본 연구는 AI 서비스 실패에서 사람들의 책임지각이 어떻게 이루어지는지를 서비스 제공자가 인간과 AI일 경우를 비교하고, 그 관계가 도덕적 기반에 따라 달라지는지 확인하고자 하였다. 연구 1에서는 머지않은 미래에 보편화될 디지털 심리치료의 실패 상황을 가정하여 서비스 제공자에 따른 책임 주체를 확인하였다. 연구 결과, 심리치료 실패 상황에서 치료자가 인간이냐 AI이냐 여부가 서비스 제공자와 사용자의 책임을 지각하는데 영향을 미치는 것으로 나타났다. 예상대로, AI에 의해서 서비스 실패가 발생했을 경우, 인간치료자인 경우보다 서비스 제공자의 책임을 더 낮게 평가했으며, 사용자의 책임에서도 인간보다 AI치료자인 경우에 책임을 더 낮게 평가하였다. 더하여 서비스 제공자와 사용자 책임지각에서 도덕적 기반의 하위요인인 공평함과 규범존중의 조절효과가 유의하였는데, 전반적으로 공평함과 규범존중의 가치를 평균 이상으로 높게 평가하는 사람들은 심리치료 실패가 인간인지 AI인

지 여부가 서비스 제공자 및 사용자의 책임지각에 크게 영향을 미치지 않았다. 연구 2에서는 국가 기관의 업무 중, 실제 발생하였던 AI 세무 서비스 실패 사건을 활용하여 책임 주체를 확인하였다. 연구 결과, 연구 1과 달리 서비스제공자뿐 아니라 책임 주체의 주효과가 유의하였는데, 세무 서비스 실패의 책임 주체를 정부나 서비스 제공자보다 사용자 책임으로 지각하는 정도가 현저히 낮았다. 연구 1과 마찬가지로 AI가 실패한 경우, 인간이 실패한 경우보다 서비스 제공자의 책임을 낮게 평가하였으며, 유의하지는 않았으나 정부의 책임을 높게 평가하는 경향이 나타났다. 또한, 도덕적 기반의 하위요인 중 규범존중의 가치를 평균 이하로 평가한 사람들과 달리 높게 평가한 사람들은 AI가 제공한 세무 서비스가 실패했을 때도 서비스 제공자의 책임을 인간인 경우만큼이나 높게 지각하였다.

연구 1과 연구 2 모두에서 서비스 제공자가 인간일 때에 비해 AI일 때 서비스 제공자의 책임을 낮게 평가하는 것으로 나타났는데, 이는 예상 가능한 결과이며 선행연구 결과와도 일치한다(Belanche et al., 2020; Maninger et al., 2022). 다만, 표 1과 표 2를 살펴보면 참가자들은 AI의 책임에 대해서도 보통 이상의 책임이 있다고 평가한 것을 알 수 있다(연구 1: M=4.52, 연구 2: M=5.41). 이러한 결과는 사람들이 AI에게도 책임을 물을 가능성을 시사하는 것으로, 이론적 배경에서 제시한 세 번째 어려운 질문인 “AI의 책임과 권리”의 문제 중 적어도 책임에 대해서는 AI의 책임을 인정하는 태도를 지니고 있다고 해석할 수도 있다. 물론 본 연구에서의 측정은 합리적이고 깊은 사고를 측정한 것이 아니고 직관적이고 빠른 반응을 측정한 것이다. 선행연구에 따르면, 도덕적 판단은 정서에 따라 달라질 수 있다(Wheatley & Haidt, 2005; 조자의, 한성열, 2008). 즉, 다른 방식으로 질문한다면 다른 결과가 나올 수도 있으므로, 본 결과를 바로 AI에게 권리와 책임이 있다는 주장의 근거로 제시하기는 어렵다. AI에게 책임과 권리가 존재하는지에 대한 질문은 추후 연구에서 직관적인 측정뿐만 아니라 합리적 사고 측정을 포함한 연구를 통해 다뤄질 필요가 있다.

본 연구의 또 다른 발견점은 민간영역에서의 AI서비스와 공공영역에서의 AI서비스 실패에 대한 책임을 사람들이 다소 다르게 생각한다는 점이다. 주요 연구문제가 아니어서 결과에는 보고하지 않았으나 AI인지 사람인지 여부와 관계없이 민간과 공공서비스 실패의 책임지각 차이를 비교해 보았을 때, 민간영역의 심리치료보다 공공영역의 세무 서비스에서 실패가 발생했을 때, 서비스를 제공한 주체에 대해 평균적으로 많은 책임을 지각하는 것으로 나타났다{$F(1, 189)=34.75$, $p<.001$, $\eta^2=.16$}. 관련 기관의 책임에서는 민간영역과 공공영역에서의 차이가 나타나지는 않았다. 사용자에 대한 책임지각은 심리치료보다 세무 서비스 실패에서 낮아졌는데{$F(1, 189)=4.80$, $p<.05$, $\eta^2=.03$}, 이는 민간영역에서는 해당 서비스를 선택할 수 있으나 공공영역에서는 본인의 선택하기보다는 시스템적으로 주어지는 경우가 많기 때문에 사용자의 책임성을 낮게 지각한 것으로 보인다. 한편, 세무라는 공공서비스를 다룬 연구 2에서는 연구 1과 달리 AI가 서비스한 세무 행정 실패에 대해 인간세무사의 실패보다 정부(기관)의 책임을 높게 평가하는 경향이 나타났다. 서비스 실패에 대한 책임 순서만을 본다면 연구 1과 연구 2 모두에서 AI인 경우는 기관(정부), AI, 사용자이며, 인간이

서비스한 경우는 인간, 기관(정부), 사용자 순으로 민간영역과 공공영역의 차이는 없었다. 이는 AI의 서비스 실패를, 서비스를 제공한 주체인 AI보다 관련된 회사의 책임으로 지각한다는 선행연구의 결과(Belanche et al., 2020)와 일치하는 것으로, 인간서비스와는 달리 AI 서비스의 경우 사람들은 관련 기관 책임을 가장 크게 생각하는 것으로 보인다. 다만 본 연구에서는 세무 서비스(연구 2)에서만 기관(정부) 책임 평가에서 조건 간 유의한 차이가 발견되었는데, 이는 사람들은 민간영역의 AI서비스보다 공공영역의 AI서비스에서 관련기관(정부)의 책임을 더 크게 물을 가능성을 시사한다고 할 수 있다. 따라서 AI에게 책임을 물을 수 있는지, 그리고 얼마나 책임이 있는지에 대한 질문에 대해 AI의 서비스가 공공영역에서 이루어졌는지, 아니면 민간영역에서 이루어졌는지에 따라 다른 답이 도출될 수 있다.

아울러 본 연구는 도덕적 기반의 하위요인 일부가 서비스 제공자에 따른 책임판단에 영향을 미친다는 것을 발견하였다. 하위 요인 중 공평함, 규범존중의 조절효과를 확인하였으며, 이는 AI의 서비스를 판단할 때 개인이 가지고 있는 도덕적 가치가 영향을 미친다는 것을 의미한다. 본 연구에서 도덕적 기반의 하위 요인 중 위해(harm)만이 유의하지 않은 결과를 보였는데, 이는 서비스 실패 자체가 피해자에게 손상을 입히는 것이 명백하기 때문에 위해를 개인적으로 중시하든 중시하지 않든 이미 발생한 명백한 손상에 대한 책임지각에는 별다른 개인차가 나타나지 않았던 것으로 보인다. 따라서 만약 시나리오 실패가 공평성위반이나 규범위반을 포함하였다면, 다른 결과가 나왔을 수도 있으므로 추후에는 위해 뿐 아니라 공평성이나 규범존중 요소가 포함된 시나리오를 사용한 연구가 필요하다.

그럼에도 불구하고 본 연구 결과에서 또한 Haidt(2008)가 설명하였던 도덕적 기반의 문화 보편적인 특징을 보인다. 구체적으로, 도덕적 기반 중 위해와 공평함은 개인의 권리 보호, 즉 개인적 가치와 관련되어 있는 반면, 내집단, 규범존중, 순수는 공동체 결속과 관련된 집단적 가치와 관련되어 있다. 실제로 선행연구들은 위해와 공평함은 진보적 정치성향, 내집단, 순수, 규범존중은 보수적 정치성향과 관련되어 있음을 밝히고 있다(Haidt, Graham, 2007; 이재호, 조긍호, 2014; 정은경, 정혜승, 손영우, 2011). 본 연구에서 조절변인으로 나타난 공평함은 개인적 가치, 규범존중은 집단적 가치에 해당하며 동시에 그 방향성도 모두 상기한 도덕적 가치가 낮은 사람들이 서비스 제공자가 AI인지, 인간인지에 따라 다른 책임판단을 보이는 것으로 나타났다. 이러한 결과는 두 가지의 함의를 가진다. 첫 번째로는 AI의 서비스 실패에 대한 책임판단에는 개인적 가치지향인지 집단적 가치지향인지가 중요한 것이 아니라 도덕적 가치 자체에 대한 절대적 관심 혹은 인식 정도가 중요할 가능성을 시사한다. 따라서 추후 연구에서는 도덕적 기반과 같은 가치를 중점으로 다룬 이론 이외에 도덕성을 다룬 다른 이론을 추가하여 통합적으로 접근할 필요가 있다. 예를 들면, Blasi(1993)은 도덕적 자아 모델(Moral self model)을 주장하며 도덕적 정체성(Moral Identity)을 개인의 도덕적 관심으로 이해하는 것이 가능하다고 설명하였다. 도덕적 정체성에 관한 선행연구에 따르면, 도덕적 정체성이 높은 사람은 친사회적 조직행동을 더 자주 보였으며(Winterich, Aquino, Mittal, & Swartz, 2013), 개인의 도덕적 정체성이 높을수록 기업의 사회적 책임(CSR)을 기준

으로 조직과의 일치성, 매력도를 판단하는 것으로 나타났다(신두은, 손영우, 2024), 이처럼 윤리와 도덕에 관한 중요성이 강조되고 있는데, 추후에는 개인의 도덕성에 관한 풍부한 이해를 위해 도덕적 관심 정도를 반영한 변인을 사용한 연구가 필요할 것으로 보인다. 두 번째로는, 도덕적 가치(공평함과 규범존중)를 높게 평가하는 사람들은 서비스 제공자가 AI인지 인간인지 여부와 관계없이 책임 주체(사용자와 서비스제공자)의 책임을 지각하는데 비슷하게 엄격하다는 점이다. 이는 도덕적 가치를 중요하게 여기는 사람들은 과정보다는 결과에 훨씬 초점을 맞추고 있어서 윤리적 문제가 발생한 상황에서는 누가 문제유발 행위를 했느냐와 같은 과정보다는 문제 상황이라는 결과에 대한 책임을 다소 무조건적으로 강하게 물을 가능성을 시사한다. 따라서 과거 선행연구들이 대부분 '어떤 도덕적 가치를 지녔는가'에 대한 것에 초점을 맞추었다면 앞으로는 '도덕적 가치를 얼마나 중시하는가'에 대한 연구도 활발하게 이루어져야 할 필요가 있겠다.

본 연구는 AI가 제공하는 서비스의 실패에서 AI제조사나 관련 기관뿐만 아니라 실제 AI의 책임성 정도를 측정함으로써 'AI에게 권리와 책임이 존재하는가?' 라는 어려운 질문을 상향식 접근 방법을 통해 탐색하였다는 점에서 학문적 의의가 있다. 본 연구는 AI가 제공하는 서비스 실패에서의 책임성을 인간의 경우와 비교하여 제시하였는데, 생각보다 사람들은 AI 자체에 대해서도 윤리적 책임이 있다고 생각하는 것으로 나타났다. 이러한 결과는 AI의 책임과 관련한 법제화 관련 논의에 기여할 수 있을 것이다. 이와 함께 도덕심리학 접근을 통해 AI서비스에 대한 도덕적 책임을 판단하는 데에는 어떠한 도덕적 가치인지보다는 도덕적 관심 정도가 더 큰 영향을 미칠 가능성과 도덕적 가치를 중시하는 사람들이 오히려 AI의 책임을 더 강하게 인정할 가능성을 제시하였다는 데에도 본 연구의 의의가 있다. 마지막으로 AI의 서비스 실패가 민간영역에서 발생하는지, 아니면 공공영역에서 발생하는지에 따라 책임판단이 달라질 수 있음을 제시한 것도 본 연구의 기여점이라고 할 수 있다.

상기한 학문적 의의에도 불구하고 본 연구는 다음과 같은 한계를 지닌다. 첫째, 본 연구에서는 AI를 컴퓨터와 같은 기계 이미지로만 사용하였는데, 최근에는 인간의 모습을 한 AI 로봇 개발도 활발하게 이루어지고 있으며 이러한 의인화가 AI 판단에 영향을 미친다는 연구도 보고되고 있다(Laakasuo, Palomaki, & Kobis, 2021). 따라서 추후 연구에서는 AI 의인화 정도가 AI 책임판단에 어떤 영향을 미치는지도 연구할 필요가 있다. 둘째, 본 연구에서는 심리치료와 세무 서비스 실패 시나리오만이 사용되었는데, 최근에는 채용, 해고, 재판, 의료, 금융 등 다양한 분야에서 AI가 사용되고 있는 만큼 더욱 다양한 분야에서의 AI 서비스와 그 책임판단에 대한 연구가 필요하다. 마지막으로 본 연구에서는 AI 서비스가 어떤 과정을 통해 도덕적 책임판단과 연결되는지, 즉 매개변인을 전혀 고려하지 못하였는데, 추후에는 선행연구에서 제시된 도덕 쌍 변인(Wegner, & Gray, 2016)과 같은 매개변인에 대한 연구도 이루어져야 할 것으로 보인다.

참고문헌

과학기술정보통신부 (2024. 05. 13). 2023 인터

넷이용실태조사.
https://www.msit.go.kr/bbs/view.do?mId=99&bbsSeqNo=79&nttSeqNo=3173621

국토교통부 (2020. 05. 01). 자율주행 자동차 상용화 촉진 및 지원에 관한 법률 시행 규칙.
https://www.law.go.kr/lsInfoP.do?lsiSeq=263929&efYd=20240710#0000

김계숙, 안현철 (2023). 무엇이 AI 프로젝트를 성공적으로 이끄는가?. 지능정보연구, 29(1), 327-351.

김도연, 조민기, 신희천 (2020). 상담 및 심리치료에서 AI 기술의 활용: 국외사례를 중심으로. 한국심리학회지: 상담 및 심리치료, 32(2), 821-847.

김승수 (2022. 10. 12). 자율주행차 시대 도래... 자동차보험 어디까지 왔나. 대한경제.
https://www.dnews.co.kr/uhtml/view.jsp?idxno=2022101213431024605 48

김용주 (2016). 인공지능(AI) 창작물에 대한 저작물로서의 보호가능성. 법학연구, 27(3), 267-297.

김원일, 윤현식 (2021). 금융 챗봇 서비스의 사용 의도에 대한 질적 탐색. 디지털융복합연구, 19(11), 181-199.

김진우 (2021). 인공지능 시스템의 책임능력-전자인 제도의 도입 필요성에 관한 논의를 중심으로-. 중앙법학, 23(4), 7-51.

뉴시스 (2023. 05. 18). '천국서 같이 살자'는 AI 챗봇 말에 극단적 선택한 남자. 문화일보.
https://munhwa.com/news/view.html?no=20230518MW0649571843 98

박휴용 (2022). 탈인본주의적 AI 윤리란 무엇인가?: '윤리적 AI'를 중심으로. 컴퓨터교육학회 논문지, 25(6), 71-83.

번가남 (2024). 인공지능이 영화 제작을 대체할 가능성에 관한 연구-오픈 AI의 챗 GPT와 소라를 중심으로. 한국과학예술융합학회, 42(3), 121-134.

백경희 (2024). 의료영역에서의 인공지능기술의 활용에 관한 법제적 고찰-원격의료 및 재택의료 활성화를 중심으로. 법학논문집, 48(1), 89-114.

서유진 (2023. 12. 30). 이혼당하고 목숨 끊고, 내각은 총사퇴...네덜란드서 AI가 벌인 짓, 중앙일보.
https://v.daum.net/v/20231230050101977

서윤경 (2023). 생성 인공지능 창작과 저작권의 쟁점들. 인공지능인문학연구, 15, 189-216.

신두은, 손영우 (2024). CSR 진정성이 구직자의 조직매력도에 미치는 영향: 가치일치성의 매개효과와 도덕적 정체성의 조절효과. 한국심리학회지: 문화 및 사회문제, 30(2), 101-120.

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

오세중 (2024. 07. 23). 강민수 국세청장 "불편부당 자세로 세무조사 엄정히 집행할 것". **머니투데이**.
https://news.mt.co.kr/mtview.php?no=2024072309370583569

윤상오 (2018). 인공지능 기반 공공서비스의 주요 쟁점에 관한 연구: 챗봇 (ChatBot) 서비스를 중심으로. 한국공공관리학보, 32(2), 83-104.

윤석찬 (2024). 소위 '약한 인공지능 (AI)'영역

에서의 민사책임에 관한 연구. 민사법학, 171-201.

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

이소정, 남혜정 (2024. 09. 26). 국내 첫 '심야 자율주행택시' 오늘부터 강남서 운행. 동아일보. https://www.donga.com/news/article/all/20240926/130104737/1

이수경 (2024). 인공지능의 민사책임에 대한 소고. 민사법학, 107, 225-261.

이재호, 조긍호 (2014). 정치성향에 따른 도덕판단기준의 차이. 한국심리학회지: 사회 및 성격, 28(1), 1-26.

이창민 (2018). 인공지능의 형사책임. Law & Technology, 14(2), 41-59.

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

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

정은경, 정혜승, 손영우 (2011). 진보와 보수의 도덕적 가치 판단의 차이: 용산재개발사건을 중심으로. 한국심리학회지: 사회 및 성격, 25(4), 93-105.

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

조자의, 한성열 (2008). 행위자의 화 (火) 가 한국대학생의 도덕적 판단에 미치는 영향. 한국심리학회지: 사회문제, 14(1), 47-75.

조한상, 이주희 (2016). 인공지능과 법, 그리고 논증. 법과 정책연구, 16(2), 295-320.

최원석, 최성곤 (2021). 자율주행 자동차 기술 동향에 관한 연구. 컴퓨터정보통신연구, 29(1), 27-34.

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

Andersson, G. (2002). Psychological aspects of tinnitus and the application of cognitive-behavioral therapy. *Clinical Psychology Review, 22*(7), 977-990.

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

Belanche, D., Casaló, L. V., Flavián, C., & Schepers, J. (2020). Robots or frontline employees? Exploring customers' attributions of responsibility and stability after service failure or success. *Journal of Service Management, 31*(2), 267-289.

Bickmore, T., Gruber, A., & Picard, R. (2005). Establishing the computer‐patient working alliance in automated health behavior change interventions. *Patient Education and Counseling, 59*(1), 21-30.

Blasi, A. (1993). The development of identity: Some implications for moral functioning. In G. G. Noam & T. E. Wren (Eds.), *The Moral Self*(pp.99-122). Cambridge, MA: MIT Press.

Chowdhury, R. M. (2019). The moral foundations of consumer ethics. *Journal of Business Ethics, 158,* 585-601.

Feinberg, M., & Willer, R. (2013). The moral

roots of environmental attitudes. *Psychological Science, 24*(1), 56-62.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health, 4*(2), e19.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health, 5*(4), e64.

Gollings, E. K., & Paxton, S. J. (2006). Comparison of internet and face-to-face delivery of a group body image and disordered eating intervention for women: a pilot study. *Eating Disorders, 14*(1), 1-15.

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,* 366-385.

Gulec, H., Moessner, M., Mezei, A., Kohls, E., Túry, F., & Bauer, S. (2011). Internet- based maintenance treatment for patients with eating disorders. *Professional Psychology: Research and Practice, 42*(6), 479-486.

Gulec, H., Moessner, M., Mezai, A., Kohls, E., Tury, F., & Bauer, S. (2011). Internet -based maintenance treatment for patients with eating disorders. *Professional Psychology: Research and Practice, 42*(6), 479-486.

Hagerty, A., & Rubinov, I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:*1907.07892.

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. (2008). Morality. *Perspectives on Psychological Science, 3*(1), 65-72.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133*(4), 55-66.

Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The Innate Mind, 3,* 367-391.

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

K. Buchholz. (2023. 07. 07). Threads Shoots Past One Million User Mark at Lightning Speed. *Statista.* https://www.statista.com/chart/29174/timetoone-million-users

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, 13*(7), 1679-1688.

Litz, B. T., Engel, C. C., Bryant, R. A., & Papa, A. (2007). A randomized, controlled proof-of-concept trial of an Internet-based, therapist-assisted self-management treatment for post traumatic stress disorder. A*merican Journal of Psychiatry, 164*(11), 1676-1684.

Maninger, T., & Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports, 5,* 100154.

Nowakowska, I., Duda, E., & Szulawski, M. (2024). Heading for the better future with my company: Work related prosocial intentions as a function of moral foundations and consideration of future consequences. *Journal of Community & Applied Social Psychology, 34*(4), e2825.

Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., & Mermelstein, R. (2011). Health behavior models in the age of mobile interventions: Are our theories up to the task?. *Translational Behavioral Medicine, 1*(1), 53-71.

Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., & Mermelstein, R. (2011). Health behavior models in the age of mobile interventions: Are our theories up to the task?. *Translational Behavioral Medicine, 1*(1), 53-71.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3), 417-457.

Waytz, A., Dungan, J., & Young, L. (2013). The whistle blower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology, 49*(6), 1027-1033.

Wegner, D. M., & Gray, K. (2016). *The Mind Club: Who thinks, what feels, and why it matters.* New York: Viking.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* W. H. Freeman and Company.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological science, 16*(10), 780-784.

Winterich, K. P., Aquino, K., Mittal, V., & Swartz, R. (2013). When moral identity symbolization motivates prosocial behavior: The role of recognition and moral identity internalization. *Journal of Applied Psychology, 98*(5), 759-770.

논문 투고일 : 2025. 01. 13
1 차 심사일 : 2025. 02. 26
게재 확정일 : 2025. 04. 14

한국심리학회지 : 문화 및 사회문제
Korean Journal of Culture and Social Issues
2025, Vol. 31, No. 2, 245～268.

Perception of Responsibility for AI Service Failures: The Moderating Effect of Moral Foundations

Ju Hyun Seong **Eun Kyoung Chung**

Department of Psychology, Kangwon National University

As AI usage is expanding for human convenience, ethical discussions on this issue remain insufficient. This study examined whether AI can bear responsibility and how people perceive it to take responsibility. In specific, by comparing AI service failures to those of humans, we focused on how the perceived responsibility is for each agent (service provider, organization, or user) and tested the moderating role of moral foundations. In Study 1, where a psychological treatment failure scenario was given, results showed that according to the service provider (AI or human) responsibility perceptions of the service provider, organization, and user varied. Fairness and respect for norms moderated the relationship of service provider type and user responsibility perception. In specific, when these values were rated low, users felt less responsible for AI failures than human failures. Study 2, involving a scenario on the government tax failures, found that responsibility was attributed the highest for the government, followed by service providers and users. AI failures led to lower provider responsibility and higher government responsibility. Respect for norms also showed moderationed perceptions, with lower values further reducing the perception of provider responsibility under AI conditions compared to that of human. These findings contribute to the understanding of whether AI can be held morally responsible and how moral judgment differs depending on the nature of the agent. Limitations and implications of the study are also discussed.

Key words : Artificial Intelligence(AI), Perception of responsibility, Moral Foundations, Digital Therapeutics, Tax Services