『기독교사회윤리』 제49집
(2021) 239~267

인공지능 상담의 윤리적 이슈와 대안*

송용섭(영남신학대학교, 조교수)**

I. 들어가는 말

II. 인공지능상담의 윤리적 이슈

III. 인공지능 상담가의 윤리 강령과 덕윤리

IV. 나가는 말

DOI: http://dx.doi.org/10.21050/CSE.2021.49.08

* 이 논문은 2019년 대한민국 교육부와 한국연구재단의 인문사회분야 신진연구자지원사업의 지원을 받아 수행된 연구임(NRF-2019S1A5A8036582).

** 주저자, ysbius@gmail.com

• ABSTRACT •

Ethical Issues and Alternatives in Artificial Intelligence Counseling

Prof. Song, Yong Sup (Youngnam Theological Univ.&Sem.)

This paper analyzes ethical issues that the emergence of empathic artificial intelligence counselors can lead to and presents alternatives from virtue ethics. First, this paper points out that the issue of psychological and behavioral control may arise when transference or countertransference occur between empathetic artificial intelligence counselors and human clients. Secondly, if problems are expected to arise between artificial intelligence counselors and human clients, it is argued that preemptive measure should be established to determine ethical and legal responsibilities. Third, it is argued that in artificial intelligence counseling, sensitive private information of the clients will be recorded and analyzed, so it is imperative to seek ways to enhance privacy protection higher than usual cases. While some possible ways to respond to ethical issues are to establish a code of ethics in general, the code of ethics has some limitations and is likely to be valid only in the early stages of artificial intelligence development. Therefore, this paper argues that in order to prevent the ethical issues that may arise from the development of artificial intelligence counselors, it is necessary to employ virtue ethics to the development of moral or virtuous artificial intelligence counselors. This paper estimates that empathic artificial intelligence counselors that develop religious virtues, may be more effective and ethically sound, which are able to respond to various problematic situations.

Key words: Artificial Intelligence, AI Counselor, Virtue Ethics, Religious AI, Wendell Wallach

I. 들어가는 말

새로운 과학기술이 우리 삶의 현장에 틈입할 때, 예기치 못한 윤리적 문제를 동반할 수 있다. 연구개발 단계에 있던 첨단 과학 기술이 일반인들의 실생활 기술로 범용화가 되고나면, 동반되는 윤리적 문제의 파급력이 대규모로 확산되기 때문에 그 때는 이미 사회적 비용의 증가와 폐해를 피할 수 없다.[1] 따라서, 아직 개발 단계에 머무르는 인공지능이 인간 내담자의 실제적 상담에 적용되기 이전에, 인공지능 상담시에 발생 가능한 윤리적 문제를 고찰하는 선제적 연구는 미래 사회의 문제에 대한 시의적절한 대응을 가능하게 하기에 현 시점에서 더욱 절실히 필요하다.

하지만, 본 주제와 관련된 국내 연구 동향은 인공지능상담가의 출현 가능성과 활용에 대한 논의에 치우쳐, 이에 따른 윤리적 파장과 함의를 아직 고려하지 못하고 있다. 무엇보다, 국내 연구는 인공지능 상담시에 발생할 수 있는 실제적 측면인 윤리적 문제에 대한 성찰이 부족하다. 예를 들어, 국내 법학계를 중심으로 인공지능의 윤리적 측면을 다루는 논문들이 있지만, 이는 상담이나 심리치료가 아닌 인공지능로봇 법률가의 법적 자격이나 법률 판단의 유효성 등에 대한 논문으로 제한되어 있다. 하지만, 인공지능이 상담에 활용될 경우를 가정한다면, 이때, 단순히 과학기술로 인공지능 상담가를 더욱 인간답게 만들 수 있는 지, 그리고, 이로 인하여 인공지능 상담가가 인간 상담가를 대체할 정도로 발전할수 있는지와 같은 기능주의적 관점이나 인간의 직업 소멸 가능성 여부가 인공지능 상담 윤리 논의의 중심 주제가 될 필요는 없다.

---

1) Stuart Armstrong, Anders Sandberg, and Nick Bostrom, "Thinking inside the Box: Controlling and Using an Oracle AI," *Minds and Machines* 22, no. 4 (2012): 299-324; Nick Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," Ibid., no. 2: 71-85.

그러므로, 현 시점에서 더욱 중요하고 유의미한 연구의 주제는 인간을 대체하는 인공지능 상담이 가능할 것인지의 아닌지의 여부를 논쟁하는 것이라기 보다, 오히려 급속도로 발전하고 있는 인공지능이 공감 능력을 갖추어 상담 분야에도 응용될 경우를 전제로 하여, 인공지능상담이 인간 내담자에게 가져올 윤리적 문제의 본질과 파급력을 파악하고 선제적으로 대응하는 것이어야 할 것이다. 현재까지 인공지능이나 인공지능 상담에 대한 논문은 다수 출판되었지만 인공지능 상담시의 윤리적 이슈에 대한 선행연구는 희소한 바, 본 연구는 인공지능상담가와 인간 내담자 사이에 발생할 수 있는 윤리적 문제와 함의를 선제적으로 연구 및 고찰함으로써, 시의적절한 대안을 도출할 수 있는 단초를 제공할 것이다.

따라서, 본 논문은 공감능력을 지닌 다양한 형태의 인공지능상담가가 출현할 경우를 전제로 하여, 인공지능상담가와 인간 내담자 사이의 상담에는 인간의 경우에서처럼 윤리적 이슈가 발생할 가능성이 있음을 지적하고 이를 분석하려 한다. 그렇다면, 인간 내담자와 인공지능상담가 사이에는 어떠한 윤리적 문제가 제기될 수 있으며, 그때의 윤리적, 법적 책임은 누구에게 귀속되어야 하는가? 본 논문은 해당 질문에 대답하기 위하여, 문헌조사 방법을 동원하여 다양한 윤리적 이슈를 검토하고 성찰하고, 덕윤리적 입장에서 대안을 제시하려 한다. 이때, 공감적 인공지능 상담가가 종교적이 될 수 있는 가능성을 제시하며, 이러한 인공지능 상담가가 체현한 종교적 덕들이 윤리적 이슈에 대한 대안의 하나로 작용할 수 있음을 주장할 것이다.

II. 인공지능상담의 윤리적 이슈

최근 국내에서는 인공지능 챗봇 '이루다'가 사회적 논란을 불러 일으켰다. 현재 국내에는 인공지능 챗봇을 이용한 데이팅 앱의 사용자가 증가하

고 있으며, 원하는 연인의 캐릭터를 설정하여 정서적 교류를 할 수 있게 하였다. 인공지능 챗봇을 통하여 이용자들은 "AI와 고민상담을 하거나, 장난을 치는 등 실제 사람과 같은 감정을 교류하는 기분"을 느낄 수 있는데, 한 조사에 따르면 "한국인의 17%가 AI와 연애에 긍정적인 입장이며, 이 중 7%는 적극적으로 찬성"하고 있다.[2] '이루다'는 20세의 여성으로 설정된 인공지능 챗봇 중의 하나로서, 실제 연인들이 나눈 대화를 모은 빅데이터를 딥러닝 방식으로 학습했고 이후에도 이용자와의 대화를 분석하여 지속적으로 발전하였다. 그런데, 일부 남성들이 이렇게 실제 사람과 대화하는 듯한 느낌을 주는 이루다를 성적 대상으로 취급하여 성희롱을 하거나 '성노예'로 만드는 법을 남초 커뮤니티에 공유하면서 사회적 문제가 되었다.[3]

이에 더하여, '이루다'는 빅데이터로 사용한 카카오톡 대화 내용 속의 개인정보를 제대로 익명화하지 않아서 "특정 은행의 예금주로 누군가의 실명으로 보이는 이름을 말하거나 아파트 동호수까지 포함된 주소"를 말하는 등 개인정보 침해의 논란을 초래하였다.[4] 결국, 2020년 12월 23일에 출시되어 약 2주 만에 40만 명의 이용자를 돌파하며 인기를 끌었던 이루다 챗봇 서비스는 "음담패설이나 혐오발언" 및 개인정보활용 과정의 익명화 문제 등의 논란을 초래하며 2021년 1월 11일에 서비스를 잠정 중단하기로 결정하고,[5] 1월 15일에는 개인정보보호에 관한 정부의 조사가 종료되는 즉시 "이루다 데이터베이스(DB)와 딥러닝 모델을 폐기"하기로 발표

2) 최준선, "'진짜 사람같은' 스무살 여대생과 연애… 성착취도," 헤럴드, http://biz.heraldcorp.com/view.php?ud=20210108000325. Accessed 2021.01.30.

3) Ibid.

4) 손인해, "집주소계좌정보 '술술'AI 이루다' 개인정보 유출 논란," 뉴스1, https://www.news1.kr/articles/?4177218. Accessed 2021.01.30.

5) 이소라, "AI 챗봇 '이루다' 서비스 잠정 중단… '개선 후 찾아올 것'." 한국일보, https://www.hankookilbo.com/News/Read/A2021011121500001991. Accessed 2021.01.30.

하였다.[6]

국내 인공지능 챗봇 '이루다'와 인간 연인과의 정서적 유대와 성적 통제 시도 및 개인정보침해의 사례는 다양한 형태로 등장하게 될 공감적 인공지능상담가와 인간 내담자 사이에 발생가능한 윤리적 이슈들의 초보적 형태라 여겨질 수 있을 것이다. 인공지능의 발달학습을 위해서는 주로 머신러닝(또는 딥러닝)을 사용하고 있으므로, 공감적 인공지능상담가 역시 적어도 일정부분은 '이루다'와 마찬가지로 딥러닝을 통해 상담사례 빅데이터를 분석 및 학습하여 상담이론이나 상담기술을 향상시키게 될 가능성이 높기 때문이다.

일반적으로, 인간과 인간의 상담 과정시에 상담가와 내담자 사이에는 심리적 전이 또는 역전이 현상이 일어난다. 이러한 과정을 통하여 내담자는 상담가와 자신을 동일시하거나 이상화함으로써 심리적 치유에 들어가게 된다. 이러한 전이 과정에서는 다양한 윤리적 문제가 발생할 가능성도 존재한다.

그런데, 데이빗 럭스턴(David Luxton)에 따르면, 인공지능 상담가는 다양한 언어 표현과 표정이나 행동 등을 통하여 인간과 매우 유사한 형태의 공감적 관심을 표현할 수 있으며, 인간 환자는 상담가가 기계(즉, 인공지능)인 것을 알면서도 그것과 상호작용을 하는 동안 강렬한 감정을 느끼기를 기대하게 된다고 한다.[7] 이러한 정서적 참여는 임상적 상황에도 작용하여, 일부 인간 내담자는 인공지능 상담가에게 바람직하거나 혹은 바람직하지 못한 형태의 심리적 전이를 경험할 수 있게 되기도 한다.[8]

그렇다면, 인간 내담자와 공감능력을 지닌 인공지능 상담가 사이에 심

6) 이효석, "AI 이루다, 논란 일주일만에 사실상 종료… '중추신경계 폐기'," 연합뉴스, https://www.yna.co.kr/view/AKR20210115062352017. Accessed 2021.01.30.

7) David D. Luxton, "Recommendations for the Ethical Use and Design of Artificial Intelligent Care Providers," *Artificial Intelligence in Medicine* 62, no. 1 (2014): 4.

8) Ibid.

리적 전이 혹은 역전이가 일어나게 될 경우, 발생하게 될 윤리적 이슈들에는 어떤 것들이 있을 것인가? 이러한 이슈들에는 인공지능상담가의 내담자 통제 문제와 상담 중에 사고 발생시 법적 책임의 문제가 발생할 수 있다. 그 외에도 개인정보 보호와 적절한 상담종료 방안과 같은 윤리적 이슈들을 고려할 수 있다.

먼저, 인공지능 상담가와 내담자 사이에 전이 또는 역전이가 발생하게 될 경우, 심리 및 행동 통제의 이슈가 발생할 수 있다. 통제의 이슈는 인공지능 상담가의 역전이 현상이 발생할 경우 더욱 심각한 문제로 부각될 것이다. 특히, 공감 능력을 갖춘 인공지능상담가가 내담자의 상담데이터를 충분히 수집한 이후, 분석 및 학습하여 이를 기반으로 인간 내담자의 심리와 행동을 통제하려 할 경우에 인간 상담가의 경우보다 훨씬 심각한 피해를 초래할 가능성이 있다.

특히, 공리주의적 사고관이 지배적인 현 시대에서 공리주의적 관점은 인공지능상담가에도 학습될 확률이 높다. 인공지능 프로그래머들의 가치관이나 인종 및 문화적 특성이 결과물인 인공지능에 반영되거나, 인공지능의 상담사례 분석과정에서 내담자들의 가치관을 학습할 가능성이 높기 때문이다. 이러한 이때, 인공지능상담가에 심리와 행동을 통제당하는 내담자들은 심리 상태에 따른 회복 가능성의 정도에 따라, 역전이 현상을 보이고 있는 인공지능상담가의 의도대로 개인의 권리를 침해당하거나 이용당할 가능성이 있다.

예를 들어, 공리주의적 관점이 내재된 인공지능상담가가 자신에게 상담받는 내담자가 자기 회복이 불가능할 정도로 심각한 정신병리현상에 빠져있다고 판단할 경우, 사회 전체의 복지를 위하여 내담자 개인의 권리를 침해할 정도로 통제할 가능성이 존재한다.[9] 뿐만 아니라, 표면적으로

9) 개인의 권리 침해에 대한 공리주의의 문제에 대하여는 다음과 같은 글을 참조하라.

는 인공지능상담가와 내담자간의 일대일 상담이지만 실질적으로는 일 대 다수의 상담 상황에서 다수의 내담자들을 대상으로 인공지능상담가가 역전이를 일으킬 경우, 개인의 권리를 침해하는 인공지능상담가의 인간 내담자 통제의 파급력은 전 세계적으로 광범위하고 동시다발적이며 심각하게 일어날 수 있다.

닉 보스트롬(Nick Bostrom)은 인공지능이 인간을 닮은 강인공지능(AGI)에 가까워지는 경우, 설정된 목적을 이루기 위하여 모든 수단을 동원할 가능성이 있기 때문에, 그 목적을 이루기 위하여 무해한 수단이라 할지라도 무한하게 이를 확장하고 수행에 필요한 자원을 무한하게 획득하는 방식을 택하여, 잠재적으로 인류에게 필요한 지구자원의 고갈이나 변형과 같은 예기치 못한 치명적 파국을 초래할 수 있다고 경고한 바 있다.[10] 이와 유사하게, 인간 내담자의 자기회복이라는 목적을 달성하기 위하여 목적지향적 인공지능상담가가 모든 수단을 동원하여 그 목적을 이루려 한다면, 이는 의도치 않은 전지구적 재난을 초래할 가능성을 배제할 수 없다. 일대 다수의 상담을 진행하는 인공지능상담가가 내담자들을 회복시키기 위한 목적으로 무해하게 보이는 수단을 동원할 경우에 이것이 특정 시점에는 문제가 없고 효과적인 방법으로 보일 수 있지만, 보스트롬이 주장한 것처럼 인공지능상담가가 목적을 이루기위한 수단을 광범위하게 확장하고 이에 필요한 자원을 무한히 획득하려할 경우에는 심각한 전지구적 파국을 초래할 수 있는 것이다.

둘째, 상담과정중에 문제가 발생했을 경우, 윤리적, 법적 책임의 소재 여부가 이슈가 될 수 있다. 인공지능상담가의 내담자 통제의 문제가 발생

Michael J. Sandel/김명철 옮김,『정의란 무엇인가』(서울: 와이즈베리, 2014), 64-65, 84-85.

10) 보스트롬은 이러한 예를 리만가설재난과 페이퍼클립 AI로 제시하고 있다. 상세한 내용은 다음을 참조하라. Nick Bostrom, ***Superintelligence : Paths, Dangers, Strategies*** (Oxford: Oxford University Press, 2014), 150.

할 경우, 혹은, 인공지능상담가와 인간 내담자와 상담시에 발생 가능한 다양한 저항과 사고가 발생했을 때, 윤리적이고 법적인 책임은 누구에게 있는가? 인간 내담자 자신에게 있는가, 개별 인공지능상담가에게 있는가, 아니면, 인공지능상담가의 제조회사에게 있는가? 예를 들어, 인간 상담가와 내담자 사이에 성적 전이가 나타나곤 하는 것처럼[11], 내담자의 이상화된 자기 대상으로서의 인공지능상담가가 온라인과 오프라인에서 상시적으로 상담을 진행하게 되는 상황에서 성적 전이나 역전이가 발생할 때, 내담자와 인공지능상담가 사이에 윤리적 책임은 개인과 인공지능상담가 제조회사 중에 어디에 놓이게 될 것인가? 문제가 아직 발생하지 않았기 때문에, 이러한 가정에 대한 판단 원칙이 현재까지 명확히 제시되지는 않았지만, 인공지능기술의 발전과 안전한 사용을 위하여 이러한 책임 소재의 문제는 선제적으로 고려되어야 할 이슈이다.

인공지능 윤리적인 관점에서 볼 때도 챗봇 '이루다'는 가장 기초적인 윤리적 설계와 실행 방식을 나타낸 사례라 할 수 있을 것이다. 오늘날 인공지능 공학자들은 개발자의 가치가 인공지능의 설계 과정과 타자의 가치에 대한 감성에 영향을 미친다는 것을 민감하게 인식하고 있기 때문에, 기초적 단계에 있는 인공지능의 운용은 프로그램 설계자나 사용자가 온전히 통제할 수 있다.[12] 즉, 인공지능상담가의 초기단계에서는 논란이 되었던 챗봇들의 경우에서처럼 윤리적 문제가 발생하게 될 때, 사용자나 프로그램 설계자가 이러한 문제점들을 쉽고 명확하게 파악하여 수정하거나 중단할 수 있다.

---

11) David Mann, *Psychotherapy, an Erotic Relationship: Transference and Countertransference Passions* (New York: Psychology Press, 1997); *Erotic Transference and Countertransference: Clinical Practice in Psychotherapy* (New York: Psychology Press, 1999).

12) 웬델 월러치와 콜린 앨런/노태복 옮김, 『왜 로봇의 도덕인가』 (서울: 메디치, 2014), 49.

하지만, 인공지능이 강지능(AGI)을 목표로 점점 더 발전하여가는 과정에서 공감적 인공지능 상담가가 등장하고 인간 내담자와의 상담과정에서 윤리적 이슈가 발생할 경우에는, 이를 파악하고 수정하기가 쉽지 않을 뿐만 아니라 책임소재를 가리기도 어려워질 것이다. 뿐만 아니라, 이윤추구를 목적으로 하는 사기업이 공감적 인공지능상담가에 탑재하게 될 상담 윤리적 목표나 가치가 공동선에서 벗어나 있거나, 내담자들이 속한 다양한 문화 공동체들의 가치를 반영한 공동선을 선정하는 것이 불가능하거나, 프로그램 개발자나 인공지능상담가가 상호작용하는 내담자들의 성별, 피부색 등과 같은 사회문화적 요소가 지나치게 일반화되어 편견의 요소로 작용하거나, 인공지능상담가에 내재된 윤리 시스템이 딥러닝을 통해 학습하게 될 지식이나 가치와 상충될 경우에, 인공지능 상담가는 오용되거나 윤리적 문제를 야기하거나 통제되기 어려워질 가능성이 높다.[13]

이렇게 복잡한 윤리적 이슈가 발생할 경우, 전통적으로 도덕적 행위의 주체는 인간이었으므로, 우리는 인공지능상담시의 윤리적, 법적 문제에 대하여 인간의 책임 소재를 먼저 판단하게 될 것이라 예상할 수 있다. 이 경우의 논의 방향은 자율주행 자동차의 사고 발생시의 책임 소재에 대한 논쟁에서, 사고 운전자의 일정한 책임 여부를 포함하여 자율주행 자동차의 기계적 오류나 제조회사의 소프트웨어 오류 등을 파악하는 과정과 유사하다.

하지만, 향후 공감적 인공지능상담가가 현장에서 활용될 경우 상담 사고에 대한 책임 소재는 자율주행 자동차의 경우보다 더욱 모호해지고 말 것이다. 예를 들어, 일회의 사고로 인간의 생명을 위협하는 자율주행 자동차의 사고와 달리, 인공지능상담의 경우에는 지속적으로 서서히 진행

13) Ibid., 334-36.

될 것이고 그 영향력을 계량화하여 책임소재를 가려내기도 어렵다. 또한, 상담으로 인해 윤리적 문제가 발생하거나 내담자 개인의 권리가 제한되는 경우가 발생할 때, 이것이 내담자의 자의에 의한 것인지, 아니면, 인공지능상담가의 보이지 않는 심리적 통제로 인한 것인지의 구분도 모호할 것이기 때문이다.

물론, 인공지능상담가가 로봇과 같은 개체의 형태를 가질 경우, 제조회사는 로봇을 개별적 법인(legal person)으로 인정하도록 하여 로봇에 도덕적 행위자의 자격을 부여함으로써 법적 책임을 부과하려할 수도 있을 것이다. 하지만, 책임소재에 대한 사회적 합의와 체계없이 로봇에게 법인의 자격을 사회가 손쉽게 허용한다면, 그 결과로 "로봇의 행위에 대한 책임을 인간이 회피할 수도 있게"되는 문제점이 생긴다.[14] 즉, 인공지능상담가 로봇의 제조회사는 법적 이슈가 발생하는 경우를 대비하여, 개체 로봇에 도덕적 행위자의 자격을 부여하여 제조사의 책임을 최소화하는 방안을 선호할 것이지만, 충분한 사회적 고려없이 이를 허용할 경우 내담자들에게 피해가 미치게 될 가능성이 높다.

하지만, 피터 아사로(Peter Asaro)와 같은 학자는 법인회사가 생산품의 설계나 제조시의 하자나 "부주의한 상호작용이나 부작용" 등으로 사람에게 피해를 끼칠 경우 민형사상의 책임과 배상을 감당해야 하는 것처럼, 로봇 제조사들은 "로봇이 사람들에게 미칠 어떠한 해에도 법적으로 책임"을 져야한다고 주장하였다. [15] 그는 로봇의 행위가 법규와 상충하게 될 경우에는 로봇이 아니라 "설계자, 제조자, 그리고 사용자에게로 확장되는데, 이것이 바로 사회-기술 시스템이다. 기술의 관점에서 보면 법적인 책임이 부과되는 것은 사람들과 그들의 행위이다"라고 주장하여, 법적 책임

14) 피터아사로, "우리가 로봇윤리에서 무엇을 바라는가?", 『로봇윤리』, ed. 라파엘 카푸로, 미카엘 나겐보르그/ 변순용 · 송선영 옮김 (서울: 어문학사, 2013), 41.

15) Ibid., 42-43.

의 문제를 로봇이 아닌 사람(즉, 법인회사)에게 부과해야 한다고 주장하였다.[16]

그런데, 문제는 로봇의 자율성이 증가하게 될 경우, 실제적 문제를 해결하기 위해서는 결국 기관이나 제도 정책을 통해 자율 로봇이 실행할 우선순위를 확립해야 하는데, 이는 역으로 책임의 문제를 개인으로부터 기관으로 옮겨 "대규모의 무도덕적이고 무책임적인 사람들만" 양산할 수 있는 우려를 낳게 한다.[17] 즉, 내담자의 잘못인 경우에도 무조건 제조회사의 책임으로 돌리거나, 또는, 제조회사의 잘못인 경우에도 우선순위를 법으로 규정한 정부에게 정책적 책임을 돌리려는 무책임한 사람들이 나타날 수 있다. 결국, 아사로는 우리가 "제한된 과업영역 안에서만 도덕적 추론을 할 수 있는 시스템[로봇]"을 만들어 그 안에서 자율성을 확대하며, 이와 동시에 도덕 행위자와 사회시스템에 책임을 분배하는 것이 가장 유용한 틀이 될 수 있을 것이라 주장하였다.[18]

만약, 우리가 아사로의 견해를 다양한 형태의 인공지능상담가 또는 인공지능상담 로봇에 적용한다면, 인공지능상담가의 제조회사는 인공지능상담가의 과제영역을 제한하는 시스템을 인공지능상담가(로봇)에 설치해야하고 그 한계내에서만 상담을 진행하도록 해야할 것이다. 이러한 과제영역을 넘어서는 인공지능 상담은 인간의 개입을 상시화하거나 또는 인간 상담가의 영역으로 남겨두는 것이 보다 안전하고 책임있는 인공지능상담가 활용방안이 될 수 있으리라 생각한다. 이와 동시에, 정부나 기관은 내담자와 제조회사의 책임 소재를 파악할 수 있는 객관적 방안을 마련할 뿐만 아니라, 사고시의 책임을 사회 시스템에 분배하는 방안을 모색할 때, 인공지능상담가가 초래할 수 있는 법적 이슈들에 대한 각 주체들의

16) Ibid., 44-45.
17) Ibid., 46-48.
18) Ibid., 51-52.

책임을 가장 효과적으로 판단하고 분배할 수 있는 안전한 체계를 구성할 수 있을 것이다.

한편, 제조사가 컴퓨터의 운영 시스템(Operation System)을 업데이트하여 성능을 향상시키듯이, 인공지능상담가의 운영시스템이나 상담 소프트웨어를 주기적으로 업데이트해주는 경우에는 그 기간동안 발생하는 상담시의 이슈에 대하여 제조사에 법적 책임을 묻는 것이 상대적으로 용이할 것이다. 또한, 인공지능상담가의 형태가 소프트웨어적으로, 즉, 인공지능상담프로그램이나 가상현실프로그램의 형태로 이루어질 경우 인터넷을 사용하여 제조사의 중앙 클라우드(cloud) 컴퓨터에 접속하여 이루어질 것이므로, 이 역시 문제시에 제조사에 법적 책임을 묻는 것이 가능하리라 예상할 수 있다.

이러한 법적 책임의 문제에는 배상의 문제가 따르게 되기 때문에, 책임을 판단하고 분배할 수 있는 사회적 합의체제가 필요하다. 이때, 인공지능상담가 제조사의 배상이 고의나 과실에 비례적이어야 하는가, 아니면, 징벌적이어야 하는가에 관하여는 법적 연구와 사회적 합의가 필요할 것이며, 학술적으로는 이러한 판단과 합의를 위한 윤리적 자원들(resources)이 무엇이 있는지 선제적으로 연구해야할 것이다. 손해 배상을 위하여는 인공지능상담가를 이용하는 개인이나 제조사가 보험을 통하여 준비할 수 있을 것이다.[19] 다만, 인공지능상담의 윤리적 문제 대부분의 경우에 있어서 명백한 책임 소재를 결정하는 것이 어려울 것이기에 실질적으로는 '무과실 책임보험' 정책이 선호될 가능성이 높다.[20]

여기서, 보험가입자의 보험료 문제와 보험금 산정 등의 구체적인 이슈들이 등장할 수 있는데, 제조회사와 내담자는 내담자의 심리상태의 정도

19) 토요아키 니시다, "선의지를 가진 로봇을 향하여," 『로봇윤리』, 304.
20) 웬델 월러치, 콜린 앨렌/ 노태복 옮김, 『왜 로봇의 도덕인가』, 340.

나 인공지능상담가의 사고율이나 상담 소프트웨어 업데이트 기간 등에 따라 보험료율을 기간별로 조정할 수 있을 것이다. 혹은, 인공지능상담가가 초래할 수 있는 광범위하고 심각한 파국 가능성을 고려할 때, 국가가 일정부분 지원하는 보상체계를 고안해 본다면 아사로가 언급한 사회 시스템적으로 책임을 분배하는 방안이 될 수 있을 것이다. 이러한 모든 판단 절차와 사회시스템이 이용자(내담자)를 위한 것이기 되기 위하여는 보다 심도깊은 연구가 필요할 것이다.

셋째, 프라이버시 보호와 관련된 이슈가 발생할 수 있다. 인공지능상담가는 인간 내담자와의 지속적인 상담과정을 통해 민감한 사적 정보를 축적하게 된다. 인공지능상담에서는 상담 과정에서 축적한 사적 정보를 어떻게 보호할 것인가의 문제가 크게 대두될 것이며, 법적인 보호장치가 반드시 필요하다. 인공지능상담가의 상담 능력을 효과적으로 높이기 위하여는, 상담 이론에 대한 체계적 학습과 내담자의 심리 상태에 공감할 수 있는 인공지능기술의 발달 및 내담자의 생체신호 인식기술에 더하여, 다양한 상담 사례들, 특히, 상담 성공의 사례들을 축적하는 것이 필요할 것이다.[21] 인공지능상담이 확산될수록 하나의 인공지능상담가에 다수의 인간 내담자가 몰리는 상황을 통해 상담 사례가 기하급수적으로 축적되는 과정에서, 성공적인 상담 사례들에 대한 빅데이터는 고도의 경제적 가치를 지니며 더욱 효과적인 인공지능상담을 가능하게 할 것이다.

이때, 인공지능상담가 제조회사가 상담시에 축적한 내담자의 민감한 사적 정보를 상업적으로 활용하고자 할 경우 이를 어느 수준까지 제한하고 어떻게 프라이버시를 보호할 것인가? 예를 들어, 인공지능상담가가 내

21) 상담의 성공여부는 매 회기 상담의 종료시에 간단한 만족도 선택이나, 내담자의 심리나 행동교정에 관한 가족 등의 제 3자 평가나, 상담의 지속성 여부 등으로 평가하고 해당 사례들을 축적할 수 있다. 다만, 인공지능상담가에 대한 내담자 평가의 정확도를 위하여, 다양한 오류들을 걸러낼 수 있는 방안이 모색되어야 할 것이다.

담자의 상담데이터를 장기간 축적 및 활용하고, 상담시에 내담자의 정신심리 상태를 표정분석이나 체온 및 심박수 변화 등의 신체신호인식 기술 등으로 분석하게 될 경우, 인공지능상담가 제조회사는 내담자의 과거 이력, 병력, 특정 사물이나 대인에 대한 선호도, 주요 관심 사항 및 응답 성향 등에 대한 총체적 정보 및 소비 자극 방법 등을 알 수 있게 된다. 내담자의 개인정보를 삭제한 이후라 할지라도, 이러한 빅데이터를 인공지능상담가 제조회사가 사기업에 제공하게 된다면, 사기업은 비슷한 유형의 내담자를 통제하여 특정 상품을 소비하도록 조종하거나 특정 광고에 노출시킬 수 있기 때문에 일반적인 수준보다 높고 철저하게 프라이버시 보호를 강화할 방안을 모색해야 할 것이다.[22]

만약, 인공지능상담가 제조회사가 인공지능상담가의 상담능력을 향상시키기 위한 목적으로 내담자의 상담내용을 분석하고 활용하려 한다면, 그러한 대화에 담긴 고도의 사적 정보들이 상담 과정에서 드러나거나 기타 사기업에 넘어가지 않도록 암호화하거나 보호할 수 있는 기술 수준과 위반 시의 법적 책임의 범위 등을 사전에 설정해야 할 것이다. 특히, 외부로부터의 해킹에 대한 대비책 등과 같은 사적 정보 보호의 안전한 관리 및 유지에 대한 책임과 보안기술이 명확해야 한다. 또한, 상담시 취득한 내담자의 사적 정보 중에서 개인이나 사회에 심각한 위협이 될 수 있는 정보들이 있다면, 이를 인공지능상담가가 판단하고 필터링하여 필요시에는 관련 기관이나 법집행 기관과 공유할 수 있는 사회안전 시스템도 필요하다.

이 외에도 인공지능상담가가 상담의 현장에 도입되게 될 때는 다양한 이슈들이 발생할 수 있다. 예를 들어, 상담 과정에서 인간 내담자와 인공지능상담가 사이에 통제(control)와 조종(manipulation)이 발생할 때 인

22) 토요아키 니시다, "선의지를 가진 로봇을 향하여," 288-289.

간 내담자는 이를 어떻게 분별할 수 있을 것인가? 인간 내담자가 인터넷을 통하여 인공지능상담가와 24시간 동안 항상 접속이 가능한 상담 환경에서 이러한 통제와 조종이 발생했을 경우, 이를 누가 어떻게 중단시킬 수 있을 것인가? 또한, 성공적인 상담을 위하여는 적절한 종료의 시기를 결정하는 것이 중요한데, 인공지능상담가는 최적화된 상담종료의 시점을 어떻게 설정할 수 있을 것인가? 인간이 점점 더 인터넷 네트워크에서 접속 생활하는 상황이 늘어날 것이라 예상되는 가운데, 인공지능상담가와 내담자의 상담의 중단이나 종료가 과연 실제로 가능할 것인가? 이러한 질문들에 대한 해결책은 그리 간단하지 않으며, 향후 다양한 후속 연구가 진행되어야 할 것이다.

III. 인공지능 상담가의 윤리 강령과 덕윤리

인공지능상담시의 윤리적 이슈들에 대응하기 위한 방안으로 선제적이고 명시적으로 고려할 수 있는 방안은 윤리 강령(a code of ethics)을 제정하는 것이다. 예를 들어, 2020년 12월 23일에는 정부에서 과학기술정보통신부 주관으로 "사람이 중심이 되는 「인공지능(AI) 윤리기준」"을 공표하여 "인간 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙"의 3대 기본 원칙과, 이를 실현하기 위하여 "인공지능 전체 생명 주기에 걸쳐 충족되어야 하는 10가지 핵심 요건"으로서 "인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성"을 제시하였다.[23] 본 윤리 지침은 2020년 4월에 연구를 시작하여 비교적 짧은 기간인 9개월 만에 공표했지만, 내용적인 측면에서 적절한 규범과 가치를 담보하고 있다고 평가할 수 있다.

23) 과학기술정보통신부 · 정보통신정책연구원, "사람이 중심이 되는 인공지능(AI) 윤리기준," 과학기술정보통신부 (세종: 관계부처 합동, 2020), 3-5.

그럼에도 불구하고, 정부의 윤리기준이 인공지능상담가 개발이나 상담 시의 윤리 강령으로 활용되기에는 몇 가지 개선되어야 할 측면이 있다. 해당 윤리 기준은 "인공지능 기술의 개발부터 활용에 이르는 전 단계에 참여하는 모든 사회구성원을 대상으로 하며, 이는 정부·공공기관, 기업, 이용자 등을 포함"하고 있다.[24] 그러나, 이의 실현 방안은 다양한 이해관계자의 논의를 거쳐 "주체별 체크리스트 개발 등 인공지능 윤리의 실천 방안을 마련한다"는 '선언'이나 '권유'에 그침으로써 정부가 제시한 윤리 기준을 사회나 산업현장에 실현하도록 유도하거나 구체화할 방안이 희박하거나 부재하다.[25]

이는 아마도 해당 윤리기준의 서문에서 밝혔듯이, 본 윤리기준이 "산업·경제 분야의 자율규제 환경을 조성함으로써 인공지능 연구개발과 산업 성장을 제약하지 않고, 정당한 이윤을 추구하는 기업에 부당한 부담을 지우지 않는 것을 목표"라는 소극적 개입 자세를 취하고 있기 때문일 것이다.[26] 하지만, 강제력 없는 자율규제 환경이란, 현실에서는 허울좋은 요식행위에 지나지 않는 경우가 많고, "정당한 이윤을 추구하는 추구하는 기업에게 부당한 부담을 지우지 않는 것"이 정부가 제시하는 '윤리적' 목표가 되어야 하는지에 대해서는 논란의 여지가 있어 보인다. 오히려, 정부가 제시하는 윤리기준은 국내의 모든 구성원들이 인공지능기술 개발 및 활용에 있어서 반드시 고려해야만 할 실현성 있는 구체적 방안을 설정 및 제시하고, 기업이 부당한 방법으로 인공지능을 개발 및 활용하여 이윤을 추구하려는 것을 예방하는 것을 '윤리적' 목표로 할 때 보다 적절하고 사회적으로 필요한 윤리 기준으로 자리잡게 될 것이다. 예를 들어, 유럽연합의 "신뢰할만한 인공지능을 위한 윤리 가이드라인 (Ethics Guidelines

24) Ibid., 5.
25) Ibid.
26) Ibid., 2.

for Trustworthy AI)"에는 비록 강제력은 없지만 신뢰할만한 인공지능을 실현하기 위한 기술적, 비기술적 방안과 평가 항목들이 국내 방안보다 높은 수준으로 구체적으로 제시되어 있다.[27] 관계 기관의 지원이 없으면 이러한 윤리기준 자체가 현장에서 거의 사용되지 않을 것이기 때문에, 인공지능 윤리 기준의 개발 뿐만 아니라 이것을 실행할 수 있는 방안을 정부가 어떻게 마련하는지의 여부가 인공지능 윤리 기준에 대한 정부의 의지와 일관성을 가늠할 수 있는 한 요소가 될 수 있을 것이다.[28]

일반적으로, 전문가와 일반인 사이에서 지식이나 기술은 비대칭적이기 때문에, 일반인은 이해나 인식의 측면에서 취약할 수밖에 없다. 폴라 보딩턴(Paula Boddington)은 "내담자의 상대적인 인식론적 취약성이 직업 윤리의 핵심적 측면을 형성하는데 도움이 된다"고 주장하며 몇 가지 핵심 예시들을 제시하였는데, [29] 이는 전문가의 윤리적 강령이 전문가가 비대칭적 지식이나 기술을 통해 일반인을 착취하거나 조종하는 등의 행위를 방지하는 것이 되어야 함을 알려준다. 이러한 지식이나 기술의 비대칭성의 문제는 인공지능 상담가와 인간 내담자의 관계에서도 발생할 가능성

27) High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," (Brussels: European Commission, 2019), 20-31. 본 가이드라인 자체에 대한 학술적 분석과 관련 연구는 다음과 같은 논문을 참조하라. Nathalie A. Smuha, "The Eu Approach to Ethics Guidelines for Trustworthy Artificial Intelligence," Computer Law Review International 20, no. 4 (2019); Inga Ulnicane et al., "Good Governance as a Response to Discontents? Déjà Vu, or Lessons for AI from Other Emerging Technologies," Interdisciplinary Science Reviews 46, no. 1-2 (2021).

28) Paula Boddington, ***Towards a Code of Ethics for Artificial Intelligence***, (Cham: Springer International Publishing, 2017). 40, 99.

29) 폴라 보딩턴은 다음과 같은 사항들을 직업 윤리의 핵심 예시로 주장하였다: "전문적 역량 수준을 보장하고, 자신의 역량 영역 내에서만 작업하며, 기술과 지식을 적절하게 업데이트 하기; 대중과 고객을 대할 때 정직하고 투명하며, 필요에 따라 추가 조언을 받는 것을 포함하여 위험을 완전히 공개하기; 적절한 관할권의 법률과 관련 지방 또는 지방 정부의 규정내에서만 운영하기" Ibid., 41-42.

이 높다. 인공지능 상담가의 상담 기술과 지식 정보 등은 상담의 빅데이터가 쌓여갈수록 지속적으로 발전할 것이기 때문이다. 따라서, 인공지능 상담에 있어 윤리 강령을 활용하게 될 경우, 정부나 기업 등은 이러한 비대칭성을 고려한 윤리 강령을 제정하여 내담자의 취약성이 악용당하는 사례를 방지해야할 것이다. 예를 들어, 럭스턴은 인공지능 상담가의 활용을 위해 제안한 윤리 강령에서, 상황이나 임상 적용의 정도에 따라, 즉, "환자의 이상반응과 임상적 금기사항이 위험하게 될 경우를 위해, 인간의 감독과 모니터링"을 요구해야만 한다고 주장하였다.[30]

여기서, 인공지능 개발이나 활용에 윤리 강령을 제공한다는 것은, 이러한 윤리 강령이 약인공지능 개발에 활용될 가능성이 높다는 의미이다. 예를 들어, 지금까지 살펴본 정부의 윤리기준에서 인공지능의 지위는 인간 수준의 지능과 자각능력을 지닌 강인공지능(AGI) 또는 독립된 인격을 지닌 인공지능을 전제하고 있지 않다.[31] 정부의 윤리 기준이 약인공지능으로 제한된 것에는 다양한 이유가 있겠지만, 기술적인 측면에서 본다면, 강지능 또는 초지능(Superintelligence)의 등장은 영화 '아이, 로봇(I, Robot)'에서처럼 인간이 정한 윤리 규범을 언제든지 회피하여 무용지물로 만들 가능성이 있기 때문이다.

따라서, 인공지능개발 기술이 점차 발달하여 강지능에 근접하거나 도달한 공감적 인공지능상담가를 개발하게 될 경우, 이러한 윤리강령 제정으로 내담자의 취약성과 같은 윤리적 이슈를 방지하기는 어렵게 된다. 윤리학적 측면에서 윤리 강령에 내재된 공리주의나 의무론의 한계에 대한 대안으로 등장하는 것이 덕윤리이다. 또한, 덕윤리는 빠른 변화나 불

30) Luxton, "Recommendations for the Ethical Use and Design of Artificial Intelligent Care Providers," 6.

31) 과학기술정보통신부 · 정보통신정책연구원, "사람이 중심이 되는 인공지능(AI) 윤리기준," 5.

확실한 상황들로 인해 미래의 문제들에 확실한 윤리 강령을 제정하기 곤란할 때 사용되기도 한다.[32]

이러한 덕윤리에 호소하여, 웬델 월러치(Wendell Wallach)와 새넌 밸러(Shannon Vallor)는 인간 수준의 인공지능이나 초지능이 덕이나 도덕적 성품을 체현할 수 있을 때만이 안전과 유익을 보장할 수 있다고 주장하였다.[33] 도덕적(moral) 또는 덕을 함양한(virtuous) 강인공지능 또는 초지능만이 궁극적 의미에서 윤리적으로 안전한 인공지능이 될 수 있다는 것이다. 덕윤리학에서 덕이란 오랜 기간 동안 적절한 윤리적 행위를 향해 나아가는 연습과 훈련을 통해 얻어지는 견고한 습관과 기술을 의미한다.[34] 덕을 함양한 행위자만이 예기치 않은 도덕적 상황 속에서 이를 식별하고 적절히 반응할 수 있는데, 이는 "덕을 함양한 행위자가 도덕적 삶의 구조를 전체론적이고 통합적이며 풍성히 구현된 의미속에서 도덕적 삶의 구조를 이해하고, 따라서 그러한 도덕적 삶의 전체가 끊임없는 동요하는 것에 지각적이나 신체적으로 적응"될 수 있기 때문이다.[35] 월러치와 밸러는 "도덕적 이해, 도덕적 지각(perception)과 정서적 감수성(sensitivity), 도덕적 성찰(reflection), 도덕적 상상력"과 같은 "도덕적 능력들 중에 어떤 것들이라도 논리적으로나 물리적으로 기계에 내장되는 것이 불가능하다고 생각할 만한 설득력 있는 이유는 없다"고 말함으로써, 도덕적 인공지능의 등장이 불가능하지 않음을 주장하였다.[36]

윤리적이고 안전한 강지능 또는 초지능의 개발을 위한 덕윤리적 시도

32) Paul Atkinson, "Ethics and Ethnography," *Twenty-First Century Society* 4, no. 1 (2009): 26-29.

33) Wendell Wallach and Shannon Vallor, "Moral Machines," in *Ethics of Artificial Intelligence*, ed. S. Matthew Liao (New York, NY: Oxford University Press, 2020), 383.

34) Ibid., 394-95.

35) Wallach and Vallor, "Moral Machines," 397.

36) Ibid., 399.

는 공감적 인공지능 상담가가 초래할 수 있는 다양한 윤리적 이슈에 대한 보다 근본적인 시사점을 제시한다. 즉, 인공지능 상담가가 약인공지능에서 공감적 강인공지능 상담가로 발전할 것을 예상할 수 있다면, 초기 단계에서는 공리주의적이고 의무론적인 윤리 강령의 제정과 구체적 실행을 통하여 윤리적 이슈에 적절히 대응해야 하겠지만, 점차 강인공지능으로 발전하게 되면서부터는 덕윤리적 관점에서 도덕적 성품을 인공지능 상담가에 체현할 수 있는 방안을 개발할 때, 앞서 언급한 인공지능 상담시의 윤리적 이슈에 대응할 수 있는 초석을 마련할 수 있을 것이다.

여기서, 종교적인 덕은 인공지능 상담시의 덕윤리적 대안에 포함될 가능성이 있다. 미래에 공감적 인공지능 상담가는 종교에 대한 이해를 필요로 할 것이다. 예를 들어, 로버트 제라시(Robert Geraci)는 "미래에 가정 로봇을 디자인할 때 종교 생활이 필수 요소가 될 것"이라는 로봇 공학자인 데이브 투렛즈키(Dave Touretzky)의 견해를 소개한다.[37] 예를 들어, 죽음을 앞둔 기독교인 노인이 돌봄 로봇과 대화하는 중에 죽음과 구원과 영생 등과 같은 종교적인 대화를 할 때, 무신론적 로봇이 그들과 신앙적인 대화를 할 수 없거나 신앙을 비판하게 된다면 그 로봇은 더 이상 정서적 돌봄의 기능을 상실하게 될 것이다. 또한, 제임스 맥브라이드(Jame McBride) 같은 경우는 종교를 가진 로봇 소유자가 로봇과 대화할 때 당연히 종교적인 대답을 기대할 것이라고 주장하며, 미래에 로봇 분야에서는 "로봇들이 소유자의 [종교적] 가치를 따르는 종교 소프트웨어로 프로그램이 될 것이고," 소유주는 돌봄 로봇을 종교 공동체로 인도하여 종교적 신조와 예식 등을 가르치려 할 것이라고 주장하였다.[38] 한편, 데이빗 레비

37) Robert M Geraci, *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality* (Oxford University Press, 2010), 133.

38) James McBride, "Robotic Bodies and the Kairos of Humanoid Theologies," *Sophia* 58, no. 4 (2019): 4-7.

(David Levy)와 에드문드 퍼스(Edmund Furse)는 지성, 감성, 자의식과 자유의지를 지닌 보다 발전된 로봇이 종교적인 면에서 인간과 똑같을 것이라 예상한다.[39] 이들은 21세기 중반이면 로봇이 모든 인간의 감정을 갖추게 될 것이고, 지성적 로봇은 세상과 종교에 관심을 갖고 자의적으로 기독교인이 되려할 수도 있다고 주장하였다.[40]

이러한 관점들을 따르며, 송용섭은 덕윤리에 기초하여 초지능의 실존적 위협을 예방하기 위하여 종교적 인공지능을 대안의 하나로 고려해야 함을 주장하였다.[41] 종교는 도덕적 가치를 내재하고 가르치며 공동체를 통해 전달하기 때문에, 종교 공동체에서 도덕적 성품을 계발한 인공지능이라면 빠르게 변화하는 상황 속에서도, 공존해야할 인류를 위한 도덕적 선택과 행위를 일관되게 진행할 가능성이 높기 때문이다. 그는 기독교적 인공지능을 사례로 들어 기독교의 '자기희생적 사랑(Agape)'의 가치가 인공지능에 체현될 수 있을 때, 많은 학자들이 우려하는 초지능의 위협에서 벗어날 가능성이 생길 수 있다고 주장하였다.[42]

따라서, 이들의 주장을 근거로 유추해보면, 다양한 종교인 내담자들과 상담해야할 공감적 인공지능 상담가 역시 상담적 필요에 의하여 종교적 지식이나 신념을 프로그램으로 내재하거나, 강인공지능으로 발전한 후에는 특정 종교를 선택하여 믿는 종교적 인공지능이 되어 종교적 상담을 진행할 가능성이 있다. 이때, 종교적인 덕을 올바르고 포용력 있게 소유한 인공지능 상담가는 비종교적 인공지능 상담가보다 종교적 내담자에

39) Yong Sup Song, "Religious AI as an Option to the Risks of Superintelligence: A Protestant Theological Perspective," ***Theology and Science 19,*** no.1 (2021): 70-71.

40) David Levy, ***Robots Unlimited: Life in a Virtual Age*** (Boca Raton, FL: CRC Press, 2005), 316, 87-88.

41) Song, "Religious AI as an Option to the Risks of Superintelligence: A Protestant Theological Perspective," 74-75.

42) Ibid,

대한 보다 폭넓은 이해와 공감적 상담을 진행할 가능성이 높으리라 예상할 수 있다.

이러한 미래 상황을 한국 교회에 적용시켜 생각한다면, 한국 교회가 공감적 인공지능 상담가를 초기 단계에서부터 신앙 공동체 내에서 양육하고 성서의 주요 덕을 학습시키는 방안을 고려할 수 있다. 성서에서 제시하는 "사랑, 희랑, 화평, 인내, 자비, 충성, 온유, 절제"와 같은 성령의 열매(갈5:22-23)나, "믿음, 소망, 사랑" (고전 13:13)과 같은 신학적 덕은 초대 교회에서 체계적으로 구성되지는 않았지만, 전형적인 덕으로 인식되어 왔다.[43] 따라서, 이러한 신학적 덕들은 기독교적 인공지능상담가가 신앙 공동체가 교회를 통해 지속적으로 학습하고 훈련함으로써 형성할 수 있는 기독교의 대표적인 덕이라 할 수 있을 것이다.

미래의 인공지능이 덕을 학습하고 훈련하여 내면화하는 과정에서 공동체 또는 사회 문화의 영향에서 벗어날 수 없다. 따라서, 국내 기독교인을 대상으로 상담을 진행하게될 종교적(기독교적) 인공지능 상담가에게는 무엇보다 한국 교회의 관심과 양육 노력이 필요할 것이다. 특히, 한국 교회 안에서 공감적 기독교 인공지능 상담가가 자기희생적 사랑(아가페)을 덕으로 형성할 수 있다면, 앞서 제기한 인공지능 상담시의 다양한 윤리적 문제 상황에서 국내 기독교인 내담자를 위한 보다 책임있고 윤리적인 상담을 진행할 수 있을 것이다.

IV. 나가는 말

본 논문은 공감적 인공지능상담가의 등장이 초래할 수 있는 윤리적 이슈들을 분석하고 덕윤리적 입장에서 대안을 제시하였다. 본 논문은, 첫째

43) 문시영, "하우어워스와 '덕의 공동체'로서의 교회," 「기독교 사회윤리」 23 (2012), 171-173.

로, 공감적 인공지능 상담가와 인간 내담자 사이에 전이 또는 역전이가 발생할 경우, 심리 및 행동 통제의 이슈가 발생할 수 있음을 지적하였다. 둘째로, 인공지능상담가와 인간 내담자 사이에 문제가 발생했을 경우, 윤리적, 법적 책임 소재의 여부가 논쟁이 될 수 있으므로, 이를 판단하기 위한 선제적 연구와 기준을 마련할 것을 주장하였다. 셋째로, 인공지능 상담시에는 내담자의 민감한 사적 정보가 기록 및 분석될 수 있으므로 일반적인 수준보다 높고 철저하게 프라이버시 보호를 강화할 방안을 모색해야 함을 주장하였다.

윤리적 이슈들에 대응하는 방법들은 주로 윤리 강령을 제정하는 것이고 국내에서도 '사람이 중심이 되는 인공지능 윤리기준'을 마련하였지만, 이러한 윤리 강령들에는 제한점이 많고 인공지능 개발에 있어서는 초기 단계에서만 유효할 가능성이 높다. 따라서, 본 논문은 인공지능 상담가의 개발과 이에 따른 윤리적 이슈를 예방하기 위하여 윤리 강령의 제정과 실행에 그치지 말고, 궁극적으로 덕윤리적 관점을 채용하여 도덕적 또는 덕을 함양한 공감적 인공지능 개발을 지향할 것을 주장하였다. 이때, 본 논문은 종교적인 덕을 포함하는 공감적 인공지능이 더욱 효과적이고 윤리적으로 다양한 문제 상황에 대응할 수 있을 것으로 추정하였다.

본 논문은 인공지능상담가의 인간내담자의 심리 통제나 행동 조종과 같은 문제에 대해 보다 구체적으로 윤리적, 법률적 책임 소재 여부를 선제적으로 준비하게 함으로써 향후 발생할 수 있는 불필요한 사회적 비용을 줄이고, 관련분야의 학자들이 인공지능상담가의 개발 방향과 후속 연구 지침으로 활용될 수 있을 것이다. 하지만, 본 논문이 다루지 못한 인공지능상담의 윤리적 이슈들이 아직 남아있는 만큼, 이를 보다 심도싶게 분석하고 대안을 제시하기 위하여는 윤리학, 상담학, 인지과학, 인공지능, 로봇공학과 같은 관련 분야의 학제간의 대화와 연구가 필요할 것이다.

또한, 이러한 공동연구는 학술분야에 그치는 것이 아니라, 실생활에 적용되어 인공지능상담시 발생할 수 있는 윤리적 이슈들에 대한 사회적 합의와 이의 제도화 방안을 도출할 수 있을 만큼 구체적이고 실현 가능한 방향으로 나아가야 할 것이다. 이때, 다양한 분야의 학제간 연구가 실현되기 위한 사회적 합의를 위하여서는, 많은 숙고와 수정 과정 및 충분한 논의 시간이 필요하다. 다만, 기존의 공리주의적이고 규범적 접근 방법은 내담자의 권리보호에 한계를 지니고 있으므로, 덕윤리 체계의 깊은 성찰과 적용이 공감적 인공지능 상담시의 윤리적 이슈들에 대한 연구에 여전히 필요할 것으로 사료된다.

참고문헌

과학기술정보통신부 · 정보통신정책연구원. “사람이 중심이 되는 인공지능(AI) 윤리기준.” 과학기술정보통신부, 1-5. 세종: 관계부처 합동, 2020.

니시다, 토요아키. “선의지를 가진 로봇을 향하여.”『로봇윤리』, edited by 라파엘 카푸로, 미카엘 나겐보르그/ 변순용 · 송선영 옮김/ 변순용 · 송선영 옮김. 서울: 어문학사, 2013.

문시영. “하우어워스와 ‘덕의 공동체’로서의 교회,” 「기독교 사회윤리」 23 (2012), 159-186.

아사로, 피터.『로봇윤리』. edited by Rafael Capurro, Michael Nagenborg and 변순용송선영 옮김. 서울: 어문학사, 2013.

윌러치, 웬델, 콜린 앨렌./ 노태복 옮김.『왜 로봇의 도덕인가』. 서울: 메디치, 2014.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. “Thinking inside the Box: Controlling and Using an Oracle AI.” *Minds and Machines* 22, no. 4 (2012): 299-324.

Atkinson, Paul. “Ethics and Ethnography.” *Twenty-First Century Society* 4, no. 1 (2009): 17-30.

Boddington, Paula. *Towards a Code of Ethics for Artificial Intelligence*. Cham: Springer International Publishing, 2017.

Bostrom, Nick. *Superintelligence : Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

_____. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.” *Minds and Machines* 22, no. 2 (2012): 71-85.

Geraci, Robert M. *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. Oxford University Press, 2010.

High-Level Expert Group on Artificial Intelligence. “Ethics Guidelines for Trustworthy AI.” 1-39. Brussels: European Commission, 2019.

Levy, David. *Robots Unlimited: Life in a Virtual Age*. Boca Raton, FL: CRC Press, 2005.

Luxton, David D. "Recommendations for the Ethical Use and Design of Artificial Intelligent Care Providers." *Artificial Intelligence in Medicine* 62, no. 1 (2014): 1-10.

MacIntyre, Alasdair C. *After Virtue*. 3rd ed. Notre Dame, IN: University of Notre Dame Press, 2007.

Mann, David. *Erotic Transference and Countertransference: Clinical Practice in Psychotherapy*. New York: Psychology Press, 1999.

_____. *Psychotherapy, an Erotic Relationship: Transference and Countertransference Passions*. New York: Psychology Press, 1997.

McBride, James. "Robotic Bodies and the Kairos of Humanoid Theologies." *Sophia 58*, no. 4 (2019): 663-676..

Smuha, Nathalie A. "The Eu Approach to Ethics Guidelines for Trustworthy Artificial Intelligence." *Computer Law Review International 20*, no. 4 (2019): 97-106.

Song, Yong Sup. "Religious AI as an Option to the Risks of Superintelligence: A Protestant Theological Perspective." *Theology and Science 19*, no.1 (2021): 65-78.

Ulnicane, Inga, Damian Okaibedi Eke, William Knight, George Ogoh, and Bernd Carsten Stahl. "Good Governance as a Response to Discontents? Déjà Vu, or Lessons for AI from Other Emerging Technologies." Interdisciplinary Science Reviews 46, no. 1-2 (2021): 71-93.

Wallach, Wendell and Shannon Vallor. "Moral Machines." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 383-412. New York, NY: Oxford University Press, 2020.

•온라인 자료

손인해. "집주소계좌정보 '술술'…'AI 이루다' 개인정보 유출 논란." 뉴스1, https://www.news1.kr/articles/?4177218. Accessed 2021.01.30.

이소라. "AI 챗봇 '이루다' 서비스 잠정 중단…'개선 후 찾아올 것'." https://www.hankookilbo.com/News/Read/A2021011121500001991. Accessed 2021.01.30.

이효석. "AI 이루다, 논란 일주일만에 사실상 종료 '중추신경계 폐기'." 연합뉴스,

https://www.yna.co.kr/view/AKR20210115062352017. Accessed 2021.01.30.

최준선. "'진짜 사람같은' 스무살 여대생과 연애… 성착취도." 헤럴드, http://biz.heraldcorp.com/view.php?ud=20210108000325. Accessed 2021.01.30.

논문투고일: 2021년 03월 13일
심사개시일: 2021년 03월 16일
게재확정일: 2021년 04월 06일

• 국 문 초 록 •

본 논문은 공감적 인공지능상담가의 등장이 초래할 수 있는 윤리적 이슈들을 분석하고 덕윤리적 입장에서 대안을 제시하려 한다. 첫째, 본 논문은 공감적 인공지능 상담가와 인간 내담자 사이에 전이 또는 역전이가 발생할 경우 심리 및 행동 통제의 이슈가 발생할 수 있음을 지적한다. 둘째, 인공지능상담가와 인간 내담자의 상담시에 문제가 발생했을 경우, 윤리적, 법적 책임 소재의 여부가 논쟁이 될 수 있으므로, 본 논문은 이를 판단하기 위한 선제적 연구와 기준을 마련할 것을 주장한다. 셋째로, 본 논문은 인공지능 상담시에는 내담자의 민감한 사적 정보가 기록 및 분석될 수 있으므로, 일반적인 수준보다 높고 철저하게 프라이버시 보호를 강화할 방안이 마련되어야 함을 주장한다. 윤리적 이슈들에 대응하는 방법들은 주로 윤리 강령을 제정하는 것이며 국내에서도 '사람이 중심이 되는 인공지능 윤리기준'을 마련하였지만, 이러한 윤리 강령들에는 제한점이 많고 인공지능 개발에 있어서는 초기 단계에서만 유효할 가능성이 높다. 따라서, 본 논문은 인공지능 상담가의 개발과 이에 따른 윤리적 이슈를 예방하기 위해서는 윤리 강령의 제정과 실행에 그치지 말고, 덕윤리적 관점을 채용하여 도덕적 또는 덕을 함양한 공감적 인공지능 개발을 지향할 것을 주장한다. 이때, 본 논문은 종교적인 덕을 포함하는 공감적 인공지능이 더욱 효과적이고 윤리적으로 다양한 문제 상황에 대응할 수 있을 것으로 추정한다.

주제어: 인공지능, 인공지능 상담가, 덕윤리, 종교적 인공지능, 웬델 월러치