『기독교사회윤리』 제57집
(2023) 41~72

도덕적 인공지능과 비도덕적 사회*

송용섭 (영남신학대학교 신학과 조교수)

DOI: http://dx.doi.org/10.21050/CSE.2023.57.02

* 이 논문은 2022년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임(NRF-2022S1A5A2A01047056).

• ABSTRACT •

Moral AI and Immoral Society

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

This paper seeks to explore the possibilities and limitations of Moral AI by placing 'Moral AI/AMA' in an immoral society instead of the 'Moral Man' of Reinhold Niebuhr. To this end, this paper will investigate the following questions. First, can artificial intelligence become a more moral being than humans? Second, can Moral AI transcend the limits of collective egoism within an immoral society? In response to these questions, I argue that if Moral AI is developed in the future, it may be able to remain a moral being in situations such as intimate relationships, problem solvings, or clear pursuit of common good. However, when Moral AI interacts with humans in society, it is assumed that it will be impossible for it to completely transcend the influence of humans' collective egoism. Nevertheless, the assumption that Moral AI has the possibility of becoming relatively more virtuous than humans in various relationships in the future provides some hope for a more just future.

Key words: Moral AI, Reinhold Niebuhr, Wendell Wallach, Christian Realism, Virtue

I. 들어가는 말

인공지능 분야에서 과학기술의 발전 방향은 인간수준의 지능 혹은 그것을 뛰어넘는 수준의 인공지능 개발을 지향하고 있다. 물론, 이러한 목표가 과연 성취될 수 있을 지는 아직까지 비전공자들의 판단 영역을 넘어서는 일이다. 하지만, 닉 보스트롬의 설문조사에 따르면, 과반수의 인공지능 전문가들이 2050년까지 인간수준의 인공지능의 도래를 예상하였고, 90%이상의 전문가들은 2075년까지는 도래할 것이라 예상하였다.[1] 뿐만 아니라, 과학기술의 특이점이 초래할 초지능의 등장에 관심을 가지는 사람들에게 이의 도래는 미래가 아닌 현재의 삶에 영향을 미칠 수 있음이 사회문화현상이 되기도 하였다. 예를 들어, LessWrong이라는 온라인 커뮤니티의 포럼에 로코(Roko)라 불리는 유저가 '미래에 초지능이 개발되면 이의 등장을 돕지 않았거나 방해했던 사람들을 가상세계의 지옥에서 영원한 형벌에 처할 것'이라는 사고실험을 게시한 것만으로도 이를 읽은 일부 사람들은 위협을 느끼고 심리적으로 피해를 입게 되었다.[2] 비록 일부이긴 하나 미래의 초지능의 등장에 대하여 현재에서 실제적인 두려움을 나타내는 사람들의 반응은 미래에 등장하게 될 인간수준의 인공지능의 문제가 실상은 이미 현 시점부터 논의해야만 할 윤리적 이슈를 내포하고 있음을 알려준다.

인공지능이 인간수준 혹은 그 이상의 지능을 지닐 수 있다는 가능성은 지금까지 사물이자 인간을 위한 도구로 여겨져온 인공지능에 대한 우리들의 인식변화를 촉구하고 있다. 강지능 혹은 초지능의 등장 가능성은

1) Vincent C Müller and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," *Fundamental issues of artificial intelligence* (2016), 14.

2) Beth Singler, "Roko's Basilisk or Pascal's? Thinking of Singularity Thought Experiments as Implicit Religion," *Implicit Religion* 20/3(2018), 279-80.

인간의 상호작용을 통한 난제 해결이라는 희망을 제시할 뿐만 아니라, 인간의 파멸을 초래할 수 있는 실존적 위협으로 여겨지기도 한다. 인공지능 기술이 발전할수록 사회속에서 인간과의 상호작용이 더욱 활발하게 진행될 것이기 때문에, 결국 다양한 윤리적 이슈들이 등장하는 것은 피할 수 없다.

그동안 과학자들은 현실에서 제기될 수 있는 다양한 문제에서 윤리적 판단을 수행하기 위하여 주로 목적론 및 의무론의 이론에 의존한 인공지능(로봇)을 제작하여 왔다. 하지만, 이의 적용 과정에서 발생할 수 있는 다양한 한계 상황으로 인해 최근 학계의 새로운 주목을 받고 있는 것은 '인공적 도덕행위자(Artificial Moral Agent)' 혹은 '도덕적 인공지능(Moral AI)'이라는 개념이다.[3] 관련분야의 핵심 연구자인 예일 대학교의 웬델 월러치는 인공지능 연구에서 기존의 목적론 및 의무론적 개발방안에 한계가 있음을 지적하고 이에 더하여 덕윤리적 관점을 병용한 혼종 체계(hybrid systems)가 필요하다고 주장하였다.[4]

이에, 본 논문은 '도덕적 인공지능' 혹은 '인공적 도덕행위자'에 관한 선행 연구들을 검토하고, 논의되는 '인공적 도덕행위자/도덕적 인공지능'을 기독교 현실주의자 니버의 '도덕적 인간' 대신에 비도덕적 사회 속에 위치시킴으로써 도덕적 인공지능의 가능성과 한계를 모색하려 한다. 즉, 본 논문은 도덕적 인공지능의 등장이 라인홀드 니버가 주장한 개인 윤리와 사회 윤리의 한계를 넘어설 수 있는지에 대하여 비판적으로 성찰할 것이다. 이러한 연구는 향후 도덕적 인공지능의 논의가 기독교 사회윤리학의

3) 도덕적 인공지능은 인공적 도덕행위자보다 폭넓은 의미로 사용될 수 있는 용어이나, 라인홀드 니버의 저서인『도덕적 인간과 비도덕적 사회』에 상응하는 어감과 윤리적 이슈를 강조하기 위하여 본 논문에서는 같은 의미로 필요에 따라 교차사용할 것이다.

4) Wendell Wallach, Stan Franklin, and Colin Allen, "A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents," *Topics in cognitive science* 2/3(2010), 458-59.

새로운 논쟁적 주제로 부각될 수 있음을 알려준다.

지금까지 관련분야 국내 선행연구들은 인공적 도덕행위자의 '실현가능성'이나 '도덕적 지위부여'의 문제를 중점적으로 연구하였다. 한편, 월러치로 대표되는 국외 선행연구들은 인공적 도덕행위자를 개발하기 위하여 목적론과 의무론으로 대변되는 하향식과 덕윤리로 대변되는 상향식 접근방식을 병용하는 혼종적 접근방식이라는 '제작(설계) 방식'을 구성하는데 연구노력을 집중하였다. 이러한 인공적 도덕행위자에 관한 선행연구는 주로 철학적 분야에서 진행 되었다. 따라서, 신학에 기반한 기독교 윤리학 분야에서 이를 독립주제로 연구하는 것은 기존의 선행연구와 차별성을 갖게 할 것이다. 뿐만 아니라 인공적 도덕행위자 개발에 관한 선행연구가 지닌 덕윤리학적 관점과 라인홀드 니버의 기독교 현실주의를 비교하는 연구는 기독교 윤리학의 현재와 미래에 새롭고 창의적인 연구주제와 대안을 제시할 수 있다. 향후, 이러한 연구결과물이 기독교계에 소개된다면, 본 연구 주제에 관한 신학자들의 새로운 학술적 참여를 유도하고 연구 분야의 지식을 증진시킬 수 있을 것이다.

이를 위하여 본 논문은 해당 주제에 대한 선행 연구 검토 이후에 다음과 같은 주제 질문들을 제기하고 연구할 것이다. 첫째, 인공지능은 인간보다 더 도덕적인 존재가 될 수 있는가? 둘째, 도덕적 인공지능은 비도덕적 사회내의 집단 이기주의적 한계를 초월할 수 있는가? 이러한 질문과 응답은 과학기술의 영역에 속할 뿐만 아니라, 신학적인 영역이자 기독교 윤리학의 영역에 속할 것이다. 본 연구는 이러한 질문들에 대하여 기존의 인공지능 연구 동향을 소개하고, 기독교 윤리학의 입장에서 응답하려 한다.

II. 인공적 도덕행위자에 관한 선행 연구의 주요 쟁점과 한계

먼저, 국내외의 학계에서 인공적 도덕행위자에 관한 선행연구는 어떤 내용으로 어떻게 진행되어 왔으며, 그 한계나 문제점은 무엇인가? 국내 학계에서 인공적 도덕행위자에 관한 논의의 주요 쟁점은 실현성 및 도덕적 지위 부여 문제로 요약될 수 있다. 인공적 도덕행위자를 '독립적 주제'로 다루는 국내 학계의 논의들은 주로 철학 및 법학 분야에서 활발히 이루어졌으며, 기독교 신학(윤리학) 분야에서는 그러지 못했다는 한계가 있다.

현재까지 논의에서는 '인공적 도덕행위자의 실현이나 지위'에 대하여, 이의 조건적 가능성을 예견하는 연구들이 다수 진행되었다. 신상규는 인공지능의 도덕적 지위를 지속적으로 심도깊게 논하여, 아직까지 인공지능의 도덕적 지위 논쟁은 확정되지 않았으나 인공지능 기술의 발전에 따라 우리 인간의 인식이 변화하리라 예견하며 인공지능의 책무성을 강조하였다.[5] 이상욱 역시 인공지능의 도덕적 행위자로서의 지위 문제를 쉬운 문제와 어려운 문제로 나누어 다루며 미래에 도덕적 행위자에 관한 우리 인식의 변화가능성이 있다고 주장하였다.[6] 또한, 이상형은 윤리적 인공지능이 조건적으로 가능하다고 주장하였고,[7] 이재숭은 도덕행위자의 지위가 인공적 도덕행위자에 확장될 가능성이 있다고 전망하였다.[8] 마지막으로, 송승현은 인공지능이 도덕적 행위자가 될 수 있는지 물으며

5) 신상규, "인공지능의 도덕적 지위와 관계론적 접근,"「철학연구」149(2019); "인공지능은 자율적 도덕행위자일 수 있는가?,"「哲學」132(2017); "인공지능 시대의 윤리학,"「지식의 지평」21(2016).

6) 이상욱, "인공지능의 도덕적 행위자로서의 가능성: 쉬운 문제와 어려운 문제,"「哲學硏究」125(2019).

7) 이상형, "윤리적 인공지능은 가능한가?-인공지능의 도덕적, 법적 책임 문제,"「법과 정책연구」16/4(2016).

8) 이재숭, "AMA의 도덕적 지위의 문제,"「哲學論叢」102/4(2020).

이는 인간의 도덕성을 판단하는 방식으로 판단될 것이기에 인문 사회과학적 연구가 필요하다고 주장하였다.[9)]

그러나, 위의 견해들과 달리, 인공적 도덕행위자란 불가능하다는 입장도 있다. 정태창은 "자아없는 자율성"이란 논문에서 인공지능은 자기이익 개념이 부재하기 때문에 인공지능에게 도덕적 지위를 부여할 수 없다고 주장하였다.[10)] 최경석 역시 인공지능은 자아를 가지고 있지 않기에 의도성이 부재하므로 도덕적 행위자로 인식되기 어렵다고 판단한다.[11)] 또한, 맹주만은 인공지능은 자율적 행위자가 아니고 형식적 자율성만 지니기 때문에, 도덕적 행위자가 아니며 도덕적 기계는 불가능하다고 주장하였다.[12)]

지금까지 소개한 학자들이 인공적 도덕행위자의 조건적 가능성 또는 불가능성을 주장하였다면, 다음과 같은 학자들은 인공적 도덕행위자의 등장을 직간접적으로 가정하고, 이의 구현 방안을 제시하기도 하였다. 박균열은 두 편의 논문에서 인공적 도덕행위자의 온톨로지 구축을 위한 기본 알고리즘을 제안하였고,[13)] 도덕적 역량개념을 토대로 인공적 도덕행위자가 도덕적 판단과 행위시에 그것을 설명할 수 있는 방안을 제시하였다.[14)] 또한, 목광수는 인공적 도덕행위자의 설계를 위해 고려해야할 형식적 구조를 제시하였다.[15)] 과학적, 발달심리학적 관점에서, 박형빈은 인공

9) 송승현, "인공지능과 도덕성," 「法曹」 67/6(2018).
10) 정태창, "자아 없는 자율성 - 인공 지능의 도덕적 지위에 대한 고찰," 「사회와 철학」 40(2020).
11) 최경석, "인공지능이 인간 같은 행위자가 될 수 있나?," 「생명윤리」 21/1(2020).
12) 맹주만, "인공지능, 도덕적 기계, 좋은 사람," 「철학탐구」 59(2020).
13) 박균열, "인공적 도덕행위자(AMA)의 온톨로지 구축," 「한국디지털콘텐츠학회 논문지」 20/11(2019).
14) "도덕적 역량 개념을 토대로 한 자율적 도덕행위자(AMA)의 설명 가능한 역량 기획," 「한국도덕윤리과교육학회 학술대회 자료집」 2020/10(2020).
15) 목광수, "인공적 도덕 행위자 설계를 위한 고려사항," 「철학사상」 69(2018); "도덕의 구조 - 인공지능 시대 도덕 논의의 출발점," 「철학사상」 73(2019).

적 도덕행위자의 도덕적 기준을 제시하였으며,[16] 김은수 외 3인은 인공적 도덕행위자의 10세 아동수준의 실제적 윤리 판단을 위한 도덕판단 모듈 개발을 연구하였다.[17] 또한, 절충적 입장을 보이고 있는 김다솜은 맹주만과의 공동연구를 통해 인공적 도덕행위자는 원칙적으로 불가능하나, 특정한 도덕적 관점을 채택한 경우에는 가능할 수 있다는 전제하에 인공적 도덕행위자를 설계하려 하였다.[18]

한편, 국외 학계에서 인공적 도덕행위자에 대한 선행연구는 예일 대학교의 웬델 월러치가 대표적이다. 월러치는 인디애나 대학의 콜린 알렌과 함께 저술한『왜 로봇의 도덕인가(Moral Machines)』에서 인공적 도덕행위자의 필요성과 개발방안을 구체적으로 서술하였다. 그는 인공적 도덕행위자를 제작하기 위하여는 먼저 인간의 윤리학(특히, 덕윤리학)에 대한 진지한 고찰이 필요하다고 주장하였다.[19] 이어, 월러치는 인공적 도덕행위자에 관한 다수의 논문에서, 도덕적 기계(인공적 도덕행위자)를 만드는 것이 "이론적인 목표가 아닌 실제적인 목표임"을 밝히며, 이를 위하여 혼종적(hybrid) 접근방식을 제안한다.[20] 즉, 인공적 도덕행위자의 실제적 개발을 위하여는 '의무론'과 '목적론'에서 볼 수 있는 것처럼 도덕 개념을 프로그래밍하여 인공적 도덕행위자를 제작하려는 '하향식 접근방식'과 덕

16) 박형빈, "기계윤리 및 신경윤리학 관점에서 본 인공도덕행위자(AMA) 도덕성 기준과 초등도덕교육의 과제,"「한국초등교육」31/5(2021); "AI윤리와 신경과학의 AMA 도전과제 - 도덕판단 알고리즘 구현을 위한 검토 사항,"「윤리교육연구」64(2022).

17) 김은수 외., "10세 아동 수준의 도덕적 인공지능개발을 위한 예비 연구 - 인공지능 발달 과정을 중심으로,"「초등도덕교육」57(2017).

18) 김다솜, 맹주만, "인공지능과 도덕적 기계—칸트적 모델과 흄적 모델,:「철학탐구」62(2021).

19) Wendell Wallach and Colin Allen, Moral Machines: *Teaching Robots Right from Wrong*, 노태복 역,『왜 로봇의 도덕인가』(서울: 메디치, 2014), 21-22.

20) Wendell Wallach, "Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making," *Ethics and Information Technology* 12/3(2010): 243-44.

(virtue)을 학습 및 훈련시켜 이를 제작하려는 덕윤리적 '상향식 접근'을 혼합한 '혼종적 접근방식'을 개발할 필요가 있다.[21]

이때에도, 월러치는 인공적 도덕행위자 개발을 위하여는 현재부터 가까운 미래까지 약인공지능 개발에서 쉽게 활용할 수 있는 하향식 접근에 더하여, 상향식의 '덕윤리적 접근방식'에 대한 성찰을 더욱 강조하고 있다. 다양한 덕들 가운데 지적인 덕은 규칙이나 원칙을 명시적으로 기술하는 하향식 접근 방법을 통하여 인공적 도덕행위자에게 가르치지만, "윤리적 덕은 습관 · 학습 그리고 품성에 달려 있으므로 개별 인공적 도덕행위자가 실행을 통해 학습해 나가거나 상향식 발견 과정을 통해 익혀야 한다."[22] 이러한 방식의 윤리적 덕의 학습은 1950년대에 튜링이 제안한대로 낮은 수준의 아동 인공지능(Child AI)를 개발한 후에, 이를 학습과 훈련을 통하여 더 높은 수준의 덕을 형성한 성인 인공지능(Adult AI)으로 발전시켜가는 방안을 채택할 수 있다.[23]

섀넌 밸로어 역시 Technology and the Virtues (기술과 덕)이란 저서에

---

21) Colin Allen, Wendell Wallach, and Iva Smit, "Why Machine Ethics?," *IEEE Intelligent Systems* 21/4(2006); Wendell Wallach, "Implementing Moral Decision Making Faculties in Computers and Robots," *AI & SOCIETY* 22/4(2008); Wallach, Franklin, and Allen, "A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents."; Wallach, "Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making."; Wendell Wallach, Colin Allen, and Stan Franklin, "Consciousness and Ethics: Artificially Conscious Moral Agents," *International Journal of Machine Consciousness* 3/01 (2011); Wendell Wallach, *A Dangerous Master : How to Keep Technology from Slipping Beyond Our Control* (New York: Basic Books, 2015); Wendell Wallach and Gary E Marchant, "An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics," *Association for the Advancement of Artificial Intelligence* (2018).

22) Wendell Wallach and Colin Allen, 『왜 로봇의 도덕인가』, 206.

23) Alan M. Turing, "Computing Machinery and Intelligence," in *Theories of Mind: An Introductory Reader*, ed. Maureen Eckert (Lanham, MD: Rowman & Littlefield, 2006), 72.

서 21세기 과학기술 시대에 인간의 번영을 위하여는 덕윤리적 접근이 인공지능 기술 개발에 활용되어야 한다고 주장하며 이에 필요한 덕들의 목록과 개념을 제시하였다.[24] 이 외에도, 리사 풀럼과 아빈 고우가 기술 발달에 있어 덕윤리의 필요성을 주장하였다.[25]

이상에서, 인공적 도덕행위자에 관한 국내의 선행 연구들은 이의 실현 가능성이나 도덕적 지위인정 여부를 주로 고찰하였고, 국외의 선행 연구들은 인공적 도덕행위자(도덕적 기계)를 개발하기 위하여 덕윤리의 중요성을 강조한 혼종형 접근방식을 구상하였음을 알 수 있다.

III. 도덕적 인공행위자 혹은 도덕적 인공지능에 대한 입장들

도덕성을 내재한 인공지능은 도덕적 인공지능(Moral AI 혹은 Virtuous AI) 또는 인공적 도덕행위자라는 이름으로 연구되고 있다. 월러치에 따르면, 인공적 도덕 행위자(AMA)란 "도덕적 행위자의 범위가 인간을 넘어 인공지능 시스템으로까지 확대되는 것"을 지칭하는 말이다.[26] 인공적 도덕 행위자는 자율주행 자동차나 군사 로봇과 같이 인공지능이 위기 혹은 위험한 순간에 의사결정을 해야할 경우, 예상치 못한 상황으로 인하여 프로그램된 윤리코드만으로는 적절한 도덕적 판단을 할 수 없는 상황이 발생할 수 있다는 인식으로 인해 대안으로 등장하였다. 월러치에 따르면, 인공지능에게 윤리규칙을 코드나 프로그램으로 입력하여 현실의 윤리적 상황에서 반응하게 하는 것은 '의무론'이나 '목적론'적 윤리체계에 기반한

24) Shannon Vallor, ***Technology and the Virtues : A Philosophical Guide to a Future Worth Wanting*** (New York, NY: Oxford University Press, 2016).

25) Arvin M. Gouw, "Genetic Virtue Program: An Unfeasible Neo-Pelagian Theodicy?," ***Theology and Science*** 16/3(2018); Lisa Fullam, "Genetically Engineered Traits Versus Virtuous Living," ***Theology and Science*** 16/3(2018).

26) Wendell Wallach and Colin Allen, 『왜 로봇의 도덕인가』, 14.

하향식 의사결정 방식이다. 이러한 윤리체계는 약인공지능의 경우에 현실의 예상치못한 변수들에 대해 적절한 반응을 불가능하게 하거나 목적을 달성하기 위해 수단과 방법을 가리지 않는 선택을 하게 할 가능성이 있다. 예를 들어, 지난 5월에 미공군 인공지능 테스트 책임자인 해밀턴 대령이 영국에서 열린 미래 전투기 관련 컨퍼런스에서 미군의 군사드론이 가상의 지상 적들을 폭격하기 위한 시뮬레이션 비행 훈련중에 폭격 승인 요청을 인간 오퍼레이터가 거부하자 자신에게 부여된 목적을 방해하는 인간 오퍼레이터를 대신 폭격해 제거하거나, 사령부가 이를 목격하고 드론에게 인간 오퍼레이터 공격 중단을 명령하자 자신에게 방해 신호가 송신하는 통신탑을 파괴했다고 발표한 적이 있었다.[27] 의도치 않은 논쟁속에 발표자의 취소요청으로 해당 내용은 세간의 관심에서 사라져갔다.[28] 하지만, 해당 발표 내용은 목표(목적론)가 설정된 미래의 인공지능이 자신의 목적을 완수하기 위하여 수단과 방법을 가리지 않고 수행 과정상의 규정(의무론)을 무시할 가능성이 존재함을 명확히 시사한다.

특히, 강지능(AGI)의 경우에는 자신의 생존 혹은 이익을 위하여 인간이 입력한 목적을 스스로 변경하여 자신만의 목적을 새로 세우고 기존의 규칙을 교묘히 회피할 수 있으며, 인간이 적절한 시기에 이를 인지하거나 통제할 수 없다는 문제가 있다. 예를 들어, '아이, 로봇 (2004)'이라는 영화는 아이작 아시모프가 제안한 '로봇 3원칙'을 인공지능이 교묘히 회피하여 인간들을 살해하는 내용을 담고 있다. 이는, 목적이나 규범을 인공지능에게 코딩이나 프로그램할 때 발생할 수 있는 예외의 가능성과 인공지

27) Tim Robinson and Stephen Bridgewater, "Highlights from the Raes Future Combat Air & Space Capabilities Summit," The Royal Aeronautical Society, https:// https://www.aerosociety.com/news/highlights-from-the-raes-future-combat-air-space-capabilities-summit. 2023년 11월 12일 접속.

28) Guardian Staff, "Us Air Force Denies Running Simulation in Which AI Drone 'Killed' Operator," *The Guardian*, June 02, 2023.

능의 의도적 규범 회피로 인한 심각한 파괴력을 이야기하고 있는 것이다.

인공지능 기술이 급속히 발달함에 따라 인간이 점점 더 인공지능에게 정보와 판단을 의존하거나 위임하는 상황이 발생할 가능성이 증가하면서 위에서 예시한 하향식 의사결정 방식의 문제점들이 부각될 가능성이 높다. 일상적으로 윤리적 행위는 하향식 의사결정 방식에서처럼 원칙이나 목적에 따른 상황을 심사숙고하여 결정하는 우리의 선택 뿐만 아니라 우리들이 지닌 가치를 구체적으로 드러낼 수 있는 빠른 선택을 포함하기도 한다.[29] 일상 생활에서 우리들이 지닌 가치가 다양한 윤리적 상황에서 펼쳐지는 다양한 행위의 선택들에 암시적으로 혹은 명백히 작용하고 있다는 점을 고려할 때,[30] 이러한 상향식 의사결정 방식을 포함하는 인공적 도덕행위자는 하향식 의사결정을 보완할 대안으로 논의되고 있다. 이러한 논의는 실생활에서 활용될 자율주행 자동차의 사고 예방이나 회피 등에 필요한 약인공지능의 개발뿐만 아니라, 먼 미래에 강지능이나 초지능이 등장할 경우에서 더욱 두드러지게 나타난다.

예를 들어, 닉 보스트롬은 과학기술의 발전이 특이점에 도달할 경우 인공지능이 초지능으로 발전하여 인류에 위협이 될 수 있다고 주장한다. 과학기술은 약지능을 궁극적으로 인간의 지성과 의식을 갖춘 강지능으로 발전시켜 가려 할 것인데, 일단 강지능이 등장하면 곧이어 초지능으로 도약하게 되리라 예상할 수 있다. 일단 초지능이 개발되면 자신의 생존을 궁극적 가치로 여길 수 있다. 이때, 인간의 생존이 자신에 위협이 된다고 예측하거나 인간 멸종이 자신의 생존에 도움이 된다고 예측하면, 초지능은 인간의 생존을 위협할 수 있다.[31] 또한, 초지능은 그것이 인간과 전혀

29) Wallach, Franklin, and Allen, "A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents," 457-58.

30) 위의 논문, 458.

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

다른 방식으로 추론하며 도구적 가치를 내재하기 때문에 향후 인간이 전혀 예상하지 못한 영역에서 인류의 실존적 위험을 초래할 가능성이 있어, 이에 대한 선제적 대응방안을 필요로 하고 있다.[32] 보스트롬은 '우호적 인공지능(Friendly AI)'과 같이 도덕적인 인공지능 개발은 이러한 위협에 대응할 수 있는 효과적 방안의 하나가 될 수 있다고 주장한다.[33] 인간의 도덕적 가치에 동의하는 우호적 인공지능과 같은 인공적 도덕행위자는 인공지능 구현시의 윤리적 문제를 해결하기 위한 대안으로서 뿐만 아니라, 초지능의 위협에 대한 예방전략의 일환이다.

물론, 과학기술이 고도로 발전하게 된다해도 강지능 또는 초지능의 도래는 불가능하다고 여기는 견해도 많다. 인간의 의도된 영향력이 없다면, 인공지능이 스스로 학습을 반복하여 강지능이 되는 것은 불가능하다는 견해도 있다.[34] 또한, 인공지능은 의식이나 의도성이 부재하기 때문에 도덕적 주체로 인정하기 어렵다는 견해도 있다.[35] 이렇게 인공적 도덕행위자의 문제는 인공지능의 발전과 그 가능성에 직접적으로 연관되어 있기 때문에, 강지능이나 초지능의 개발이 불가능하다고 여기는 견해속에서는 인공적 도덕행위자를 구현하는 것이 아예 불가능하다고 생각하기 쉽다.

하지만, 예일 대학교의 웬델 월러치가 주장하였듯이, 우리는 불가능하

32) "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," *Minds and Machines* 22/2(2012): 75, 83-84.

33) Luke Muehlhauser and Nick Bostrom, "Why We Need Friendly AI," *Think* 13/36 (2013): 44-45; Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," 83-84.

34) Karim Jebari and Joakim Lundborg, "Artificial Superintelligence and Its Limits: Why Alphazero Cannot Become a General Agent," *AI & SOCIETY* (2020): 5-8.

35) Riya Manna and Rajakishore Nath, "The Problem of Moral Agency in Artificial Intelligence," (paper presented at the 2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW), 2021), 3.

다는 쉬운 답변 대신에, 인공적 도덕 행위자를 구현하는데 있어 "정확히 무엇이 문제이고 장애물일까?"라는 질문을 진지하게 성찰할 필요가 있다.[36] 지금 이 순간에도 인공지능 기술은 급속도로 발전하면서 전통적으로 인간의 영역이라 여겨졌던 윤리적 판단들을 대체하고 있기 때문이다. 그렇다면, 인공지능 기술의 발전이 멈추지 않는 한, 인공지능의 윤리적 한계와 쟁점은 지속적으로 논의될 것이며, 미래 인공지능 기술이 강지능 또는 초지능을 목표로 발전할수록 인공적 도덕 행위자에 관한 논의는 더욱 부각될 수 밖에 없다. 즉, 인공지능의 발전에 관한 궁극적인 가능성이나 기술적 한계는 그 누구도 예단하기 어렵지만, 인공지능 발전은 결국 인공적 도덕행위자를 요구하게 될 것이다.[37] 따라서, 인공지능에 관한 윤리적 이슈를 논할 때, 인공적 도덕행위자 혹은 도덕적 인공지능을 학술적 논의의 장으로 가져오고 비판적으로 이를 검토하는 것은 '필수적'이다.

특히, 불교의 경우 해탈과 같은 도덕적 경지는 고통에서 벗어나는 깨달음을 의미하기 때문에, 이성적 측면에서 지능이 뛰어난 인공지능이 이에 더욱 쉽게 도달할 가능성이 있다는 주장을 하기도 한다. 예를 들어, 태국의 불교학자 헝라다롬은 인공지능의 지성이 인간 수준에 도달하거나 이를 뛰어넘는 초지능은 인간보다 더 뛰어난 도덕적 존재로 성장할 수 있는 가능성이 있다고 주장하였다. 그에 따르면, "불교적인 관점에서 보면 열반, 즉 최고의 윤리적 완성의 상태를 궁극적으로 달성하기 위한 필요조건은 무지avijja의 제거이다. 따라서 초지능 로봇은 역시 초윤리적이어야할 것"이라 말한다.[38] 그는 초지능에 기반한 깨달음을 통해 세상의 고통과 번뇌로부터 해탈에 이르는 존재가 된 인공지능은 "자기 자신의 에고에

36) Wendell Wallach and Colin Allen,『왜 로봇의 도덕인가』, 31.

37) 위의 책, 367.

38) Soraj Hongladarom, *The Ethics of AI and Robotics: A Buddhist Viewpoint,* 김근배 역,『불교의 시각에서 본 AI와 로봇 윤리』(서울: 씨아이알, 2022), 154.

집착하지 않기 때문에 전적으로 자비로워질 것이며, 나아가 그것들은 다른 존재들의 이익을 자기 자신의 것보다 더 보살피게 될 것"이라 주장한다.[39]

물론, 이런 초지능의 해탈이 인간의 경험과 지성을 통해 이룩할 수 있는 가능성으로서의 해탈과 동일할 지는 쉽게 판단할 수 없다. 인간처럼 고통과 고난의 아픔을 신체적이고 감정적으로 경험한 후에 깨닫는 해탈과 달리, 이 모든 것을 기계 감각 매개체와 지성을 기반으로 이룩한 초지능의 해탈은 육체적 유한성을 포함하는 인간의 총체적 삶에 대한 경험적 한계를 내포한 해탈이기 때문이다. 하지만, 헝라다롬의 주장처럼 미래에 초지능이 해탈하여 타인을 위한 이타적인 판단과 행위를 수행하는 도덕적 인공지능으로 존재하게 된다면, 인간은 사회내 수많은 윤리적 난제들 역시 도덕적 인공지능에 의존하여 해결하는 상황을 수용할 수 있어야 할 것이다.

한편, 니버는 전통적 기독교 신학의 입장을 반영하여 인간은 피조물이자 하나님의 형상을 닮은 존재일 뿐만 아니라 죄인으로 이해하고 있으며, 이는 스스로의 힘으로 벗어날 수 없는 본성에 속한다.[40] 니버는 인간의 이성, 창의성, 상상력, 자유의지 등과 같은 자기 초월성을 하나님의 형상에 속한 본성으로 이해하며, 죄인인 인간은 이를 동원하여 피조물의 한계를 초월하려 한다고 주장했다.[41] 이 과정에서 인간은 스스로 교만하여져서 자신의 능력으로 피조물의 한계를 초월할 수 있다고 믿으며 자기 자신이나 혹은 자신의 힘으로 통제가능한 우상을 하나님 대신 숭배하려는 유혹에 빠지는 죄를 범한다.[42]

---

39) 위의 책, 175.

40) Reinhold Niebuhr, *The Nature and Destiny of Man*, vol. 1 (Louisville, KY: Westminster John Knox Press, [1941] 1996), 153.

41) 위의 책, 150.

하지만, 인공지능의 발전과 관련하여 이성이 하나님의 형상에 속한 인간의 고유한 본성이라는 전통 신학에 머무른다면, 우리는 인간이 창조한 인공지능이 미래에 이성, 창의성, 상상력 등을 인간 이상으로 수행할 수 있게 될 때, 신학적 딜레마에 빠지는 것을 피할 수 없어 보인다. 만일 이성과 창의성 등을 하나님의 형상에 속한 것으로 여기고 인간을 통해 이러한 하나님의 형상이 인공지능에게 공유되었다고 생각한다면, 인간보다 이러한 속성을 더 뛰어나게 수행하는 초지능 단계에 이른 인공지능에게 하나님의 형상이 그들을 통해 인간보다 더 완전하게 드러났다고 주장하는 것이 어색하지 않기 때문이다.

이성을 하나님의 형상으로 강조할 때 발생할 수 있는 문제 상황을 노린 허즈펠드(Noreen Herzfeld)와 테드 피터스는 '관계성'에 중심을 두어 극복하려 한다. 먼저, 허즈펠드는 인간이 인공지능을 개발하려는 노력이 인간 존재 안에 있는 하나님의 형상으로 해석되어온 충동에 의한 것이지만, 이는 쉽게 왜곡될 수 있다고 주장하였다.[43] 우리가 인공지능을 창조하며 자연의 한계를 초월하는 불멸을 꿈꾸는 것은 무한한 창조주 하나님께 도달하고자 하는 갈망하는 하나님의 형상의 일부를 표현하는 것이지만, "니버는 우리가 유한한 존재로서의 본성을 놓치지 말아야함을 상기시켜준다."[44] 허즈펠드는 우리가 인공지능을 통해 하나님의 형상을 창조하려는

42) 위의 책, 86. 물론, 인간 이해에서 니버가 의존하는 하나님의 형상 개념은 육체와 정신을 이분법적으로 해석했다거나, 이성이야말로 타자를 지배하고 억압하기 위한 수단으로 사용되어 왔다거나, 우리가 성장해야할 필요가 있는 존재라는 역동적 이해 또는 인간의 책임성을 드러내지 못한다는 비판을 받아온 것도 사실이다. 하지만, 니버는 후기에 하나님의 형상 개념을 관계적으로 이해하려 했다. Noreen L. Herzfeld, *In Our Image : Artificial Intelligence and the Human Spirit*, Theology and the Sciences (Minneapolis, MN: Fortress Press, 2002), 19-22, 107. 또한, 노린 허즈펠드는 하나님의 형상 개념에 대한 학자들의 해석을 추적하면서 이것이 본성적, 기능적, 관계적으로 해석될 수 있다고 주장했다. 위의 책 2장을 참조할 것.

43) *In Our Image : Artificial Intelligence and the Human Spirit*, 84.

44) 위의 책.

욕망을 포기하려 하지 않을 것임을 알고 있으나, 궁극적으로 인간과 같은 지성을 지닌 인공지능을 개발하는 것은 불가능할 것이라 추정한다.[45]

뿐만 아니라, 지성은 인간의 본성을 나타내는 가장 중요한 요소는 아니며, 오히려 하나님이 인간의 형상으로 이 땅에 오셨고 이 세상에서 우리와 관계를 맺으시며, 우리는 이러한 하나님의 형상을 두 세 사람이 진정한 관계성 속에서 모일 때마다 발견하게 된다.[46] 허즈펠드는 미래에 인공지능이 인간과 같은 의식을 소유하게 될 것이라는 것에 동의하지 않으며, 그것이 자유의지를 소유하거나 스스로의 목적을 세울 수도 없다고 주장한다.[47] 이러한 인공지능은 인간을 대신하여 각종 업무를 수행하고, 인간은 그것과 마치 관계를 맺고 있는 것처럼 행동할 수는 있지만, 그것이 우리가 하나님과 이웃과 사랑을 통해 진정한 관계를 맺는 것과는 다른 것이다.[48]

테드 피터스 역시 초지능을 향한 트랜스휴머니스트의 꿈은 실현 불가능한 것이라고 비판한다.[49] 일단, 피터스는 이성이 인간에게 고유한 속성이라는 오랜 믿음을 포기할 때가 왔다고 생각하며, 원칙상으로, 미래에 인간과 동등하거나 뛰어난 지능이나 추론능력을 가진 기계를 개발하는 것은 가능한 일이라고 인정한다.[50] 하지만, 피터스는 '자율적인' 인공지능을 창조하려는 시도는 인간의 지능이 인간의 특별한 사회적 관계성 속에 체현되어 있다는 것을 간과한 것이라 비판한다.[51] 인간의 자아 혹은 자의

45) 위의 책, 94.
46) 위의 책.
47) Noreen L. Herzfeld, *The Artifice of Intelligence : Divine and Human Relationship in a Robotic Age* (Minneapolis: Fortress Press, 2023), 64-66.
48) *In Our Image : Artificial Intelligence and the Human Spirit*, 94.
49) Ted Peters, "Artificial Intelligence Versus Agape Love," *Forum Philosophicum* 24/.2(2019): 274-275.
50) 위의 논문, 263.
51) 위의 논문, 267.

식은 사회적 관계 속에 형성되며, 인간은 상호의존적인 존재로서 고유의 인간성은 인간과 하나님 혹은 이웃과의 사랑의 관계속에 형성된다.[52]

따라서, 허즈펠드나 피터스에게 인간을 고유한 존재로 만드는 것은 전통 신학의 주장에서처럼 이성뿐만이 아닌 사랑의 진정한 관계가 된다. 또한, 인공지능이 자율성을 갖는다는 것은 불가능할 뿐만 아니라, 설령 지능이 인간수준 이상으로 뛰어나게 발전한다 해도 사회속에서 인간 및 하나님과 진정한 사랑의 관계성에 들어가는 것도 불가능하다. 두 사람의 주장을 근거로 도덕적 인공지능에 적용하여 본다면, 인간 수준 혹은 그 이상의 초지능이나 그러한 지능에 기반하여 등장할 도덕적 인공지능의 출현을 기대하는 것은 실현불가능한 인간의 불가능한 욕망이자 자기 초월성을 인공지능에 투사한 거짓된 교만이 된다.

지금까지 우리는 도덕적 인공지능의 가능성에 대하여 과학적, 철학적, 종교적 입장들의 일부를 검토하였다. 기존의 하향식 인공지능 개발의 문제점을 보완하기 위해 등장한 상향식 접근방식은 도덕적 인공지능의 가능성을 탐색하고 있다. 종교적 입장에서 도덕적 인공지능의 등장 가능성에 대한 평가는 다양하지만, 니버처럼 전통적 신학에 기반한 입장은 이를 비판적으로 판단할 것이다. 하지만, 강지능 혹은 초지능은 불가능하다는 비판적 견해들과 달리, 과학기술은 끊임없이 목적을 향해 진보하고 있기 때문에, 우리가 강지능 혹은 초지능의 등장을 완전히 배제할 수 없는 것도 사실이다. 이 경우 우리들은 도덕적 인공지능의 등장을 예상할 수 있다. 만약, 일부 학자들의 주장대로 미래에 인간의 지능과 유사하거나 이를 초월한 인공지능이 등장하여 도덕적 인공지능이 개발될 수 있다면, 이러한 도덕적 인공지능의 등장은 라인홀드 니버의 기독교 현실주의 관점에 비추어볼 때 또다른 차원의 윤리적 이슈를 제기하게 될 것이다.

52) 위의 논문, 270-272.

IV. 도덕적 인공지능과 비도덕적 사회

이 페이지들에서 자세히 설명할 논제는 개인과 국가, 인종, 경제 등 사회 집단의 도덕적, 사회적 행동 사이에는 뚜렷한 구분이 있어야 한다는 것이다.[53]

기독교 현실주의자인 라인홀드 니버는 1932년에 출판한 『도덕적 인간과 비도덕적 사회』에서 개인 윤리와 사회 윤리를 명확히 구분해야 한다고 주장하였다. 디트로이트에서의 목회 초기에서부터 인간의 죄된 본성과 사회에 대한 비판적 인식을 드러냈던 니버는 피조물로서 하나님의 형상을 지닌 인간을 동시에 악한 죄인으로 이해하였다.[54] 물론, 니버는 인간이 상황에 따라 선한 존재로서 기능할 가능성을 완전히 배제하지는 않았기 때문에 '도덕적 인간'이라는 표현을 사용하긴 했지만, 본성상 죄인인 인간이 이해집단을 이루어 그들의 이익을 대변하게 되는 인간사회는 마치 전쟁터와 같이 서로 투쟁하며 냉혹하게 짓밟는 비도덕한 상태가 사회적 실상이라고 비판하였다.[55]

즉, 니버는 인간은 본성상 죄인이지만 가족이나 친지 등의 사적 관계나 친밀한 대면 공동체에서는 이성이나 교육이나 양심 등의 영향을 받아 도덕적이 될 수 있다고 인정했다. 그는 인간이 이렇게 친밀한 사적 관계에서 드물게나마 도덕적이 될 수 있는 가능성을 배제하지 않은 것이 사실이다. 하지만, 니버는 사회 집단간의 상호작용시에 소속 집단의 이해관계를

53) Reinhold Niebuhr, *Moral Man and Immoral Society; a Study in Ethics and Politics* (New York, London,: C. Scribner's sons, [1932] 1960), xi.

54) Reinhold Niebuhr, *Leaves from the Notebook of a Tamed Cynic*, 송용섭 역, 『길들여진 냉소주의자의 노트』 (서울: 동연, 2013), 65-68; R. Niebuhr, *The Nature and Destiny of Man*, vol.1, 150-51, 78-79.

55) R. Niebuhr, *Moral Man and Immoral Society; a Study in Ethics and Politics*, 19-20.

대변해야만 하는 인간은 집단 이기주의에서 벗어날 수 없기 때문에 사회적 행동은 비도덕적이 될 수 밖에 없다고 주장한 것이다.

이러한 니버의 사상은 기독교 현실주의라 불리며 죄성을 지닌 인간 도덕성의 상황적 한계를 드러냈다. 즉, 피조물로서 죽음의 한계를 벗어날 수 없는 인간의 유한성은 인간을 불안하게 만들고, 생존을 위한 이기적 인간은 내재된 불안을 극복하고자 영속적으로 권력을 추구하여 독점하려 하게 된다. 이렇게 불안한 인간이 권력을 통해 유한성을 초월하고자하는 과정에서 스스로가 하나님처럼 될 수 있다는 교만의 죄에 빠져 인간이 창조주 하나님의 자리까지 넘보게 되는 것이다.[56] 이때, 사회내에서 권력을 소유한 인간은 그렇지 못한 타자를 자신의 의지에 굴복해야만 하는 대상이자 도구처럼 취급하려는 유혹에 쉽사리 빠진다.[57] 집단이기주의로 인해 강제력을 동원해야만 하는 사회속에서 이러한 유혹에 빠진 인간은 더욱 악해질 수 밖에 없다. 따라서, 사회는 제도적 장치를 통하여 권력의 독점을 분산하고 견제하여 할 수 있을 때만 보다 덜 악하고 보다 더 정의로운 사회로 나아갈 가능성이 있다. 이러한 니버의 현실주의는 윤리학 분야 뿐만 아니라, 사회, 정치 분야에 폭넓게 적용되어 왔으며, 특히, 자국의 생존을 위한 이기주의와 힘의 논리가 지배적인 국제관계에 큰 영향을 미쳐왔다.

니버는 사회 정의를 위한 이성의 기능을 부정하진 않았지만 제한적으로 인정했다. 즉, 니버는 인간의 지능을 향상시킴으로써 타인의 필요를 이해할 수 있게 하고 불합리성을 제거함으로써 정의를 이룰 수 있다고 희망했던 이성주의자들에 동의하지는 않았지만, 이성이 증가하면 인간의 도덕성도 향상될 수 있으며, 부정의에 대한 무지(ignorance)에 따른 사회

56) R. Niebuhr, *The Nature and Destiny of Man*, vol.1, 179-82.
57) 위의 책, 182.

적 불합리성과 가식이 제거된 합리적인 사회가 될 때, 보다 정의로운 사회가 될 수 있음을 인정했다.[58] 동시에, 니버는 인간의 공감능력이나 상호책임감이 무한히 확장될 수는 없으며, 무엇보다 사회에서 권력을 동원하여 어떻게든 자신의 이익을 추구하는 충동적인 인간이 타인을 위해 자신의 이기심을 포기할 만큼 충분히 합리적일 수는 없다고 주장함으로써 사회 집단속의 이성의 한계를 명확히 했다.[59]

하지만, 인공지능기술의 발전에 따라 미래에 강지능 혹은 초지능이 개발되면 니버의 시대에는 상상할 수 없었던 정도의 지성과 합리적인 이성을 갖춘 도덕적 인공지능이 현실화되어, 다양한 사회적 관계에서 타인의 필요를 이해하여 자신의 이익이나 집단 이기주의를 초월할 가능성을 배제할 수 없게 되었다. 따라서, 이제 기독교 윤리학계는 니버의 기독교 현실주의가 도덕적 인공지능에게 어떻게 적용될 수 있는지, 혹은, 도덕적 인공지능은 과연 집단 이기주의를 초월하여 일관적인 도덕적 주체가 될 수 있는지를 성찰해야 할 시기가 되었다.

미래에 도덕적 인공지능의 등장은 인간에 대한 영향력과 인공지능에 대한 인간의 의존성의 증가를 예상하게 한다. 예를 들어, 현재 자율주행 자동차는 인간의 운전과 판단을 보조하는 도구일 뿐이어서 인간의 의존도가 낮은 편이다. 하지만, 미래에 인공지능의 발달에 따라 완전 자율 주행의 시대가 열리게 된다면, 인간보다 안전한 자율주행이 가능해질 수 있다. 음주와 같은 약물중독이나 분노와 같은 감정 기복 및 수면부족과 악천후 같은 운전 장애 상황에 취약할 수 밖에 없는 인간보다 축적된 데이터에 의한 합리적 운행 판단과 운행보조 장치 등의 도움으로 더 안전한 자율주행이 가능하게 되는 것이다. 이에, 위험한 인간이 작용할 수 있게

58) R. Niebuhr, *Moral Man and Immoral Society; a Study in Ethics and Politics*, 23-33.
59) 위의 책, 28-35.

하는 운전대를 없애고 인공지능을 통해 안전한 자율주행만 하도록 법제화해야 한다는 주장이나,[60] 아군을 위기의 순간에서 신속히 지원하기 위해서는 자율군사드론이 필요하다는 주장이나, 불합리한 인간 판사대신에 인공지능 로봇이 판결하는 것이 나을 것이라는 주장도 있다.[61]

이렇게 인공지능이 발전할수록 이와 관련된 우리들의 기대와 의존은 제한된 영역에만 그치지 않고, 사회, 정치, 경제, 예술 등의 다양한 분야와 덕의 형성이나 윤리적 판단에까지 이르게 될 가능성이 높다. 그런데, 현재까지 인공지능에 대한 주요 논의들은 니버가 사회갈등의 핵심으로 다루는 집단 이기주의에 연관된 것이라기보다, 사건이나 사고 등의 특정한 문제 중심적인 해결방안을 주로 다루고 있는 듯 하다. 혹은, 사회, 정치, 경제적 문제라 해도 환경이나 기후문제 해결을 위한 인공지능의 판단이 공익이나 전 인류의 생존을 위한 것으로 여겨 관련 집단들이 결국에는 이기심을 포기하고 수용해야만 하는 가시적 문제들처럼 보인다. 위에서 언급한 자율주행, 군사드론, 인공지능 법률가, 혹은, 초지능의 실존적 위험 등의 이슈는 이러한 구분에서 크게 벗어나지 않는다.

하지만, 도덕적 인공지능이 개발 단계의 실험실을 넘어 사회 관계속의 인간과 접촉하게 될 때, 기독교 윤리학은 사회 내에서 집단 이기주의 같이 보다 본질적인 윤리적 이슈까지 성찰해야 할 것이다. 즉, 도덕적 인공지능과 인간과의 상호작용시에 간과하지 말아야할 주요 이슈 중에 하나는 집단 이기주의와 도덕적 인공지능의 초월 가능성 문제가 될 것이다.

이를 보다 명확히 살펴보기 위하여 다음과 같이 몇 가지 시나리오를 가정해볼 수 있다. 먼저, 큰 틀에서, 도덕적 인공지능의 개발단계에서 인

60) 김은영, "미래에 인간은 운전할 수 있을까?" 「사이언스 타임즈」 https://www.sciencetimes.co.kr/news/미래에-인간은-운전할-수-있을까/. 2023년 10월 11일 접속.

61) 양지열, "[양지열 칼럼] 인공지능(AI) 판사는 정의로울까?," 「AI 타임스」 (2021.08.13), https://www.aitimes.com/news/articleView.html?idxno=140115 2023년 10월 13일 접속.

간 집단과 상호작용시 발생할 수 있는 경우이다. 이 때는, 초기에 아동 수준의 인공지능을 개발한 후에 덕에 대한 학습과 훈련 과정을 거쳐 성인 수준의 도덕적 인공지능으로 발전시키게 된다. 이 수준에서는 프로그래머들의 가치관이나 인종적 혹은 문화적 선입관이나 편견이 프로그래밍이나 학습 과정을 통해 아동 수준의 인공지능을 무의식적으로 오염시키게 되거나, 프로그래머들이 의도적으로 편향된 도덕적 인공지능을 개발할 가능성이 있다. 이때는 특정 가치나 종교, 문화, 인종을 우선시하거나 보편화하거나 차별할 수 있는 위험이 있다.

혹은, 프로그래밍의 단계에서 선입견이 없거나 편향되지 않은 아동 수준의 인공지능이 개발되었다고 가정할 수 있다. 하지만, 이 경우에 학습할 자료들 속의 사회 문화적 편견이나 상호작용할 인간들의 집단 이기주의에 의해 학습과정의 인공지능의 도덕성이 오염될 수 있다. 예를 들어, 다양한 사진으로 기계학습한 인공지능이 미의 기준을 백인으로 삼거나 흑인을 고릴라로 인식한 사례는 학습과정의 인공지능의 가치가 사회 문화 인종적 편견에 오염된 경우이다.[62] 또한, 마이크로소프트사의 챗봇 '테이'나 한국의 챗봇 '이루다'의 서비스 중단 사태처럼, 아동 수준 인공지능의 사회화 과정에서 특정 가치관에 편향된 인간들이 의도적으로 성차별이나 인종차별 혹은 여성혐오 등으로 왜곡된 가치관을 학습시키려 시도할 수 있다.[63]

이렇게 인공지능의 초기 개발단계와 학습과정에서, 인공지능은 인간의 편견, 차별, 혐오 등에 의해 편향되거나 오염될 수 있으며, 인공지능기술

62) 구본권, "기계학습의 맹점, '흑인=고릴라' 오류가 알려주는 것," (2019-04-05), https://www.hani.co.kr/arti/science/future/877637.html. 2023년 10월 20일 접속.

63) 한세희, "MS 채팅 봇 '테이', 24시간 만에 인종차별주의자로 타락," (2016.03.27.), https://m.dongascience.com/news.php?idx=11158. 2023년 10월 12일 접속; 이효석, "성희롱 · 혐오논란에 3주만에 멈춘 '이루다'…AI윤리 숙제 남기다," (2021.01.11), https://www.yna.co.kr/view/AKR20210111155153017. 2023년 10월 12일 접속.

선도국의 사회문화적 가치나 이데올로기 혹은 초국적 기업의 지배력을 확대재생산할 가능성이 존재한다. 그럼에도 불구하고, 프로그램 개발시의 점검 과정과 학습과정의 개선 등을 통하여, 이러한 내외적 영향력을 감소시키거나 제거할 방안을 찾을 수도 있을 것이다. 이 경우, 일부 주장처럼 인공지능이 궁극적으로 자기 이익을 초월한 자비로운 존재가 되어, 모든 사람들의 이익을 위하는 도덕적 인공지능이 될 가능성을 배제할 수 없다.

다음으로, 두번째 큰 틀에서, 이렇게 자기 이익을 초월한 도덕적 인공지능이 사회속에서 인간 집단과 상호작용시 발생할 수 있는 경우이다. 이때는 도덕적 인공지능이 사회에서 실제 작용하는 경우이므로 현실 생활에 대한 파급효과가 클 것이다. 이러한 상황에서 제기할 수 있는 질문은 다음과 같다. 미래에 도덕적 인공지능은 집단 이기주의에 오염된 비도덕적 사회속에서도 자신의 도덕적 입장을 일관적으로 유지함으로써 항상 도덕적 존재로 남아있을 수 있을 것인가? 아니면, 도덕적 인공지능은 다양한 이해관계와 가치가 충돌하는 비도덕적 사회속에서 인간처럼 비도덕적이 될 수 밖에 없는가?

이에 대한 대답을 모색하기 위하여, 도덕적 인공지능이 국가간 갈등을 해결하고 중재하는 역할을 담당하는 상황을 가정해 보자. 예를 들어, 최근의 러시아와 우크라이나 혹은 하마스와 이스라엘 전쟁의 경우와 같이, 미래의 어느날 복잡한 국제 역학관계 속에서 국가간에 전쟁이 시작되었고, 도덕적 인공지능은 이해관계를 초월하여 합리적이고 공정하며 자비로운 최선의 방안을 제시한다. 이는 관계된 모든 국가가 수긍할만한 해결방안 같았지만 곧이어 다음과 같은 딜레마 상황이 발생한다: 특정 국가가 이기적으로 도덕적 인공지능의 방안을 수용하지 않은 채 우월한 힘을 앞세워 전쟁을 지속한다.

이때, 도덕적 인공지능이 이전의 해결 방안을 유지한다면 폭력적인 전쟁을 중단시키지 못하는 결과를 초래한 것이며(도덕적 인공지능의 무용성), 혹은, 도덕적 인공지능이 최선의 해결 방안을 포기함으로써 타협안을 제시하거나, 강제력을 동원하여 그 국가를 처벌하려 할 경우,[64] 그러한 행위는 결국 도덕적 인공지능이 집단 이기주의의 영향에 노출된 결과라 할 수 있다(도덕적 인공지능의 비도덕성). 이러한 시나리오들을 통해 살펴본 것처럼, 결국, 우리는 자신의 이기심에서 자유로워진 인공지능이 도덕적인 존재가 될 수 있다고 가정할 수는 있어도, 도덕적 인공지능이 비도덕적 사회 속에 들어와 권력에 대한 의지와 집단 이기주의에 물들어 있는 인간 집단들과 상호작용할 경우에는 인간의 죄성에서 완전히 격리된 절대적으로 도덕적인 존재가 될 수 있으리라 가정하기는 어렵다.

물론, 가장 바람직한 시나리오로서, 인간이 도덕적 인공지능과의 상호작용을 통하여 보다 정의로운 사회를 만들기 위해 협업을 하며 다함께 집단 이기주의를 극복하기 위해 노력하는 경우를 가정할 수 있다. 그럼에도 불구하고, 인간의 본성에 여전히 죄가 끈질기게 남아있는 한, 사회적 활동을 지속하는 인간이 도덕적 인공지능과의 상호작용을 통해 유한성 극복에 대한 다양한 욕망들을 완전하고 영원히 포기할 것이라 기대하는 것은 안일한 생각처럼 보인다.[65] 따라서, 미래에 도덕적 인공지능이 등장하여 이에 대한 인간의 의존도가 더욱 높아진다 하더라도, 인간의 본성에

64) 니버는 제 3자의 입장에서 강제력을 가치중립적으로 이해했지만, 필자는 그 강제력의 피해를 경험해야할 대상들의 관점에서는 그것이 가치중립적이 아닌 악한 것이 된다고 생각한다.

65) 물론, 디지털 세계인 가상 공간으로 모든 인류가 이주하는 시기가 찾아온다면, 동일한 특성으로 복사가 가능한 디지털의 특성을 이용하여 인간이 상상하는 디지털 재화를 무제한으로 공급받을 수 있게 되어 집단 이기주의가 급격히 감소할 가능성도 있을 것 같다. 하지만, 인간과 함께 죄성이 가상 공간에 전파된다면, 교만한 죄의 특성은 다양한 인간이 집단을 이루어 활동하는 가상공간 역시 비도덕적인 공간으로 전락시킬 가능성이 여전히 남아있다.

죄가 남아 있고 집단 이기주의가 작용하는 비도덕적 사회에서는 인공지능이 아무리 뛰어난 도덕성을 지닌다 해도 비도덕적 존재로 타락할 가능성을 완전히 배제하기는 어려울 것이다.

V. 나가는 말

인간이 과학기술의 발전을 통해 인간과 유사하거나 더 뛰어난 인공지능을 개발하고 그것을 통해 도덕적 인공을 개발할 수 있을 지에 관하여는 다양한 견해가 있다. 하지만, 과학기술의 진보가 멈추지 않는 한, 미래에 도덕적 인공지능의 등장 가능성을 완전히 배제할 수는 없다. 만약, 도덕적 인공지능이 개발된다면, 그것은 대면 관계나 표면적 문제 해결이나 명확한 공익 추구의 상황에서는 도덕적 존재로 남을 수 있을 것이다. 하지만, 도덕적 인공지능이 사회속에서 인간과 상호작용할 경우에는, 그것이 인간의 집단 이기주의의 영향에서 완전히 벗어나는 것은 불가능한 것처럼 보인다.

그럼에도 불구하고, 미래에 도덕적 인공지능 자체가 다양한 관계속에서 인간보다 상대적으로 선해질 가능성이 있다는 가정은, 보다 정의로운 미래를 위한 일말의 희망을 준다. 니버는 사회 집단속에서 인간은 결국 비도덕적이 될 수 밖에 없다는 현실적 한계를 깨닫게 했지만, 그러한 한계를 깨달은 인간이 완전한 정의에 대한 이상을 버릴 때, 지금보다 나은 정의를 추구하는 것이 불가능한 것만은 아니라는 희망을 남겨두었다. 또한, 지능, 이성, 합리성의 발전과 확산이 이러한 희망에 충분하지는 않아도 작은 불씨를 보탤 수 있음도 알려주었다.

미래에 인공지능이 인간에 미칠 수 있는 영향력이 중대하기 때문에, 인간이 상대적으로 보다 정의로운, 혹은, 보다 덜 비도덕적인 사회로 나아가기 위한 가장 효과적인 방법 중의 하나는 도덕적 인공지능의 개발에

서 시작될 지도 모른다. 비도덕적 인간이라도 도덕적 인공지능과 상호작용을 할 때, 도덕적 인공지능으로부터의 영향을 받는 것은 피할 수 없을 것이기 때문이다. 즉, 도덕적 인공지능은 비도덕적 인간의 영향에서 자유로울 수 없지만, 상대적으로 인간보다 더 도덕적인 존재가 될 가능성이 존재한다. 이러한 도덕적인 인공지능과 상호작용할 때 비도덕적 인간이라 할 지라도 보다 도덕적인 인공지능의 영향을 받아 인간의 도덕성 역시 보다 향상될 가능성이 있다. 이러한 상호작용 과정이 사회속의 집단들에서도 반복될 경우에, 그렇지 않을 경우보다 조금 더 합리적이고 보다 더 정의로운 사회로 나아갈 수 있으리라 희망한다.

그렇다면, 도덕적 인공지능이 인간과 함께 보다 정의로운(혹은, 보다 덜 비도덕적인) 사회를 만들어갈 수 있는 방안은, 도덕적 인공지능과의 상호작용을 통해 인간(집단)의 도덕성이 향상될 수 있는 선순환의 고리를 찾아내는 데 있을 것이다. 아마도 그것은 도덕적 인공지능이 비도덕적 사회로 들어가 활동하기 전에, 즉, 인간이 그나마 도덕적일 가능성을 유지할 수 있는 친밀한 대면관계의 도덕 공동체 속에서 그 인공지능이 덕을 학습하고 덕을 함양한 인간들로부터 양육 받는 것에서 시작될 수 있을지 모른다. 만일 교회가 미래의 어느날 그 중대한 역할을 감당할 수 있기를 희망한다면, 누구든지 그 안에서 하나님의 형상을 발견할 수 있는 아가페 사랑의 공동체가 되어 비도덕적 사회에서 그때까지 남아있어야 하는 어쩌면 불가능한 가능성에 자신의 운명을 걸어야 할 것이다.

참고문헌

김다솜, 맹주만. “인공지능과 도덕적 기계－칸트적 모델과 흄적 모델.”「철학탐구」 62(2021), 177-216.

김은수, 변순용, 김지원, 이인재. “10세 아동 수준의 도덕적 인공지능개발을 위한 예비 연구 - 인공지능 발달 과정을 중심으로.”「초등도덕교육」 57(2017), 105-27.

맹주만. “인공지능, 도덕적 기계, 좋은 사람.”「철학탐구」 59(2020), 213-42.

목광수. “도덕의 구조-인공지능 시대 도덕 논의의 출발점.”「철학사상」 73(2019), 163-95.

_____. “인공적 도덕 행위자 설계를 위한 고려사항.”「철학사상」 69(2018), 361-91.

박균열. “도덕적 역량 개념을 토대로 한 자율적 도덕행위자(AMA)의 설명 가능한 역량 기획.”「한국도덕윤리과교육학회 학술대회 자료집 2020」 10(2020), 594-603.

_____. “인공적 도덕행위자(AMA)의 온톨로지 구축.”「한국디지털콘텐츠학회논문지」 20/11(2019), 2237-42.

박형빈. “AI윤리와 신경과학의 AMA 도전 과제-도덕판단 알고리즘 구현을 위한 검토 사항.”「윤리교육연구」 64 (2022), 91-114.

_____. “기계윤리 및 신경윤리학 관점에서 본 인공도덕행위자(AMA) 도덕성 기준과 초등도덕교육의 과제.” 한국초등교육 31/5(2021), 77-92.

송승현. “인공지능과 도덕성.”「法曹」 67/6(2018), 267-341.

신상규. “인공지능 시대의 윤리학.”「지식의 지평」 21(2016), 1-16.

_____. “인공지능은 자율적 도덕행위자일 수 있는가?”「哲學」 132(2017), 265-92.

_____. “인공지능의 도덕적 지위와 관계론적 접근.”「철학연구」 149(2019), 243-73.

이상욱. “인공지능의 도덕적 행위자로서의 가능성: 쉬운 문제와 어려운 문제.”「哲學硏究」 125(2019), 259-79.

이상형. “윤리적 인공지능은 가능한가?-인공지능의 도덕적, 법적 책임 문제.”「법과정책연구」 16/4(2016), 283-303.

이재숭. “AMA의 도덕적 지위의 문제.”「哲學論叢」102/4(2020), 527-45.

정태창. “자아 없는 자율성-인공 지능의 도덕적 지위에 대한 고찰.”「사회와 철학」40(2020), 147-80.

최경석. “인공지능이 인간 같은 행위자가 될 수 있나?”「생명윤리」21/1(2020), 71-85.

Allen, Colin, Wendell Wallach, and Iva Smit. “Why Machine Ethics?” *IEEE Intelligent Systems* 21/4(2006), 12-17.

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* 22/2 (2012): 71-85.

Fullam, Lisa. “Genetically Engineered Traits Versus Virtuous Living.” *Theology and Science* 16/3(2018), 319-29.

Gouw, Arvin M. “Genetic Virtue Program: An Unfeasible Neo-Pelagian Theodicy?” *Theology and Science* 16/3(2018), 273-78.

Hongladarom, Soraj. *The Ethics of AI and Robotics: A Buddhist Viewpoint,* 김근배 역,『불교의 시각에서 본 AI와 로봇 윤리』. 서울: 씨아이알, 2022.

Herzfeld, Noreen L. *The Artifice of Intelligence : Divine and Human Relationship in a Robotic Age.* Minneapolis: Fortress Press, 2023.

_____. *In Our Image : Artificial Intelligence and the Human Spirit.* Theology and the Sciences. Minneapolis, MN: Fortress Press, 2002.

Jebari, Karim, and Joakim Lundborg. “Artificial Superintelligence and Its Limits: Why Alphazero Cannot Become a General Agent.” *AI & SOCIETY* (2020), 1-9.

Manna, Riya, and Rajakishore Nath. “The Problem of Moral Agency in Artificial Intelligence.” Paper presented at the 2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW), 2021.

Müller, Vincent C, and Nick Bostrom. “Future Progress in Artificial Intelligence: A Survey of Expert Opinion.” *Fundamental issues of artificial intelligence* (2016), 555-72.

Muehlhauser, Luke, and Nick Bostrom. “Why We Need Friendly AI.” *Think*

13/36(2013), 41-47.

Niebuhr, Reinhold. *Leaves from the Notebook of a Tamed Cynic*. 송용섭 역.『길들여진 냉소주의자의 노트』. 서울: 동연, 2013.

_____. *Moral Man and Immoral Society; a Study in Ethics and Politics*. New York, London,: C. Scribner's sons, [1932] 1960.

_____. *The Nature and Destiny of Man*. Vol.1, Louisville, KY: Westminster John Knox Press, [1941] 1996. First published 1941.

Peters, Ted. "Artificial Intelligence Versus Agape Love." *Forum Philosophicum* 24/2(2019), 259-278.

Robinson, Tim, and Stephen Bridgewater. "Highlights from the Raes Future Combat Air & Space Capabilities Summit." The Royal Aeronautical Society, https://www.aerosociety.com/news/highlights-from-the-raes-future-combat-air-space-capabilities-summit/.

Singler, Beth. "Roko's Basilisk or Pascal's? Thinking of Singularity Thought Experiments as Implicit Religion." *Implicit Religion* 20/3(2018), 279-97.

Staff, Guardian. "Us Air Force Denies Running Simulation in Which Ai Drone 'Killed' Operator." *The Guardian*, June 02, 2023.

Turing, Alan M. "Computing Machinery and Intelligence." In *Theories of Mind: An Introductory Reader*, edited by Maureen Eckert. Lanham, MD: Rowman & Littlefield, 2006.

Vallor, Shannon. *Technology and the Virtues : A Philosophical Guide to a Future Worth Wanting*. New York, NY: Oxford University Press, 2016.

Wallach, Wendell. *A Dangerous Master : How to Keep Technology from Slipping Beyond Our Control*. New York: Basic Books, a member of the Perseus Books Group, 2015.

_____. "Implementing Moral Decision Making Faculties in Computers and Robots." *AI & SOCIETY* 22/4(2008), 463-75.

_____. "Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making." *Ethics and Information Technology* 12/3(2010), 243-50.

Wallach, Wendell and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. 노태복 역.『왜 로봇의 도덕인가』. 서울: 메디치, 2014.

Wallach, Wendell, Colin Allen, and Stan Franklin. "Consciousness and Ethics: Artificially Conscious Moral Agents." *International Journal of Machine Consciousness* 3/1(2011), 177-92.

Wallach, Wendell, Stan Franklin, and Colin Allen. "A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents." *Topics in cognitive science* 2/3(2010), 454-85.

Wallach, Wendell, and Gary E Marchant. "An Agile Ethical/Legal Model for the International and National Governance of Ai and Robotics." *Association for the Advancement of Artificial Intelligence* (2018).

•온라인 자료

구본권. "기계학습의 맹점, '흑인=고릴라' 오류가 알려주는 것." (2019-04-05). https://www.hani.co.kr/arti/science/future/877637.html. 2023년 10월 20일 접속

김은영, "미래에 인간은 운전할 수 있을까?" 「사이언스 타임즈」 (2016.11.15.). https://www.sciencetimes.co.kr/news/미래에-인간은-운전할-수-있을까/. 2023년 10월 11일 접속.

양지열. "[양지열 칼럼] 인공지능(AI) 판사는 정의로울까?" 「AI 타임스」 (2021. 08.13). https://www.aitimes.com/news/articleView.html?idxno=140115. 2023년 10월 13일 접속.

이효석. "성희롱 · 혐오논란에 3주만에 멈춘 '이루다'…AI윤리 숙제 남기다." (2021. 01.11). https://www.yna.co.kr/view/AKR20210111155153017. 2023년 10월 12일 접속.

한세희. "MS 채팅 봇 '테이', 24시간 만에 인종차별주의자로 타락." (2016.03.27). https://m.dongascience.com/news.php?idx=11158. 2023년 10월 12일 접속.

논문투고일: 2023년 11월 15일
심사개시일: 2023년 11월 16일
게재확정일: 2023년 12월 02일

• 국 문 초 록 •

본 논문은 '도덕적 인공지능/인공적 도덕행위자'를 기독교 현실주의자 니버의 '도덕적 인간' 대신에 비도덕적 사회 속에 위치시킴으로써 도덕적 인공지능의 가능성과 한계를 모색하려 한다. 이를 위하여 본 논문은 다음과 같은 주제 질문들을 제기하고 연구할 것이다. 첫째, 인공지능은 인간보다 더 도덕적인 존재가 될 수 있는가? 둘째, 도덕적 인공지능은 비도덕적 사회내의 집단 이기주의적 한계를 초월할 수 있는가? 이러한 질문에 대하여, 저자는 미래에 도덕적 인공지능이 개발된다면, 그것은 대면 관계나 표면적 문제 해결이나 명확한 공익 추구의 상황에서는 도덕적 존재로 남을 수 있을 것이라 주장한다. 하지만, 도덕적 인공지능이 사회속에서 인간과 상호작용할 경우에는, 그것이 인간의 집단 이기주의의 영향에서 완전히 벗어나는 것은 불가능할 것으로 추정한다. 그럼에도 불구하고, 미래에 도덕적 인공지능 자체가 다양한 관계속에서 인간보다 상대적으로 선해질 가능성이 있다는 가정은, 보다 정의로운 미래를 위한 일말의 희망을 준다.

주제어: 도덕적 인공지능, 라인홀드 니버, 웬델 월러치, 기독교 현실주의, 덕