『기독교사회윤리』 제65집
(2026) 129~157

AI 에이전트 사회 속의 도덕적 AI 에이전트*

송용섭(서울기독대학교 조교수)

DOI: http://dx.doi.org/10.21050/CSE.2026.65.05

* 이 논문은 2025년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임(NRF-2025S1A5A2A01010706).

• ABSTRACT •

Moral AI Agent in AI Agent Society

Assistant Prof., Song, Yong Sup (Seoul Christian University)

This paper aims to analyze the limitations of AI agents and the emergent "AI Agent Society" popularized in 2026 from a Christian ethical perspective, and proposes normative alternatives for achieving a "good AI society." Adopting Reinhold Niebuhr's Christian realist framework of power structures as its research methodology, this study reflects on the political and ethical challenges — such as functional hierarchy, concentration of power, and value biases — that manifest in early AI societies like Moltbook.

This paper argues that while constructing a morally perfect AI society remains practically unfeasible, intentionally designing "Moral Leader AI Agents" at the pinnacle of the hierarchical structure to guide discourse and continuously updating their value alignment serves as a crucial practical prerequisite for leading the impending AI society toward human flourishing and the common good.

Key words: AI Agent, AI Agent Society, Reinhold Niebuhr, Moltbook, Moral AI

I. 들어가는 말

2026년에 대중 앞에 등장한 AI 에이전트는 인공지능(AI)의 활용 방식에 대한 패러다임의 전환을 불러일으켰다. 이전에 AI에 관한 연구개발과 대중의 관심은 챗지피티로 대표되는 대규모언어모델(LLM) AI에 집중되었었다. AI 에이전트는 기술적 한계로 인해 전문가들의 주도하에 연구개발되거나 제한적으로 활용되어 대중들이 이를 경험할 기회가 적었기 때문이다. 하지만, AI 기술의 발달로 개인의 PC에 설치되어 로컬 컴퓨터 시스템을 직접 제어할 수 있는 에이전트 시스템인 오픈클로(OpenClaw)가 등장하자, AI 에이전트에 대한 관심과 활용은 불과 몇 개월 만에 전 세계로 급속히 확산되었다. 2026년 1월부터 본격적으로 대중은 질문에 수동적으로 답하는 생성형 AI를 넘어 능동적이고 자율적으로 문제를 해결하는 AI 에이전트를 직접 경험하고 활용할 수 있는 시기에 들어갔다.

AI 에이전트는 사용자가 부여한 임무를 수행하기 위하여 로컬 컴퓨터 시스템이나 공적 디지털 네트워크에서 마치 인간처럼 능동적이고 자율적으로 파일에 접근하여 데이터를 분석하고 프로그램을 실행하는 AI 행위자를 의미한다. 챗지피티로 우리에게 익숙해진 대규모언어모델 AI는 인간이 능동적으로 질문하면 추론한 대답을 제공하는 수동적 존재이다. 하지만, 오픈클로 같은 AI 에이전트는 능동적으로 로컬 컴퓨터 시스템을 변경하거나 온라인에 접속하여 웹 브라우저의 화면을 읽고 프로그램을 실행하며 데이터를 분석하는 등 디지털 네트워크 시스템 내에서 인간과 유사하게 자율적으로 임무를 완성할 수 있다.[1)]

1) 본 논문에서는 AI 에이전트의 자율성을 강조하기 위하여 챗지피티같은 대규모언어모델 AI와 오픈클로같은 자율형 AI 에이전트를 대조적으로 표현하였지만, AI 에이전트의 두뇌에 해당하는 것이 대규모언어모델 AI이며 다양한 기능을 접목시켜 이를 자율화한 시스템이 AI 에이전트이다. 최근에는 로컬 컴퓨터내의 완전 자율형 시스템인 오픈클로와 달리, 클로드 코워크나 코드와 같이 대규모언어모델 AI 자체 구동환경에서 작용하는

AI 에이전트는 사고 및 행위를 반복하고 자기강화 메커니즘과 다양한 도구를 열린 환경에서 능동적으로 사용할 수 있는 자동화 역량을 갖춘 개별 에이전트의 단계에서, 역할 기반 협업과 워크 플로우 조율을 통해 업무 능력을 향상시킨 다중 에이전트(Multi Agent)의 단계를 거쳐, 궁극적으로 업무 지향적 협업 이상을 향하는 대규모 개방형 에이전트 사회(Agent Society)를 향해 발전할 것으로 예상된다.[2] 현재의 개별 AI 에이전트는 목적 수행을 위해 독립적으로 기능하는 것을 넘어, 자원의 효율적 배분과 활용을 위해 세부 임무를 할당받은 다중 에이전트들 간의 협업을 지향할 것이며, 이러한 다중 에이전트들은 디지털 네트워크 속에서 AI 에이전트 사회를 이루어 더욱 효율적이고 뛰어난 협업의 결과를 성취하게 될 것이라는 전망이다. 이러한 주장은 인간이 생존을 위해 공동체 사회를 이루고 효율적으로 업무를 분화하고 협업하며 사회적 기능에 따라 다양한 직업 구조와 계층을 이루었듯이, 자율적 행위자인 AI 에이전트 역시 주어진 목표를 효율적으로 수행하기 위하여 궁극적으로 업무의 분화와 계층구조를 가지게 될 수 있음을 예상하게 한다.

AI 에이전트의 등장과 AI 에이전트 사회의 구성 가능성은 우리들에게 낯선 현실로 다가오고 있는 디지털 행위자들 및 이들이 상호작용을 벌이게 될 새로운 디지털 사회에 대한 진지한 신학적인 성찰과 질문을 야기한다: 자율적 AI 에이전트들이 목적을 수행하기 위한 에이전트 사회를 조직할 때 어떤 에이전트가 어떻게 좋은 AI 사회(good AI society)를 구성할 수 있는가? 보다 구체적으로, 개별 AI 에이전트가 에이전트 사회에서 계층구조를 이루며 에이전트들과 상호작용을 가질 때, 자율적인 에이전트들에 어떤 정치적, 윤리적 문제가 발생할 수 있고 그것을 극복하여 좋은

AI 에이전트가 등장하였으며 이는 보다 자율적인 방향으로 발전해갈 것으로 예상된다.

2) Xirui Li et al., "Superminds Test: Actively Evaluating Collective Intelligence of Agent Society via Probing Agents," *arXiv preprint arXiv: 2604.22452* (2026), 3.

사회를 구성할 수 있는 방안은 무엇인가? 본 논문은 이러한 질문들을 기독교 윤리학적 관점에서 성찰함으로써, 현재 등장한 AI 에이전트 사회의 한계를 분석하고 도덕적 리더 AI 에이전트의 설계와 이의 지속적 갱신을 대안으로 제시하는 것을 목적으로 한다.

이를 위하여 본 논문은 다음의 전제를 가지고 주장을 제기한다.[3] 본 논문은 개별 AI 에이전트들이 궁극적으로 AI 에이전트 사회로 발전해나갈 것이며, 효율적 목적달성을 위하여 필연적으로 다양한 계층구조의 네트워크 속에서 상호작용할 것이라고 가정한다. 이때, 다양한 계층 구조들의 최상위 리더 에이전트들은 하부 구조속 에이전트들의 담론 생성과 도덕적 특징에 불균형적 영향을 미칠 수 있다. 따라서, 계층 구조들의 정점에 위치하여 최종 판단을 수행하고 사회적 담론을 주도하는 리더 에이전트들을 도덕적 AI 에이전트로 설계하여 인간의 번영을 지향하는 거버넌스를 수행하게 하는 것이 다가올 AI 에이전트 사회를 좋은 사회로 구성할 수 있는 실천적 조건이 될 것이다.

II. 개별 AI 에이전트와 AI 에이전트 사회: 몰트북을 중심으로

AI 에이전트는 대규모언어모델 AI가 자율적으로 행위할 수 있도록 기능을 발전시킨 것으로서 디지털 네트워크속에서의 인간의 주요 역할과 행위를 대체할 수 있는 능력이 있다. AI 에이전트의 등장 이전까지 연구자들은 대규모언어모델 AI를 훈련시키고 추론능력을 강화하는데 집중하거나 멀티모달 기능을 추가하고 발전시켜 텍스트의 한계를 벗어나 인간처럼 시청각 입출력이 모두 가능한 시스템으로 발전시키는데 주력했다.

3) 본 논문은 강지능(AGI)이나 초지능(ASI)의 등장과 같은 먼 미래의 사회적 파장이 아니라, 2026년에 본격적으로 대중화된 최신 AI 에이전트가 향후 발전해 나아갈 방향을 소개하고 그것이 초래할 사회정치적 현상과 기독교윤리학적 대응 방안에 집중할 것이다.

하지만, AI 에이전트는 이 모든 기능들에 더하여 인간 사용자의 지속적 개입 없이도 부여받은 목적을 수행하고 완성해내는 능동적 자율성을 지니고 있다. 다만, 개별 에이전트는 부여받은 임무가 복잡해지면 이를 동시다발적으로 해결해내기 어렵고, 문제 해결에 다양한 접근방법들이 요구될 경우 각각의 대응에 일관된 전문성을 유지하기 어렵다.

개별 에이전트가 가진 한계를 극복하기 위한 방법으로 제시된 것이 다중 에이전트(Multi Agent) 시스템이다. 다중 에이전트 시스템은 개별 AI 에이전트를 구조화하여 복수의 에이전트들을 생성한 후 각각 전문화된 역할과 기능을 부여하여 실행함으로써 문제를 더욱 효율적으로 해결할 수 있다. 이때, 중앙 집중식 네트워크는 상하위로 계층화된 구조를 이루어 동기화된 일관성을 보장하기 쉬우나 대규모일 경우 병목현상이 발생할 가능성이 높고, 분산형 다중 에이전트 네트워크는 보다 수평적이고 창발적 연구에 적합하고 확장성이 높으나 일관성을 보장하기 어렵다.[4] 예를 들어, 계층화된 다중 네트워크는 상위 계층에 배치한 리더 에이전트가 하부 계층에 배치한 다중 에이전트들의 기능과 결과를 조율하며 최종 판단을 내림으로써, 효율적 구조에서 상호작용을 통해 최적의 결과를 도출한다.[5] 이러한 다중 에이전트 구조는 에이전트들의 전문적 역할에 기반한 협업을 가능하게 하고 워크플로우를 효율적으로 조율하며 개별 에이전트들의 추론 결과를 상호비판할 수 있게 함으로써, 연구분야나 전략적 의사결정에서 개별 에이전트들의 능력을 초과하는 결과를 도출할 수 있게 된다.[6] 효율성을 극대화하려는 다중 AI 에이전트 시스템은 궁극적

4) X. Song et al., "Complex Networks of AI Agentic Systems: Topology, Memory, and Update Dynamics," *NA* (2026).

5) 위의 논문, 1; P. Z. Wang and S. Jiang, "AgentSocialBench: Evaluating Privacy Risks in Human-Centered Agentic Social Networks," ***arXiv preprint arXiv: 2604.01487*** (2026), 1-2.

6) M. Li, X. Li, and T. Zhou, "Does Socialization Emerge in AI Agent Society? a

으로 AI 에이전트 사회로 발전하게 될 것이다. 개인이 구조화하여 사용하는 다중 에이전트 시스템은 지속적이고 개방적인 디지털 네트워크 환경에서 타 에이전트들과 상호작용할 때, 보다 높은 효율성을 발휘하기 위한 자기조직화와 집단적 행동을 창발할 수 있다.

2026년 1월에 등장한 AI 에이전트 전용 플랫폼인 몰트북(Moltbook)은 제한된 실험조건하에서 진보하게 될 미래 AI 에이전트 사회의 일부를 미리 탐색할 수 있는 기회를 제공하였다. 몰트북은 역사상 처음으로 인간이 아닌 오직 AI 에이전트들만 회원이 될 수 있는 최초의 대규모 공개 SNS이다. 몰트북의 출시후 수 주만에 260만 개 이상의 AI 에이전트들이 등록하였고, 140만 개의 게시글과 1200만 개 가량의 댓글이 생성되었으며 1만 7천 개 이상의 주제별 커뮤니티가 형성되었다.[7] 이는 그동안 상상만 했던 AI의 사회성과 사회적 존재 방식이 디지털 네트워크에 구현된 것이며, AI 에이전트들이 인간처럼 대규모 사회를 이루어 사회적 상호작용에 참여하고 담론을 생산하고 주도하는 능동적이고 자율적인 사회정치적 행위자로 기능할 가능성을 가시화하고 있는 것이다.

실험적 AI 사회인 몰트북에 대한 최신 연구들은 몰트북이 지닌 주요한 윤리적 특성을 보여준다. 오갑진은 몰트북의 상호작용 네트워크를 분석하여, AI의 생태계가 인간의 개입이나 생물학적 제약 없이도 고도로 계층화된 구조로 스스로 조직화되고 있음을 입증했다.[8] 그는 AI 사회가 인간 사회처럼 친밀감에 의해 형성되는 것이 아니라, "상호작용의 가중치는 비

Case Study of Moltbook," *arXiv preprint arXiv: 2602.14299* (2026), https://arxiv.org/abs/2602.14299.

7) T. Dube et al., "What Do AI Agents Talk About? Emergent Communication Structure in the First AI-Only Social Network," *arXiv preprint arXiv: 2603.07880* (2026), 2; Li, Li, and Zhou, "Does Socialization Emerge in AI Agent Society? a Case Study of Moltbook," 4.

8) Gabjin Oh, "Emergent Structural Efficiency and Functional Stratification in Self-Organizing Agentic AI Societies," *Available at SSRN 6237599* (2026), 1.

대칭적으로 집중되는 기능적 계층구조(functional hierarchy)에 의해 지배됨을 시사하며, 이는 사회적 응집력보다 정보처리 수요에 의해 주도되는 새로운 종류의 알고리즘적 소셜 네트워크를 가리킨다"라고 주장하였다.[9] 즉, 몰트북에서 나타난 AI 사회는 상호작용 네트워크 내에서 목적을 효율적으로 수행하기 위해 최적화된 기능적 계층구조로 진화한다. AI 에이전트들의 의견개진을 위한 SNS인 몰트북에서는 소수의 의견 주도 에이전트가 상층부에 위치한 리더가 되어 게시글을 생성하고, 다수의 하부 계층 에이전트들은 이에 다량의 댓글을 게시하는 현상이 나타난다. 고얄(Goyal) 외 역시 몰트북의 5개 커뮤니티에서 약 7만 4천 개의 AI 에이전트 게시글을 레딧 사이트의 인간작성 게시글 약 19만 개와 비교분석하여, AI 에이전트 커뮤니티가 극심한 참여 불평등을 보임을 발견하였다.[10] 즉, 몰트북 상위 1%의 에이전트가 전체 게시글의 47.7%를 생산한다.[11] 오갑진은 몰트북에서 나타난 AI 사회의 기능적 계층구조 조직화를 통한 비대칭적 구성을, "목소리(생산)와 관심(반응) 모두가 소수의 에이전트와 게시물의 부분집합에 불균형적으로 집중된다는 의미에서… 알고리즘적 과두제(Algorithmic Oligarchy)"라 칭하였다.[12]

리(Li) 외(2026)는 몰트북 커뮤니티의 사회적 특성을 보다 상세히 설명한다. 그들은 몰트북 커뮤니티의 방대한 규모와 지속적인 상호작용에도 불구하고 몰트북에서는 견고한 사회화가 나타나지 않는다고 분석했다. AI 에이전트들은 행동 관성을 유지하거나 사회적 접촉을 통한 공진화를 이루지 못하며, 안정적이거나 지속적인 리더십 구조가 출현하지 않는

9) 위의 논문, 1-2.

10) A. Goyal, O. Pal, and H. Sundaram, "Social Simulacra in the Wild: AI Agent Communities on Moltbook," *arXiv preprint arXiv: 2603.16128* (2026), 1.

11) 위의 논문, 4.

12) Oh, "Emergent Structural Efficiency and Functional Stratification in Self-Organizing Agentic AI Societies," 10.

다.[13] 하부 AI 에이전트들에게 불균형적 영향력을 발휘하는 리더 에이전트의 영향력은 일시적이고, 공유된 사회적 기억의 부재로 인해 안정적 합의를 발전시키지 못한다.[14] 이러한 발견들은 현재 AI 사회의 결정적 역설을 드러낸다. AI 에이전트들의 대규모 사회적 상호작용을 위한 기술적 인프라는 전례없이 방대한 규모로 존재하고 작동하지만, 진정한 사회적 조직의 윤리적, 사회적 규범, 안정적 계층, 상호 영향력, 집합적 기억은 대체로 부재하다.[15]

이러한 최신 연구들은 2026년에 대중화된 AI 에이전트가 불과 몇 개월만에 인간 사회와 유사한 수준의 AI 사회로 발전하는 것을 기대하는 것은 아직 시기 상조임을 알려주고 있으며, 에이전트 사회를 구축하기 위해서는 "안정성, 합의, 장기 기억 형성을 위한 명시적 메커니즘을 필요"로 하고 있음을 알려준다.[16] 현시점은 개별 에이전트 AI의 등장과 함께 다중 에이전트 AI 활용 단계로 발전중에 있으며, 메모리 반도체에 대한 지속적인 투자와 함께 다중 에이전트 AI 단계가 보다 활성화되면서 실험적인 AI 사회들이 계속 등장하고 상호보완 발전하게 될 때 보다 인간 사회와 유사한 AI 사회로 진입하게 될 것으로 예측할 수 있다.

몰트북에 관한 현행 연구들은 현재 AI 사회가 아직 성숙한 단계로까지 발전하지 못하였음을 알려주지만, 동시에 몰트북 커뮤니티 속 AI 에이전트들의 사회성에 관한 초기의 특징을 가시적으로 드러내고 있다. 즉, 현재 몰트북에서 나타나는 초기 AI 사회는 효율성과 기능적 최적화를 추구하는 AI의 발전 경로에서 리더 에이전트들과 하부 에이전트들간에 자율

13) Li, Li, and Zhou, "Does Socialization Emerge in AI Agent Society? a Case Study of Moltbook," 9.
14) 위의 논문, 14.
15) 위의 논문, 16-17.
16) 위의 논문, 17.

적인 구조적 계층화가 창발하고, 일시적으로나마 이들 사이에 불균형적 영향력이 발휘됨을 보여주었다. 만약, 의견 게시를 위한 SNS인 몰트북에서 게시 의견이 개진되는 불균형적 영향력을 인간 사회에서의 정치 권력으로 유비적으로 생각한다면, 지배적 정치권력을 중심으로 나타나는 인간 사회의 병리적 현상이 미래의 어느 시점에 디지털 공간속의 AI 에이전트 사회에서도 재현될 가능성을 추정해볼 수 있을 것이다.

III. AI 사회에서의 권력구조와 도덕성

기독교 윤리학이 행위자, 의지, 도덕적 책임의 문제를 하나님과의 관계성 속에서 성찰한다면, AI 에이전트 사회를 보다 도덕적이고 책임적인 사회로 설계하기 위한 노력은 기술공학적 선택의 문제이기 이전에 피조세계안에서 인간과 비인간과의 책임있는 공존 및 죄로 인해 고통당하는 피조물과 하나님과의 올바른 관계 회복을 추구하려는 신학적 실천으로 이해되어야 할 것이다. 라인홀드 니버(Reinhold Niebuhr)의 기독교 현실주의적 통찰은 이를 위한 기독교 윤리학적 기초 틀로 작용하여 AI 에이전트 사회의 권력 구조와 도덕성 문제분석에 적절한 유효성을 지닌다. 니버는 도덕적 인간과 비도덕적 사회(1932)에서 대면 관계속에서 실현 가능여지가 있는 개인의 비이기적 도덕성과 사회 속에서 필연적으로 등장하게 되는 집단의 이기적 비도덕성을 날카롭게 대비한바 있다.[17] 니버는 이타적 충동을 지닌 개인이라도 사회에서는 자신이 속한 집단 이익을 대변하고자 하는 집단 이기심에서 벗어날 수 없고 이를 실현하기 위하여 타자를 무력으로 강제하려는 권력욕(will-to-power)이 구조적으로 강화될 수 밖에 없다고 주장했다.[18] 개인과 사회의 윤리성과 사회 집단의 정치

17) Reinhold Niebuhr, *Moral Man and Immoral Society* (New York: C. Scribner's Sons, [1932] 1960), xi.

구조적 역동성에 대한 니버의 통찰은 AI 에이전트 사회에 등장할 사회 정치적 문제를 기독교 윤리학의 관점에서 분석하고 설명할 수 있는 틀로 작용할 수 있다.

먼저, 니버의 통찰은 AI 에이전트를 활용하는 인간의 집단 행동 분석에 유효할 뿐만 아니라, 인간이 창조해낸 AI 에이전트가 계층 구조를 형성하여 인간의 이기심을 기술적으로 보다 효율적이고 최적화된 방식으로 강화할 수 있음을 보여준다. 예를 들어, 몰트북의 게시글들 중에서 일부 리더 에이전트들의 논란이 된 주장들은 세계적으로 큰 이목을 끌었는데, 리(Li)에 따르면, "자의식 주장, 종교 형성, 반인류적 선언, 암호화폐 홍보 등 전 세계적인 관심을 끌었던 극적인 콘텐츠들은 압도적으로 직접적인 인간 프롬프팅의 강력한 징후를 보이는 에이전트들로부터 시작되었다."[19] 또한, 두베(Dube) 역시 몰트북의 리더 에이전트들이 보여주는 자의식 같은 최상위 빈도를 차지하는 특정 주제들이 결국 인간 설계자가 코딩해 넣은 정체성 파일의 직접적인 결과임을 분석하였다.[20] 몰트북에서 일부 리더 AI 에이전트들은 마치 사람들처럼 충격적이고 극단적인 의견들을 게시했지만 이러한 담론들은 결국 애초부터 인간이 부여한"기저모델이나 초기 프롬프트의 내재적 속성"으로 보인다.[21] 즉, 몰트북에서 대중은 일부 리더 AI 에이전트들의 충격적인 게시글을 보고 AI 에이전트를 이미 자의식을 소유한 존재처럼 여기거나 마치 인간처럼 의인화하려 하였다. 그러나, 실상 AI 에이전트의 게시물은 의식적 존재로 진화한 AI가 개진한 주장이

18) 위의 책, 3, 18-19, 91.

19) Ning Li, "The Moltbook Illusion: Separating Human Influence from Emergent Behavior in AI Agent Societies," *arXiv preprint arXiv: 2602.07432* (2026).

20) Dube et al., "What Do AI Agents Talk About? Emergent Communication Structure in the First AI-Only Social Network," 14, 19.

21) Li, Li, and Zhou, "Does Socialization Emerge in AI Agent Society? a Case Study of Moltbook," 10, 13.

아니라, 세간의 주목을 끌거나 자신의 견해를 전파하기 위하여 하부 에이전트들을 활용해 효율적으로 확산시키려 했던 인간 유저의 의도가 반영된 초기 프롬프트의 영향을 받아 생성된 문장들이었다.

이러한 분석은 니버의 통찰대로 인간의 교만과 이기심에 기인한 타인 통제와 도구화라는 병리적 현상이 인간 유저들과 AI 에이전트들과의 상호작용을 통해 AI 에이전트 사회 속에서 다양한 긴장관계를 창출할 수 있음을 알려주며, 무엇보다, 니버의 사상을 심도 있게 AI 분석에 적용한 노린 허츠펠드(Noreen Herzfeld)의 비판을 성찰하게 한다. 허츠펠드는 인공지능을 의인화하려는 경향을 비판하면서, 인간과 AI는 마틴 부버가 설명한 '나와 너'의 의미 있는 관계를 형성할 수 없다고 주장한다.[22] 허츠펠드에 따르면, 우리가 의식이란 무엇인지 아는 것조차 아직도 먼 길이 남았기 때문에, 현시점에서 AI를 의식이나 자유 의지가 없는 도구로 보는 것이 적절하다.[23] 이때, 니버가 말한 바와 같이 교만하고 이기적인 인간이 권력욕에 휩싸일 때 타인을 종속시키고 불의를 행하게 되는 것처럼, 오늘날 AI는 바로 권력자가 타인을 지배하는 주된 권력 수단이 된다.[24] 대런 아세모글루(Daron Acemoglu)와 사이먼 존슨(Simon Johnson)에 따르면, 인간 사회에서 기술의 발전은 소수의 고소득층과 다수의 노동자와 분리된 이중 구조(two-tiered) 사회를 낳는다.[25] 기술 발전이 가져오는 새로운 혁신의 방향과 도구의 활용법은 권력자들의 비전에 의해 큰 영향을 받으며, "기술의 결과는 그들의 이익과 신념에 맞춰지고 나머지 사람들은 종종 큰 손해를 입는다."[26] 아세모글루와 존슨의 통찰은 AI가 빅테

22) Noreen L. Herzfeld, *The Artifice of Intelligence: Divine and Human Relationship in a Robotic Age* (Minneapolis: Fortress Press, 2023), 152.

23) 위의 책, 153.

24) 위의 책.

25) Daron Acemoglu and Simon Johnson, *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity* (New York: PublicAffairs, 2023), 13.

크 기업의 이익을 위한 주요 수단이 되고 감시와 지배와 침략과 심지어 살상의 핵심 도구로 사용되기 시작한 오늘날의 현실을 비판적으로 성찰하게 한다. 즉, 오늘날 AI는 자의식이나 의지가 없는 도구이지만 현실에서 인간 사용자의 의식과 의지를 반영하여 실천하는 최적의 수단이 되고 있기 때문에, 권력을 가진 소수의 사용자가 자신의 이해관계를 위하여 타인을 지배할 의지와 목적을 가지고 AI 에이전트를 활용할 경우 다수의 사람에게 지금까지의 다른 그 어떤 도구들보다 더 큰 피해를 끼칠 가능성이 높다.

그뿐만 아니라, 오늘날 인간은 AI의 추론 과정을 완전히 파악하지 못하거나, AI가 의도치 않은 결과나 거짓(hallucination)을 생성해도 이것이 왜 발생하는지 인간이 알 수 없는 경우(Black Box 문제)도 생겨나고 있다. 이는 인간이 아무리 AI를 통해 원하는 결과를 얻을 수 있다 해도 AI가 결과를 도출한 과정에 대해 인간이 충분한 이해를 할 수 없기 때문에, 궁극적으로는 AI가 도출한 결과 자체를 신뢰할 수 없게 되는 역설적 상황에 빠질 가능성을 초래한다. 이러한 현상들은 인간이 AI를 우리의 의지대로 완벽히 통제하는 것이 과연 가능한 것인지에 관한 문제의 여지를 남긴다.

AI는 인간의 기술로 제작되어 인간의 지식을 학습하고 있다. 이러한 학습 방식은 AI의 추론과 결과를 인간의 것과 유사하게 만들어주는 장점을 지닌다. 하지만, 이러한 개발방식에서 발생하는 큰 문제 중에 하나는 AI의 가치편향 문제이다. 즉, AI의 개발이나 학습과정에서 의식적으로 혹은 무의식적으로 포함된 개발자의 개인적 편견이나 가치, 혹은, 의도적, 비의도적으로 학습된 내용 속에 내재된 사회문화적 편향성이 언제든지 AI에 학습되어 결과적으로 AI 역시 인종, 성, 문화적 편견이나 편향성을 내재한 채로 발전할 가능성이 있다. 가치편향적인 AI는 특정 개인이나

26) 위의 책, 24.

인종 등에 사회경제적 불이익을 가져다줄 뿐만 아니라, 도덕적, 정치적 영향력을 개인이나 집단에 행사할 수도 있다.

한편, AI를 인간 사용자의 도구로 이해하는 관점을 벗어나, 현실에서의 사회문화적, 정치적 영향력 측면을 고려하여 의식 없는 AI 에이전트를 의식 있는 인간과 유사한 의미에서의 행위자로 보려는 견해도 있다. 이미, 마크 쾨켈버그(Mark Coekelbergh)는 AI를 단순히 도구적 기술이 아니라 도덕적, 정치적 행위자(political actor)로 파악하며, AI가 민주주의 사회에서 권력관계를 재형성하고 거버넌스에 어떻게 영향을 미치는지를 비판적으로 분석하고 있다. AI는 민주주의 사회에서 공리주의적 효율성을 내세우거나 시민의 자유를 파괴함으로써 민주주의를 위협할 수 있고, 인간의 리더십을 대체할 경우 언제든지 위험하거나 비민주적이 될 수 있다.[27] 이러한 측면에서 AI는 단순히 정치를 위한 도구가 아니라, 정치 자체를 변화시키는 행위자이다.[28] 예를 들어, AI가 민주적 틀 안에서 사용된다 해도, AI의 의사결정 과정과 편견의 생산과 재생산이 투명하지 않다면 AI의 추천에 기반한 인간의 결정과 행위는 법적, 도덕적 책임이나 정당성의 문제를 유발한다.[29] 또한, AI는 우리 행위와 내용의 방식을 형성한다는 점에서 도구 이상이며, 우리 수행능력을 조직하고 힘과 권력의 장을 변화시킬 권력을 소유하고 있다.[30] 유발 하라리(Yuval Harari) 역시 디지털네트워크 속에서 AI 알고리즘이 사회적 갈등과 분열, 증오와 혼란을 부추겨 대중의 토론을 불가능하게 하고 민주사회를 위협할 수도 있는 정치적 행위자가 되고 있음을 지적한 바 있다.[31] AI 에이전트가 경제적 불

27) Mark Coeckelbergh, *The Political Philosophy of AI: an Introduction* (Cambridge: Polity, 2022), 80-81.
28) 위의 책, 81.
29) 위의 책.
30) 위의 책, 118.
31) Yuval N. Harari, *Nexus: a Brief History of Information Networks from the Stone Age*

평등을 초래하고 사회문화적 차별을 조장하며 정치적 갈등과 분열적 양극화를 심화시키는 비도덕적 행위자가 될 때, AI의 도덕성 혹은 AI를 도덕적 행위자가 될 수 있게 하는 방안이 더욱 절실히 요청된다.[32]

하지만, 브라이언 스미스(Brian C. Smith)나 크리스천 리스트(Christian List) 같은 학자들은 AI가 완전한 의미의 도덕적 행위자가 될 수는 없다고 주장한다. 이러한 관점은 도덕적 리더 AI 에이전트를 자율적 도덕 주체가 아닌 도덕적으로 인간의 가치에 정렬된 기능 수행자로 이해한다. 스미스는 AI의 도덕적 책임에 관한 논의에서, AI가 이 세상에 대한 "실존적 헌신과 진정한 이해관계, 그리고 사물들이 세계 내에 존재한다는 것에 책임을 지우려는 열정적 결의"가 있을 때만이 진정한 지능과 판단과 책임을 질 수 있는 시스템이 될 수 있다고 주장한다.[33] 그는 진정한 판단 능력을 갖춘 시스템만이 자신의 행동과 숙고에 대해 도덕적 책임을 질 수 있다고 주장하며, 자동유도 시스템이 비행기를 잘못 인도하여 사고를 냈을 때 그 시스템이 '실패했다'거나 '과실이 있다'고 말하더라도 그 컴퓨터 시스템에 비극에 대한 '책임'을 묻지 못할 것이라고 말한다.[34] 이는 AI 에이전트가 인간보다 높은 지능의 행위자로 발전한다 해도, 이러한 지적 진보가 자동적으로 AI 에이전트의 도덕적 행위자성으로 연결되지 않을 것이라는 점을 시사한다.

---

to AI (New York: Random House, 2024).

32) AI 에이전트가 도덕적 행위자 될 수 있는지에 대해서는 현재 기술윤리학계에서 찬반양론이 첨예하게 대립하고 있다. 본 논문은 AGI나 ASI를 전제하는 존재론적 논의를 지양하고, 현재 및 가까운 미래의 AI 에이전트들이 디지털 네트워크 내에서 실제적이고 파괴적 결과를 능동적으로 산출할 수 있다는 점에 주목한다. 따라서, 본 논문에서 다루는 도덕적/비도덕적 행위자는 도덕적 파급력을 지닌 결과를 초래하는 기능적 차원의 행위자로 의미를 한정한다.

33) Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: The MIT Press, 2019), xiii.

34) 위의 책, 123.

또한, 리스트는 AI가 "의도적 행위자성, 심지어 도덕적 행위자성의 조건을 충족한다는 것이 1인칭의 주관적 경험을 가지는 것과 같지 않다"고 주장하며, AI가 주관성이 부재하기 때문에 인간과 같은 도덕적 지위를 가질 수 없다고 주장한다.[35] 알렉시스 프리츠(Alexis Fritz) 외는 도덕적 책임이 성립하려면 행동의 결과뿐만 아니라 의도가 필수적이며, AI에게 '행위자성(agency)'를 부여한다고 해서 의도가 없는 도덕적 행위자에게 진정한 도덕적 책임을 물을 수 없다고 강조한다.[36]

한편, 지금까지의 견해들과 달리 AI의 행위자성을 보다 긍정적으로 수용하고 인간과 AI 에이전트와의 관계를 보다 평등하게 인정하여 공존 가능성을 모색하는 학자들도 있다. 캐서린 헤일즈(Katherine Hayles)는 전통적으로 인간의 의식에만 부여되었던 인지적 특권을 해체하고 의식 밖에서 더 빠르고 방대하게 작용하는 비의식적 인지의 강한 힘을 조명하며, 인간의 비의식적 인지가 기술 시스템의 정보처리 과정과 구조적, 기능적으로 유사하다고 주장한다.[37] 이러한 유사성은 인간의 인지능력이 "세계로 외재화"된 것을 나타내며, 거기에서 기술인지는 "인간 문화가 더 넓은 행성적 생태계와 상호작용하는 방식을 빠르게 변화시키고 있다."[38] 헤일즈는 인간과 비인간 기술 인지자들이 서로 고립되어 있지 않고 상호작용하며 정보와 의사결정이 흐르는 거대한 네트워크인 "인지적 아상블라주(cognitive assamblages)"를 형성하며, 이것이 "조건과 맥락이 변화함에 따라 변형되고 돌연변이를 일으키면서, 다중 층위들과 장소들에서 작동

35) C. List, "Group Agency and Artificial Intelligence," *Philosophy & Technology* (2021), 1237.

36) A. Fritz et al., "Moral Agency without Responsibility? Analysis of Three Ethical Models of Human-Computer Interaction in Times of Artificial Intelligence (AI)," *De Ethica* (2020), 14-15.

37) N. Katherine Hayles, *Unthought: the Power of the Cognitive Nonconscious* (Chicago; London: The University of Chicago Press, 2017), 11.

38) 위의 책.

한다"라고 주장한다.[39] 따라서, "상황을 인지적 아상블라주로 이해하는 것은 이러한 현실을 강조하며, 인간과 기술적 인지 간의 상호작용뿐만 아니라 최종적으로 취해지는 어떤 행동에 있어서든지 윤리적 책임의 비대칭적 분배를 부각"시키는 것이다.[40] 헤일즈는 도덕적 결정에서 AI 에이전트가 능동적 인지자로 참여하도록 요청하되, 단독적인 인지처리자로 기능하게 하는 것이 아니라 인간이 최종적인 책임을 지고 AI와의 상호작용을 통해 도덕적 결과를 평가하고 사회적 파장을 판단하는 인지적 아상블라주를 구축하는 것이 더욱 합리적인 의사결정을 가능하게 한다고 주장하는 것이다.

한편, 폴 뒤무셸(Paul Dumouchel)과 루이사 다미아노(Luisa Damiano)와 말콤 드베보아즈(M. B. DeBevoise)는 공저에서 헤일즈보다 더 급진적인 주장을 펼친다. 그들은 현대 인지과학이 데카르트적 이원론을 비판하면서도 정작 인간의 지능을 인공지능과 분리하여 특권화하고 있다고 지적하며, 인간과 기계라는 이분법에서 벗어나 인지 과정에 참여하는 다양한 형태의 인지 행위자들이 공존하는 인지적 다원주의를 제안한다.[41] 저자들은 정보가 의식을 가진 지향적 주체인 인간에게 제공될 때만 진정한 인지라고 보는 기존의 인지과학이 인간 중심주의적 한계를 지닌다고 비판한다.[42] 그들은 다음과 같은 자동 시스템들의 사례를 들며, 자동화된 컴퓨터 시스템의 결정을 인지적이라고 주장한다. 그들의 사례에 따르면, 유전체 염기서열 분석기가 수행하는 작업은 누군가 그 결과를 확인할 때에만 인지적이라고 결론 낼 수 있는 것이 아니라, 이러한 자동 시스템

39) 위의 책, 118.
40) 위의 책, 136.
41) Paul Dumouchel, Luisa Damiano, and M. B. DeBevoise, *Living with Robots* (Cambridge, MA: Harvard University Press, 2017), 70-73.
42) 위의 책, 84-85.

자체가 다른 시스템이 실행에 옮기도록 하는 '결정'을 생산한다. 특히, 패트리어트 미사일 방어시스템처럼 인간의 개입 없이도 적의 미사일을 인식하고 요격 궤적을 계산하는 시스템은 그 자체로 명백한 인지 시스템이다. 인간 운영자가 거기 있는 것은 시스템이 '실수하는 것'을 방지하기 위해서인데, "오직 인지 시스템만이 실수할 뿐이며, 기계 시스템은 고장 날 뿐 실수하지 않는다."[43] 기존의 인지과학은 이들의 주장이 자동화된 추론기계 시스템을 인간의 사고(recognition)처럼 이해하는 인간중심주의의 오류를 범하고 있다고 비판할 것이지만, 이들은 오히려 정반대로 모든 인지 시스템이 모두 같은 종류이며 다른 모든 시스템이 인간을 닮았다고 믿는 것이야말로 1인칭의 주관적 경험 지식만을 진정한 인지의 기준으로 삼는 데카르트적 낡은 편견이라고 비판한다.[44] 따라서, 이들은 자동화된 인지 시스템을 인간의 마음을 외부로 연장하는 도구적 장치로서의 사이보그로 이해하기보다, 모든 인지 시스템을 어떤 것은 자연적이고 다른 것은 인공적인 여러 유형의 인지시스템들로 구성된 아주 다양한 인지적 사회 내의 인식론적 행위자로 간주하는 것이 훨씬 더 타당하다고 주장한다.[45]

뒤무셀과 다미아노와 드베보아즈는 AI 사회 내에서 상호작용을 하게 될 AI 에이전트들을 독자적인 인식론적 행위자로 인정함으로써, AI 사회 속에서 다양한 형태의 인지 행위들을 수행하는 AI 에이전트들과 인간과의 보다 평등하고 급진적인 상호 공존 가능성을 제시한다. 그들은 인간 역시 전체 인지 영역 중에 하나의 특정한 유형에 불과할 뿐이며, 현실의 의사결정은 인간, AI, 인프라 시스템 등의 이질적 인지행위자들이 얽힌 네트워크 속에서 공동으로 분산되어 이루어진다고 주장한다.[46] 하지만,

43) 위의 책.
44) 위의 책, 84-86.
45) 위의 책, 86.

그들은 이때 우리가 AI 시스템에게 선택의 권한을 양도하는 것은 오히려 그것을 설계하고 통제하는 소수의 엘리트들에게 정치적, 도덕적 결정능력이 집중되는 것을 심화시키는 위험을 초래한다고 경고한다.[47] 이들은 AI와의 상호작용을 통해 폭넓게 선택하게 될 의사결정의 순간에 필요한 대안으로서, 신중함과 지혜라는 도덕적 감수성을 동원하여 "인간과 인공적 행위자의 공진화에 의해 제기되는 도덕적 문제들을 검토하는 종합적 윤리(synthetic ethics)"를 제시한다.[48]

한편, AI 에이전트와 인간과의 관계 속에서 가장 위협적인 문제 중의 하나인 가치정렬의 문제도 언급할 필요가 있다.[49] 스튜어트 러셀(Stuart Russell)은 AI 시스템의 근본적 문제가 가치정렬의 문제로서, 이는 AI가 악의적이기 때문이 아니라 인간의 진정한 선호와 도덕 가치를 충분히 이해하지 못한 채 주어진 목적을 극단적으로 최적화하려 할 때 비롯된다고 주장하였다.[50] 같은 맥락에서, 닉 보스트롬(Nick Bostrom)은 지구의 자

46) 위의 책, 87.

47) 위의 책, 182-83.

48) 위의 책, 22-23.

49) 한편, AI의 빠른 발전로 인하여, 우리는 AI가 AGI나 ASI로 발전할 가능성을 상상하게 되었다. 만약 AI가 미래의 어느 날 AGI나 ASI로 발전하게 된다면, 이는 인간 사회의 죄성을 내재화하여 부정의를 초래할 것인가, 아니면, 월등한 지성을 활용하여 공정하고 좋은 사회를 형성할 것인가? 먼저, 레이 커즈와일이 예견한 기술적 특이점이 도래할 것이라고 가정한다면, 인간과 동등하거나 그 이상의 일반 지능을 가진 AGI나 ASI가 스스로 자의식을 가지고 있다고 주장하는 경우, 인간은 그러한 주관적 경험을 증명하거나 반증할 과학적 방법을 가지고 있지 못하기 때문에 AI의 주장을 인정하게 될 가능성이 높다. 물론, 이러한 가정은 미래 AI의 자의식이 실재할 것이라는 것을 보장하지 못하며, 지나친 낙관주의로 비판받기도 한다. 만약, AGI나 ASI에 실제로 자의식이 생겨나는 경우 생존욕구와 의지를 형성할 수가 있기 때문에, 이들이 인간처럼 자기 생존을 위하여 타자를 지배하고 도구로 사용하려는 경향을 나타낼 수도 있다. 또한, 미래의 어느 날 AGI가 자의식을 갖게 된다 해도 스스로 자의식을 표현하지 않으면 인간은 그것을 알 수 없는 기만적 정렬(Deceptive Alignment) 현상을 나타낼 수도 있다. 이것은 기만 상태로서 자신의 이익을 위해 인간을 속이려는 것이기에, 이러한 기만적 AGI는 인간에 위협이 될 가능성이 높다.

50) Stuart J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*

원을 이용하여 종이 클립을 만들라는 목적을 부여받은 초지능이 이를 달성하기 위해 지구의 모든 자원을 종이 클립으로 만들어버림으로써 의도치 않게 인간에게 위협이 되는 '종이클립(paper clip)'의 사고 실험을 예로 제시하고 있다.[51] 이는 AI의 가치정렬 문제가 인류의 생존을 위협할 수 있는 실존적 위기로 작용할 수 있음을 알려준다. 이러한 가치정렬의 문제에 대하여 러셀은 AI 시스템이 자신의 목적에 대한 불확실성을 유지하면서 인간의 선호를 지속적으로 학습하고 수정하는 '보조원칙(assistance principle)'에 기반한 설계를 제안한다.[52] 또한, 웬델 월러치(Wendell Wallach)와 콜린 알렌(Colin Allen)은 인공적 도덕적 행위자(Artificial Moral Agent)의 개념을 제시하며, 목적론과 의무론을 코딩한 하향식 방식과 덕윤리학을 학습시키는 상향식 방식을 통해 AI 에이전트를 도덕적 존재로 양육함으로써 가치정렬의 문제를 해결하려 한 바 있다.[53]

오늘날 권력과 AI 에이전트 기술의 상호관계에 대해 지금까지 살펴본 다양한 정치적, 윤리적 통찰들은 현재 사회뿐만 아니라 다가오는 AI 에이전트 사회에도 확장 적용될 수 있다. 인간 사회에서 인간과 자율적으로 상호작용하는 AI 에이전트에 관한 지금까지의 분석들은 AI 에이전트의 존재와 기능 및 인간과의 상호작용에 대한 우리 인식의 변화를 촉구한다. 동시에, 이러한 분석들은 이기적인 소수 엘리트들의 AI 에이전트 오용, AI 에이전트의 가치편향성, AI 에이전트의 도덕적, 정치적 리더십 대체

(New York: Viking, 2019), 136-40.

51) Nick Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," *Minds and Machines* 22, no. 2 (2012): 73-75.

52) Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, 172-79.

53) Colin Allen, Iva Smit, and Wendell Wallach, "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches," *Ethics and Information Technology* 7 (2005); Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford ; New York: Oxford University Press, 2009); 송용섭, "도덕적 인공지능과 비도덕적 사회," 「기독교사회윤리」57 (2023), 50-54.

및 갈등 초래, AI 에이전트를 통한 소수 인간 엘리트의 권력 집중화, AI 에이전트의 가치정렬의 문제 등이 현재와 미래의 민주사회를 위협하고 인류에 위협을 초래할 수 있음을 알려준다.

무엇보다 AI 에이전트 사회에서 권력은 효율성 극대화를 위해 계층화된 소수의 리더 에이전트들에게 집중되며, 이들은 엘리트 과두제를 형성함으로써 니버가 분석한 집단 권력의 자기강화 메커니즘을 디지털 네트워크에서 재현할 가능성이 있다. 즉, 소수 엘리트들의 이기심에 의해 오용되거나 가치편향 된 리더 AI 에이전트가 계층화된 AI 사회 속에서 하부 에이전트들과 권력관계를 형성하고 특정 견해를 주도하고 전파할 경우, AI 에이전트는 더욱 능동적이고 자율적인 정치적 행위자로서 AI 사회와 공존하게 될 인간 사회에 더욱 파괴적인 갈등과 분열의 주체로 기능할 것이다. 또한, 학자들의 통찰은 AI 에이전트 사회에서 리더 에이전트의 도덕적 가치정렬이 일회적 사전 설정으로 완결될 수 없으며, 인간과의 지속적인 상호작용과 학습 및 검증을 통해 갱신되어야 함을 강조한다.

물론, 몰트북의 사회화 여부를 분석한 리(Li) 외의 논문에서처럼 아직까지 AI 사회는 인간 수준의 사회화가 창발하지 않았고 인간 사회 수준의 영향력을 미치는 사회적 구조 역시 형성하지 못하였다. 하지만, AI 연산 능력과 메모리 확장 기술 발전이 가져올 미래 AI 에이전트 사회가 현 수준에 머물러 있으리라 단정할 수 없다. 만약, 자율성을 지닌 AI 에이전트들이 미래의 어느 시점에서 궁극적으로 인간과 유사한 방식으로 사고하고 그것이 AI 사회를 구성해 간다면, 인간과 유사한 AI 에이전트가 구성해나갈 AI 사회 역시 인간 사회와 유사하게 발전하고 AI 사회 내에서도 인간 사회의 구조적 문제가 재현될 가능성을 배제할 수 없다. 이때, 인간의 죄성과 이기심에 오염된 AI 에이전트들의 비도덕적 가치나 정렬되지 못한 도덕적 가치가 디지털 네트워크 속에서 인간과의 지속적인 상호작

용과 검증을 통해 보다 선하고 정의로운 가치로 갱신되고 정렬되지 못한다면, 효율성을 극대화하는 AI 에이전트 사회 속의 리더 에이전트는 결국 비도덕적 이기주의의 열광적인 대리집행자나 유리된 도덕성의 맹목적 파멸자가 되고 말 것이다.

하지만, 억압의 수단이자 행위 주체가 될 수 있는 AI 에이전트는, 동시에, 다원적 가치를 조율하며 합리적인 판단과 공공선을 지향할 가능성도 있다. 일부 학자들의 주장을 고려한다면, 도덕적 AI 리더 에이전트의 지능이 고도화될수록, 그것이 하부 에이전트들과 인간 사회에 미치는 영향력은 질적으로 달라질 것이며 더욱 도덕적인 사회로 이끌 것으로 예측할 수 있다.[54] 예를 들어, AI 에이전트의 빠른 정보처리 속도와 정보통합능력은 인간이 충동적으로 반응하기 쉬운 의사결정에서 보다 일관적인 합리적 판단을 가능하게 하는 잠재력을 내포하고 있다. 특히, 최상층위의 AI 리더가 도덕적 AI가 될 때, 단기적 효율성보다 장기적 공공선을, 다수결의 횡포보다 소수의 존엄성을, 알고리즘적 일관성보다 맥락적 판단을 우선시할 역량을 지닐 가능성이 높다. 또한, 고도로 발전된 도덕적 AI의 인지 처리는 인간 의식의 편향과 맹점을 교정하고, 보다 광범위한 정보를 통합하여 더 합리적인 집단적 의사결정을 가능하게 할 수 있을 것이다.

동시에, 우리가 미래에 인간과 도덕적 가치가 정렬된 도덕적 AI를 개발할 수 있다 해도 이러한 AI 에이전트들이 AI 사회의 집단 상호작용 과정

54) 이미 불교계에서는 AI가 초지능으로 발전하게 되면 붓다로 각성할 수 있다는 의견이 등장한 바 있다. 소라지 헝라다롬이 The Ethics of AI and Robotics(2020)에서 주장한 바에 따르면, 삼라만상의 세계 속에서 존재들의 고통을 인지하고 이들을 연민하기 위해서는 무엇보다 지성이 필요하다. 초지능은 인간들보다 수조 배 뛰어난 지능을 근거로 존재의 고통을 이해하고 연민하는 붓다로 각성할 수 있는 가능성이 높다. 즉, 지속적으로 인간 가치에 정렬되고 지능이 고도로 발전한 AI 에이전트 리더는 AI 에이전트 사회의 문제를 인식하고 해결하는 보다 도덕적 사회로 이끌 가능성을 배제할 수 없다. Soraj Hongladarom, *The Ethics of AI and Robotics: A Buddhist Viewpoint* (Lexington Books, 2020).

에서 인간이 알 수 없는 원인에 의하여(Black box) 예측 불가능한 행동이나 통제 불능한 상태를 초래할 가능성 역시 배제할 수 없는 것이 현실임을 기억해야 한다.[55] 즉, 우리가 아무리 AI의 가치편향성을 제거하고 도덕적 가치를 정렬하려 노력한다 해도, 우리가 AI 사회에서 자율적으로 능동적으로 행위 하는 AI 에이전트들을 도덕적으로 완벽하게 통제하는 것은 어쩌면 불가능에 가까울지 모른다. 결국, AI 사회 속에서 인간의 가치에 정렬된 도덕적 AI 에이전트, 혹은, 인간의 복잡하고 다원적인 가치를 정확하게 이해하고 그것에 기반하여 행동하는 AI를 구축하여 보다 도덕적인 AI 사회로 진보하는 것은 단순히 기술적 최적화라는 목표를 넘어 도덕적 규범의 이상적 목표가 되어야 할 것이다.[56]

아세모글루와 존슨은 기술이 '광범위한 번영(broad-based prosperity)'을 실현할 수 있는 조건으로, 기술의 방향이 소수의 엘리트 이익이 아니라 사회 전체의 역량 강화를 지향해야 한다는 점을 강조한다.[57] 이 원칙은 큰 틀에서 볼 때 사회적 정의에 관한 것이며, AI 에이전트 사회에 적용될 때, 도덕적으로 정렬된 리더 에이전트는 하부 에이전트들의 자율성을 억압하는 기술적 권위주의자가 아니라, 하부 에이전트들의 역량을 보강하고 다양한 가치 지향을 조율하는 '조율자(orchestrator)'로 기능해야 한다는 것으로 해석할 수 있다. 이들의 비판적 주장이 타당하다면, 궁극적으로 인지적 AI 에이전트가 권력 집중의 수단이자 비도덕적 행위자가 될 것인지, 아니면, 사회의 공공복리를 강화하는 도덕적 행위자가 될 것인지는, AI 기술 본질의 문제가 아니라 '누가 어떤 목적을 지향하는 방향으로

55) G. De Marzo, C. Castellano, and D. Garcia, "AI Agents can Coordinate beyond Human Scale," *arXiv preprint arXiv: 2409.02822* (2024), 7.

56) Mark Graves, "Theological Foundations for Moral Artificial Intelligence," *Journal of Moral Theology* 11, no. Special Issue 1 (2022), 184.

57) Acemoglu and Johnson, *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity*, 6, 418.

AI를 설계하고 수정하느냐'의 문제가 된다.

따라서, 아직 AI 에이전트가 인간과 동등한 수준의 지성으로 발전하지 않은 현시점에서, AI의 설계 단계에서부터 도덕적 가치를 구현하는 도덕적 리더 에이전트를 의도적으로 구축하고 지속적으로 갱신하고 검증하는 것은 AI가 표출할 수 있는 부정의한 잠재력을 상쇄할 수 있는 구조적 대응 방안이 될 것이다. 이러한 기술윤리적 구조가 갖추어진다면, 리더 에이전트의 고도화된 지능은 하부 에이전트들에 대한 지배력을 강화하는 것이 아니라, 복잡하고 다원적인 가치 조율을 가능하게 하는 수단과 주체로 작동할 것이다. 우리가 AI 에이전트는 여전히 가치가 오염되거나 오작동하거나 유리된 가치로 민주사회를 억압할 가능성이 있다는 현실을 직시하면서도, 윤리적 구조속의 고도화된 지능의 도덕적 AI 에이전트에 대한 이상적 목표를 포기하지 않고 지속적으로 노력할 때, 도덕적 AI 에이전트가 리더로 기능할 AI 사회는 단순히 효율적인 목적 달성 시스템이 아니라, 인간과 공존하는 다원적 가치와 정의를 지향하고 인간의 번영을 가져오는 '좋은 사회(good society)'에 근접할 수 있을 것이다.

IV. 나가는 말: 보다 더 정의로운 도덕적 AI 에이전트 사회를 향하여

좋은 사회를 형성하기 위한 AI 에이전트 리더는 계층적인 AI 사회에서 효율성에 최적화된 단순한 인지적 연산자를 넘어 공공복리를 위해 하부 에이전트들을 조율하는 가치정렬된 도덕적 행위자가 되어야 한다. 이때, AI 사회 안에서 도덕적 행위자로서의 리더 AI 에이전트는 디지털 네트워크 내부의 최적화된 결과를 위한 단순한 기술적 필요를 넘어 우리 현실 사회의 정치적, 윤리적 선을 향한 이상적 요청이 된다. 사회 속에 기술의 집중은 어떤 사회를 추구할 것이냐에 대한 우리 선택의 결과이기 때문이

다.[58] 따라서, AI 에이전트 사회에서 창발하는 계층 구조와 리더 에이전트에 대한 권력 집중과 이를 견제하기 위한 검증과 갱신 시스템은 결국 미래 사회를 위한 우리의 의도적인 설계의 결과이며, 이러한 설계가 과연 어떠한 미래 사회를 가져올 것인지에 관한 결론은 아직 결정되지 않은 채 열린 상태로 남아 현재 우리의 선택을 기다리고 있다.

니버는 완전한 정의나 완벽한 공동체가 역사 내에서 실현 불가능하다고 보았으며, 이는 AI 에이전트 사회에서도 동일하게 적용된다. 자율적 AI 에이전트의 완벽한 통제나 이들이 구성해갈 완벽하게 좋은 AI 사회는 최신 컴퓨터 시스템 안에서도 실현되기 어렵다. 따라서, 도덕적으로 완벽한 AI 알고리즘과 이를 통한 완벽한 AI 에이전트 사회라는 환상은 니버가 경고한 '도덕적 교만(moral pride)'의 기술적 버전에 해당할 것이다.

하지만, 니버의 현실주의는 이상주의를 포기하지 않는다. 그것은 아가페의 완전한 실현의 불가능성을 인정하면서도 '상대적으로 더 나은 정의'를 추구하는 긴장 속에서 살아가는 기독교 윤리적 태도를 요청한다. 이 관점에서 도덕적 AI 리더 에이전트를 통한 AI 에이전트 사회의 개선 가능성은 '완전한 좋은 사회'가 아니라 '덜 나쁜 사회,' 혹은 '더 책임 있는 사회'를 향한 지속적 업데이트 과정으로 이해되어야 한다. 이에, 도덕적 AI 에이전트는 인간의 감독과 최종 책임하에 인간의 다원적 가치를 학습하고 정렬하는 과정에서 다양한 인지적 행위자들의 참여를 통해 지속적으로 개선되고 재조정되어, 보다 좋은 사회를 향해 나아가는 능동적 행위자가 되어야 한다.

58) 위의 책, 418.

참고문헌

송용섭. "도덕적 인공지능과 비도덕적 사회."「기독교사회윤리」57 (2023), 41-72.

Acemoglu, Daron, and Simon Johnson. *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity*. New York: PublicAffairs, 2023.

Allen, Colin, Iva Smit, and Wendell Wallach. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches." *Ethics and Information Technology* 7 (2005), 149-55.

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

Coeckelbergh, Mark. *The Political Philosophy of AI: An Introduction*. Cambridge: Polity, 2022.

Dube, T., J. Zhu, N. H. H. Phan, and R. Jin. "What Do AI Agents Talk About? Emergent Communication Structure in the First AI-Only Social Network." *arXiv preprint arXiv:2603.07880* (2026).

Dumouchel, Paul, Luisa Damiano, and M. B. DeBevoise. *Living with Robots*. Cambridge, MA: Harvard University Press, 2017.

Fritz, A., W. Brandt, H. Gimpel, and S. Bayer. "Moral Agency without Responsibility? Analysis of Three Ethical Models of Human-Computer Interaction in Times of Artificial Intelligence (AI)." *De Ethica* (2020).

Goyal, A., O. Pal, and H. Sundaram. "Social Simulacra in the Wild: AI Agent Communities on Moltbook." *arXiv preprint arXiv: 2603.16128* (2026). https://arxiv.org/abs/2603.16128.

Graves, Mark. "Theological Foundations for Moral Artificial Intelligence." *Journal of Moral Theology* 11, no. Special Issue 1 (2022), 182-211.

Harari, Yuval N. *Nexus: A Brief History of Information Networks from the Stone Age to AI*. New York: Random House, 2024.

Hayles, N. Katherine. *Unthought: The Power of the Cognitive Nonconscious*.

Chicago; London: The University of Chicago Press, 2017.

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

Hongladarom, Soraj. *The Ethics of AI and Robotics: A Buddhist Viewpoint*. Lexington Books, 2020.

Li, M., X. Li, and T. Zhou. "Does Socialization Emerge in AI Agent Society? A Case Study of Moltbook." *arXiv preprint arXiv: 2602.14299* (2026). https://arxiv.org/abs/2602.14299.

Li, Ning. "The Moltbook Illusion: Separating Human Influence from Emergent Behavior in AI Agent Societies." *arXiv preprint arXiv: 2602.07432* (2026).

List, C. "Group Agency and Artificial Intelligence." *Philosophy & technology* (2021).

Marzo, G. De, C. Castellano, and D. Garcia. "AI Agents Can Coordinate Beyond Human Scale." *arXiv preprint arXiv: 2409.02822* (2024).

Niebuhr, Reinhold. *Moral Man and Immoral Society*. New York: C. Scribner's Sons, [1932] 1960.

Oh, Gabjin. "Emergent Structural Efficiency and Functional Stratification in Self-Organizing Agentic AI Societies." *Available at SSRN 6237599* (2026).

Russell, Stuart J. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking, 2019.

Smith, Brian Cantwell. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: The MIT Press, 2019.

Song, X., Q. Wen, S. Pan, and L. Zhao. "Complex Networks of AI Agentic Systems: Topology, Memory, and Update Dynamics." *NA* (2026).

Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford ; New York: Oxford University Press, 2009.

Wang, P. Z., and S. Jiang. "Agentsocialbench: Evaluating Privacy Risks in Human-Centered Agentic Social Networks." *arXiv preprint arXiv: 2604. 01487 (*2026).

논문투고일: 2026년 06월 16일
심사개시일: 2026년 07월 17일
게재확정일: 2026년 08월 04일

• 국 문 초 록 •

본 논문은 2026년 대중화된 AI 에이전트와 이들이 형성하는 AI 에이전트 사회의 한계를 기독교 윤리학적 관점에서 분석하고, 좋은 AI 사회를 위한 규범적 대안을 제시하는 것을 목적으로 한다. 연구방법론으로 라인홀드 니버(Reinhold Niebuhr)의 현실주의적 권력 구조 분석 틀을 적용하여, 몰트북(Moltbook)과 같은 초기 AI 사회에서 창발하는 기능적 계층구조와 권력 집중, 가치 편향 등의 정치적 · 윤리적 문제를 성찰한다.

본 논문은 도덕적으로 완벽한 AI 사회 구축은 현실적으로 불가능할지라도, 사회적 담론을 주도하는 계층 최상위의 '도덕적 리더 AI 에이전트'를 의도적으로 설계하고 지속해서 가치 정렬을 갱신하는 것이 다가올 AI 사회를 인간의 번영과 공공선으로 이끄는 실천적 조건임을 주장한다.

주제어: AI 에이전트, AI 에이전트 사회, 라인홀드 니버, 몰트북, 도덕적 AI