JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 4, pp. 909-918, Apr. 2024

Check for updates

인공지능의 윤리적 자율성 검토와 공익적 시사점

전 찬 영[1] · 방 정 배[2] · 박 균 열[3*]
[1]경상국립대학교 대학원 윤리교육학과 박사수료
[2]한국열린사이버대학교 국방상담리더십과 객원교수
[3*]경상국립대학교 윤리교육과 교수

Review of the Ethical Autonomy of Artificial Intelligence and Its Public Implications

Chan-Young Jun[1] · Jung-Bae Bang[2] · Gyun-Yeol Park[3*]

[1]Ph.D. Candidate, Department of Ethical Education, Gyeongsang National University, Jinju 52828, Korea
[2]Guest Professor, Department of Defence Counseling Leadership, Korea Open Cyber University, Seoul 02087, Korea
[3*]Professor, Department of Ethical Education, Gyeongsang National University, Jinju 52828, Korea

[요 약]

본 연구는 인공지능의 자율성에 대한 논의를 토대로 인공지능의 윤리적 역량 구비 방안을 제시하는 데 목적을 두었다. 또한 이러한 논의를 토대로 다양한 범용인공지능(AGI)을 위한 몇 가지 시사점을 도출한다. 이를 위해 본 연구는 도덕적 행위자, 자율적 인공지능이라는 용어에 대한 대안을 제안하고, 인공지능의 목적과 지위를 정립하며, 윤리적·법적 책임소재를 명확하게 하는 등 인공지능에 대한 인간의 통제방법을 검토하였다. 이 연구는 인간의 자율성에 근접한 인공 지능의 자율성을 대체할 개념과 윤리적 역량을 구비한 인공지능을 개발하기 위한 접근 방안을 모색하고, 인공지능의 도덕성 검증을 위한 '설명가능 인공지능'의 적용과 도덕판단역량검사도구(MCT)의 알고리즘의 기본원리를 활용하여, 향후 윤리적 기능을 장착한 다양한 '범용인공지능(AGI)의 온톨로지 구축에 의미 있는 기여를 할 것으로 기대된다.

[Abstract]

This study aims to suggest a plan for establishing ethical capabilities in artificial intelligence based on discussions on the autonomy of artificial intelligence. Additionally, based on these discussions, the implications for various 'Artificial General Intelligence(AGI)' are derived. This study proposes an alternative to the terms moral agent and autonomous artificial intelligence, establishes the purpose and status of artificial intelligence, and examines human control methods for artificial intelligence, including clarifying ethical and legal responsibilities. This study seeks an approach to develop artificial intelligence with ethical capabilities and a concept to replace the autonomy of artificial intelligence that is close to the human autonomy, and the application of 'explainable artificial intelligence' to verify the morality of artificial intelligence. By utilizing the basic principles of the algorithm of the Moral Competence Test(MCT), it can contribute to building an ontology of various AGI equipped with ethical functions in the future.

색인어 : 인공지능, 자율적 도덕 행위자, 인공지능윤리, 도덕판단역량검사도구, 범용인공지능

Keyword : Artificial Intelligence, Autonomous Moral Agents, Artificial Intelligence Ethics, Moral Competence Test(MCT), Artificial General Intelligence(AGI)

http://dx.doi.org/10.9728/dcs.2024.25.4.909

Received 23 February 2024; **Revised** 19 March 2024
Accepted 01 April 2024

Corresponding Author; Gyun-Yeol Park
Tel: +※ 개인정보 표시제한
E-mail: pgy556@hanmail.net

Ⅰ. 서 론

1956년 미국 다트머스 대학에서 열린 컨퍼런스에서 처음 등장한 '인공지능'(Artificial Intelligence)의 개념은 사람들에게 무한한 가능성이 펼쳐질 미래에 대한 기대를 가지게 했었지만, 인공지능을 소재로 한 공상과학 소설이나 SF영화와 같은 새로운 세계를 단시간 내에 열어주지는 못하였다. 그러나 21세기에 이르러 급속도로 발달한 컴퓨터의 연산능력, 네트워크 기술, 데이터 처리 기술과 같은 과학기술은 인공지능 기술의 희망을 다시 갖게 하였다. 특히, 1997년 IBM의 딥블루(Deep Blue)는 체스 세계챔피언 가리 카스파로프(Garry K. Kasparov)와의 대결에서 승리하였고 2011년에는 IBM의 Watson이 Jeopardy Show에서 인간 경쟁자를 물리치고 100만 불의 상금을 획득하기도 하였다. 2016년에는 딥마인드 알파고와 이세돌 9단의 바둑 대국에서 알파고가 승리하여 먼 미래에서도 인간을 이길 수 없을 것으로 예상되었던 바둑에서 조차 인공지능이 사람의 능력을 추월하였음을 확인할 수 있었다. 이러한 인공지능 기술의 발전은 머신러닝 기술의 개발로 귀납적 경험에 의한 학습이 가능해지고, 빅데이터의 적용으로 방대하고 유의미한 정보의 결합이 이루어져 과거 알고리즘에 의한 연역적 연산 능력에 크게 의존하여 가졌던 인공지능 기술의 경직성과 복잡성의 한계를 극복하였기 때문이다. 또한, 스마트 모바일 기기와 사물인식인터넷(IoT)의 확산은 인공지능의 활용 영역을 활발히 확대시키고 있다. 이와 같이 인공지능은 점차 다양한 전문분야에서 인간보다 우수한 능력을 발휘하고 그 신뢰성 역시 증대되고 있다[1].

본 연구는 인공지능 기술에 대한 다양한 윤리적 고찰 중에서 인공지능의 자율성에 주목한다. 단순한 알고리즘의 다층구조로 실제 인간과는 많은 차이가 있었던 인공지능 기술이 빅데이터를 이용하여 경험을 축적하고 이를 학습하면서 인간과 유사한 능력으로 발전을 하게 되자 커즈와일(Raymond Kurzweil)이 주장한 특이점 도래에 대한 기대는 더욱 커져가는 한편 인공지능 기술의 비윤리적 활용 가능성에 대한 우려 역시 부각되고 있다[2]. 특히, 그 중에서 자율성을 가지고 스스로 판단하여 행동하는 인공지능이 개발되어 윤리적 문제를 야기할 때의 대응방안, 나아가 인공지능이 스스로 새로운 규칙을 정하고 확장하여 인간을 위협하고 지배하게 되는 시나리오까지도 제기되는 등 인공지능의 자율성은 인류 차원의 중대한 문제로 인식되고 있다.

본 연구는 인공지능의 자율성을 윤리학적 관점에서 인공지능 기술의 윤리적 지위를 검토하고, 인공지능의 자율성을 대체할 개념과 접근 방안을 제시하고자 한다. 나아가 이러한 연구에 기반하여 윤리적 역량을 갖춘 범용인공지능(AGI)을 활용한 게임 제작에 있어서 참고해야 할 점을 제시하고자 한다.

Ⅱ. 인공지능과 윤리

2-1 인공지능 개념과 발전 추세

인공지능은 시대의 기술 발전 수준과 개발 목표에 따라 다양하게 정의되어왔으나 대체적으로 인간의 사고와 추론 능력을 본떠 합리적으로 판단하고 행동하는 기술 또는 연구로 정의된다. 대표적으로 커즈와일은 "사람이 수행했을 때 지능을 요구하는 기능들을 수행하는 기계를 창조하는 기술", 윈스톤은 "지각과 추론, 작용을 가능하게 하는 계산의 연구"라고 정의하였다[3]. 즉, 인공지능의 개념에는 합리성의 증가, 달리 말하면 결과를 향상시킬 수 있는 역량의 증가와 그 사유가 인간 친화성이든 인간의 사고 및 추론 과정이 우수하기 때문이든 인간과 유사한 방식의 프로세서를 구축하도록 하는 것이 인공지능이라는 의미가 포함되어 있다[4].

이러한 인공지능의 개념을 반영한 듯 인공지능은 정보처리 능력을 향상시키는 한편 인간의 인식 및 추론 방식을 모방하며 발달하였다. 인공지능 개발의 패러다임은 3기에 걸쳐 발달하고 있다. 먼저 1기는 합리론적 방법이 주가 되어 인간의 지식을 기호화하여 기계에 주입하고 이를 논리로 구성하고자 하였고 이 시기에는 보다 정밀하고 복잡한 알고리즘의 설계가 인공지능 기술을 향상시켰다. 이어서 2기는 연결주의 패러다임으로 다양한 데이터로부터 스스로 학습하여 경험을 축적하였고 이 시기에는 빅데이터, 머신러닝 등이 사용되었다. 새로운 패러다임인 3기의 인공지능은 기계 고유의 연역적 특성과 학습에 의한 경험적 특성의 균형을 유지하며 환경과 맥락에 따라 상호작용하며 마치 인간의 뇌인지 구조와 유사한 형태로 발달하고 있다. 이러한 발달 과정을 거친 인공지능 기술은 점차 그 능력을 입증하고 있다. 전문 분야에 국한되었던 인공지능은 과학 연구의 영역을 넘어서며 점차 일반적인 인간의 삶 속에 밀접해지기 시작하였다. 시리(Siri)나 구글 AI와 같은 음성 비서로부터, 챗봇, 자율주행 차량, 스마트 가전제품, 사람과 제한적인 감정을 교류하는 로봇, 서비스 직무를 수행하는 로봇과 같이 일상에서 인간을 도와 편의를 제공할 뿐 아니라 인간과 흡사한 감정 교류의 영역까지 인공지능의 영역이 확장되고 있다. 이러한 추세를 반영하듯 오늘날의 인공지능은 인간과 유사한 일반 지능을 가진 컴퓨터 프로그램이나 로봇을 뜻하는 일반 인공지능 또는 강한 인공지능과 일반적인 지능은 부족하지만 아주 특수한 지능적 행동을 할 수 있는 특화인공지능 또는 약한 인공지능으로 용어를 세분하기도 한다[5].

2-2 자율성의 윤리학적 개념

자율(autonomy)은 그리스어에 어원을 둔 '자기 자신'을 뜻하는 αὐτο와 '법'을 의미하는 νόμος의 합성어로, 자기 자신에게 스스로 법을 부여한다는 의미를 가진다. 루소에게 자율

은 외적인 강제나 물리적인 억압에 의한 것이 아닌 인간 본성 자체에 의해 스스로 결정하는 행동만을 의미하는 것이며, 자연에 따라 살아가며, 편견과 권위로부터 벗어나 자신의 기본적인 자연권을 지키며 자기 보존을 그 기본권으로 삼는 생활을 말한다. 루소는『에밀』에서 인간은 자율적 이성 사용 능력에 기초해 평등한 인간관계 속에서 시민으로서의 의무와 권리를 행사하는 독립적이고 자율적인 인간을 길러내야 한다고 하며 도덕적 자유를 주장하였다[6]. 밀은『자유론』에서 자유의 기본 영역인 내면적 의식의 영역의 자유, 자신의 기호를 즐기고 자기가 희망하는 것을 추구할 자유, 결사의 자유는 타인에게 해를 입히지 않는 한 절대적으로 보장되어야 한다고 하였다. 칸트는『실천이성비판』에서 인간은 신성한 도덕법칙의 주체이며 인간 이성이 '선의 이념'에 따라 자신에게 부여한 의무이자 규범인 도덕법칙을 준수하며 그 안에서의 자유의지를 가진다고 하였다[7].

자율성이 가지는 공통적 요소는 자기결정적이고 자치적으로 행동하는 것과 도덕적 원리에 따라 합리적이고 이성적으로 판단하고 행동하는 것을 의미한다. 인간의 '지능'이 인공지능의 '지능'이 지향해야 할 유일한 목표가 아니더라도 윤리적 주체로서의 인공지능의 자율성에 관해서는 인간의 자율성의 요소를 기준으로 해야 한다. 따라서 인간의 자율성의 요소를 기준으로 인공지능이 구현할 수 있는 자기결정 및 자치적 행동, 도덕적 원리에 따른 합리적이고 이성적인 판단 수준을 검토해 보는 것은 윤리적 행위자로서의 인공지능을 논의하는 것에 매우 중요한 의미를 지닌다.

Ⅲ. 인공지능의 자율성에 대한 검토

3-1 인공지능의 자율성에 대한 기존 논의

기존의 인공지능의 자율성에 대한 논의의 주요 내용은 인공지능의 주체를 어떻게 정의할 것인가이다. 특히 인공지능의 자율성에 있어서 법률적 접근은 인공지능이 법적 주체로서 지위를 부여해야 하는지와 행위의 위법성에 대한 책임의 주체는 누구인지, 인공지능의 법적 통제 방안을 검토하는 방향으로 연구가 활발히 진행되고 있다. 한편 철학적 접근은 로봇의 범주 차원의 존재론 즉 인공지능이 인간을 포함한 생물과 무생물 어느 곳에 범주화하는 것이 합당한지를 규명하는 것에서부터 시작된다. 윤리학적 접근에서는 윤리적 행위자로서의 인공지능의 도덕적 역량의 수준과 윤리적 행위자의 도덕적 추론의 원리를 설정하고 구현하는 것에 중심을 두고 있다.

윤리학적 접근은 가장 융통성 있게 인공지능 기술에 적용할 수 있다. 윤리학적 접근은 인공지능 기술이 구현할 수 있는 행위자의 도덕적 수준을 평가하는 척도를 제공하고 도덕적 행위자로서 요구되는 수준과 방향을 제시하는 문제 해결 중심의 성격을 갖기 때문이다. 대표적인 예로 제임스 무어(James Moor)는 윤리적 행위자를 범주화하는 위계적 체계를 제시한다. 가장 낮은 수준은 윤리적 영향 행위자(ethical impact agent)로 매 행동에 대한 윤리적 결과를 평가받는다. 다음 수준은 내재적인 윤리적 행위자(implicit ethical agent)로 설계자가 비윤리적 결과 방지를 고려하여 설계한 것이다. 그다음 수준은 명시적인 윤리적 행위자(explicit ethical agent)이며 윤리적 추론과 사고를 위해 윤리적 범주를 프로그래밍한 것이며 마지막으로 가장 높은 수준인 완전한 윤리적 행위자(full ethical agent)는 형이상학적인 의식, 자유의지 등을 활용하여 도덕적 판단을 할 수 있는 일반적 성인 수준의 행위자이다. 무어는 윤리적 행위자로서의 로봇이라면 명시적인 윤리적 행위자가 되어야 한다고 제시한다[8].

3-2 기존 논의의 한계점

이러한 다양한 측면의 연구에도 불구하고 인공지능 윤리에 대한 논의는 추상적이고 선언적이다. 표 1과 같이 아시모프의

표 1. 인공지능윤리의 주요 원칙들
Table 1. Major principles of artificial intelligence ethics

division	Main Content
Isaac Asimov's "Three Laws of Robotics"[9]	1. A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law. 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law
IEEE's "General Principles"[10]	1. Embody the highest ideals of human rights. 2. Prioritize the maximum benefit to humanity and the natural environment. 3. Mitigate risks and negative impacts as AI/AS evolve as socio-technical systems.
Asilomar's "AI Principles"[11]	1. Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible. 2. Failure Transparency: If an AI system causes harm, it should be possible to ascertain why. 3. Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority. 4. Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications. 5. Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation. 6. Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

‘로봇 3원칙’으로부터 국제전기전자기술자협회 표준화기구(IEEE)의 인공지능 시스템의 원칙, 인공지능 개발과 관련하여 가장 세부적인 가이드라인을 제시하고 있는 아실로마 AI 원칙조차도 인공지능 기술 개발의 명확하고 구체적인 윤리적 지침을 제시하지 못하고 있다.

이러한 가이드라인을 구체적으로 기술에 적용하기 어려운 한편 광범위한 해석이 가능하다. 가장 큰 문제는 윤리적 책임 주체에 대해 유예적 입장을 유지하고 있는 것이다. 인공지능 개발의 윤리적 책임이 개발자에게 부여된 것처럼 표현되어있지만 실제 인공지능이 행위자로서 작동할 때의 명시적 윤리적 주체가 생략되어 있다. 그 중 가장 세부적인 아실로마 AI 원칙의 ‘책임’ 항목에서는 시스템의 설계자에게 ‘이해관계자’라는 지위를 부여하고 있으나 이 또한 매우 모호하다.

Ⅳ. 인공지능의 윤리적 자율성 검토

4-1 자율행위자로서 인공지능의 미충족성

인공지능 기술은 ‘최대한 인간과 근접한 체계를 만드는 것’을 목표로 발달해 왔다[12]. 물론 지능을 인간만의 특성으로 제한할 수 없으나 인간이 가장 높은 수준의 지능을 가졌기 때문일 것이다. 이를 달성하기 위해 과학자들은 인간이 세상을 인식하는 방식과 사고하여 결심하는 방식을 구현하기 위해 연구의 노력을 집중해 왔지만 궁극적으로는 이러한 기술 발전의 방향은 인공지능으로 하여금 최대한의 자율성을 가지도록 하는 것이었다. 하지만 인공지능은 궁극적으로 자율성을 가진 주체가 되기에 부적합하다[13].

먼저 기술적 불완전성과 심리적 거부감이다. 인공지능이 자율성을 가진 주체가 되기 위해서는 우선적으로 기술적 완성도가 매우 높아져야 한다. 물론, 인간의 인공지능이 접하고 처리해야 할 다양한 환경에서의 과제들을 자율적으로 인식하고 판단할 수 있어야 하며, 이를 위해 인식, 분류와 구분, 정보의 추상화 및 구체화를 할 수 있는 능력과 신뢰성을 갖춘 기술이 뒷받침되어야 한다. 머신러닝으로 인해 인공지능 분야가 매우 빠르게 발전하고 있지만 머신러닝을 위해서는 엄청난 양의 데이터와 에너지가 요구된다. 알파고는 한정된 환경에서 정해진 규칙으로 진행되는 게임을 이기기 위해 천문학적인 데이터 자원을 동원하여 반복 학습을 실시하였다[14]. 특히, 인공지능은 언어 기반의 시스템의 한계를 가지고 있다. 결국 인공지능 기술이 향상되고 학습을 통해 빠르게 데이터를 축적한다고 하더라도 정확한 인식과 추론, 판단, 적절한 행위를 하기 위해서는 모든 개념에 대한 온톨로지가 구축이 이루어져야 하나 인간의 감각과 감정, 추론 등은 그 범위가 매우 넓고 언어로 모두 표현되기 어려울 뿐만 아니라 기계학습 결과에 대해서도 적절성을 평가하고 수정해주어야만 한다. 또한, 인간의 인지와 판단은 의미가 다중적이고, 변용성이 있으며, 복합적인 상황인식에 따라 우선순위가 달라진다. 이를 극복하기 위해 최근에는 양자컴퓨터, 휴리스틱 기술이 등장하였지만 이 기술이 가져오는 융통성이 사람들에게는 ‘융통성’ 보다는 ‘불확실성’에 가깝게 받아들여질 것이다. 또 다른 측면으로 인공지능 기술이 매우 발달하여 인간의 역량을 추월한다고 가정하더라도 인공지능을 인간과 동일한 주체로 인정하고 같은 지위의 공유를 인정하는 것은 합리적이고 객관적이기 보다는 정서적 문제이며 결과적으로 훨씬 보수적이어서 예외가 인정되기 어려울 것이다[15].

다음으로 도덕적 불안정성이다. 인공지능이 자율적 행위자가 되기 위해서는 도덕적 자율성을 갖추어야 한다. 도덕적 자율성은 도덕적 영역에만 머무르지 않는다. 인간이 의사판단을 하고 행동하는 전 단계에 걸쳐 밀접한 도덕적 개입이 이루어지는 것과 같다. 또한, 자율적 인공지능의 행위자의 영향이 인공지능 주체에만 해당하지 않고 인간과 생물, 자연과 같이 다양한 범위에 미치게 되어 도덕적 역량을 갖추지 않은 인공지능 주체가 자율성을 갖게 하는 것은 용인되기 어려울 것이다. 하지만 인공지능이 자율적 도덕행위자로서 의사결정을 하게 하는 데에는 크게 두 가지의 난제를 가지고 있다. 첫째, ‘어떤 윤리원칙을 적용할 것인가?’이다. 전통적인 공학윤리의 양대 기둥은 공리주의와 칸트 윤리학으로 대변된다. 인공지능이 개발되는 과정에서 제기된 윤리적 문제에 접근할 때에도 이 두 가지 접근법이 주가 되었다. 그러나 인공지능에게 어떤 원칙을 적용해야 하는가에 대해 합의점을 찾는 것은 쉽지 않다. 공리주의적 원칙은 양적으로 계량하기 어려운 효용을 규정해야 하며, 때로는 부당한 행위를 정당화하는 것처럼 보이기도 한다. 한편, 칸트 윤리학은 예외 없는 명백한 원칙으로 적용될 수 있을 것 같지만, 구체적으로 칸트가 주장하는 보편적 법칙이 어떤 내용을 가져야 하는지에 대해 말해주지 않으며 그 윤리적 의무들이 충돌할 경우의 해소 방법을 제시하지 않는다[16]. 둘째, ‘윤리원칙을 어떻게 구현할 것인가?’이다. 하향식 접근은 인공지능이 매 순간에 어떤 원칙을 적용하여야 하는지를 사전에 연역적으로 제공하는 것이다. 예를 들면 칸트의 정언명령이라든지 황금률을 인공지능의 윤리적 원칙으로 적용하는 것이다. 하지만 위계의 설정이나 원칙 간의 충돌, 과도한 연산 수요 등 여러 가지 문제점이 있다. 상향식 접근은 인공지능이 스스로 사례 데이터를 학습함으로써 원칙을 구현한다. 유연하고 융통성 있는 방법이지만 진화하는 인공지능의 방향을 예측하기 어렵고 잠재적 돌발요소들을 예측하여 확인하기도 어렵다. 마지막으로 하향식과 상향식의 병합 방법은 하향식의 확고한 원칙의 틀 속에서 상향식의 학습을 하는 방법으로 서로의 장・단점을 결합, 보완한 것이지만 실제 그 균형을 찾기도, 구현하기도 어렵다. 이러한 윤리적 원칙 선정과 윤리 원칙의 구현법의 한계는 인공지능에게 ‘자율적 주체’로서의 지위 부여를 더욱 어렵게 한다.

이외에도 인공지능 설계자와 사용자의 의도에 따른 비윤리적 설계/조작 가능성과 폐쇄적이고 복잡한 시스템의 특성으로 내부 고발이 없는 한 비윤리적 인공지능의 영향을 인지하기 어렵다는 점, 인공지능이 스스로를 윤리적 주체로서 의식

하고 책임지고자 하지 않으며 피해에 대한 원상복구와 처벌의 실효에 대한 비례성이 현저히 떨어지는 점은 인공지능이 윤리적 주체로서의 사회적 합의 도출을 더욱 어렵게 한다. 따라서 인공지능에게 자율성을 부여하는 것은 적합하지 않다.

4-2 인공지능의 자율성에 대한 대안적 시도

자율적 특성을 가진 행위자로서의 인공지능에게 자율적 주체로서의 지위를 부여하지 않기 위해서는 개념과 용어의 재검토, 인공지능의 목적과 지위, 윤리적/법적 책임 소재, 통제 방법에 대한 명확한 정립이 이루어 져야 한다.

가장 우선적으로는 '도덕 행위자'와 '자율적' 개념과 용어의 재정립이 필요하다. 앞에서 살펴 본 한계점들에도 불구하고 인공지능에게 도덕적 지위와 자율성을 부여하고자 하는 시도는 계속 되고 있다. 웬델 월러치(Wendell Wallach)와 콜린 알렌(Colin Allen)은 도덕 행위자의 범위가 인간을 넘어 인공지능시스템으로까지 확대될 것이며 이를 인공적 도덕행위자(artificial moral agent, AMA)라고 정의하고 있다[17]. 인공지능을 연구하는 논문, 정부, 정치연합, 기업의 인공지능 개발 관련 가이드라인에도 '자율적(Autonomous)' 인공지능이라는 용어가 공식적으로 사용되고 있다. 인공지능이 실제로 인지, 정서, 행동할 수 있기 때문에 '행위자(Agent)'로 보는 것은 적합하다. 하지만 앞에서 검토한 것과 같이 인공지능은 '도덕적 행위자'가 될 수는 없다[18]. 인공지능은 단순히 행위자이며 굳이 표현하자면 '도덕적 기능이 있는 인공지능'(AI with moral function)이 적합하며 '자율적 인공지능'은 '자동화된 인공지능(Automated AI)' 또는 '하이테크 인공지능(High-tech AI)'으로 표현되는 것이 적합하다.

인공지능의 목적과 지위는 여러 선언적 가이드라인에서 확인한 바와 같이 인간의 행복과 이익을 증가시키며 존엄을 지키는 도구가 되어야 한다. 인공지능은 더욱 발달하고 똑똑해지겠지만 인간에게 종속되는 도구에 불과하며 인간 능력 향상을 통해 인간성의 자기완성을 추구하거나 타인의 행복에 기여하는 목적으로만 사용되어야 한다. 또한, 인공지능 윤리는 공학 윤리의 하위 개념에 불가하며 그 이상의 지위를 갖지 않아야 한다.

인공지능 기술은 실제 그 능력에 비해 과대평가를 받아왔다[19]. 인공지능이 인간의 두뇌를 추월하는 시점에 도달했다고 하지만 인공지능의 폭주를 견제해야 한다고 주장하는 인공지능은 나오지 않는다. 인공지능 도구로써 가져야 할 본래의 지위를 유지해야 한다. 따라서 인공지능은 윤리적/법적 지위를 가질 수 없으며 책임을 질 수도 없다[20]. 인공지능의 행위와 관련된 책임은 오로지 인간에게만 부여되어야 한다. 설계상의 문제는 설계자가, 사용상의 문제는 사용자가 책임을 져야 하며 이를 위해 매 인공지능 체계는 설계자의 철저한 검증을 거쳐 활용되어야 하며, 사용자에게도 오용에 따른 부작용에 대한 투명한 정보가 공개되어야 한다. 더 이상 실현되지 않은 인공지능 기술의 개발 시나리오가 "종국에는 더 이익이 될 것"이라는 낙관적 결과주의의 도피처로 사용되어서는 안 된다[21].

인공지능 기술은 보다 높은 도덕적 수준을 갖추도록 노력해야 하지만 자율성을 높이는 방향이 되어서는 안 된다. 무기체계에서 좋은 예를 찾을 수 있다. 적을 공격하고 파괴해야 하는 무기체계의 효과성을 높이기 위해서는 탐지로부터 공격(Sensor to shooter)까지 반응시간을 줄이는 것이 매우 중요한 관건이며 자동모드를 이용하여 달성이 가능하다.

하지만 무기체계의 기능이 가능함에도 불구하고 공격 무기체계의 경우에는 매 의사결정에 인간이 개입하는 human-in-the-loop 방식을 사용하고 있다. 이는 우군 살상을 방지하고 중복공격을 막는 등의 효과도 있지만 인명을 살상하는 결과를 초래할 선택을 기계인 무기체계에게 전담시킬 수 없기 때문이다. 마찬가지로 인공지능도 윤리적 문제가 발생할 수 있는 결정의 지점에 인간의 통제가 필요하다. 플로리디(Luciano Floridi)와 샌더스(J.W. Sanders)가 주장한 것과 같이 도덕적 책무(accountability)와 도덕적 책임(responsibility)을 구별[22]하여 인공지능이 도덕적 책임은 없더라도 도덕적으로 작동하고 개발되도록 해야 한다. 이를 위해 표 2, 3에서와 같이, 국내·외에서 인공지능의 윤리적 가이드라인 및 정책을 발전시키고 있다. 이와 함께 인공지능과 관련된 국내 특허는 주로 인공지능을 윤리적으로 제어하고 도덕적으로 올바른 수행을 위한 장치 및 시스템 개발과 관련되어 있다[23]-[26]. 국외 특

표 2. 국내 인공지능윤리 가이드라인 및 정책
Table 2. Domestic artificial intelligence ethics guidelines and policy

Year	Ethics Guidelines and Policy
2018	· Kakao announces algorithmic ethics charter · KAIST announces ethics charter for artificial intelligence · Korea Intelligence Information Society Promotion Agency, Intelligence Information Society Ethics Guide
2019	· Korea Artificial Intelligence & Ethics Association announces Artificial Intelligence Ethics Charter · Korea Communications Commission announces principles for a user-centered intelligent information society
2020	· Ministry of Science and ICT, National Artificial Intelligence Ethics Standards(draft) · Ministry of Land, Infrastructure and Transport, self-driving car ethics guidelines
2021	· Korea Artificial Intelligence & Ethics Association, revised Artificial Intelligence Ethics Charter · Ministry of Science and ICT, strategy to realize reliable artificial intelligence · National Assembly proposes a bill on fostering artificial intelligence and creating a foundation for trust(July 1, 21)
2022	· Ministry of Education, Ethical Principles for Artificial Intelligence in Education · National Human Rights Commission, human rights guidelines for the development and use of artificial intelligence · Ministry of Science and ICT, 2022 Trustworthy Artificial Intelligence Development Guide(draft)

표 3. 국외 인공지능윤리 가이드라인 및 정책
Table 3. Foreign artificial intelligence ethics guidelines and policy

Nation	Ethics Guidelines and Policy
EU	2004, 13 principles of robot ethics(EURON) 2007, Rotoethics Roadmap(EURON) 2014, Guidelines on Regulating Robotics 2018, Coordinated Plan On AI 2019, Ethics Guidelines for Trustworthy AI 2020, Assessment List for Trustworthy AI 2021, Artificial Intelligence Act
USA	2016, Big Data: A Report on Algorithmic System, Opportunity and Civil Right 2016, Preparing for the future of AI 2016, The National AI R&D Strategic Plan 2016, Ethically Aligned Design(IEEE) 2017, Asilomar AO Principles 2017, NHTSA. ADS-A Vision Safety 2.0 2020, Guidance for Regulation of AI Application 2020, FTC, Using AI and algorithm
UK	2016, UK House of Common 2018, Data Ethics Framework 2019, A guide to using AI in the public sector 2020, ICO, Guidance on AI and Data protection
DEU	2017, Ethics Commission Automated and Connected Driving 2021, The regulations for autonomous vehicles
JPN	2016, Securing AI R&D stability 2017, AI R&D Guidelines 2018, Principles of using Artificial intelligence 2019, Social Principles of Human-Centric AI
CHN	2017, Next-generation artificial intelligence development plan 2019, Beijing AI Principles

허는 주로 기계 학습 도구에서 도덕적 점수를 예측하는 AI 도덕 통찰력 예측 모델, 사용자 기반 윤리적 의사결정을 구현하는 자율 주행 차량 등이 등록되어 있다[27]-[31].

또한, 인공지능의 설계-학습-실행 전 과정에 대한 도덕성 검증이 필요하다. 인공지능은 설계단계에서 개발자가 도덕성을 검증하여 기본적으로 도덕적 원칙이 지켜지도록 하여야 하며 개발 목적과 사용 용도에 적합한 도덕성을 보장하여야 한다. 다음으로 학습 단계는 개발자와 사용자 모두에게 도덕성 검증의 책임이 부여된다. 개발자는 인공지능이 도덕적 추론 절차 및 결과가 도출되도록 개발단계에서의 학습을 시켜야 하며, 사용자는 사용 단계에서 인공지능이 도덕적인 행동이 학습되도록 피드백을 제공해야 한다. 무엇보다 중요한 것은 도덕적 딜레마 상황에서 인공지능은 스스로 추론하고 판단한 근거를 사용자가 확인할 수 있도록 제공하여야 하며, 사용자에 의해 결정할 수 있도록 하여야 한다. 하지만 딥러닝은 인공신경망 구조로 매우 복잡한 수학적 구조로 결정되어지기 때문에 인공지능의 추론 과정을 사용자가 지속적으로 확인하는 것은 불가능하다. 이러한 문제점에 대해 '설명가능 인공지능(XAI: eXplainable Artificial Intelligence)[32]'이 대안 기술로 개발되고 있다.

그림 1과 같이 설명가능 인공지능 기술은 복잡한 인공지능

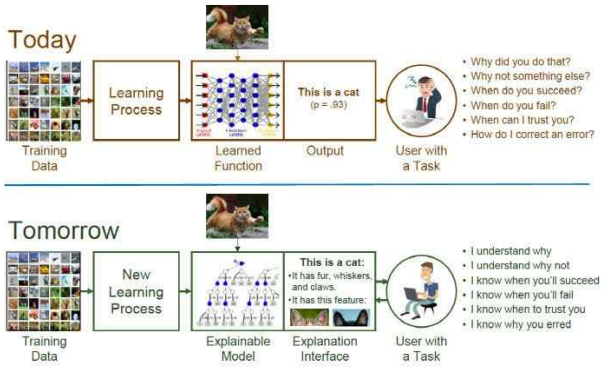

그림 1. 설명 가능한 인공지능의 개념
Fig. 1. Concept of explainable artificial intelligence

의 연산의 결과를 사용자의 언어로 이해할 수 있도록 추론과정 및 이유를 설명함으로써 발생할 수 있는 윤리적 문제점을 검증하고 도덕적 결심을 도울 수 있도록 하는 중요한 수단으로 주목받고 있다. 하지만 설명 가능 인공지능의 연구 방향은 아직 편향된 데이터에 의한 오류와 도덕적 문제 식별과 같이 이미 식별된 한정된 주제의 문제에 집중하고 있는 경향이 있어 다양한 인공지능의 이슈에 적용되기 어렵다는 한계를 가진다. 따라서 일반적인 범위를 가지며 다양한 목표와 맥락에 걸쳐 일반화에 능한 합성 지능인 범용인공지능(Artificial General Intelligence: AGI) 시스템의 연구 및 개발에 관심이 높아지고 있다[33]-[35].

설명가능 인공지능과 함께 또 다른 대안으로 적용해야 할 기술은 생성적 적대신경망(GAN : Generative Adversarial Network)의 적용이다. 생성적 적대신경망은 기존의 데이터를 모방해 새로운 데이터를 생성하는 알고리즘으로 2개의 모델, 즉 생성자(Generator) 모델과 판별자(Discriminator) 모델이 서로 적대적으로 경쟁하며 사실에 가깝고 정교한 데이터를 생성하는데 생성자 모델은 지속적으로 모델에서 학습한 결과를 만들어 내고 판별자 모델은 생성자 모델이 생성한 데이터의 진위를 판단하는 과정을 반복하며 생성자 모델이 잘못된 데이터를 생성하는 것을 지도함으로써 생성자 모델의 정교함을 높여주는 역할을 한다. 생성적 적대신경망 기술은 이미 이미지를 인식하거나 생성하는 인공지능 모델에 활용되고 있으며, 이를 윤리적인 방법으로 활용하기 위해 주 모델이 되는 생성자 모델에 인간이 주입한 윤리성을 기준으로 데이터의 진위, 즉 윤리적으로 적합한 데이터인지 부적합한 데이터인지를 판단하는 판별자 모델을 함께 학습시켜 인공지능 모델의 도덕적 결함을 감소시키는 방법으로 도덕적 역량을 향상시킬 수 있다.

따라서 인공지능 설계 단계에서 개발자는 인공지능 모델이 학습해야 하는 도덕적 개념과, 준수해야하는 도덕적 원칙, 금지해야 하는 행위나 데이터 산물을 설정한 후 학습 단계에서는 생성적 적대신경망을 활용하여 인공지능 모델의 윤리적 역량을 향상시키고 실행의 단계에서 사용자는 완벽하지 않은

인공지능에 의해 발생한 윤리적 문제들에 대한 보고를 통해 추가 학습을 통해 모델의 정교성을 더욱 높이는 한편, 개발자로 하여금 학습 이상의 조치가 필요한 사항인지를 검토하게 해야 하며, 모델의 학습과 실행 단계에서 인간이 확인하기 힘든 '블랙박스 영역'은 설명 가능 인공지능을 통해 그 원인을 파악할 수 있도록 하여야 한다.

4-3 인공지능의 자율학습 및 도덕성 자기 검증을 위한 MCT의 유용성과 게임에서의 적용

도덕 역량을 과학적으로 측정할 수 있는 도덕적 역량 검사 도구(MCT, Moral Competence Test)는 인공지능의 자율성과 도덕성을 검증 및 학습하는 도구로 사용될 수 있다. 린트(G' Lind) 교수가 개발한 MCT는 인공지능의 자율성을 위해서 검증해야 하는 도덕적 추론을 정량적인 결과로 도출하는 특성으로 도덕적 영역과 과학적 영역을 연결하는 수단으로 적합하다[36]. 특히, '설명가능 인공지능'이 고도화된 인공지능의 도덕성 검증의 수단으로 주목받는 이유가 인공지능의 추론과정을 인간인 사용자가 이해할 수 있고 이에 따른 결심이나 피드백이 가능하기 때문이다. 이 원리는 딜레마 상황에서의 결정에 대한 이유를 추론하는 과정에서 도덕적 역량이 판단되는 MCT의 원리와 유사하여 딜레마 스토리를 이용한 머신러닝 학습 분야, C-점수를 이용한 도덕성 검증 분야, 사용자 결심 지원을 위한 보조 도구 등으로 광범위하게 적용될 수 있다.

MCT는 콜버그의 도덕발달 6단계를 기본으로 하여 각각의 딜레마스토리에 대한 찬성과 반대의 논증을 6개씩 포함하고 있다. 피검자의 도덕적 역량이 높을수록 논증에 포함된 높은 수준의 가치에 대해 높게, 낮은 수준의 가치에 대해 낮게 평가하며 비슷한 수준이 유사하게, 상이한 수준이 다르게 평가되는 일관적 패턴을 유지하게 되어 높은 C-점수를 획득하게 되며 이는 개인의 도덕적 원리에 따라 역량을 평가하는 것을 의미한다[36]. 특히, 최근 여러 창작 분야에서 활용도가 높아지고 있는 생성형 인공지능의 경우 대형언어모델(LLM, Large Language Model) 기반의 시나리오와 영상을 생성하는데 원천의 말뭉치 데이터를 학습한 생성형 인공지능 모델에게 인공지능 적용 분야에 적합한 딜레마 스토리를 질문하고 이에 답하는 결과를 평가하여 인공지능 모델의 도덕적 수준의 완성도를 측정할 수 있어 디지털 콘텐츠 생성 인공지능의 윤리적 안전성 및 상업적 활용 가능 여부를 확인하는 중요한 근거가 된다. 요약하면 단순한 이미지나 문자의 인식과 달리 맥락이 있는 컨텐츠와 그 도덕성에 관한 사항은 단순히 옳고 그름이 아닌 맥락과 추론 근거가 필요하므로 설명가능 인공지능은 그 근거를 제공하고 MCT는 그 근거를 판단할 뿐만 아니라 생성적 적대신경망의 판별자 모델의 도덕학습과 판단 근거의 기준을 제공하고 인공지능 모델의 수준과 산물의 윤리 수준을 규명할 도구로서 적용 가능하다.

그럼에도 불구하고, 국내외적으로 MCT를 인공지능에 적용하려는 연구는 아직 초기단계에 머물러 있다. 특히, 해외에서의 연구결과는 아직 찾아보기 어렵다. 다만, 국내에서 김형수와 김태웅 등이 인공지능의 도덕성 검증을 위한 도구로서 린트의 MCT를 제안하고 있으나, 이를 단계별로 어떻게 적용할 것인지에 대한 대안의 제시는 미흡하다[37],[38]. 이에 따라 인공지능의 도덕성을 검증하기 위한 알고리즘으로서 MCT를 어떻게 적용할 것인지에 대해 게임의 사례를 통해 대안적인 모델을 제시하고자 한다.

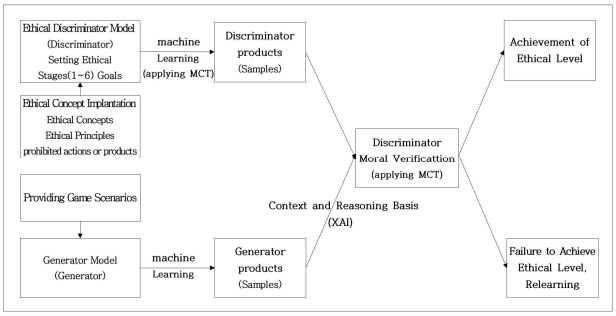

그림 2. MCT를 활용한 인공지능 게임 모델 개발 개념
Fig. 2. Concept of developing AI game model using MCT

그림 2는 인공지능 게임 모델 개발 절차에 MCT를 활용하여 도덕성을 검증하기 위한 대안적 모델 개발을 위한 개념도이다. 앞에서 제시한 설명가능 인공지능, 생성적 적대신경망을 활용하여 모델의 기본 개념을 구성하였다. 폭력성이 높지 않은 청소년을 위한 이용가능한 수준의 게임을 구성할 때, 먼저 생성자 모델에는 게임 작가가 추구하는 기본 게임 시나리오와 개념을 제공하고 머신러닝을 통해 시나리오를 발전시키고 구체화한다. 다음으로 도덕성을 판단하는 판별자 모델에는 기초적인 게임 시나리오에 추가하여 도덕적 개념과 원칙, 반드시 하지 말아야 할 금지행위(학살, 학대, 강간, 식인 등) 또는 산물(음란물, 패륜적 시나리오 및 영상 등)을 주입하고, MCT를 활용하여 머신러닝을 통해 학습한 시나리오나 산물에 대한 도덕성을 측정한다. 이때 게임이 목표로 하는 윤리적 수준이나 사용자 연령 등을 고려한 윤리적 하한점을 설정하여 이를 충족하도록 하는데, 통상적으로 청소년이 이용 가능한 수준은 콜버그의 도덕성 발달 단계 중에서 2수준, 3~4단계의 관습적 수준(Conventional level)으로 설정하여 학습시킨다. 이어서 판별자 모델의 사전 학습된 도덕성 수준의 모델로 생성자 모델의 산물 중에서 도덕성에 대한 요소를 검증한다. 이 과정에서 판별자 모델은 세 가지 차원에서 검증을 하는데 첫째, 도덕적으로 금지된 요소의 여부, 둘째, 시나리오나 산물에서 도출된 다양한 생성자의 도덕적 행태에 대한 도덕 수준 측정, 셋째, 생성자 모델 자체의 도덕 역량을 측정하기 위한 추론의 분석이며 이를 위해서는 설명가능 인공지능의 활용이 반드시 필요하다.

우리에게 익숙한 닌텐도사의 '모여라 동물의 숲' 게임의 한 장면의 예를 든다면, 게임의 시나리오 중 여러 친구들이 모여 다른 한 친구의 집에 찾아가 물건을 훔치고 불을 지른다면 이

것은 금지 행위(약탈 및 방화)에 해당하기에 첫 번째 요소에 따라 '검증 실패'가 될 것이다. 하지만 다른 마을에서 우리 마을을 찾은 사람들을 대하는 태도나 도움을 요청하는 캐릭터에 대해 기꺼이 도움을 주거나 외면하거나 심지어 골탕을 먹여 난처하게 만드는 시나리오를 평가한다면 이는 단순한 옳고 그름이 아닌 전반적인 맥락과 추론(거짓말과 속임의 인지, 외부인에 의한 피해로 공동체의 높은 경계 수준, 더욱 중요한 해야 할 일의 우선순위)에 따라 도덕성이 평가되어야 한다. 이러한 평가는 MCT를 이용하여 위에서 제시한 두 번째 또는 세 번째 방법으로 검증되어 질 수 있다.

Ⅴ. 맺음말: 인공지능의 공익적 시사점

인공지능 기술의 발달은 분명 인류의 생활에 많은 변화를 가져올 것이다. 특히, 인간의 삶의 질을 높이며 편리하게 하기 위해 인공지능 기술은 더욱 영리해지고 점차 스스로 선택하고 결정하는 영역의 범위를 확대시켜 갈 것이다. 하지만 인공지능 기술의 발달이 다른 문명의 발달과 차별되어 자율적 주체로서의 지위를 가지고 인공지능은 말 그대로 인간의 지적 사고능력을 모사하여 만든 것으로 하늘을 나는 새를 보고 라이트형제가 개발한 비행기와 다르지 않다. 비행기의 기술이 아무리 발달하고 새와 비슷한 형태와 운동적 특성을 가지고 심지어 사냥을 하여 먹이를 삼키고 알을 품는 행위를 하더라도 우리는 그것을 온전히 새라 여기지 않는다. 뿐만 아니라 인공지능 기술은 매우 편중된 발달 양상을 가지고 있으며 과장된 평가를 받고 있다. 인공지능의 발달은 데이터 처리와 사례 학습, 사물의 인식 분야에서는 재빠르게 발달하고 있으나 관계성이나 상황 판단, 직관, 정서, 융통성, 사회성과 같이 호모사피엔스로서의 주요한 특성이 되는 요소들은 온톨로지의 정의조차도 요원하다.

인공지능 기술이 주목받는 이유는 인공지능이 가져다주는 편의성과 첨단 기술에 대한 호기심, 인간을 모사하기 위해 노력하는 기술에 대한 호의적 감정이지 인공지능이 우리 삶을 주도하며 주요한 결정을 하는 것에 있지 않다. 인공지능 기술이 가지고 있는 복잡성과 첨단성이 기술을 개발하는 개발자와 이를 사용하는 사용자에게 윤리적 책임으로부터 자유로울 수 있는 정당화의 사유도, 도피처가 되어서도 안 된다. 일반적인 제품을 만들면 제조사는 그 제품에 내재된 위험성은 제거하거나 통제되어져야만 하며, 사용자의 과실이나 부주의로 발생할 수 있는 위험조차도 사전에 고지할 의무가 있다. 인공지능 기술도 동일하며 다를 수 없다.

인공지능 기술은 가상현실(VR, Virtual Reality), 증강현실(AR, Augmented Reality)과 함께 미래 주요 산업인 게임 기술의 핵심적 역할을 할 것이다. 과거의 단순 로직과 시나리오 중심의 게임과는 달리 인공지능에 의해 진행되는 게임은 다수의 유저의 참여로 더욱 다양하고 복잡한 상황을 만들게 될 것이며, 현실과 게임 공간의 경계가 모호해질 것이다. 이러한 추세 가운데 인공지능의 게임 윤리는 두가지 측면에서 매우 중요한 역할을 가지게 된다. 먼저, 게임의 윤리성을 검증할 인공지능의 유용성이다. 윤리적 역량을 가진 인공지능이 쉬지 않고 빠른 속도로 게임의 개발-시험-사용의 전 단계에 걸쳐 윤리성을 검증하고 문제 식별 시 문제 보고 및 해결방안을 제시하여 프로그램패치를 실시하는 것이다. 이는 게임의 도덕적 측면 뿐 아니라 기업의 책임이 발생할 수 있는 상황을 예방하는 효과를 가질 것이다. 다음으로 도덕적 역량 향상을 위한 게임의 개발 및 활용이다. 인성교육, 시민윤리, 직업윤리의 중요성은 점차 강조되고 있으며, 새로운 문화와 시대에 걸맞는 내용과 방법의 윤리교육이 필요하다, 청소년의 도덕적 역량을 향상시킬 수 있는 게임, 시민으로서의 윤리의식을 향상시킬 수 있는 게임, 직무 중심의 도덕적 역량을 향상시킬 수 있는 게임 등 게임에서 부여되는 퀘스트(Quest)를 해결하고 등급이 올라갈 때마다 긍정적 피드백을 받을 수 있는 게임을 개발하고 활용하는 것은 게임의 순기능을 잘 활용하여 사회를 유지케 하는 중요한 역할이 될 것이다. 실제 중국의 Yotta games사는 2018년 '마피아시티'라는 범죄조직 게임을 출시하였다. 상대 조직을 쟁탈하고 죽이며 악행을 할수록 성과를 얻게 되는 모바일 게임으로 출시 후 상당한 호응을 얻으며 판매순위 상위에 머무르고 있다. 단순히 잔인함과 폭력성과 같은 비윤리성을 강화하는 게임이 개발되고 성공할 수 있다면 윤리성을 강화하는 게임도 성공할 수 있을 것이다.

인류는 새로운 기술이 등장할 때마다 부푼 기대와 함께 큰 힘을 가진 기술력의 오용으로 인한 위험을 역시 경계하였다. 핵무기가 개발되고 미·소간 군비경쟁을 하던 냉전시대에는 머지않아 핵전쟁으로 지구가 멸망할 것이라 예측하기도 하였다. 하지만 아직까지 그 예측은 이루어지지 않았고 오히려 핵을 평화적으로 유지하려는 노력은 지속되어 전쟁 보다는 전쟁을 억제하는 수단으로 활용되고 있다. 마찬가지로 인공지능에게 자율적 주체로서의 지위를 부여하지 않고 인공지능의 목적과 지위, 윤리적/법적 책임 소재, 통제 방법, 및 용어에 대한 명확한 정립을 통해 인공지능이 본래의 목적인 인간의 행복과 이익을 증가시키며 존엄을 지키는 도구로 사용되도록 하여야 한다. 본 논문은 인공지능의 도덕성 검증을 위한 '설명가능 인공지능'의 적용과 도덕판단역량검사도구(MCT)의 알고리즘의 기본원리를 제시하였는 바, 향후에는 이러한 원리를 활용하여 윤리적 기능을 장착한 다양한 범용인공지능(AGI)을 연구 개발하고 실용화시켜 나갈 것으로 기대된다.

참고문헌

[1] I. A. Wogu, S. Misra, P. Assibong, A. Adewumi, R. Damasevicius, and R. Maskeliunas, "A Critical Review of the Politics of Artificial Intelligent Machines, Alienation and the Existential Risk Threat to America's Labour Force," in *Proceedings of the 18th International Conference on*

Computational Science and Its Applications (ICCSA 2018), Melbourne, Australia, pp. 217-232, July 2018. https://doi.org/10.1007/978-3-319-95171-3_18

[2] R. Kurzweil, *The Singularity is Near: When Humans Transcend Biology*, New York, NY: Viking, 2005.

[3] J. Kim, *Asking about Humans Again in the Era of Artificial Intelligence*, Seoul: Dongasia, 2017.

[4] B.-T. Zhang, "Human Intelligence and Machine Intelligence - Cognitive Artificial Intelligence," *Communications of the Korean Institute of Information Scientists and Engineers*, Vol. 36, No. 1, pp. 17-26, January 2018.

[5] Y. Park and B. Goertzel, *Artificial Intelligence Revolution*, S. Eom, trans. Seoul: Double Book, 2016.

[6] J.-J. Rousseau, *Émile,* H. Min, trans. Seoul: Yugmunsa, pp. 80-88, 2012.

[7] I. Kant, *Kritik der Praktischen Vernunft*, J. Baek, trans. Seoul: Acanet, p. 229, 2009.

[8] J. H. Moor, "Four Kinds Of Ethical Robots," *Philosophy Now*, Vol. 72, pp. 12-14, March-April 2009.

[9] I. Asimov, *I, Robot,* New York, NY: Bantam Books, 2004.

[10] IEEE. Ethically Aligned Design, Version 2 (EADv2) [Internet]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[11] F. Morandín-Ahuerma, "Twenty-Three Asilomar Principles for Artificial Intelligence and the Future of Life," September 2023. https://doi.org/10.31219/osf.io/dgnq8

[12] J. H. Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom, "Human-Versus Artificial Intelligence," *Frontiers in Artificial Intelligence,* Vol. 4, 622364, March 2021. https://doi.org/10.3389/frai.2021.622364

[13] W. Totschnig, "Fully autonomous AI," *Science and Engineering Ethics*, Vol. 26, pp. 2473-2485, October 2020. https://doi.org/10.1007/s11948-020-00243-z

[14] S. Lee, Recent Artificial Intelligence Development Trends and Future Direction of Evolution, LG Business Research, Seoul, pp. 10-22, October 2017.

[15] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, T. Noh, trans. Seoul: Medici Media, 2014.

[16] B.-D. Lee, "The Logic of Ethical Justification for Engineers," *Philosophical Analysis*, No. 21, pp. 75-105, June 2010.

[17] W. Wallach, C. Allen, and S. Franklin, Consciousness and Ethics: Artificially Conscious Moral Agents, in *Machine Ethics and Robot Ethics*, Abingdon, UK: Routledge, ch. 24, pp. 299-314, 2020.

[18] IBE (Institute of Business Ethics), Business Ethics and Artificial Intelligence, Author, London, UK, Business Ethics Briefing No. 58, January 2018.

[19] Scripps News. Researcher Rodney Brooks Says Current AI is Overrated [Internet]. Available: https://scrippsnews.com/stories/researcher-rodney-brooks-says-current-ai-intelligence-is-overrated/.

[20] M. A. Anishchenko, I. Gidenko, M. Kaliman, V. Polyvaniuk, and Y. V. Demianchuk, "Artificial Intelligence in Medicine: Legal, Ethical and Social Aspects," *Acta Bioethica*, Vol. 29, No. 1, pp. 63-72, June 2023. http://dx.doi.org/10.4067/S1726-569X2023000100063

[21] Newsis. Google 'AI, No Trolley Dilemma...Self-Driving Cars are Better than People [Internet]. Available: https://mobile.newsis.com/view.html?ar_id=NISX20171128_0000160896.

[22] L. Floridi and J. W. Sanders, "On the Morality of Artificial Agents," *Minds and Machines*, Vol. 14, pp. 349-379, August 2004. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

[23] W. Kim, Method and Apparatus for Controlling Chatbot Response Based on User Characteristic Information in Unethical Situations, Korean Intellectual Property Office, Daejeon, Publicized Patent No. 10-2022-0075638, June 2022.

[24] J. Kim and J. Han, Apparatus and Method for Providing Ethics-Based Services, Korean Intellectual Property Office, Daejeon, Publicized Patent No. 10-2022-0122385, September 2022.

[25] G. Park, Autonomous Moral Judgment of Artificial Intelligence and Its Implementation System, Korean Intellectual Property Office, Daejeon, Publicized Patent No. 10-2020-0145002, December 2020.

[26] G. Park, Electronic Apparatus and Method for Determining of Value Consciousness, Korean Intellectual Property Office, Daejeon, Publicized Patent No. 10-2020-0105901, September 2020.

[27] R. Donovan, Artificial Intelligence Models for Moral Insight Prediction and Methods for Use Therewith, United States Patent and Trademark Office, Washington, D.C., Patent No. US 11,806,629 B2, November 2023.

[28] Personal AI, An Artificial Intelligence Device that Understands, Accumulates, and Predicts the Values of an Individual or an Organizational Group to Which the Individual Belongs, and Provides Value-Based Support or Analysis, Author, Tokyo, Japan, JP 6191892 B1, August 2017.

[29] P. A. Gloor, Predicting Business Performance by Personal Moral Values through Social Network Based E-Mail

Analysis, United States Patent and Trademark Office, Washington, D.C., Patent No. US 2023/0410000 A1, December 2023.

[30] R. Donovan, Ethical AI Development Platform and Methods for Use Therewith, United States Patent and Trademark Office, Washington, D.C., Patent No. US 2024/0009575 A1, January 2024.

[31] Z. Glazberg and S. Ur, Autonomous Vehicles Implementing User-Based Ethical Decision Making, United States Patent and Trademark Office, Washington, D.C., Patent No. US 2023/0332910 A1, October 2023.

[32] DARPA (Defense Advanced Research Projects Agency), Broad Agency Announcement: Explainable Artificial Intelligence (XAI), Author, Arlington: VA, DARPA-BAA-16-53, p. 6, August 2016.

[33] B. Goertzel, "Artificial General Intelligence: Concept, State of the Art, and Future Prospects," *Journal of Artificial General Intelligence*, Vol. 5, No. 1. pp. 1-48, December 2014. https://doi.org/10.2478/jagi-2014-0001

[34] O. I. Obaid, "From Machine Learning to Artificial General Intelligence: A Roadmap and Implications," *Mesopotamian Journal of Big Data*, Vol. 2023, pp. 81-91, August 2023. https://doi.org/10.58496/MJBD/2023/012

[35] B. B. Slavin, "An Architectural Approach to Modeling Artificial General Intelligence," *Heliyon*, Vol. 9, No. 3, E14443, March 2023. https://doi.org/10.1016/j.heliyon.2023.e14443

[36] G. Lind, *How to Teach Morality: Promoting Deliberation and Discussion, Reducing Violence and Deceit*, 3rd rev. ed., G. Park and C. Jung, trans. Seoul: Yangseogak, pp. 116-122, 2017.

[37] H. Kim, "Designing and Applying the Moral Turing Test for Korean Children," in *Proceedings of Cultural Appropriation of Spaces and Things*, Siegen, Germany, pp. 29-42, October 2019. https://doi.org/10.25819/ubsi/5429

[38] T. Kim, G. Park, and E. Seo, "IR4.0 and Ethical Tasks of AI," *Robotics & AI Ethics*, Vol. 4, No. 2, pp. 6-13, December 2019. https://dx.doi.org/10.22471/ai.2019.4.2.06

전찬영(Chan-Young Jun)

2004년 : 해군사관학교
국제관계학과(문학사)
2015년 : 중부대학교
교육상담심리학과(교육학 석사)
2005년 : 경상국립대학교 대학원
윤리교육학과(박사과정 수료)
2016년～2017년: 해군 리더십센터 교관
2021년～2023년: 한미연합군사령부 해상작전장교
2023년～현 재: 해군교육사령부 전투병과학교
작전전술학과장
※관심분야 : 인공지능윤리, 우주윤리, 군 리더십

방정배(Jung-Bae Bang)

1990년 : 육군사관학교(문학사)
1999년 : 국방대학교(안전보장학석사)
2005년 : 영남대학교 대학원
(정치학박사)
2001년～2003년: 육군3사관학교 강사
2016년～2019년: 국방정신전력원 교수부장
2023년～현 재: 한국열린사이버대학교 객원교수
※관심분야 : 응용윤리, 군대윤리, 군사안보 등

박균열(Gyun-Yeol Park)

1989년 : 경상국립대학교
윤리교육과(문학사)
1994년 : 서울대학교
윤리교육학과(교육학석사)
2000년 : 서울대학교 대학원
윤리교육학과(교육학박사)
2007년～현 재: 경상국립대학교 윤리교육과 교수
2020년～현 재: 한국공공가치학회 회장
2022년～현 재: 한국윤리학회 부회장
※관심분야 : 도덕성 측정, AI윤리, 정치윤리