Brain, Digital, & Learning https://doi.org/10.31216/BDL.2020.10.1.077
2020, Vol. 10, No. 1, 077-092.

Ethics of Artificial Intelligence and Moral Education*

Kuk Hyeon Kim[1)]
[1)]Korea National University of Education

인공지능 윤리와 도덕과 교육

김국현[1)]
[1)]한국교원대학교

<ABSTRACT>

This paper discusses the necessity of an ethical approach to Artificial Intelligence in moral education and outlines the educational content of AI ethics in moral education. The main contents of this paper are as follows. Firstly, in relation to the necessity of an ethical standpoint on artificial intelligence ethics or machine ethics, robot ethics, it discusses the continuity or similarity, discontinuity or difference, and complementarity between AI as artificial moral agents and human being. Categories of AI ethics are divided into ethics on AI, ethics of human intelligence, AI ethics for human beings or ethics for coexistence of human beings and artificial intelligence. And then moral educational implications are drawn by arranging the research results related to the morality, or ethical design of artificial intelligence as an artificial moral agent. Secondly, the trend of education on artificial intelligence ethics and the issues of moral education are discussed. problems, phenomena, issues, ethical dilemma that are dealt with as educational contents in the education on AI ethics are suggested. And then, what issues, problems, and phenomena can be selected and organized as educational contents, in others words, what learning theme will be included in AI ethics education in Korea's school moral education are discussed from the standpoint of moral curriculum revision.

Keywords : AI, ethics of AI, moral agent, moral education, curriculum, reflection

I. Introduction

본 논문은 인공지능이 인간 본성의 고유한 능력 특성으로 생각되어 온 도덕적 인지, 감정, 행동 능력과 유사성, 또는 어떤 측면에서는 우월성을 보인다는 점에서 인공지능 윤리에 대한 전향적 성찰을 토대로 인간 윤리를 향상할 방안과 특히 학교 도덕 교육에서 인공지능에 대한 윤리학적 접근과 인공지

* 이 논문은 제26차 한중 윤리학 국제학술대회에서 발표한 논문을 수정, 보완한 것이다.
Received February 27, 2020; Reviewed March 6, 2020 Corrected March 10, 2020; Accepted March 13, 2020

능 윤리에 관한 학습 방안을 논의한다.

인간 윤리와 인공지능 윤리의 상보성 및 생산적 상호작용에 중점을 둔 본 논문의 문제의식은 신경과학 또는 뇌과학의 기술적(descriptive) 정보를 학교교육과 관련 처방적(prescriptive)으로 활용하려는 (Lee, 2019) 다음 몇 가지 선행연구들과 맥락을 같이한다. 정치적 성향 및 도덕성과 뇌 구조의 상호의존성에 대한 신경과학적 접근(Ha, 2016), 사회적 행동의 기초에 대한 fMRI 기술을 활용한 뇌과학적 접근(Lee & Kwon, 2016), 디지털 콘텐츠 중독 뇌 영역 및 특성 정보에 근거한 뇌 기반 디지털 콘텐츠 중독 예방책 마련 연구(Kwon, Eom, & Kwon, 2014), 수학불안에 대한 신경과학적 접근(Ma et al., 2013) 등이 그러한 예다.

이러한 문제의식을 바탕으로 본 논문은 인공지능의 발달이 인간 윤리가 변화하는 데 미치는 영향을 도덕성, 도덕적 행위자 개념을 중심으로 살펴보고 학교 도덕교육에서 학생들의 인공지능 윤리 학습을 위한 교육내용 요소를 교육과정 맥락에서 제안한다. 이런 맥락에서 선행연구에서 학교 도덕교육의 교육내용으로서의 인공지능 윤리 유형, 윤리적 쟁점과 현상 등을 추출한다.

인공지능에 관한 다양한 학문 분야에서 이루어진 연구 성과와 비교할 때 윤리학적 관점의 연구, 특히 학교 도덕교육의 인공지능 윤리에 대한 접근 필요성, 방향, 과제에 관한 연구는 적실성이 있으며 긴급한 과제이다. 도덕교육은 인공지능이 일상화된 환경에서 살아가는 학생들이 도덕철학과 도덕심리학 측면에서 인공지능 윤리 문제에 접근하여 인공지능 윤리에 대한 스키마를 발달시키도록 촉진해야 한다. 이를 위해 인공지능에 대한 윤리교육에서 다룰 도덕성과 지식의 구조에 대한 이론 틀을 마련하고 앞으로 이루어질 교육과정 개정에 앞서 논의 결과를 개정안에 반영할 방안을 미리 논의해야 한다.

이러한 문제의식을 토대로 본 논문은 인간의 윤리와 인공지능 윤리의 본질이나 특성 비교, 인공지능의 도덕적 행위성(moral agency)과 윤리적 쟁점과 결과, 그리고 인공지능의 윤리적 설계 측면에서 기계의 도덕성 발달과 관련한 윤리적 고려사항을 다룬다. 그리고 인공지능과 기계학습의 옳고 그름에 관련한 쟁점과 사례조사 및 분석을 중시하는 학습과정에 초점을 맞춘다.

본 논문은 인공지능이라는 근본적인 기술 변화가 낳은 결과에 대한 윤리 이론과 행위 문제를 다룬다. 먼저, 인공지능 윤리의 개념화와 이론화와 관련해, 도덕성과 그 발달을 중심으로 인공지능과 인간윤리의 관계 특성을 파악하고 인공지능 윤리 유형을 설정해 인간 도덕성 발달의 본질, 다양성, 한계와 이상 등을 분명히 하는 데 중점을 둔다. 그래서 인간 윤리와 인공지능 윤리의 관계, 도덕적 행위성에서 인간과 인공지능의 특성, 인공지능 윤리의 유형 및 범주 문제를 주로 논의한다. 그리고 논의 결과를 토대로 학생들이 삶의 시공간에서 생생하게 경험하는 인공지능의 윤리적 쟁점을 깊이 학습하도록 촉진하는 교육에 대해 논의한다.

II. Ethics of Artificial Intelligence and Human Being

1. Morality of AI and Human Being

1) Moral Agency

인공지능 윤리에 대한 논의에서 중심은 인간의 도덕적 행위성(moral agency)을 모사한 알고리즘 설계와 적용, 즉 준도덕성(quasi-morality)을 토대로 인공 도덕적 행위자를 만드는 것이다. 도덕적 행위성은 행위자가 자기 행동의 결과를 이해하며 자기 행동을 선택할 권리가 있다는 것을 전제한다. 그러나 도덕적 행위자가 되기 위해서 반드시 옳고 그름을 느낄 수 있어야 하는 것은 아니며 단지 옳고 그름의 차이만 알 수 있으면 된다. 사이코패스가 살인을 잘못으로 느끼거나 후회할 때에만 행위 책임성을 물을 수 있는 것이 아니라 사회가 살인을 그릇된 행동으로 본다는 사실만 인식하면 행위 책임성을 물을 수 있는 것과 같은 맥락이다(Kaplan, 2015).

그러나 인공지능 시스템이 규칙을 준수하는 것만으로 도덕적 행동을 보장할 수 있는가? 카플란은 그렇지 않다고 설명한다. 행동 규칙은 더 긴요한 목적이 발생하면 행위자가 규칙을 스스로 위배하거나 변경할 것이라는 추정을 근거로 한다. 인간은 환경의 변화를 살펴

그것에 맞게 행동을 수정할 수 있는 기계를 만들 수 있다. 또는 기계 스스로도 그렇게 진화할 수 있다. 그러나 따라야 하는 행동 수정 원칙이 무엇인가는 해결되지 않는다. 특히 적용할 규칙이 없거나 더 긴박한 윤리적 문제를 해결하기 위해 규칙을 어겨야 하는 경우 참조할 더 깊은 행위의 원칙이 필요하다. 이러한 점에서 인공지능의 행동 지침으로 기능할 더 명확하며 실행할 수 있는 도덕 이론이 필요하다(Kaplan, 2015).

의식적 추론, 자유의지, 도덕적 책임은 도덕적 행위자의 속성으로서 강조되어 왔다. 이러한 속성 외에 인공지능의 도덕적 행위성과 관련하여 상호작용성, 자율성, 적응성이 속성으로서 제기된다. 상호작용성은 행위자가 환경과 영향을 주고받으며 행동하는 능력이다. 자율성은 상호작용에 대한 직접적 반응 없이도 어느 정도의 복잡성과 환경과 분리하는 능력이다. 적응성은 행위자가 자기 경험에 따라 작동방식을 스스로 학습하는 능력이다(Wallach & Allen, 2009).

도덕적 행위자는 도덕적 기준에 따라 자기 행동을 통제하고 행동 결과를 평가할 수 있어야 한다. 이는 선천적이고 주관적인 감각에 따라 옳고 그름을 판단하지 않아야 한다는 것을 강조한다. 인공지능은 특정 환경에서 도덕적으로 적절한 측면을 감지할 능력이 충분하고 행동을 선택할 수 있으므로 도덕적 책임의 조건을 충족하는 도덕 행위자가 될 수 있다(Kaplan, 2015).

인공지능의 도덕적 행위성을 논의할 때 인간과 인공지능의 본질적인 차이를 고려할 필요가 있다. 인간 유기체는 생화학적 플랫폼에서 그리고 사고 능력은 정서적인 뇌에서 출현했다. 그러나 인공지능은 논리적 플랫폼에서 개발된다. 인공지능의 이 특성은 윤리적 문제와 관련한 인공지능의 강점을 함축한다. 인공지능은 인간의 판단과 결정에 개입하는 편견, 고려의 제한성, 확증편향 등에서 상대적으로 자유롭다는 점에서 인간의 고려보다 일관되고 우월한 결과를 낳을 수 있다(Wallach & Allen, 2009).

인공 도덕 행위자 설계에서 중심 문제는 추상적 가치를 인공지능의 제어 구조에 구현하는 방법이다. 그래서 옳고 그름과 관련한 상황에서 적합한 행위 대안을 모색하는 행위자로서 인공지능을 생각하고 상황 제약 사항들, 갈등을 해결하는 올바른 공식을 찾는 실제적 접근 방법을 모색한다. 그래서 인공 도덕 행위자가 개인이나 문화의 윤리체계와 관습의 세부적 차이를 잘 다루기를 기대하는 것은 현실적이지 않다. 컴퓨팅 관점에서 윤리학의 도덕 원리는 양립할 수 없는 행동과정을 제시하거나 제시하지 못할 수 있기 때문이다. 그리고 핵심적인 윤리적 원리는 임의의 행동이 미칠지 모르는 무제한의 결과 때문에 컴퓨팅이 불가능할 수 있다. 그래서 문화적으로 다양한 인공 도덕 행위자를 인정할 수 있다(Wallach & Allen, 2009).

2) Moral Competencies

도덕적 역량(이하에서 윤리적 역량과 바꾸어 쓸 수 있는 의미로 사용)은 좋음과 옳음, 정당함과 부당함, 옳고 그름을 추론하며 옳고 그름의 구분에 따라 자신의 선택, 판단, 지침, 행동에 책임을 지는 능력이다. 그러나 로봇이 안락사, 양육권, 소수자 우대정책 등 중요한 윤리적 결정에 어떤 의미로든 책임을 질 수 있는가라는 인공지능의 윤리적 책임 문제가 제기된다. 또, 자신에게 중요한 판단과 조언을 다른 사람이 공감해주길 욕구하는 사람들의 경향을 생각하면 인공지능의 성능이 아무리 향상해도 책임 수행이 적합하지 않거나 부당한 상황이 여전히 남는다(Suskind & Suskind, 2015).

인간 윤리의 본성에 대한 이론들에서 레스트(J. Rest)의 도덕성 4요소 모델은 인간 윤리의 관점에서 인공지능의 윤리적 역량을 고려하는 이론 틀 중 하나이다. 슈리베르와 매시락은 도덕적 인식 또는 민감성, 도덕적 판단 또는 추론, 도덕적 의도 또는 동기화, 실행력과 품성 및 행동을 강조하는 레스트의 도덕성 이론을 지식, 기술, 태도와 능력으로 범주화하여 윤리역량 개념 모형을 제시한다. 도덕성의 요소는 인간의 이상적 윤리적 의사결정 요소를 말한다. 윤리적이라는 용어는 관련된 역량의 본질을 규정한다(Schrijber & Maesschlak, 2015).

도덕적 역량이 발달한 사람은 특정 상황에서 상충하는 가치들이 문제가 되는 때를 감지한다. 레스트가 도덕적 민감성 요소의 세 가지 측면으로 기술한 자기 행동이 타인에게 미치는 영향 인식, 가능한 해결책 및 타인에게 미치는 영향 인식, 공감과 역할채택

은 상황을 윤리적인 것으로 규정하는 데 필요한 기능과 능력이다. 도덕적 행위자는 다른 논증 유형, 즉 규칙과 절차와 관련한 논증, 행동이 타인에게 미치는 결과와 관련한 논증, 행동이 자기 자신에게 미치는 결과와 관련한 논증을 익히 알고, 문제를 해결하는 단 하나의 방법만 있는 것이 아니라는 점을 받아들이려는 융통성이 있으며, 특정 상황에 다른 시각들로 거의 자동으로 접근한다. 이기적 고려에 기초한 행동이 항상 비윤리적인 것만은 아니다. 공직자가 시민의 선물을 받기 거부하는 것이 자신의 중립성에 영향을 미치리라고 생각하거나, 공적 신뢰를 해치는 비윤리적 행동이라는 인식을 만들기 때문이 아니라 처벌을 두려워하는 것 때문이라면 그 행동은 윤리적일 수 있다. 그러나 그것이 곧 윤리적 역량을 보여주는 것은 아니다(Schrijber & Maesschlak, 2015).

3) Moral Development

대부분 인간은 능동적 학습과 수동적 학습 혹은 두 가지의 조합으로 학습한다. 학습효과는 교육자, 학습 시기, 학습 방법, 학습 과정에서의 감정, 다른 것에 대한 고려 같은 많은 요소에 따라 다르다. 학습의 정황은 특히 중요한 영향을 미친다. 어떤 것을 학습한 것을 정황과 결부하지 못하면 그것을 이해하기 어렵다. 기계는 사람과 같이 학습한다. 수동적 학습에서는 새로운 규칙이나 관련된 정보가 기계의 프로그램에 영향을 미친다. 인간의 새로운 지식 기억 과정은 이전에 프로그래밍된 기본적인 유전자 프로그램의 변경이나 개선 과정이다. 기계도 실험을 통해 능동적으로 학습할 수 있다(Warwick, 1997).

인공 도덕 행위자 문제는 기계가 현실 세계에 반응하는 올바른 성향을 어떻게 구현하는가의 문제이다. 이는 도덕적 계산이 아닌 도덕성 발달 과정의 복잡성과 가변성, 즉 개인이 자기의 실재를 구축하고 자아, 타인, 환경을 인식하고 일상생활의 도덕적 문제들을 해결하는 방식에 초점을 두는 도덕심리학의 문제이다(Wallach & Allen, 2009).

인공지능의 도덕 발달 문제에서 중요한 것은 누구의 도덕이나 어떤 도덕을 인공지능에 구현할 것인가이다. 윤리적 서브루틴에 관한 논의는 행동을 어떻게 구현할 수 있을지에 관한 개념 설정을 제시하는 것으로 알고리즘이나 소프트웨어 코드가 윤리적 지식을 효과적으로 표현하는 구성요소, 그리고 윤리 이론과 도덕적 행동의 인지, 정서 측면의 관련성을 정교하게 이해해야 한다. 기계윤리에 대한 접근 방법은 윤리 이론을 컴퓨터 프로그램으로 구현할 수 있는지의 관점에서 평가해야하기 때문이다(Wallach & Allen, 2009).

도덕규범을 프로그래밍할 때 도덕규범의 내용과 형식에 대해 의견이 일치하기는 매우 어렵다. 컴퓨터 윤리 연구는 인공 도덕적 행위자 창조를 목표로 선험적 도덕원칙을 선택해 도입하고 의무에 기초한 규범윤리 원칙을 존중하는 시스템을 구축하는 하향식 접근법을 활용한다. 이와 달리 상향식 접근법은 기계학습 알고리즘에 관한 가능한 한 많은 사례를 입력하는 방식이다. 그러나 기계가 인간처럼 사회가 허용하는 도덕원칙을 익혀 실행하거나 도덕원칙을 분명하게 표현하도록 보장할 수 없다는 약점이 있다. 사례기반 추론 연구법은 가장 유사한 사례를 참조해 윤리적 문제를 해결하도록 하는 방식이다. 그러나 기계의 인간 공감 개념 및 능력을 익힐 가능성 면에서 한계가 있다(Kaplan, 2015).

로봇에게 옳고 그름을 가르치려면 경우의 수별로 도덕적으로 옳거나 그른 것을 입력해야 하지만 이는 현실적이지 않다. 인간의 도덕 모델을 정립할 수 있다면 로봇의 두뇌에 탑재할 수 있을 것이다. 또 로봇과 인간이 공유할 수 있는 도덕 체계를 구축할 수 있다면 인간의 윤리도 향상할 수 있다. 따라서 로봇이 어떤 상황에서든 스스로 윤리적 판단을 할 수 있는 구조를 만들 필요성이 제기된다. 이런 주장은 로봇에 도덕 알고리즘의 차원 1을 이기, 차원 2를 신뢰, 차원 3을 이타와 자기희생, 차원 4를 관용성과 다양성의 도덕적 차원으로 설정하자고 제안한다. 여기서 도덕적 차원 4는 자기편인지 아닌지 판별할 때 공통 규율과 개별 규율을 매우 엄격히 구별한다. 개별 규율을 제외하고 공통 규율만을 기준으로 사용한다. 또 어떤 대상을 자기편으로 판별하지 않았을 때의 대응 방식인 무시, 회피, 공격을 무시와 재시도, 회피, 방어라는 더 관대한 방식으로 바꾸어 쓸 수도 있다. 이러한 제안은 이러한 설계 방식을 통해 인간도 이상적인 윤리적 역량을 발달시킬 수 있다는 점을 함축한다(이시훈 역, 2015).

2. Relational characteristics of AI Ethics and Human Ethics

인간과 인공지능, 또는 인간 윤리와 인공지능 윤리의 관계는 연속성이나 유사성, 단절성이나 불연속성, 상보성 유형으로 구분할 수 있다. 이런 유형 구분은 인공지능에 대한 다음 관점들을 근거로 한다. 첫째, 데닛(D. Dennett) 등의 강한 인공지능 관점으로 인간 뇌를 디지털 컴퓨터로, 의식을 컴퓨터 프로그램으로 생각하면서 뇌가 하는 모든 것을 컴퓨터에 복사할 수 있다고 생각한다. 둘째, 설(J. R. Searle) 등의 약한 인공지능 관점으로 컴퓨터가 다른 사물의 동작을 시뮬레이션 하는 것처럼 인간 마음도 시뮬레이션 할 수 있다고 믿는다. 셋째, 펜로즈(R. Penrose) 등은 컴퓨터가 인간 마음을 적절히 시뮬레이션 할 수 없다고 믿는다(Warwick, 1997).

1) Continuity or Similarity

인간의 궁극적 질문은 인간의 본성이 무엇인가이다. 인간성은 여러 가지 특징의 조합이다. 그리고 그 대부분은 다른 동물뿐 아니라 기계와 공유한다. 이런 점에서 생명의 연속성을 보인다, 인간은 인간성 전체로서만 고유하다. 인간성의 본질은 인간의 모든 속성이 뒤얽힌 인과적 결합이다. 인간의 본성은 전체의 문제이지 개별적인 특징 문제는 아니다(Mazlish, 1993).

연속성은 인간과 현재의 인공지능이 옳고 그름에 관한 인간의 사고, 감정, 행위의 비일관성이라는 불완전한 도덕성을 공유한다는 데 근거한다. 인간의 도덕적 직관에 관한 연구 결과들은 인간의 윤리적 사고와 판단 및 결정은 감정을 기반으로 하는 직관과 이성적 추론의 정당화가 병행하면서 도덕적 행동을 산출한다는 것을 밝혀왔다. 상황 종속성은 인간 윤리의 비일관성을 설명한다. 개인의 사고, 감정, 행동은 어느 정도의 일관성이 있으나 특정 상황에 의해 제한되는 경향이 있다.

인공지능의 사고 능력으로서 이성이 인간의 고유한 특성이 아니라 감정이 인간의 고유한 특성이며 인공지능은 감정을 느낄 수 없다는 주장이 존재한다. 그런데, 감정은 동물도 있으므로 이성과 마찬가지로 인간의 고유한 특성이 아니라는 주장도 가능하다. 사고는 감정, 미래에 관한 생각이 있어야 가능한 의지, 동기를 수반하고 감정과 의지는 사고를 수반한다. 인간 삶에서 그것들은 분리되지 않는다(Mazlish, 1993). 또 윤리는 인간에게 고유한 것인가에 대한 다른 견해들이 있다. 윤리도 정도의 문제이다. 윤리는 의식에서 나오고 자기에 대한 지식을 강화하고 잘못을 저지르면서 성장한다. 윤리는 필연적으로 선택의 문제이다. 특별한 의미에서만 윤리는 인간에게 고유하다고 할 수 있다(Mazlish, 1993).

기계는 인간과 유사한 방법으로 행동하고 그래서 기계의 감정을 표현한다는 생각도 가능하다. 인간은 기계의 느낌이나 내적 상태를 프로그래밍할 수 있다. 그런데 인간의 감정도 부모에 의해 생물학적으로 프로그래밍된 것이다. 따라서 기계는 감정을 느낄 수 없다는 말은 과학의 관점에서 볼 때 틀린 것이다. 인종이나 피부색이 다른 인간은 감정이나 느낌이 없다는 역사적 편견이 있었고 적어도 어떤 사람의 감정은 다른 사람과 같지 않다는 생각이 있었다. 그런데 기계는 인간과 똑같지는 않아도 감정, 자기 의지, 의식 등을 표현한다(Warwick, 1997).

2) Discontinuity or Difference

인간과 비인간의 경계는 어디인가라는 물음은 인간은 기계가 어느 정도까지 자기를 대신해야 만족하는가, 그리고 어디까지 맡길 의향이 있는지 묻는 것이다(Watson, 2016). 인간이 어떤 인공지능 윤리를 원하는지 결정할 때 다음 질문이 도움이 된다. 첫째, 물리적 세계와 디지털 세계는 같은 것으로 취급해야 한다. 물리적 세계에서 비난받을 것이면 디지털 세계에서도 마찬가지로 비난받아야 한다. 둘째, 인공지능이 인간처럼 지적으로 행동하거나 윤리적으로 행동할 가능성을 과장하는 데서 나타나듯이, 다수의 집단지능이 개인의 지능보다 뛰어나다는 신화를 깨야 한다(Watson, 2016). 기계와 인간은 지능의 형태가 다르다. 기계는 인간과는 다른 측면에서 지능적일 수 있다. 기계가 인간의 독특한 특징을 보이지 않는다고 해서 지능적이지 않다고 말하는 것도 옳지 않다(Warwick, 1997).

인간과 인공지능의 이질성은 기술 측면에서 알고리즘 규칙과 도덕 규칙의 비연계성, 행위 측면에서

공학(자)와 윤리학(자)의 단절, 도덕성 측면에서 자기 의식성, 의사소통, 감정, 기억과 성찰적 정체성 등에서의 단절을 말한다. 인간 윤리의 핵심은 도덕성이 요소와 도덕성 발달로서 도덕심리학 이론과 경험연구 결과를 통해 나타난다. 레스트의 도덕성 4요소 모델은 인간의 도덕적 행동을 낳는 심리과정 중심으로 인간 윤리를 설명하는 대표적 관점이다.

매튜스(E. Matthews)는 설(J. Searle)의 강한 인공지능 논제, “이행된 프로그램이 그 자체만으로 마음을 갖게 되는 것”, 즉 충분히 정교한 프로그램을 이행하는 컴퓨터는 “마음을 가졌다”는 논제를 비판한다(Matthews, 2005). 그러나 매튜스는 컴퓨터가 일반적으로 인간이 수행하는, 인간 마음과 관계한다고 말할 수 있는 과제를 수행할 수 있다 해도, 컴퓨터가 “마음을 가졌다”고 말할 수 있는 것은 아니라고 설명한다. 마음을 가졌다는 것은 어떤 높은 수준의 지적 능력이 있다는 것을 넘어서 자신의 감정, 감각, 소망, 희망, 목적을 갖는 것도 포함한다. 이 때문에 인간 마음의 높은 수준의 지적 능력들조차도 인간 인격의 다른 차원과 별개로 이해될 수 없다는 이유에서이다. 어떤 활동에 참여하고 목적을 갖는 것은 뇌 자체가 아니라 온전한 인간이라는 이유에서이다(Matthews, 2005). (Matthews, 2005).

스키마는 현재 사건을 빨리 해석하는 데 사용될 수 있는 획득된 지식으로 정보를 조직하고 저장하는 정신의 지도이다. 스키마의 힘을 완전히 이해하기 위해서는 개인이 기억하는 사건에 수반된 감정을 고려해야만 한다. 스키마는 기대와 감정을 불러일으킨다. 기억에 관한 개인의 사고는 감정을 일으킬 수 있고 개인의 감정적 경험은 옛 기억을 일으킬 수 있다. 개인은 현재 사건에 옛 스키마에 저장된 정보를 이용할 것이고 다음에 일어날 일을 예견하는 데 도움이 될 수 있다. 이것은 좋기도 하지만 적절하지 않을 때도 있다(Siegel, 2010).

둘째, 인간과 인공지능의 자기의식 능력은 차이가 있다는 주장이 있다. 그러나 공학자는 컴퓨터의 자기 복제를 자의식의 한 가지로 보기도 한다. 다음으로, 인간과 로봇의 의사소통 방법이 전혀 다르다고 보기도 한다. 이러한 입장에서 로봇과 인간이 별개이며 로봇을 인간처럼 만드는 것이나 그 반대는 헛되다는 것이다. 로봇에게는 인간이 가지는 개인 또는 종으로서의 역사적 경험이 거의 또는 전혀 없으며 가족사회화 경험, 정치적, 경제적, 사회적 상호작용이 없다(Mazlish, 1993).

기계는 인간지능의 논리적 기능을 더 빠르게 더 효율적으로 복제할 수 있다. 그러나 감성지능, 공감, 도덕적 상상력, 즉 정보가 불완전하거나 목표들이 상충하는 맥락에서 판단을 내려야 하는 영역은 기계의 능력 밖이므로 인간에게 당분간 의존해야 한다(Rumsey, 2016). 이런 입장에서 인간 윤리 영역에 기계가 전적으로 참여하게끔 하는 연구를 우려하며 금지하자는 주장이 제기된다.

인간의 도덕성은 인간 생명의 유한성에서 비롯한다. 인공지능이 영생하며 자기의식과 감정을 가지는 것으로 가정한다면 인간윤리와 인공지능 윤리는 연속적일 수 있다. 그러나 아시모프의 Bicentennial Man에서 로봇 NDR-13 모델은 자신의 부속물을 유기물로 대체하고 양전자 두뇌가 쇠퇴하도록 허용해 죽어갈 때만 인간으로 인정된다. 이러한 점에서 영생하는 기계가 완전히 도덕적 존재가 되는 것은 인공도덕 행위자에게 기대하는 바와 거리가 멀다(Wallach & Allen, 2009).

그리고 인간의 도덕성은 의식뿐 아니라 무의식에도 의존한다. 스콰이어와 켄델에 의하면(Squire & Kandel, 2009), “무의식 상태에 있다는 이유로 이런 형태의 기억은 인간 경험의 수수께끼를 만들어 낸다. 이 기억에 의식적 기억이 접근할 수 없으면서도 과거 사건들에 의해서 형성된 기질 습관, 태도, 기호가 들어있다. 이것은 인간 행동과 정신에 영향을 미치며 정체성의 근원”이다. 무의식적 본성 때문에 인공지능은 비서술 기억(nondeclarative memory)을 아직은 분석, 처리, 저장하지 못한다. 감정은 가치가 포함된 결정을 좌우하는 지극히 중요한 요소이다. 감정은 믿음과 의사결정에서 결정적인 역할을 하며 기억을 고착한다. 인공지능은 감정에서의 한계로 인간의 의사결정 과정을 인간과 같은 속도로 모방하지 못한다(Rumsey, 2016).

인공지능은 문화 다양성을 반영하는 데 한계가 있을 수 있다. 문화는 기억과 의미에 광범위한 기틀을 제공한다. 문화는 매우 보수적으로 행동 습관, 믿음, 가치, 지식을 유지한다(Rumsey, 2016). 그런데 인공지능은 문화 맥락 처리에서 한계가 있을 수 있

다. 그래서 갈등해결 역량을 위한 인공지능을 설계할 때 비교문화 관점이 필요하다. 예를 들어, 권력거리(power distance)가 큰 국가에서는 상하 계급의 권력 차이가 매우 크며 부하들이 상사에게 도전하는 것은 적절하지 않다고 여긴다. 그래서 이런 문화에서 대인관계능력과 리더십은 보편적인 역량체계의 역량과 다르다. 대인관계능력은 보편적인 역량체계의 직접적 갈등해결 행동지표와 달리 상급자의 체면을 살리기 위해 중재자를 통해 간접적으로 갈등을 해결하는 방안이 좀 더 긍정적 효과를 낳을 수 있다. 리더십 역량의 경우 권력 거리가 큰 문화에서는 상황에 따라 달라지는 유연한 리더십 유형보다는 지시적 리더십 유형 선호도가 높을 수 있다. 따라서 국가나 지역의 문화적 규준을 정확히 반영하기 위해서 기존의 역량체계를 상황에 맞게 수정할 필요가 있다(Taylor, 2007).

3) Complementarity

인간은 인공지능과의 관계에서 인간과 기계의 네 번째 불연속을 거치고 있다. 기계와 비교해 인간이 특권적 위치에 있다는 무의식적 가정이 틀렸다는 것이 드러나면서 네 번째 불연속을 넘어서는 형이상학적 의식 및 사고방식의 변화가 필요하다는 것이 드러났다. 인간을 기계와 분리해 생각할 수 없다. 그리고 인간과 기계가 완전히 다르다는 생각을 유지하기 어렵다. 인간과 기계의 불연속성은 인간과 기계를 같은 원리로 설명할 수 있고 물질이 진화해 복잡한 형태의 유기체가 되고 생각하는 기계의 구조를 이룬다는 것을 깨닫게 되었기에 이제 간극이 더 빨리 메워진다. 인간과 기계가 아무 차이가 없다는 주장은 인간이 다른 동물들과 똑같다고 우기는 것과 마찬가지로 어리석은 일이다. 인간과 기계의 차이는 정도의 문제이다(Mazlish, 1993).

인간과 기계의 상호 연관성에 접근할 때 인간의 삶에 던지는 의미를 바르게 계속 이해하는 것이 중요하다. 인간이 동물과 기계 양쪽에 기원을 둔 독특한 존재이며 동물의 성질과 기계의 성질이 인간 본성에 녹아 있음을 느낄 수 있어야 한다(Mazlish, 1993). 인간의 진화는 점점 더 문화, 즉 인간의 제2의 본성을 중심으로 전개된다. 기계는 인간 문화의 주요 부분이며, 문화의 한 부분으로서 기계는 인간이 만든 것이지만 그 차제가 유용성을 가진 것으로 보인다. 인간의 본성을 완전하게 이해하려면 인류 진화의 새롭고 복잡한 방식을 이해해야 한다(Mazlish, 1993). 가능한 진화형태로서 인공지능은 이제 인간의 가능성이다(Mazlish, 1993). 기계에 관한 이론은 인간을 이해하는 데 유용하며 그 반대 역시 마찬가지이다. 인간 뇌를 이해하는 것은 인공지능 연구에 빛을 준다(Mazlish, 1993).

3. Categorization of AI Ethics

Pana는 인공지능 윤리를 컴퓨팅윤리, 전산윤리, 기계윤리, 글로벌 정보윤리로 구분하고 그 의미를 다음과 같이 정의한다(Pana, 2006). 컴퓨팅윤리는 인공지능 환경에서 정보를 처리하고 전송하는 다양한 직업의 인터넷 작업자 윤리로서 소프트웨어 속성 보호, 사용자 신원 및 친밀감 보장, 네티켓 공유 및 보존 문제를 윤리적 쟁점으로 삼는다. 전산윤리(Computational Ethics)는 이론 및 실제의 도덕적 문제를 결정하는 토대를 고려하는 윤리로서 전산 방법을 활용해 도덕 이론을 연구해 진화하는 모델링과 시뮬레이션을 통해 의미 있는 사회적 결정의 도덕적 영향을 예측한다. 기계윤리(Machine Ethics)는 컴퓨터 자체에 관한 윤리로서 철학을 배경으로 하는 일반인과 AI 전문가가 협력해 도덕적 기능을 하는 지능적 기계의 철학적(존재론, 가치론, 실용주의, 윤리학) 토대를 연구한다.

본 논문은 인공지능 윤리의 유형을 인공지능에 대한 윤리(Ethics on AI), 인공지능의 윤리(Ethics of Al), 인공과 인공지능의 공진화윤리(Coexistence ethics of Human Being and AI) 구분한다. 다음과 같은 근거 때문이다. 첫째, 인공지능의 발전에 따라 인간의 윤리도 변화하므로 도덕적 행위자로서의 인간 향상을 위해 필요한 인간과 인공지능 행위자의 관계, 윤리적 쟁점과 초점을 분석하기 위해서이다. 둘째, 인공지능 윤리 학습 주제를 범주화하고 교육과정 내용 체계 개발, 교과서 조직, 인공지능에 관한 도덕과 교수·학습 실행 가능성과 효과를 높이기 위해서이다.

1) Ethics on AI

인공지능에 대한 윤리는 인간이 인공지능을 대하는 윤리적 관점으로 인간의 인공지능 사용, 관리, 경영 윤리로 표현할 수도 있다. 인공지능을 통한 진보로 기술 측면에서 무엇이 가능한지를 고려하나 그것보다는 이상적 사회의 특성, 특히 인간존엄성이라는 윤리적 가치 관점에서 진보에 주목하는 것이 필요함을 중시한다. 인공지능에 대한 인간의 윤리의 요소로서 다음 몇 가지 들 수 있다.

첫째, 인간의 삶의 의미, 어떻게 살 것인가, 공정과 정의 실현 등 자동화가 가능하지 않은 문제(Watson, 2016: 336-338)에 대한 인간 자율성과 비판적 사고이다. 외주화의 도덕적 위험은 자기 행동의 결과를 예측하는 속도를 넘어선다는 것, 그리고 지식이 사용되는 방식에 책임을 지지 않으려 한다는 데 있다(Rumsey, 2016). 인공지능은 주관적 판단이 필요한 활동이나 영역에서 효과적으로 기능하지 못할 가능성이 크다. 인공지능이 이런 영역이나 활동에 관련한 물음에 주는 답은 인간이 찾는 답보다 탁월하고 직관력 있는 현명한 것으로 받아들여지지 않을 것이다(Kaplan, 2015).

알고리즘에서 모든 것은 이분법적 판단의 대상이다. 인공지능도 삶의 의미 발견에 도움이 되는 다양한 정보를 줄 수 있을 것이다. 그런데 삶에 의미를 발견하는 근거가 되는 관심사 모두가 건강하고 도덕적으로 가치 있는 것은 아니라는 점에서 인간 삶의 의미를 찾는 데 있어서 인공지능의 역할을 긍정적으로 평가하기는 어렵다. 인공지능이 개인마다 다른 삶의 의미 문제에 답을 줄 수는 없기 때문이다. 예를 들어, 증오는 어떤 사람들의 삶에 의미를 줄 수 있다. 격렬한 증오는 공허한 삶에 의미와 목적을 줄 수 있다. 삶의 의미를 추구할 때 인간의 지성과 정서는 결합해야만 한다. 그리고 현상에서 한걸음 물러나 성찰하는 것이 필요하다(Noddings, 2016).

둘째, 인간은 언제, 어디서, 왜 인공지능을 사용할지 사회적 합의를 형성해야 한다.(Watson, 334-336) 아시모프 로봇 3원칙 수정안은 지침 계획안이 될 수 있다. 원칙 1에서 인간 보호에는 신체 건강뿐 아니라 장기적인 정신 건강도 포함된다. 한편, 적과 싸우는 전투 로봇은 인간이 감독한다고 전제한다면 특정 상황에서는 인간에게 위해를 가하도록 허용해야 한다. 단, 적의 목표가 죄 없는 민간인의 대량학살이라면 원칙을 수정할 필요성이 있다. 원칙 2에서 사회적 합의를 통해 인간의 명령에 순위를 설정해야 한다. 그리고 사고가 발생하거나 인간이 다치거나 죽을 때 책임 있는 행위자가 운영자, 감독, 프로그래머, 설계자, 제조업자 중 누구인지를 명확히 해야 한다. 인간은 기계는 의식을 지니기 전까지 책임을 져야 한다(Watson, 2016). 로봇 원칙 3은 물리적 로봇을 염두에 두었다. 그러나 현대 로봇공학은 소프트웨어도 연구 대상으로 삼는다. 따라서 원칙 3은 데이터 청렴성 유지도 포함해야 한다. 아주 심각한 일이 벌어졌을 때 모든 기록은 전부 공개되어야 한다. 데이터에도 의무와 책임이 있다. 소프트웨어의 명령을 받는 기계처럼 모든 프로그램은 공공의 감시를 받아야 한다. 그래도 가까운 미래에도 인간이 책임을 져야 한다.

셋째, 자신과 타인의 문화 다원성을 반영한 인공지능 설계이다. 의료 인공지능 IBM Watson for Oncology는 미국인에 맞게 개발한 것으로 다른 국가 특히 아시아인 환자를 대상으로 인종 특수성을 고려하지 못한다. 인종 차는 왓슨과 국내 의료진의 의사결정에 차이를 생성한다(최윤섭, 2019).

넷째, 인간 자신을 존중하는 태도이다. 자동화 과정에서 비행기 사고로 사망한 사람 수는 크게 줄었다. 1962-1971년 100만 명당 133명이 사망했으나 2002-2011년에는 100만 명당 2명으로 줄었으나 조종사들의 숙련도는 하락했다. 자동조종 기능에 과도하게 의존하면서 조종사의 상황 인식 저하, 전문지식과 반사 신경 감퇴, 집중력 저하, 수동비행기술 약화가 나타났고 사고의 원인이 되었다(최윤섭, 2019). 인공지능이 인간 자신의 저하시킬 위험성은 기술 발전과 시장 수요가 맞물리는 것이 시간문제라는 점과 관련된다(Colvin, 2015). 교육 인공지능과 관련하여 과제 보고서를 채점하는 소프트웨어와 과제물 작성 소프트웨어 모두 존재하고 둘 다 발전하고 있다. 그런데 과제물 작성 소프트웨어가 채점 소프트웨어를 만족시키는 방향으로 최적화하고 결국 모든 과제물의 점수가 A로 채점되는 상황에 이를 수 있다. 이는 학생과 교사 모두에게 부정적 결과를 낳을 것이다(Colvin, 2015).

다섯째, 인간 스스로 자신의 인공지능에 대한 태도를 성찰하고 통찰을 얻는 태도이다. 인공지능 악용, 인공지능 통제권 포기나 인공지능에 모든 결정을 위임하는 것, 인공지능이 우리가 원하는 것 대신 우리가 요구하는 바를 그대로 줄 것이라는 기대는 우려할 만하다. 인공지능은 평가함수를 설계한 사람이 누구라도 그에게 봉사한다. 이점은 인공지능의 민주주의 문제와 관련된다(Domingos, 2015).

여섯째, 자신과 타인 그리고 기계에 대한 공감이다. 공감은 인간과 인간의 관계를 통해 진화했다. 따라서 공감은 인공지능의 본성이 아니며 인공지능에서 얻을 수 없다. 공감을 키우는 간단하고 효과적인 방법은 놀이 같은 인간 상호작용의 확대이다. 공감이 줄어들고 자기애가 증가하는 현상은 다른 인간과 어울려 놀 기회가 거의 없을 때 예측되는 결과이다(Colvin, 2015). 공감의 토대 중 하나는 실행기능(executive function)인데 지속적 상황 주시, 목표와 계획의 능동적 유지, 주의를 분산하는 자극을 억제한다. 즉 상호작용능력을 포함한 기본 기능을 관리 및 조정하는 능력으로 어려운 문제를 해결하고 문제를 발견해서 고치고 복잡한 활동을 계획하고 어려운 결정을 내리고 장기적이며 이로운 목표를 위해 순간적인 충동을 억제한다. 이 기능은 직접적 대화를 통해 서로를 알아보거나 상대 생각을 아는 활동 등을 통해 발달한다(Colvin, 2015).

일곱째, 인간에 대한 진실성과 사랑이다. 이야기는 인간의 전유물이 아니다. 인공지능도 이야기를 만들 수 있다. 그런데 효과적인 이야기하기는 인간적인 상호작용이 필요하다. 화자가 어떤 사람인지 알고 그 사람과 직접 관계함으로써 그의 진실성을 알며 신뢰를 형성한 경우 이야기는 청자에게 큰 영향을 미친다. 이야기의 영향력은 화자와 청자의 신뢰관계 형성에 달려있다는 것이다. 그래서 청자는 화자가 누구인지 평가하고 믿을만하지 결정하고 그의 진정한 열정을 판단하기 전에는 움직이지 않는다. 어떤 사람이 이야기하는가가 매우 중요하게 작용한다는 것이다. 돌턴(G. W. Dalton)과 동료들의 변화에 대한 연구 결과는 훌륭한 정보원으로부터 받는 피드백이 제대로 수용된다는 것을 밝혔다. 평판이 나쁜 사람의 피드백은 신뢰를 받지 못한다는 것이다. 피드백 제공자에 대한 존중이 없으면 피드백은 변화를 일으킬 수 없다(Colvin, 2015).

2) Ethics of AI

인공지능의 윤리는 인공지능 자체에 관련한 윤리적 쟁점으로 설계자 윤리이거나 초인공지능의 경우에는 기계 자체의 도덕성과 관련된다. 인공지능의 윤리는 무행위(inaction) 문제, 편견 없는 설계 문제, 인공 도덕 행위자의 책임과 관련한다.

첫째, 인공지능의 무행위의 윤리적 문제이다. 컴퓨터는 임의성에 기대지 않고 선택할 수 있다. 증거의 경중을 가리고, 지식과 전문성을 적용하고, 불확실성에서 결정을 내리고, 위험을 무릅 쓰고, 새로운 정보에 기초해 계획을 수정하고 자기 행동의 결과를 관찰하고 기호 추론을 통해 추론하고 인과관계에 대한 깊은 이해가 없는 상태에서도 기계학습 기술을 적용해서 행동 방향을 정하는 것 등 직관으로 불릴 능력을 발휘할 수 있다(Kaplan, 2015). 그러나 인공 도덕 행위자의 무행위는 윤리적으로 살펴볼 필요가 있다. 자기 책임능력을 고려하는 공학자들은 로봇이 어려운 문제나 상황을 직면하면 우선 작동을 멈추고 인간이 문제를 해결할 때까지 기다려야 한다고 생각한다. 행위와 무행위 사이의 도덕적 구별이 있다고 해도 인공 도덕 행위자의 설계자는 좋거나 옳은 행위를 외명하고 무행위를 무작정 선택하면 안 된다. 서비스 로봇에게 규정된 의무와 과제, 예를 들면, 돌봄이 필요한 사람에게 몇 시간에 한 번씩 약을 가져다주어야 한다는 책임에 어긋나기 때문이다(Watson, 2016).

둘째, 편견 없는 결정이다. 정교한 판단이 필요한 업무들과 관련해 알고리즘의 편견 없는 결정은 인간에 대해 비교 우위를 보인다. 인공지능은 점점 더 인간이 처리할 수 없는 복잡한 결정을 믿고 맡길 수 있는 존재가 되어간다. 인공지능은 복잡한 결정을 내려야 하는 인간에게 단순히 원자료를 넘기는 존재가 아니다. 이는 2008년의 금융시장 붕괴의 촉매였다는 점에서 잘 나타난다. 기업의 고위 의사결정 과정에 인공지능이 이용되고 어떤 경우에는 인공지능의 판단에 따르는 의존 현상이 앞으로도 확산할 것이다(Cameron, 2017).

셋째, 인간의 윤리적 관습과 행동을 이해하고 따르는

설계이다. 강한 인공지능 윤리란, 윤리적으로 타당한 방식으로 행동하도록 인공지능을 설계해야 하는 것으로 표현할 수 있다. 어떤 사람이 자율주행자동차 지붕에 올라타는 것 같은, 기기를 만든 사람이 예측하지 못한 상황이 발생한다면 자율주행자동차는 어떻게 해야 할까? 운행을 계속하면서 어떤 상황에서든 인간을 위험에 처하게 해서는 안 된다는 단순한 원칙을 추론해서 실제로 올바른 판단을 내리게 될 수도 있다. 로봇이 갖추어야 할 행동적 요구는 도덕적 측면을 넘어서 사회적 측면까지 확장된다. 인공지능은 어떤 행동이 필요하며 사회적으로 어떤 의미가 있는지 전반적으로 의식할 수 있는 로봇이므로 인공지능 시스템이 인간의 관습과 행동을 반드시 이해하고 따르도록 설계해야 한다(Kaplan, 2015).

넷째, 인공지능의 정신범죄(mind crime)를 예방하는 것이다. 정신범죄는 특히 도덕적으로 고려해야 할 이해관계를 포함하는 프로젝트 수행 실패 상황으로 인공지능이 도구적 이유로 한 행동의 결과로 나타날 수 있는 부작용이다. 기계 초지능은 도덕적 지위를 가지는 내부 과정을 창조할 수 있다. 예를 들면 인공지능은 인간 정신을 아주 상세하게 시뮬레이션하면서 자각할 수 있으며 인간 심리와 사회를 잘 이해하기 위해 자각 있는 인간 모델을 수조 개 만드는 상황에서, 인간 모델들은 가상공간에 놓여 다양한 자극에 노출되고 각각에 대한 반응을 연구할 수 있다. 일단 한 인간 모델의 정보적 유용성이 소진되고 나면 그것을 파괴할 수도 있다. 이러한 관행이 높은 수준의 도덕적 지위를 가진 존재, 예를 들면 인간 모델이나 여러 종류의 지각 있는 지성체에 적용될 때 이 행위는 학살에 견줄 수 있다. 따라서 인간 모델이나 디지털 지성체에게 죽음과 고통을 주는 도덕적 문제가 될 수 있다(Bostrom, 2014).

다섯째, 인공지능을 설계한 사람의 책임윤리이다. 인공지능은 뇌처럼 자동적이고 규칙 지배적이고 결정적 도구지만, 인간은 결정하는 데 자유롭고 개인적으로 책임 있는 행위자이다. 책임은 사람들이 상호작용할 때 발생한다. 개인적 책임은 공적 개념이다. 즉 책임은 당신이 타인의 행동에 그리고 타인이 당신의 행동에 가지는 개념이다. 책임은 뇌에 부여하는 것이 아니라 인간에게 부여하는 것이다. 책임은 규칙을 따르는 인간 동료들에게 요구하는 도덕적 가치이다. 책임의 문제는 사회적 선택의 문제이다. 사회 규칙 안에서 만들어진 책임이라는 개념은 뇌의 신경구조 안에는 없다(Gazzaniga, 2005).

인공지능 윤리 문제는 인간이 기술을 의인화하는 경향에서 비롯된다. 기계가 가지고 있지 않은 능력을 기계에 부여하는 데 비롯되는 피해를 말한다. 인간 도덕 행위자에게 바라는 정도의 지능과 의도를 실현하는 기술이 부족한 상태에서 그런 능력을 기계에 부여하는 시도는 위험하며 인간이 책임을 지지 않게 할 수 있다(Wallach & Allen, 2009).

좋은 인공 도덕 행위자가 무엇을 의미하는지 밝히는 문제는 설계자가 어떤 목적을 가져야 하는지 명확히 하는 것이다(Wallach & Allen, 2009).

좋음은 설계자와 사용자가 부여한 특정 목적을 기준으로 측정된다. 자동화 시스템에 요구되는 좋은 행동은 쉽게 특정될 수 없다. 좋은 자동 행위자는 인간에게 해를 주지 않고서는 행동을 취할 수 없다면 인간 감독자에게 알려야 할까? 이런 의미에서 좋음에 대한 언급은 윤리의 영역에 들어서게 된다. 좋은 도덕 행위자란 자신의 행동이 미칠 수 있는 해나 의무 태만 가능성을 인식해야 하며 그러한 지식에 비추어 바람직하지 않은 결과를 피하거나 줄이는 행동을 선택할 수 있어야 한다. 하지만 컴퓨터가 도덕적 결정을 하거나 윤리적 판단을 한다는 것은 아니다. 컴퓨터가 취하는 행동의 윤리적 중요성은 그 속에 프로그래밍한 규칙에 내재한 가치에서 전적으로 비롯된다(Wallach & Allen, 2009).

여섯째, 인간 개인의 안전과 프라이버시 보장, 윤리적 행동 실천 가능성 향상을 보장하는 설계이다. 기계 윤리와 컴퓨터 윤리의 교차점에 놓인 중요한 사안은 웹에 떠도는 데이터 수집 로봇인데, 프라이버시 기준을 거의 또는 전혀 존중하지 않고 정보를 뒤진다. 컴퓨터로 정보를 복제하는 것은 수월해 지적 재산권에 대한 법적 기준이 침해되므로 저작권법에 대한 재검토가 요구되었다. 인터넷 아카이브에서 사용되는 자료수집 봇은 자신들이 모으는 자료의 도덕적 중요성을 평가할 수 없다(Wallach & Allen, 2009).

3) Coexistence Ethics of Human Being and AI

인간을 위한 인공지능 윤리는 인공지능과 인간의 공존, 그리고 인공지능을 통해 인간 자신이 도덕적

정체성과 주체성을 회복하는 데 필요한 윤리이다. 인간을 위한 인공지능 윤리는 인간의 도덕적 행위성과 정체성 향상이라는 좋은 결과를 낳을 수 있는 공진화와 공존윤리에 초점을 둔다. 인공지능을 사용해 인간의 이타주의적인 이해와 관용을 발달시킬 수 있다. 인간의 도덕은 선호, 제약, 무책임의 도덕성이다. 인간의 도덕성과 인간 윤리는 기계 윤리의 모델이 될 수 없다. 이제 보편적이고 전적으로 발명된 과학기술 기반 그리고 그런 의미에서 인간과 기계 모두가 구현할 수 있는 인공 윤리를 구성할 때이다(Pana, 2006). 인간을 위한 인공지능 윤리로 다음 몇 가지 들 수 있다.

첫째, 인공지능은 자전거나 철도 등 다른 발명품보다 더 근본적 수준에서 인간 삶에 영향을 끼친다. 인공지능이 더 잘, 더 빠르게, 더 낮은 비용으로 보편화하면서 사회의 부가 증대된다면 부의 공정한 분배 방법을 고민해야 한다(Kaplan, 2015).

둘째, 인간이 자기를 성찰적으로 이해하는 것이다. 인공 도덕 행위자 설계는 도덕 행위자인 인간 자신 그리고 윤리 이론의 속성을 이해하는 데 활용된다. 일부 기본적인 도덕적 결정은 컴퓨터에 구현하기 쉬울 수 있으나 더 어려운 도덕적 난제를 다루는 기술은 현재 기술을 넘어선다. 인간의 능력을 로봇에 구현하는 데 필요한 단계적 절차를 세세하게 밟아가며 도덕적 결정이 내려지는 방법에 따라 사고하는 연습은 인간의 자기이해 과정이다. 이런 점에서 인공지능은 자아의식에도 영향을 미친다(Wallach & Allen, 2009; Kaplan, 2015).

기계윤리는 인공 도덕 행위자 구현과 관련한 철학적, 실제적 문제만큼이나 인간의 의사결정에 관련된다. 인공 도덕 행위자에 관한 성찰 및 제작 실험을 통해 인간이 어떻게 작동하고, 인간의 어떤 능력이 기계 설계에 구현될 수 있으며, 동물이나 인간이 창조한 새로운 형태의 지능과 구별되는 인간의 속성을 깊이 생각할 수밖에 없다. 인공지능이 마음에 관한 철학 분야의 새로운 탐구 방향을 이끈 것처럼 기계윤리는 윤리학의 새로운 탐구 방향을 이끌 수 있다. 로봇 연구 및 AI 실험실은 도덕적 의사결정 이론을 인공적인 시스템에서 검사하는 실험실이 될 수 있다. 인공 도덕 행위자 제작에 관한 윤리적 고찰이 필요한 것은 인공지능이 이미 일상 활동의 윤리적 영역으로 들어왔기 때문이다.(Wallach & Allen, 2009)

파나는 인공 도덕 행위자의 특성을 논의하면서 인간과 인공지능의 공존 윤리를 제안한다(Pana, 2006). 인간 윤리는 인간 도덕성의 본질과 주로 인공지능 행위자의 몇 가지 특성에서 도출된 기계 윤리와 인간의 근본적 불일치 때문에 기계윤리의 효율적 모델이 될 수 없다. 인공 도덕적 행위자는 복잡성 수준, 특수한 기능, 어느 정도의 자유가 필요하다는 주된 이유로 다른 필수 자질들이 부여되는 개별 실체로 생각해야 한다. 인공 도덕적 행위자는 a)개별 실체(복잡한, 전문적인, 자율적이거나 자기결정적이며 예측할 수 없는 것), b)개방적, 자유롭게 행동하기도 하는 시스템(결정의 특수한, 유연한, 휴리스틱 기제와 절차), C)문화적 존재: 자유로운 행동은 인간(자연적) 또는 인공 존재의 행동에 문화적 가치를 준다, d)수업과 교육에 개방적인 시스템, e)상태표현과 생활표현이 있는 실체, f)도덕지능(moral intelligence) 같은 다양한 또는 여러 가지 형식의 지능이 있는, g)자동성(automatism), 지능, 신념(인지와 정서의 복합물)이 있는, h)성찰(reflection)하는(도덕적 삶은 의식적 활동만 아니라 영적 형식이다), i)실제나 가상 공동체의 구성 요소/구성원으로 다루어질 수 있거나 다루어져야 한다.

기계윤리는 어려운 문제, 즉 도덕적 자유에 관한 실천 명령에 이론 요구 조건을 결합하는 문제를 해결해야 한다. 인간 행동에서 도덕성은 도덕적 자유를 실천하는 것을 가정한다. 이렇게 가정하고 도덕 규범을 적용할 때만 인간 행동은 문화적 가치가 있을 수 있다. 또, 기계의 경우 자유와 책임의 상호관계는 매우 제한적인 것으로만 기능해야 한다. a)책임은 자유의 조건이다, b)자유는 책임의 원인이다. 더 명확히, (b)에서, 책임의 정도는 자유의 종류와 수준에 달려 있고 그것이 결정하기도 한다.

따라서 기계윤리는 결정된 행동 영역과 상황에 도덕규범을 적용할 때 선택의 자유를 보장할 필요성을 고려해야 한다. 인간은 기계에 자유를 허용할 때 발생하는 위험을 감수하고, 기계가 자신의 자유도를 증가할 가능성도 수용해야 한다. 기계의 행동은 인간뿐 아니라 기계의 결정, 그리고 다른 기계들의 결정에도 의존한다. 이는 인간과 기계가 적용할

수 있는 인공 윤리로, 둘은 더 나은 공통의 윤리로 자연과 인공 사이에서 만날 수 있다(Pana, 2006).

파나는 도덕적 체계를 구조적 수준의 위계를 포함하는 매우 다양한 형태의 지능으로 제시한다. 도덕적 체계는 사회생활에 꼭 필요하다. 도덕적 체계에서 살기 위해서 우리는 다양한 형태의 지능을 다양하게 결합해 사용해야 한다. 지적 인공 행위자를 위한 도덕 규칙을 구현하는 개념적 및 기술적 방법을 찾을 때 도덕적 삶의 복잡성을 인정해야 한다.

III. Approaches to AI Ethics in Moral education

1. Learning AI Ethics in Moral Education

도덕교육에서 인공지능 윤리를 학습한다는 것은 인간성의 모사로서 인공지능을 봄으로써 인공지능을 통해 인간 자신의 도덕적 정체성을 탐구하는 것이라 할 수 있다. 인공지능의 도덕은 인간의 도덕을 모사하는 것이라는 입장에서 인간이 창조한 인공지능은 인간의 거울이다. 인공지능이 인간 삶을 식민지화했다고 표현할 만큼 기원이 짧아도 모든 곳에 침투했기 때문이다.

앞으로 학생들은 수명 연장, 일의 변종 증가, 여가 생활을 점점 더 많이 여가 생활을 점유할 지적 기계 인터페이스에 대한 참여, 기술에 대한 인간의 참여에 끝없이 적응해야 한다. 자기창조 및 재창조 능력, 스스로 시간을 의미 있게 활용하는 능력, 정신적 삶, 다양한 종류의 관계능력, 변화에 대처하고 다루며 즐기는 능력 등의 기능을 학습해야 할 것이다. 이 기능들은 학제적 역량, 감성 지능, 깊은 관계 능력을 요구할 것이다. 인간의 창의성과 공감 능력은 인공지능이 지배하는 작업 환경에서 직업의 중심요소가 될 것이다(Cameron, 2017).

미국의 인공지능 데이터 기업 Figure Eight가 발간한 2017 *Data Scientist Report*에는 “인공지능 윤리와 관련해 중요하게 생각하는 쟁점이 무엇인가?”라는 물음의 응답 결과가 실려 있다. 응답자들은 ‘기계학습에 프로그램화된 인간의 선입견과 편견’, ‘전쟁에서의 인공지능과 자동화 사용’, ‘보편적으로 합의된 도덕규범의 프로그래밍 불가능성’, ‘인간 노동의 기계 대체’, ‘건강관리(비밀유지, 예측, 진단)와 의료의 기밀성, 인공지능과 피할 수 없는 유전병 예측 가능성’, ‘자율주행 자동차와 트롤리 문제, 안전’ 순서로 쟁점을 들었다.

<Appendix 2>은 인공지능 및 인공지능 윤리에 대한 선행연구를 근거로 학교 도덕교육에서 인공지능 윤리를 교수·학습할 때 활용할 수 있는 교육내용을 추출하고, 교육내용을 도덕성 4구성 요소와 도덕교육의 핵심가치(core values)와 관련해 제시한 것이다.

2. Moral Dilemma as educational contents for learning AI Ethics

다음은 인공지능 선행연구에서 인용한 추출한 인공지능 윤리의 도덕적 딜레마 예이다. 학생들의 발달 수준과 흥미, 도덕성 발달 초점 목표에 맞게 재구성해서 활용할 수 있다.

- 마을 주차장(도심에 있는 시립도서관 등)에는 차를 계속 옮기면서까지 장기 주차하는 차량이 많지 않으리라는 계산에서 두 시간 무료 주차할 수 있다. 그러나 차에 스스로 이동 주차하는 기능이 탑재되면 어떻게 될까? 내가 소유한 로봇이 영화관의 매표소에 나 대신 줄을 서도 관습적으로나 윤리적으로 문제가 없을까?
- 차량 1대가 간신히 다닐 수 있는 좁은 다리를 내 차가 건너는데 건너편에서 아이들이 가득 탄 통학차가 갑자기 다리에 들어섰다. 차량 2대가 동시에 지날 수 없으므로 한 대는 양보해야 한다. 통학차에 탄 아이들을 위해 양보하는 결정을 할 차를 사는 것이 옳은가? 새 차를 고를 때 연비를 고려하듯이 자율주행차를 살 때 차량의 공격성이 중요한 고려사항이 될 가능성은 없는가?
- 나의 로봇 여러 대를 물건을 사 오라고 보냈다면, ‘1인당 1개’라는 문구를 어떻게 해석해야 할까? 내 로봇이 나를 위해 거짓말을 해도 괜찮을까? 추수감사절 저녁 식사

때 미성년 딸아이에게 포도주를 따라주라고 로봇에게 지시한다면 로봇은 나를 경찰에 신고해야 할까?

· 개를 산책시키는 로봇이 '잔디밭에 들어가지 마시오.'라는 지시를 지키기 위해서 개에게 물어뜯기는 아이를 구하지 않고 그대로 서 있어야 하는가?

· 심장마비로 쓰러진 사람이 목숨이 위태로워 한시 바삐 병원에 데려가야 한다면 자율주행차는 신호등이 빨간색일 때마다 멈추어 신호가 바뀌기까지 기다리며, 또 안전 속도를 준수하며 가야 하는가?

· 도로에 갑자기 개가 뛰어들면 개를 피하려고 앞지르기가 금지된 중앙선을 넘어야 할까?

· 자율주행차가 차에 탄 사람의 목숨을 구하기 위해 길가의 개를 치고 지나갈 수 있다. 그런 경우는 상대적으로 분명하게 선택할 수 있다. 그러나 나이 든 부부와 한 무리의 아이 중 한쪽을 쳐야 한다면 어떻겠는가? 앞자리에 앉은 아이와 뒷자리에 앉은 아이 중 한 명을 죽음으로 몰고 가야 하는 선택에 처한다면 어떤 결정을 내려야 할까?

· 인간보다 엄청나게 뛰어난 지능을 가져 다른 기계만이 그 기계를 이해할 수 있는 진화한 기계는 지루해하고 우울증을 느끼고 피해망상증 환자가 될 수 있나?

다음은, 인공지능 윤리와 관련한 도덕적 추론 학습을 위한 가상 이야기다.

① 열 추적 미사일 딜레마

발사된 열 추적 미사일은 적 전투기의 열원인 엔진 열을 추적해 전투기에 접근해 전투기에 닿자마자 폭발해 승무원을 죽인다. 미사일을 로봇이라면, 법칙 1을 위배한다. 이때 우리가 할 수 있는 것은 무엇일까? 미사일이 적전투기를 폭파하기 전에 확성기로 '전투기가 곧 폭발하니 비행기에서 탈출하라'고 경고하는 '인간 검출기'를 미사일에 탑재해야 할 것이다. 미사일은 적 조종사가 탈출하기 전까지 전투기를 폭파할 수 없다. 그런데 불행히 적 조종사는 전투기가가 음속보다 빨라 미사일의 경고를 들을 수 없다. 경고를 듣는다고 해도 조종사는 탈출 후에야 전투기를 폭파한다는 것을 알고 있으므로 탑승한 채로 있을 것이다. 이는 아시모프의 법칙을 지킬 때 일어나는 어리석은 상황이다(Warwick, 1997).

② 아시모프 로봇 윤리의 책임성 문제

A가 소유한 지능 로봇을 상상해보자. A는 B와 싸우고 있고 로봇은 주인 A가 B를 죽이는 것을 방관한다. 로봇은 개입할 힘이 있으나 개입하지 않기로 한다. 로봇이 인간이라면 많은 국가의 법률상 로봇은 B를 죽인 A와 마찬가지로 유죄이다. 로봇은 침묵함으로써 살인에 동의했다. A는 어떤 희생을 치른다 해도 주인을 보호하도록 로봇에게 가르치거나 로봇을 프로그래밍 한다. 로봇은 사실상 경호원으로 행동한다. A가 C를 만나고 C가 A를 심하게 공격한다면 로봇은 아시모프의 법칙 1을 어기는 행동이라도 C를 죽일 수 있다. 만약 로봇이 C가 A를 살해하지 못하도록 하지 못하면 임무를 충실히 수행하지 않는 것이다. 그러나 동시에 의무를 다하는 것이기도 하다. 로봇은 승산이 없는 상황에 놓인다. 이러한 문제에서 로봇이 어떤 행동을 하던 간에 누군가는 죽을 수 있다. 현실에서 로봇을 통제할 법칙은 없다. 아시모프의 법칙은 허구이다(Warwick, 1997).

IV. Conclusions and Implications

본 논문은 미래 예측 불가능성은 간과한 영향력, 부적절한 정보, 인간의 오류 가능성, 우발성, 복잡한 기술이 상호 영향을 미치고 새로운 가능성을 유발하는 성질 등이 원인이며 전문가나 준전문가의 미개입 상태에서 실용적 전문성을 공급하는 모형을 사용할 때 윤리적 제약은 시스템이 더욱더 유능해

지기 전에 사회적 논의가 필요하다는 데 동의한다.

인공지능 윤리와 인간윤리에 대한 논의는 궁극적으로 인간 자신을 돌아보는 사고 과정이다. 인간은 인공 도덕 행위자를 통해서 인간의 윤리적 역량이 보편적으로 취약함을 깨닫고, 기계와 인간 윤리의 다름을 통해 인간의 감정과 의식이 다름을 더 깊이 이해하게 된다. 그리고 도덕교육 차원에서 인공지능의 증강현실에 힘입어 실제에 더 가깝게 도덕성을 발달시킬 방법을 모색할 수 있다. 또 인공지능에 탑재된 비교문화 차원의 윤리적 경험 사례들을 통해 인공지능 윤리에 대한 추론과 의사결정 능력을 실질적으로 향상하는 교육 방안을 구안할 수 있다.

앞으로 학교 도덕교육에서는 도덕성, 도덕적 행위성, 도덕 발달, 인공지능 설계를 위한 알고리즘을 중심으로 인공 도덕 행위자 이해, 인공지능 윤리의 범주와 쟁점, 인공지능 윤리를 통한 인간과 인공지능 윤리 향상 방안에 대한 교수·학습이 이루어져야 한다. 인공지능 윤리는 학생들에게 삶의 조건이며 해결할 현실 문제이기 때문이다. 체계적인 인공지능 윤리교육의 설계와 실행을 위해서는 도덕·윤리 교사, 학생, 윤리교육 전문가, 윤리적 인공지능 알고리즘 설계자들의 인공지능 윤리 교육 요구 분석, 학생들의 흥미에 적합한 학습 주제 선정과 조직, 다양한 학습 자료 개발, 교사와 학생의 인공지능 윤리교육 요구 및 실태 등에 대한 경험연구가 필요하다.

References

Anderson, M., & Anderson, S. L. Eds. (2011). *Machine Ethics.* New York: Cambridge University Press.

Bostrom, Nick (2014). Super Intelligence. Jo S. J. trans. *Super Intelligence.* Seoul: KKachi Books.

Cameron, Nigel M. (2017). Will Robots Take Your Job? Go H. S. trans. (2018). *Robots and Your Job*, Seoul. Eum Publishing.

Choi T. S. (2019). *Medical AI*, Seoul: Cloudnine.

Colvin, Geoff(2015). *Humans Are Underrated*, Shin D. S. trans. (2016). *Humans have been underestimated*, Seoul: Hansmedia.

De Schrijber, A., & Maesschlak, J. (2015). A New Definition and Conceptualization of Ethical Competence. In Terry K. cooper & Donald C. Menzel, Eds.. *Achieving Ethical Competence for Public Service Leadership.* New York: Routledge.

Domingos, Pedro (2015). *The Master Algorithm: How the quest for the Ultimate Learning Machine Will Remake Our World*. Kang H. J. trans. (2016). *Master Algorithm*. Seoul: Business Books.

Folkman, Joseph R. (2006). *The Power of Feedback*. Lee J. H. trans. (2007). *The Power of feedback*. Seoul: Bookfolio.

Gazzaniga, M. (2005). *The Ethical Brain*, Kim H. Ytrans. (2009). *The Ethical Brain*, Seoul: Bada Books.

Ha, J. B. (2016). Neuroscientific Correlates of Political Orientation and Morality. *Brain, Digital, & Learning, 6*(3).

Kaplan, Jerry (2015). *Humans NEED NOT APPLY*, Shin D. S. trans. (2016). *Humans NEED NOT APPLY*, Seoul: Hans media.

Kim, K. H. (2019). Analysis of needs for moral education curriculum and direction for curriculum improvement. Paper presented at the Summer Conference of The Korean Society for Moral & Ethics Education.

Kwon Seung-Hyuk, Eom, Jeung-Tae, & Kwon, Yong-Ju (2014). Brain regions associated with digital contents addiction: Focus on brain-imaging for internet and video game addiction. *Brain. Digital, & Learning, 4*(1). 11-20.

Lee, S. H. trans.(2019). *How to embed moral engine in AI*. Seoul: Cloudnine.

Lee, Sang Hee (2019). Criticism of educational philosophy on educational neuroscience. *Brain, Digital, & Learning, 9*(2). 49-57.

Lee, Yeong-Ji, & Kwon, Yong-Ju (2016). Human sociality in a view of brain activation: Focused on social pain and pleasure. *Brain, Digital, & Learning, 6*(1). 21-29.

Ma, Min-young, Yang, Eun-kyung, Lee, Gyeo-re, Baek, Mi-kyung, Cho, Young-min, Cho, Hyun-jin, & Shin, Jae-hong (2013). Necessity of

neuroscientific approaches for the study on mathematics anxiety. *Brain, Digital, & Learning, 3*(2). 1-10.

Matthews, Eric (2005), Mind: Key concepts in philosophy, Ahn S. C. trans.. What does mind mean to you: philosophical reflection on mind., Seoul: Peolchim Books.

Mazlish, Bruce(1993), The fourth discontinuity: The co-evolution of humans and machines, Kim H. B. trans (2001), The Fourth Discontinuity, Seoul Science Books.

Nick Bostrom (2014). S*uperintelligence: Paths, Dangers, Strategies.* Jo S. J. trans. (2017), *Superintelligence.* Seoul: KKachi Books.

Noddings, Nell (2016). *Peace Education*, New York: Cambridge Press.

Pana, Laura (2006). Artificial Intelligence and Moral intelligence, *Triple C*, *4*(2), 254-264.

Rumsey, Abby Smith (2016). *When We are No More-How Digital Memory Shaping Our Future*, Kwak H. S. trans. (2016). *Memory disappearing era*, Seoul, Uknow Books.

Siegel, Judith P. (2010). *Stop Overreacting: Effective Strategies for Calming Your Emotions.* Lee Y. N. trans (2017). *Emotional Control*. Seoul: Sigmapress.

Squire, L. R. & Kandel, E. R. (2009). *Memory: From Mind to Molecules*. Jeon, D. H. trans. (2016). *Secret of memory*, Seoul: Haenamu.

Susskind, Richard & Susskind, Daniel (2015). *The Future of the Professions: How the technology transform the Work of Human Experts.* Wee D. S. trans (2016), *Era of the Fourth Industrial Revolution, the Future of Professions.* Seoul: Wiseberry.

Taylor, Ian (2007). *The Assessment and Selection Handbook*, Kim M. S., & Jo. I. C trans. (2014). *Handbook of Evaluation Center*. Seoul: Sigmapress.

Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong.* No. T. B. trans. (2016). *Why is the moral of robot*. Seoul: Medichi.

Warwick, Kevin (1997). *March of the Machine*, Institute of System control at KAIST trans. (1999). *March of Machine*, Seoul: Hanseung Publisher.

Watson, Richard (2016). *DIGITAL vs HUMAN*, Bang J. I trans (2017). *For those who are afraid of the age of AI*, Seoul: Wonderbox.

[저자의 소속과 직위]
김국현 : 한국교원대학교, 교수, 단독저자

<Appendix 1> Theoretical framework of ethical competencies

Categories	Knowledge	Skills	Attitude & Abilities
규칙 기준	법, 윤리강령, 규칙/절차	규칙 적용	규칙 중요성 이해
윤리적 민감성	조직과 사회에서의 지위	윤리 관점에서의 상황 정의 다른 해결 방법 모색	공감, 역할채택
윤리적 판단력	자기, 타인, 규칙에 미치는 결과	다른 도덕적 논증 사용	융통성: 자기, 타인, 규칙에 미치는 결과로 제한하지 않음
윤리적 동기와 품성	자기보다 타인과 규칙에 미치는 결과 중시	자기 행동 선택 시 타인과 규칙에 미치는 결과 우선	자율성 자아강도 (도덕적 용기)

<Appendix 2> Learning topics for AI ethics in moral education

Moral Values	Learning Themes	Ethical Competencies
성실	· 로봇에게 옳고 그름을 가르칠 수 있을까? · 로봇에게 '옳고 그름'을 가르치려면 어떻게 해야 할까? · 인공지능이 어떤 가치를 증진할 수 있을까, 또는 증진해야 할까? · 로봇을 인간 삶에 받아들이면 소중한 인간적 가치들을 약화하고 인간성을 타락시키게 될까? · 인간은 컴퓨터가 도덕적으로 중요한 결정을 내리기 정말 원하는가?	윤리적 민감성 윤리적 동기화
책임	· 도덕적 의사결정을 기계화하려고 시도한다면 인간에게 어떤 결과가 생길까? · 기계를 지능적 도덕 행위자로 만들려고 하는 시도는 잘못된 것일까? · 인공지능은 인간을 대체할 수 있는가? · 인공지능이 낸 사고는 누가 책임을 져야하는가?	윤리적 판단력
정의	· 아시모프의 3원칙은 로봇의 윤리적 행동을 보장할 수 있는가? · 인공지능도 범죄를 저지를 수 있을까? · 인공지능의 윤리적 판단은 인간의 판단보다 우월한가? · 인공지능을 통해 증가한 사회적 부는 어떻게 분배되어야 할까?	윤리적 판단력
배려	· 기계를 윤리적으로 배려해야 하는가? · 로봇과 사랑에 빠질 수 있을까? · 기계를 향한 사랑이 인간을 향한 사랑을 뛰어넘을 수 있을까? · 편견이 없는 알고리즘을 만들기 위해 설계자나 일반 시민이 무엇을 할 수 있는가? · 인간과 컴퓨터가 정말 똑같은 방식을 쓰고 있을까, 아니면 인간이 세상과 상호작용하고 배우는 방법에는 뭔가 다른 점이 있는 걸까? 만일 최종 결과로 같은 행동이 나타난다면 그런 구별에 과연 의미가 있을까?	윤리적 민감성