『人文科學』 제130집 2024년 4월
DOI http://doi.org/10.23017/inmun.2024.130..133

인공지능 딜레마, 철학교육에서 다루기*

홍예리**

I. 들어가며: 인공지능 윤리교육, 철학교육, 인공지능 딜레마[1)]

논문을 시작하기에 앞서, 본 논문에서 다루는 용어와 범주에 대한 몇 가지 언급이 필요할 것이다. 우선 본 논문이 가진 문제의식의 출발점은 인공지능 윤리교육이었다. 코켈버그(2020)는 “AI 윤리는 기술 변화와 그것이 개인의 삶에 미치는 영향뿐 아니라 사회와 경제의 변화에 관한 것”이라고 말한다(20).

* 본 논문의 초고는 2024년 2월 17일, 한국철학교육학회 2024 동계학술대회에서 발표한 바 있다.
** 이화여자대학교 이화철학연구소 연구원

1) 투고 당시 본 논문의 I장은 본 논문에서 다루고자 하는 인공지능의 범주에 대해서만 논의하고 있었으나, 심사위원 중 한 분께서 I장의 논의가 필수적이지 않다고 지적해 주셨다. 다른 두 분의 심사위원도 주로 I장에서 언급한 개념 정의와 관련하여 미흡한 점을 지적해 주신 것으로 미루어 보아, 본 논문의 문제의식 및 다루고자 하는 용어와 범주가 투고본의 I장에서 명확히 드러나지 않은 것으로 판단되었다. 따라서 I장에서 불필요한 부분을 삭제하였으며 본 논문의 전반적인 문제의식 및 본 논문에서 혼재되어 다루어지는 세 가지 용어 및 범주에 대한 설명을 제시하는 방향으로 I장 전체를 수정하였다. 중요한 통찰력을 제공해 주신 세 분의 심사위원께 모두 깊이 감사드린다.

인공지능은 자율주행차, 자율 무기, 추천 알고리듬, 콤파스(COMPAS) 시스템과 같은 법원의 의사결정 등에 이미 일상적으로 사용되고 있으며, 그러한 사용은 우리 사회와 경제를 바꾸어 놓고 있다. 이러한 상황에서, 인공지능과 관련하여 윤리적인 교육의 필요성은 이미 널리 받아들여지고 있다.

코켈버그의 『AI 윤리에 대한 모든 것』(2020)을 번역한 신상규, 석기용(2023)은 <옮긴이의 말>에서 "국내에는 AI 윤리와 관련한 책들이 이미 여럿 나와 있다"고 인정하면서도, "대학에서 AI 관련 윤리를 가르치면서 느낀 한 가지 아쉬움은, AI 윤리 중 일부 제한된 문제만을 다루거나 교양 차원에서 가볍게 읽을 만한 책은 많지만 관련 쟁점을 모두 다루면서도 그 서술이 정확하고 체계적인 책은 많지 않다"고 한다(239). 옮긴이들의 말처럼, 인공지능 윤리와 관련된 수업은 일부 주제만을 다루거나 여러 주제를 교양 차원에서 가볍게 다루는 두 종류로 나누어지는데, 이러한 현상은 인공지능 윤리 수업이 진행되는 방식과도 연관이 있다.

김효은(2020a)에 따르면 인공지능과 관련된 윤리를 가르치는 방식은 크게 두 가지로, 수강생들이 공학적이고 기술적인 측면을 실습하면서 그 안에 내재된 형태로 윤리적인 고려를 하게 만드는 것과, 인공지능이 적용되는 다양한 분야의 문제들을 토론하거나 자신의 입장을 세워보는 방식으로 다루는 것이 있다. 다시 말하면, 인공지능 윤리 교육은 "프로그래밍 실습이 포함된 수업과 그렇지 않은 수업"으로 나누어질 수 있다(홍예리, 2023, 45). 프로그래밍 실습을 하는 수업을 공학에 내재된 형태의 수업이라고 할 수 있을 것이고, 이러한 수업에서는 인공지능 윤리와 관련된 문제 중 알고리듬 편향, 투명성, 설명 가능성, 공정성 등을 다룬다.[2] 한편 프로그래밍 실습이 없는 수업에서는 다양한 분야에서 인공지능과 관련된 윤리적 문제를 다룬다. 옮긴이들이 "AI와 관련

2) 김효은(2023)의 "공정한 인공지능(Fair Artificial Intelligence)" 수업의 경우, 도덕적 로봇을 만드는 수업이 포함되어 있고, 인공지능의 편향, 설명가능성, 공정성 개념을 중심으로 구성되었다. 광주과학기술원의 MOOC 강의인 "인공지능 로봇의 윤리" 수업(김건우, 2021)도 데이터와 프라이버시 문제, 편향과 차별, AI 로봇 규율을 위한 대응 방안 및 AI 로봇의 구현 방법 등을 주로 다룬다.

한 여러 철학적, 윤리적 쟁점을 가장 포괄적이고 체계적으로 다루고 있다"고 평가한 코켈버그의 책(2023, 239)에는 두 종류의 수업에서 모두 다루는 주제들이 잘 드러나 있는데, 프로그래밍 실습을 하는 인공지능 윤리 수업에서 주로 다루는 주제로는 프라이버시, 투명성, 설명 가능성, 편향, 공정성 등의 개념이 있고, 프로그래밍 실습이 꼭 필요하지 않은 주제로는 인간과 인공지능 사이의 관계, 인공지능의 기술적 및 존재론적 본성, 인공지능의 도덕적 지위, 불평등, 일의 미래, 인공지능과 기후 변화 및 에너지 문제와의 관계 등이 있다.

인공지능과 관련된 윤리적 이슈를 가르치는 수업을 설계할 때, 두 방식 중 어느 하나의 방식이 절대적으로 옳은 것은 아니다. 프로그래밍 실습을 한다면 수강생이 실제로 프로그래밍의 과정에서 윤리적인 측면을 고려해 볼 수 있다는 장점이 있지만, 반대로 기술적인 이해 자체가 수업의 문턱을 높일 수 있다. 프로그래밍 실습을 하지 않는다면 다양한 주제에 대해서 토론할 시간적 여유가 더 많겠지만, 한편으로는 기술적인 이해가 선행되지 않아서 윤리적인 주제와 관련해서도 무엇이 문제인지 이해하기 어려운 상황이 발생할 수 있다. 또 두 방식이 양립불가능하여 반드시 어느 하나를 택해야 하는 것도 아니다. 공학에 윤리가 내재된 방식의 수업을 일부 포함하면서 더 넓은 범위의 주제들을 다룰 수도 있다. 추가로, 필자는 기존의 두 가지 방식에 덧붙여 인공지능 윤리에서 다루는 관련 이슈들에 대하여 스스로 질문을 형성하는 탐구공동체적 성격이 가미될 필요가 있다는 주장을 펼친 바 있다(홍예리, 2023).

이때 '스스로 질문을 형성하는 탐구공동체적 성격'을 제공해 주는 것이 바로 본 논문에서 이야기하는 철학교육이다. 철학교육은 마치 국어교육, 수학교육, 과학교육처럼 하나의 교과 과목으로서의 철학교육을 이야기하는 것이 아니다. 그러므로 교과교육으로서 윤리교육과 명확하게 선을 긋는 수업이 아니다. 또한 역사 순으로 철학 이론의 발전 과정을 다루는, 즉 철학사적인 수업을 말하는 것도 아니고, 특정 철학자의 입장을 분석하여 이해하는 것을 목표로 하는 수업도 아니다. 대신, 본 논문에서 이야기하는 철학교육은 매튜 립맨(Matthew Lipman)이 철학교육 방법론인 '어린이를 위한 철학(Philosophy for

Children, 이하 P4C)'에서 말하는 철학교육이다.[3] P4C 방법론에서는 철학교육에 참여하는 교수자와 수강생이 하나의 탐구공동체가 되는 것을 중시한다. 전통적인 교수자와 수강생 대신, 토론의 진행자(facilitator)와 탐구공동체를 이끌어가는 일원이 있을 뿐이다. 즉, P4C 방법론에 따르면 철학교육은 철학이라는 학문을 가르치기보다 '철학함' 자체, '철학을 하는 방법'을 가르치는 것이며, 스스로 질문을 던지고 탐구공동체가 서로의 질문을 공유하는 것을 중심으로 한다. 이러한 의미의 철학교육을 인공지능 윤리교육에 적용할 경우, 기존의 인공지능 윤리교육에서 다루는 여러 주제를 메타적으로 생각하고 질문을 던지는 단계로 나아갈 수 있을 것이다. 예를 들어, 자율주행차에 관한 수업에 P4C 방법론을 적용한다면, 단순히 트롤리 딜레마에서 어느 쪽을 택하는지 편을 정하거나, 찬성과 반대로 나누어 토론하거나 해결책을 모색하는 것을 넘어서서, 운전이라는 행위가 우리의 삶에서 가지는 의미에 대해서까지 메타적으로 생각해 볼 기회를 제공할 것이다.[4]

그리고 인공지능 윤리 수업이 다룰 수 있는 여러 주제 중, 중요한 가치들이 상충하며 그중 하나를 택해야 하는 진퇴양난의 상황을 나타낸다는 점에서 '딜레마'라는 표현으로 묶일 수 있는 주제들이 있다. 이러한 주제들은 서로 다른 분야로 흩어져 있고, 가치의 상충과 진퇴양난 상황을 다룬다는 공통점이 있지만 그 모든 주제가 '딜레마'라는 이름으로 묶여서 다루어지는 것은 아니다. 예를 들어, 앞서 옮긴이들이 많은 범주의 주제를 다룬다고 언급한 코켈버

3) P4C가 '어린이를 위한 철학'의 줄임말이기 때문에 어린이만을 대상으로 하는 것으로 오해할 수 있다. 그러나 P4C 방법론과 함께 립맨이 개발하고 발전시킨 미국 뉴저지 몽클레어 주립대학교의 IAPC 커리큘럼을 살펴보면, 어린이뿐 아니라 중고등학생에 해당하는 청소년도 대상으로 하여 설계되었다. 또한 "질문을 만들어 내는 과정을 중시하는 P4C의 토론 수업 진행 방식은 어린이뿐 아니라 다양한 연령대의 사람들에게 적절하게 조절될 수 있고, LMS와 같은 현실적인 온라인 도구를 이용하는 방식으로도 변형될 수 있다." (홍예리, 2023, 69) 따라서 본 논문에서 언급하는 철학교육 및 P4C 방법론은 특별한 나이나 학년에 제한되지 않고 철학 토론 수업에 전반에 일반적으로 적용할 수 있는 방법론을 의미한다.

4) 구체적으로 P4C 방법론을 인공지능 윤리교육에 적용한 토론 수업으로의 응용 방법은 필자의 다른 논문(홍예리, 2023)을 참조하기를 바란다.

그의 저서(2020)는 명시적으로 인공지능과 관련된 딜레마 상황을 다루고 있지 않다. 변순용과 이연희(2020)는 자율주행차와 관련된 트롤리 딜레마를 다루고 있지만(192-212), 자율주행차와 유사하게 딜레마 상황을 유발하는 다른 주제를 딜레마로 분류하여 다루지 않는다.[5] 한편 인공지능과 가치 연구회의 저작(2021)에서는 기존 공학 윤리의 관점에서 본 이득과 윤리 간의 딜레마(22-23), 전자 인격 부여의 딜레마(47-48), 자율주행차와 관련된 트롤리 딜레마(55-71), 투명성 개념과 관련된 딜레마(86-91), 킬러 로봇과 관련된 의사결정의 딜레마(94-95) 등 많은 주제를 '딜레마'라는 이름 하에 다루고 있다.

이러한 상황을 기반으로 하여, 본 논문은 인공지능 윤리 수업과 관련하여 등장하는 딜레마를 세 가지 종류로 정리하고, 세 종류의 딜레마를 철학교육적으로 어떻게 다룰 수 있을지 고민하고자 한다.

II. 인공지능 딜레마: 세 가지 종류

딜레마는 다른 말로 진퇴양난의 상황이라고도 하며, 두 개의 판단 사이에 끼어 어느 쪽도 결정할 수 없는 상태를 말한다. 인공지능과 관련된 딜레마는 여러 가지가 언급되지만, 크게 세 종류로 나눌 수 있다. 첫째는 구체적인 상황에서 특정한 윤리적 딜레마가 제시되는 경우, 둘째는 인공지능 윤리와 관련된 기본적인 개념 중 특정 개념이 본질적으로 딜레마를 발생시키는 경우, 마지막은 인공지능의 개발 전반과 관련된 딜레마다. 이 장에서는 각각의 딜레마가

5) 변순용과 이연희(2020)의 경우, 9장에서 창작하는 인공지능이 가져올 긍정적인 전망과 노동이 사라짐으로써 인간의 본질을 훼손할 위험에 대해서 언급하고 있으며(167-168), 15장에서는 킬러 로봇과 관련하여 효율성과 윤리성의 문제를 다루면서, 킬러 로봇이 인간 병사보다 더 효율적인지, 또 감정 없이 더 윤리적인 결정을 내릴 수 있는지에 대한 논의를 담고 있다(269-270). 16장에서는 소셜 로봇이 인격적인 대화 상대자가 될 수 있는지, 실제 사람과 동일한 외형의 로봇을 허용해야 하는지를 다룬다(283-287). 이러한 주제는 모두 II장에서 소개할 상황별 딜레마로 분류될 수 있으나, 해당 저작에서는 특별히 '딜레마'로 분류되고 있지는 않다.

어떤 내용인지 살펴볼 것이다.

1. 상황별 딜레마

인공지능과 관련된 딜레마는 주로 구체적인 상황에서 주어지는 윤리적 선택과 관련된 경우가 많다. 이들 딜레마는 특정한 상황에서 주어진 두 가지 가치 중 무엇을 선택할지 묻는 형태이다.

가장 많이 언급되는 인공지능 윤리 관련 딜레마는 자율주행차와 관련된 딜레마다. 김은경, 이영준(2023)에 따르면, 고등학교 인공지능 기초 교과서 8종 모두에서 자율주행차와 관련된 딜레마를 다루고 있다. 이 딜레마는 철학에서 기존에 공리주의적 선택과 관련하여 널리 알려진 트롤리 딜레마(Trolley Dilemma)를 변형시킨 것이다. 전형적인 트롤리 딜레마는 브레이크가 고장 난 탄광 전차가 소수의 사람과 다수의 사람들 중 한쪽을 희생시킬 수밖에 없는 상황에서 어느 쪽을 선택할지 묻는다. 이 딜레마 속 탄광 전차가 자율주행차로 바뀌면, 실제로 자율주행차가 현실에서 마주하는 선택 상황에서 무엇을 선택하도록 훈련할지와 관련된 딜레마가 된다. 나아가 딜레마 상황은 단순히 소수와 다수 중 어느 쪽을 희생시킬지 비교하는 것을 넘어서, 희생당하는 쪽에 다양한 특성을 추가하여 다양한 가치를 평가하게끔 하는 방향으로 응용될 수 있다.

트롤리 딜레마를 자율주행차와 관련된 것으로 응용한 연구 중 가장 대표적인 것은 MIT 미디어 랩의 모럴 머신(Moral Machine, www.moralmachine.net)이었다.[6] 모럴 머신에서는 자율주행차와 관련하여 13가지 윤리적 딜레마 상황을 제시하였다. 공통적인 상황은 다음과 같다. 각 방향 1차선뿐인 도로에서

6) 심사위원 중 한 분께서 모럴 머신에 대한 소개 대신 김은경과 이영준의 연구(2022)와 같은 선행연구에 관한 내용으로 대체할 것을 권하셨다. 그러나 김은경과 이영준(2022)은 초등학생의 인공지능에 대한 인식에 모럴 머신이 부정적인 영향을 미쳤다는 연구인 반면, 본 논문은 인공지능과 관련된 딜레마에 메타적인 질문을 던지게 하는 것을 목적으로 하므로, 방향성이 조금 다르다고 판단하였다.

자율주행차의 브레이크가 고장 났고, 콘크리트 장벽에 충돌하면 자율주행차의 탑승객이 모두 사망한다. 제시되는 상황은 모두 전형적인 트롤리 딜레마처럼 희생자의 숫자도 고려 대상이지만, 몇 가지 기준이 추가된다. 첫째, 자율주행차 안에 탑승객이 있는 경우와 없는 경우로 나눌 수 있다. 탑승객이 있는 경우, 탑승객을 보호할지 또는 외부의 대상을 보호할지 결정해야 하고, 탑승객이 없는 경우, 외부의 여러 대상 중 어느 쪽을 보호하거나 희생해야 할지 결정해야 한다. 둘째, 맞은편 차선 쪽으로 방향을 튼다는 능동적인 개입을 선택할지 아닐지에 따라서도 충돌의 결과는 달라진다. 능동적인 개입을 택할 때 희생되는 대상과 그렇지 않은 대상이 달라지기 때문이다. 셋째, 보행자가 무단횡단 중인지 아닌지도 고려 대상이다. 보행자가 무단횡단 중이라면 보행자의 잘못이 더 크다고 판단하겠지만, 보행자가 초록색 보행 신호에 길을 건너고 있다면 윤리적 딜레마는 더 커질 것이다. 이외에도 모럴 머신은 탑승객과 보행자의 성별·나이·체력·사회적 가치관 및 종까지도 고려하여 13개의 선택지 세트 중 하나를 고르게 설계되어 있다.

다음으로, 자율주행차의 의사결정뿐 아니라 인공지능이 사용되는 다양한 상황에서 제시되는 딜레마가 있다. 김은경, 이영준(2023)은 기존의 고등학교 인공지능 기초 교과서 8종에서 모두 인공지능 딜레마를 다루고 있지만 특히 자율주행차와 관련된 딜레마에 지나치게 편중된 점을 지적하면서 다양한 인공지능 딜레마 개발이 요구된다고 주장한다. 이들은 특히 트롤리 딜레마를 변형시킨 자율주행차 딜레마가 인간을 해치는 극단적인 사례만을 반복적으로 제시하여 교육적으로도 부적절하다는 점을 지적한다. 이러한 문제의식을 바탕으로, 이들은 다음과 같이 9가지의 구체적인 인공지능 윤리 딜레마 문항을 제시한다.[7)]

---

7) <표 1>은 김은경과 이영준(2023)이 표 4, 5, 6으로 구분한 것을 본 논문에서 편의상 하나로 합쳐놓은 것이다. 해당 딜레마 문항들은 http://aiethics.kr/에서 확인할 수 있다.

〈표 1〉 김은경, 이영준(2023)이 개발한 9가지 인공지능 윤리 딜레마

순서	관련 기술	상충하는 가치
1	AI 스피커	노약자 안전 vs. 사생활 침해
2	안면인식 CCTV	사회 안전 vs. 인권/사생활 침해
3	경찰 로봇	사회 안전 vs. 로봇으로 인한 사람의 상해 가능성
4	딥페이크 기술	정서적 유용성 vs. 범죄 악용 가능성
5	도시 청소 로봇	효율성 vs. 노인 일자리 감소
6	드론 배달 & 택배	효율성 vs. 배달 & 택배 관련 일자리 감소
7	킬러 로봇	방어적 목적 vs. 인공지능의 살인 방조
8	아티스트 인공지능	(예술 작품으로서의) 가치 인정 vs. 가치 불인정
9	제미노이드	정서적 유용성 vs. 인간 고유의 영역 침해

9가지 딜레마는 현재 다양한 분야에서 개발 중이거나 상용화가 이루어지는 인공지능 관련 기술을 중심으로 하여, 각 기술마다 일반적으로 중요하게 여겨지는 두 가지 가치가 충돌하는 상황으로 개발되었다. 1~3번은 사회적 공공선과 인간의 기본적인 권리 중 어느 것을 선택하는지와 관련된 딜레마다. 구체적으로 1번은 사회적 공공선과 사생활 보호, 2번은 사회적 공공선과 인권, 3번은 사회적 공공선과 인간의 신체 보호권이 침해될 가능성이 충돌하고 있다. 예를 들어, AI 스피커는 집 안의 상황을 감지하다가 노약자가 위급한 상태일 때 119에 신고할 수도 있지만, 사생활 침해의 우려가 있다. 4~6번은 사회적 공공선과 기술 합목적성이 충돌하는 사례로, 좋은 목적으로 개발된 기술이지만 악용되거나 예상치 못한 사회적 부작용이 생길 우려가 있는 경우다. 예를 들어, 딥페이크 기술로 이미 고인이 된 가까운 사람의 영상을 만들 경우, 슬픔을 달랠 수 있다는 정서적 유용성이 있지만, 같은 기술이 범죄에 악용될 수도 있다. 5번과 6번 기술은 제대로 작동한다면 효율성 측면에서 효과를 기대할 수 있지만 관련되는 일자리가 줄어들 것이라는 우려가 있다. 7~9번은 기술 합목적성과 인간 존엄성이 상충하는 경우로, 좋은 의도에서 개발되었지만 인간의 본질을 약화하거나 손상할 우려가 있는 경우이다. 9번의 경우, 김은경, 이영준(2023)은 “부모와 똑같은 목소리와 모습을 한 돌봄 로봇을 개발”하는 상

황을 제시하였으나 필자는 이 사례가 넓게 보면 제미노이드(Geminoid)에 속하는 것으로 보았다.[8] 딥페이크 기술과 마찬가지로, 제미노이드도 정서적인 측면에서 일부 유용할 수 있지만, 인간과 제미노이드 사이의 경계가 어디이며 인간의 본질이 무엇인지에 관한 질문을 제기하고, 나아가 인간의 본질을 흐릿하게 만들 수도 있다.

이처럼 트롤리 딜레마를 자율주행차의 의사결정에 적용한 한 가지 종류의 딜레마에서도 상황에 따라 여러 가지 응용 가능성이 있으며, 자율주행차를 제외하더라도 오늘날 개발되었거나 개발 중인 여러 분야의 인공지능 기술과 관련하여 다양한 종류의 딜레마가 존재한다. 또한 앞서 제시된 딜레마 외에도 여러 가지 응용 및 변형을 통해서 또 다른 딜레마 상황을 설정할 수 있을 것이다.

2. 투명성의 딜레마

인공지능 윤리에서 대표적으로 거론되는 핵심 개념 중 한 가지는 투명성(transparency)이다. 최소한의 규칙만 주어지면, 인공지능은 스스로 문제 해결 방식을 찾는 것을 목적으로 학습할 수 있지만, 바로 이 인공지능의 학습 과정 및 문제 해결 방식을 찾는 의사결정 과정을 외부에서 알 수 없다. 사용자뿐 아니라 개발자도 이 모든 과정을 정확히 알 수는 없기에, 인공지능의 의사결정 과정은 블랙박스(blackbox)라고 불린다. 즉, 인공지능이 어떤 과정을 거쳐서 학습하고 의사결정을 하는지 그 속이 깜깜해서 들여다볼 수 없다.

이 블랙박스 문제를 해결하기 위해서 제시된 개념이 투명성 개념이다. 투명성 개념에 따르면, 블랙박스이기 때문에 들여다보기 어려운 인공지능의 학습 및 의사결정 과정에 대해서 인간이 그 과정을 되짚어 볼 수 있도록 투명하게 설명이 주어져야 한다. 인공지능 윤리에서 강조되는 또 다른 개념인 설명가능성(explicability) 개념도 투명성 개념에서 비롯된다(인공지능과 가치 연구

8) 제미노이드는 쌍둥이를 뜻하는 'Gemini'와 휴머노이드(humanoid)의 합성어이며 특정한 실존 인물과 닮게 만든 인공지능 로봇을 일컫는다.

회, 2021, 87). 자신의 절차와 과정을 스스로 설명할 수 있는 인공지능을 개발하는 것이 투명성을 충족시키기 위한 한 가지 방법이기 때문이다.

인공지능 윤리에서 투명성 개념은 중요하다. 인간에게 막대한 영향을 미치는 인공지능의 학습 과정 및 의사결정 과정을 인간이 알고 있어야 그 방향을 올바르게 이끌 수 있고, 예상치 못한 문제가 생겼을 때 무엇이 원인이었으며 누가 책임을 져야 하는지 밝혀낼 수 있기 때문이다. 또한 투명성 개념은 이미 발생한 사고가 다시 발생하지 않도록 인공지능을 수정하거나 추후 인공지능과 관련하여 발생할 사고를 예방하는 데에도 필요한 개념이다.

그런데 인공지능과 가치 연구회(2021)에 따르면, 투명성 개념은 그 개념 자체가 딜레마를 발생시킬 여지를 가지고 있다. 가장 쉽게 떠올릴 수 있는 첫 째 딜레마는 정보 보호와 관련된 딜레마다(ibid. 90-91). 인공지능의 학습 및 의사결정 과정을 추적하면서 해당 인공지능이 학습한 모든 데이터가 공개될 수도 있고, 해당 인공지능을 개발한 기업 고유의 정보가 노출될 수도 있다. 인공지능의 학습 데이터에는 민감한 개인 정보가 포함되어 있을 수도 있고, 해당 개인은 자신의 정보가 인공지능 학습에 이용되는 것을 동의하였을지라도 그 내용이 인공지능 의사결정과정의 투명성을 위해서 공개되는 것은 거부하였을 수도 있다. 또한 기업 입장에서는 핵심적인 기업 비밀이 누설되지 않는 선에서 투명성을 추구해야 한다는 딜레마가 발생한다.

둘째, 인공지능의 발전 방향과 관련된 딜레마다(ibid. 86-87). 인공지능이 발전할수록 그 내부의 구조적인 알고리즘과 학습 및 의사결정 과정은 더욱 복잡해지기 때문에, 투명성을 확보하기가 더 어려워진다. 특히 오늘날 대규모 데이터를 학습하여 스스로 문제 해결 방식과 답을 만들어 내는 생성형 인공지능의 경우, 데이터가 방대하고 그 과정이 너무나 복잡해서 인간이 모든 것을 모니터링하는 것은 쉽지 않다. 즉, 인공지능이 발전하여 점점 더 많은 데이터를 학습하고 의사결정이 정확해질수록, 투명성을 만족시키기는 어려워지는 것이다. 이런 측면에서 볼 때, 투명성 개념은 본질적으로 인공지능의 발전 방향과 상충하는 개념이다. 투명성을 확보하는 것에 집중하면 인공지능의 발전

이 더뎌질 수 있고, 인공지능의 발전을 우선시하면 투명성을 확보하기가 어려워지기 때문이다.

이 문제를 해결하기 위해서 설명가능한 인공지능(Explainable Artificial Intelligence, XAI)을 개발하거나 인공지능의 학습 및 의사결정 과정을 또 다른 심층 신경망에 연결해서 기록을 남기는 해결책이 제시되고 있지만, 여기에도 몇 가지 문제가 더 얽혀 있다(ibid. 87-89). 우선, 기술적인 측면에서 인공지능의 모든 의사결정 과정을 설명하도록 만드는 것은 비현실적이기도 하고 불필요한 일이기도 하다. 또한 투명성 개념을 지나치게 강조하다 보면 저장해야 하는 데이터의 양이 늘어나기도 하고, 전반적인 유지에 대한 부담이 늘어날 수도 있다는 점도 고려해야 한다. 게다가, 보다 근본적이고 철학적인 차원에서, 과연 설명이 무엇인지 명확한 정의가 필요하다는 문제도 있다. 인공지능이 어떤 과정을 거쳐서 의사결정을 했는지 아무리 상세한 기록을 남겨도 사용자 관점에서 납득할 만한 설명이 아니라 단지 해석의 여지가 열려 있는 정보일 가능성이 있다. 또한 인간에게는 의사결정과정의 투명성이 그다지 엄격하게 요구되지 않지만, 인공지능에만 엄격하게 투명성을 요구한다는 점에 대해서도 문제 제기가 이루어질 수 있다. 인간도 본인이 어떤 판단을 왜 내렸는지 그 과정의 단계를 스스로 낱낱이 알기는 어려운 경우가 많기 때문이다.

마지막으로, 투명성 개념이 필요하긴 하지만 과연 어느 정도까지 우선시되어야 하는 개념인지에 대한 고민도 남아있다. 투명성 개념은 인공지능 윤리의 주요 개념 중 책무와도 관련이 있다. 인공지능 윤리에서는 책임(responsibility)과 책무(accountability) 개념이 구분된다(김효은, 2019, 57-59). 책임은 어떤 사고 또는 사건이 발생한 이후에, 즉 사후적으로 문제가 되는 권리를 침해하거나 보장한 사람에게 처벌이나 보상을 부여하는 것이다. 한편 책무는 영어 표현인 'accountability'가 잘 드러내듯, 어떤 사고 또는 사건에 대해서 어떻게 그러한 결과가 나오게 되었는지에 대한 설명도 포함하는 개념이다. 한 사람에 의해서 다른 사람의 권리가 침해당하면, 그 사람이 잘못한 것이므로 책임을 지고 처벌을 받거나 보상을 하면 된다. 반면 인공지능과 관련하여

인간이 피해를 보는 일이 발생하였을 때, 책임보다 책무 개념이 더 필요하게 된다. 그 이유는 인공지능이 의사결정을 내리는 과정에서 왜 그렇게 결정하게 되었는지 설명이 필요하고, 그러한 설명이 주어져야 궁극적으로 책임 소재를 명확하게 가릴 수 있기 때문이다. 이런 관점에서 볼 때, 의사결정과정이 불투명한 블랙박스라는 오늘날 인공지능의 특징이 결국 투명성, 설명가능성, 책무라는 개념을 요구하게 되고 이 개념들이 모두 같은 맥락에서 얽혀 있다는 점을 알 수 있다.

그런데 투명성을 최우선의 가치를 놓고 인공지능의 의사결정과정을 낱낱이 밝힌다고 해도, 책무가 밝혀진다는 점이 보장되지는 않는다(인공지능과 가치 연구회, 2021, 101-102). 앞서 투명성은 정보 보호나 인공지능의 발전 방향과 딜레마 관계에 있다고 하였는데, 정보 노출을 감수하고 발전을 더디게 해서라도 투명성 개념을 중시해야 한다면, 그 이유는 바로 책무 때문일 것이다. 하지만 모든 것을 감내한다 해도, 책무가 자동으로 밝혀지는 것은 아니다. 앞서 설명가능한 인공지능과 관련해서도, 아무리 자세한 기록을 남겨도 여전히 해석의 여지가 남아있는 정보 덩어리에 불과할 뿐 납득할 만한 설명의 수준에 이르지 못할 수도 있다고 하였다. 마찬가지로, 투명성을 높이기 위해서 인공지능의 의사결정과정의 단계마다 아무리 자세한 기록을 남긴다고 해도, 여전히 어느 부분에서 잘못된 것이며 왜 특정한 결과가 발생하였고 책임 소재가 어디에 있는지를 찾기 어려울 수 있다. 또, 인공지능의 의사결정과정은 아무런 문제가 없었지만 애초에 그 인공지능이 학습하도록 주어진 데이터 세트가 편향으로 가득한 자료였을지도 모른다. 즉, 책무 개념은 투명성 개념이 요구되는 이유이자 일종의 목표 중 하나이지만, 그 목표 달성을 위해서 투명성 개념이 과연 어느 정도로 우선시되어야 하는지에 대한 의문이 제기될 수 있다.

정리하자면, 오늘날 인공지능의 의사결정과정이 불투명한 블랙박스이므로 투명성은 인공지능 윤리에서 반드시 요구되는 개념이다. 투명성은 인간이 인공지능의 의사결정과정에 대한 통제권을 가지는 방법의 일환으로서도 필요하고, 인공지능과 관련한 사건 사고에 있어서 책무를 밝히는 데에도 필수적이

다. 하지만 애초에 불투명한 의사결정과정을 투명하게 만들어야 하므로, 여러 어려움을 발생시키는 개념이기도 하다.

3. 인공지능 발전과 방향 모색: 죄수의 딜레마

마지막 딜레마는 트리스탄 해리스가 "딜레마"라고 표현하는 것이다. 전직 구글의 디자인 윤리학자였던 그는 2020년 다큐멘터리 영화 ≪소셜 딜레마(The Social Dilemma)≫ 및 인터페이스 디자이너 에이자 라스킨(Aza Raskin)과 함께 한 2023년 강연 ≪인공지능 딜레마(The A.I. Dilemma)≫에서 "딜레마"라는 표현을 사용하였다. 이들이 주장하는 인공지능의 딜레마는 특정한 상황이나 개념에 국한된 것이 아니다. 이 딜레마는 본 논문에서 정리한 세 종류의 딜레마 중 가장 큰 맥락의 딜레마인데, 해리스와 라스킨이 우려하는 것은 오늘날 거대 언어 모델을 기반으로 한 생성형 인공지능의 발전 속도와 방향 전반에 대한 것이기 때문이다.

현재 인공지능 개발에 뛰어든 기업은 매우 많고, 세계적인 규모의 기업들이 인공지능 개발에 열을 올리고 있다. 2024년 1월 9일에 열린 CES(국제전자제품박람회, the International Consumer Electronics Show)에서도 단연 화두는 인공지능이었다(아주경제, 2024; 파이낸셜 뉴스, 2024). 세계 최대 규모의 정보통신기술 융합 전시회인 이 행사에서, 국내 기업인 삼성도 자사의 인공지능 가우스(GAUSS)를 소개하였다. 인공지능은 이제 특별한 것이 아니라 거의 대부분의 생활가전에 탑재되어서, 똑똑하지 않은 가전제품은 도태되는 시대가 되었다.

≪인공지능 딜레마(The A.I. Dilemma)≫에서 해리스와 라스킨은 우리가 이제 인공지능과의 첫 번째 접촉을 넘어서서 두 번째 접촉 단계로 들어섰다고 진단한다. 첫 번째 접촉은 우리의 관심과 집중을 사로잡는 SNS를 통해서였고, SNS 뒤에는 추천 알고리즘이라는 인공지능이 작동하고 있었다. 이 종류의 인공지능은 마치 『모모』의 회색 신사처럼 우리 인생에서 시간을 훔쳐 갔고, 끊

임없는 타인과의 비교를 통해 무엇이 우리에게 중요한지 잊게 했다. 두 번째 접촉은 언어처리모델인 GPT로 대표되는 생성형 인공지능의 등장이다. GPT는 언어에 초점을 맞춘 언어처리모델이었지만, 2024년 1월 출시한 삼성의 새로운 스마트폰 갤럭시 24에 탑재된 가우스는 언어와 이미지 등 여러 개 분야를 넘나드는 멀티모달 생성형 인공지능이다. 언어처리모델의 등장 이후 멀티모달 생성형 인공지능의 등장은 인간이 만든 모든 문화적 산물은 언어를 기반으로 하였기 때문에 가능하였다. 인류 문명의 모든 산물은 언어로 치환될 수 있고, 인공지능이 자연 언어를 학습할 수 있게 되면서, 언어처리 및 학습 능력은 곧 인류 문명 전체를 학습하는 것으로 이어졌다.

해리스와 라스킨은 거대 언어 데이터를 기반으로 하여 등장한 이러한 종류의 인공지능을 "골렘 인공지능(GLLMM AI, Generative Large Language Multimodal AI)"이라고 부른다.[9] 이들의 주장에 따르면, 골렘이 다른 인공지능보다 특히 위험한 것은 언어 능력을 갖췄고 그 발전 속도가 기하급수적으로 빠르기 때문이다. 2018년 GPT는 인간과 비교할 수 있는 수준의 문제해결력을 갖추지 못했다. 2020년에서야 본격적인 발전을 보여 4살 정도의 문제해결력을 보였다. 그리고 2022년 1월에는 7살 수준이었지만, 같은 해 11월에는 9살 수준으로 발전하였다. 인공지능이 전공자 수준의 수학 문제를 해결할 확률에 대해서, 2021년 전문가들은 정답률 52%까지 발전하는 데 4년이 걸릴 것으로 전망하였다. 하지만 1년도 지나지 않아, 인공지능은 정답률 50%를 넘는 데에 성공했다(Harris & Raskin, 39:40). 이미 여러 멀티모달 생성형 인공지능은 텍스트 수준의 학습을 넘어서서 이미지, 영상, 음성 등 여러 종류의 자료를 학습하고 있다.

해리스와 라스킨이 생성형 인공지능으로 인류가 얻게 될 이점을 무시하는

9) 골렘(Golem)은 서양 이야기 속 거인 괴물을 말한다. 판타지 소설이나 게임에서 지하 감옥이나 보물을 지키는 괴물로 등장하고, 돌이 뭉쳐서 만들어진 거인의 모습으로 나타난다. GLLMM은 "Generative Large Language Multimodal"의 약자이지만, 이를 "Golem"으로 읽는다는 것은 그만큼 이러한 종류의 인공지능이 거대한 괴물 같은 가능성을 가지고 있으므로 주의해야 한다는 의도가 담긴 표현이다.

것은 아니다. 무조건 두려워하면서 디스토피아적인 전망만 강요하는 것도 아니다. 인공지능 개발을 아예 폐기하자는 것도 아니다. 다만 이들은 우리가 개발 속도에 제동을 걸고, 잠시 시간을 내서 "실제로 일어나고 있는 일(what is happening)"과 "일어나야만 하는 일(what needs to happen)" 사이의 차이에 대해서 생각해 보고, 이 골렘을 앞에 놓고 우리가 어느 방향으로 나아가야 할지 논의해야 한다는 점을 강조한다(Harris & Raskin, 1:02:30). 골렘이 발전하는 속도는 너무 빠른 반면, 우리는 아직 첫 번째 접촉으로 인한 부작용, 즉 SNS와 추천 알고리즘으로 인한 여러 문제도 다 해결하지 못했다. 항상 그랬듯이, 법이나 윤리적 가이드라인이 발전하는 속도는 기술의 발전 속도보다 느리다. 우리는 골렘이 개발되는 속도를 좇아서 골렘을 활용하고 골렘의 의사결정에 의존하지만, 골렘을 규제할 법과 철학은 골렘의 발전 속도를 따라가지 못하고 있다.

그리고 바로 여기에서 딜레마가 등장한다. 우리는 잠시 멈추어서, 골렘이 우리를 어디로 이끌지 생각해 볼 시간을 내야 하지만, 그 시간을 낼 수 없다. 골렘이 성장하는 속도에 그저 끌려갈 뿐이다. 해리스와 라스킨에 따르면, 상황이 이렇게 된 것은 인공지능의 개발에 대한 무한 경쟁 때문이다. 이들은 인공지능 개발 경쟁을 일종의 군비경쟁에 비유한다. 도태되기 싫다는 두려움 때문에 모두가 발전 속도에 제동을 걸지 못하고 엄청난 자원과 에너지를 쏟아부으면서 경쟁의 속도를 더 빠르게 만들고 있다는 점에서, 인공지능 개발 경쟁은 군비경쟁과 유사하다. 실제로 이들은 강연 중 인공지능 개발 속도를 늦출 때 미국이 중국에 뒤처지게 되는 것에 대한 많은 사람의 우려도 언급한다(Harris & Raskin, 59:51). 그러나 이들은 인공지능 개발 경쟁이 핵무기 개발 경쟁보다 더 무서운 상황이라고 본다. 핵무기는 학습을 통해서 자신을 스스로 강하게 만들지 못하지만, 인공지능은 그게 가능하기 때문이다. 인공지능 개발이 일종의 군비경쟁이 되는 상황을 표로 나타내면, <표 2>와 같다.[10]

10) A와 B는 각각 글로벌 기업이나 개별 국가로 생각할 수 있다. 해리스와 라스킨의 강연에서 이러한 표가 등장한 것은 아니지만, 이들의 강연에서 A와 B는 미국과 중국에 해당한다.

〈표 2〉 인공지능 개발 가속과 휴지(休止) 및 성찰의 딜레마

	B의 인공지능 개발 가속	B의 인공지능 개발 휴지(休止) 및 성찰
A의 인공지능 개발 가속	무한 경쟁	A만 발전, B는 도태
A의 인공지능 개발 휴지(休止) 및 성찰	B만 발전, A는 도태	모두가 고민하여 인공지능의 발전 방향을 찾음

A와 B가 인공지능 개발을 잠시 멈추고 모두가 고민하여 인공지능의 발전 방향을 찾는 것이 가장 이상적인 선택지이다. 그러나 현실은 상대방만 발전하고 자신은 도태될 것이 두려워서 양측 다 개발을 선택하고, 무한 경쟁 상태에 놓인다.

어디서 많이 본 것 같은 딜레마 아닌가? 자율주행차와 관련된 딜레마가 철학에서의 트롤리 딜레마와 본질적으로 같은 것처럼, 인공지능 개발을 둘러싼 군비경쟁은 결국 오래된 죄수의 딜레마(Prisoner's Dilemma)와 같다. 수학자 존 내시(John Nash)가 게임 이론에서 제시한 이 딜레마의 내용은 다음과 같다. 공범 관계로 추정되는 두 명의 용의자가 경찰서에 잡혀 있다. 이들을 각각 불러서 자백할 기회를 주는데, 경우의 수는 네 가지 있다.

〈표 3〉 죄수의 딜레마의 선택지

	용의자 B의 자백	용의자 B의 묵비권 행사
용의자 A의 자백	A, B 모두 3년 형	A 석방, B 10년 형
용의자 A의 묵비권 행사	A 10년 형, B 석방	A, B 모두 1년 형

<표 3>을 보면, 가장 좋은 선택지는 용의자 A와 B가 모두 묵비권을 행사하여 각자 1년 형만 받는 것이다. 하지만 각자 자신의 이익을 위해서 행동하고 상대의 행동을 확신할 수 없으므로, 양쪽 모두 자백하는 쪽을 택해서 둘 다 3년 형을 살게 된다.

군비경쟁은 개별 국가 간 협정이 이루어지기 어렵고 상대의 행동을 확신할 수 없으므로 자신에게 이득이 되는 쪽으로 결정하지만 결과적으로 모두가 손해를 본다는 점에서 본질적으로 죄수의 딜레마와 같다. 인공지능 개발도 마찬가지다. 각 기업 간, 그리고 국가 간 어느 쪽도 속도를 늦출 수 없다. 해리스와 라스킨은 강연의 시작과 마무리에서 다음과 같이 말한다.

> 기술이 권력을 부여하면, 경주가 시작됩니다. 여러분이 서로 협력하지 않으면, 그 경주는 비극으로 끝납니다. 하지만 단 한 명의 선수도 비극으로 끝날 경쟁을 혼자서 멈출 수는 없습니다. (Harris & Raskin, 06:42)

> 이제 우리 모두에게 달려 있습니다. 새로운 기술을 개발한다는 것은 새로운 책임을 진다는 것을 뜻합니다. 그 기술이 새로운 종류의 책임을 밝히고, 언어와 철학, 법을 만들어 낼 텐데, 이 모든 것은 그냥 생기는 것이 아니기 때문입니다. 그리고 그 기술이 권력을 부여한다면, 경주가 시작됩니다. 우리가 이 경주에서 협력하지 않으면, 경주는 비극으로 끝날 것입니다.[11] (Harris & Raskin, 01:04:26)

죄수의 딜레마는 상대의 행동에 대한 확신이 없다는 전제 하에 개별 의사결정 주체 수준에서 가장 이성적이고 합리적인 선택을 하지만, 그 선택이 결과적으로 모두에게 최선의 선택이 아닌 상황을 가져오는 딜레마이다. 즉, 개별 의사결정 주체 수준에서 덜 합리적으로 보이는 선택지를 택해야 더 큰 맥락에서 보았을 때 최선의 선택을 할 수 있다.

이 마지막 딜레마에서 상충하는 가치는 두 가지 측면에서 이야기할 수 있

11) 원문: “If that technology confers power, it will start a race. And if you do not coordinate, the race will end in tragedy. There’s no one single player that can stop the race that ends in tragedy.” ... “Now it is up to us collectively. When you invent a new technology, it’s your responsibility as that technology is to help uncover the new class of responsibilities, create the language, the philosophy, and the laws, because they’re not going to happen automatically. If that tech confers power, it’ll start a race, and if we do not coordinate, that race will end in tragedy.”

다. 첫째, 개발과 발전의 경주에서의 승리와 그 승리로 얻을 수 있는 권력, 그리고 인류의 미래를 위한 숙고라는 두 가지 가치가 상충한다. 둘째, 개별 의사결정 주체 수준의 합리성과 인류 공동체 전체 수준 차원의 합리성이다. 전자의 합리성에 따라 추구하는 것이 해리스와 라스킨이 "실제로 일어나고 있는 일"이라고 표현한 것에 해당하고, 후자의 합리성에 따라 추구해야 하는 것이 "일어나야만 하는 일"에 해당한다고도 볼 수 있다. 첫째 측면과 둘째 측면은 곧 연결되는데, 개별 의사결정 주체의 차원에서 볼 때 합리적으로 우선시해야 하는 가치는 개발과 발전의 경주에서 이기는 것과 그에 따르는 보상이고, 공동체 전체의 차원에서 합리적으로 우선시해야 하는 가치는 인류의 미래를 위한 숙고이다. 두 가치가 단순히 눈앞에 병렬적으로 제시되었을 때, 우리는 후자가 더 중요한 가치라고 쉽게 답할 것 같지만 개별 의사결정 주체의 차원에서는 다른 의사결정 주체들의 행동에 대한 확신이 없으므로 선택이 쉽지 않다.

현시점에서 인공지능 개발의 속도를 잠시 늦추고 인류의 미래를 논의해야 한다. 하지만 이러한 상황을 알고 있다고 해도, 속도를 늦추면 경쟁에서 도태될 위험이 있기 때문에 개발과 발전을 멈출 수 없다. 그럼에도 불구하고 멈춰서 숙고하지 않는다면, 골렘이 어떤 부작용을 가져올지 아무도 모른다. 이게 가장 마지막 종류이자 가장 큰 맥락의 딜레마다.

III. 인공지능 딜레마와 철학교육

지금까지 인공지능과 관련된 딜레마를 세 가지 종류로 정리해 보았다. 상황별 딜레마는 특정한 상황에서 어떤 윤리적 가치를 선택할지의 문제와 관련된 딜레마였고, 투명성의 딜레마는 인공지능 윤리의 대표적인 개념 중 투명성 개념이 다른 가치와 상충하면서 발생하는 딜레마였다. 마지막 딜레마는 인공지능 개발 및 발전 경쟁의 현 상황과 관련된 딜레마로, 군비경쟁처럼 인공지능 개발을 멈추지 못하다가 인류가 인공지능에 대한 통제권을 잃어버릴 것에

대한 우려에서 비롯되었으며 본질적으로 죄수의 딜레마와 같은 딜레마였다.

이제 인공지능과 관련된 딜레마를 놓고 철학교육에서는 두 가지를 물어야 한다. 왜 이들 딜레마를 철학교육에서 다루어야 하며, 과연 어떻게 다룰 수 있을 것인가?

1. 왜 철학교육에서 다루어야 하는가?

인공지능과 관련된 딜레마들은 기존에 철학교육이 아닌 여러 맥락에서 다루어지고 있다. 앞서 언급하였다시피, 상황별 딜레마는 고등학교 인공지능 기초 교과서에 이미 포함되었다. 투명성의 딜레마는 기존의 인공지능 윤리교육에서 다루어지고 있고, 군비경쟁의 딜레마는 인공지능 산업 전반의 발전과 관련하여 논의될 수 있다. 즉, 세 종류의 딜레마는 현재 서로 다른 맥락에서 다루어지고 있다. 하지만 철학교육에서 인공지능과 관련된 딜레마만을 분석하여 따로 다루면서 메타적인 성찰을 목표로 하는 경우는 적다. 그러면 지금처럼 서로 다른 맥락에서 이러한 딜레마를 다루는 것으로 과연 충분할까?

그렇지 않다. 물론 다양한 맥락에서 서로 다른 딜레마를 다루는 것도 중요하지만, 그것만으로는 부족하다. 첫째, 모든 딜레마는 가치의 상충을 내포한다. 세 가지 딜레마를 보면 무언가 공통점이 보인다. 어떤 두 종류의 가치가 우리에게는 모두 중요한데, 이 중에 무엇을 택할지가 문제가 되는 것이다. 상황별 딜레마에서는 사회적 공공선과 인간의 기본권, 사생활 보호 등이 충돌하고 있었다. 투명성의 딜레마에서는 투명성이라는 가치와 정보 보호 또는 인공지능의 발전이라는 가치가 충돌한다. 군비경쟁의 딜레마에서는 기업이나 국가라는 단위의 합리성과 인류 공동체 전체의 합리성이 상충한다. 결국 딜레마는 둘 다 중요한 어떤 가치들 사이에서, 하나의 가치를 택하면 다른 가치를 포기해야 하는 상황에서 발생한다.[12)]

12) 한 심사위원께서 본 논문에서 이야기하는 인공지능 딜레마와 인공지능 공학에서 사용하는 "트레이드 오프(trade-off)" 개념과의 관계에 대해서도 고찰해 볼 것을 제안해 주셨다. 소중한

가치의 상충이라는 상황의 본질을 이해하고 그러한 상황에서 올바른 판단을 내리는 훈련을 하기 위해서, 개인은 단순히 각 딜레마를 개별적으로 이해하고 마치 퀴즈를 풀듯 두 선택지 중 하나를 고르는 것 이상의 이해가 필요하다. 우선, 이 모든 상황이 각각 나름대로 중요한 가치들이 상충하는 상황이라는 점을 메타적으로 이해해야 한다. 다음으로, 개인은 여러 중요한 가치 중 자신이 어떤 가치를 우선시하며 그 이유는 무엇인지 알아야 한다. 즉, 가치들 사이에 비중을 둘 줄 알아야 하고, 저울질할 줄 알아야 한다. 물론 그 저울질의 결과가 영원히 고정되는 것은 아니고, 개인이 살아가면서 겪는 많은 경험과 숙고를 통해서 수정될 수 있다. "그냥", "남들이 그렇다고 하니까", "그게 재미있어 보여서"와 같은 이유가 아니라, 자신 나름의 합리적인 과정인 추론을 거쳐서 어떤 가치를 다른 가치보다 더 우선시할지 결정해야 한다. 또한 그러한 과정에는 여러 가치 사이의 차이, 서로 다른 선택지와 결과, 자신의 우선순위와 타인의 우선순위를 비교해 보고 서로의 입장 차이를 이해하는 절차가 포함될 것이다. 나아가 이러한 과정에는 자신이 의미 있게 여기는 것은 무엇이며 왜 그것을 의미 있게 여기는지에 대해서 시간과 에너지를 들여서 스스로 밝혀내는 자기 탐구의 작업이 필요하다. 이 과정은 넓게 보면 전반적인 자아 성찰의 시간일 것이다. 이처럼 메타적인 숙고의 과정은 인공지능만을 다루는 교과목에서는 깊이 있게 배우기 어렵고, 결국 철학적인 탐구로 돌아가야만 배울 수 있는 능력이다.

둘째, 인공지능과 관련된 딜레마는 트롤리 딜레마나 죄수의 딜레마처럼 고전적으로 다루어지던 딜레마와 본질적으로 유사하지만, 한 가지 큰 차이점이 있다. 바로 그 선택의 결과가 우리의 현실과 곧장 연결된다는 것이다. 따라서 우리는 길잡이가 더욱 절실하게 필요한 상황이다. 앞서 살펴본 딜레마가 우리의 상상 속에만 머무른다면, 이러한 딜레마에 대해서 깊이 생각할 필요가 없을지도 모른다. 그저 재미있는 사고 실험이라고 생각하고 넘어가도 될 것이

제안에 감사하며 시간과 분량의 관계상 이는 추후의 과제로 남겨놓도록 하겠다.

다. 하지만 세 종류의 딜레마 모두 더 이상 우리의 상상 속에 존재하지 않는다. 이미 현실에서 자율주행차는 레벨3와 레벨4가 언급되고 있다. 지금 당장 문을 열고 길에 나섰을 때 도로에 자율주행차가 가득한 것은 아니지만, 이미 그 정도로 자율주행차가 상용화된 시점에서 트롤리 딜레마를 고민하기 시작한다면 늦을 것이다.

이러한 딜레마는 갑자기 등장한 것이 아니다. 마법처럼 현실 세계로 뛰쳐나오긴 했지만, 갑자기 등장한 괴물은 아니다. 동화책 속 괴물이 현실에 등장했다면, 우리는 괴물을 다루는 방법을 찾기 위해서 동화책을 다시 찾아볼 것이다. 마찬가지로, 철학은 여러 가치가 상충하여 일으키는 딜레마를 어떻게 다룰지 오래전부터 이야기해 왔다. 철학이 마치 문제집의 정답지처럼 답을 알려주는 것은 아니지만, 인공지능과 관련된 딜레마도 그 원류인 철학을 찾아보아야 우리에게 필요한 길잡이를 얻을 수 있을 것이다.

2. 어떻게 철학교육에서 다룰 것인가?

전 세계적으로 가장 널리 활용되고 있는 철학교육 방식은 매튜 립맨(Matthew Lipman)은 철학교육 방법론인 '어린이를 위한 철학(Philosophy for Children, 이하 P4C)'이다. P4C 탐구공동체로 이루어지는 철학 수업에는 두 가지 특징이 있다. 첫째는 토론의 촉발제로서 내러티브를 중시한다는 점이고, 둘째는 토론 및 탐구의 주제로서 질문 자체를 형성하는 과정을 중시한다는 점이다. 이 장에서는 인공지능과 관련된 딜레마가 P4C 방법론의 두 가지 특징을 만족시킬 수 있다는 것을 보여주고, 나아가 구체적으로 인공지능 딜레마에 초점을 맞춘 14차시 커리큘럼을 제안할 것이다.

1) P4C 방법론으로 인공지능 딜레마 다루기

P4C 탐구공동체에서 인공지능과 관련된 딜레마는 토론 및 탐구를 시작하는 내러티브의 역할을 할 수 있다. 딜레마는 그 자체로 특정한 상황을 내포하

고 있으며, 상황을 응용하여 내러티브를 제공할 수 있기 때문이다. 나아가 어떤 종류의 딜레마를 다루더라도, 철학교육에서 인공지능과 관련된 딜레마를 다룬다면 해당 딜레마 상황에서 어떤 질문을 만들어 낼 수 있는지 수강생들이 충분히 고민할 시간을 줄 수 있다.

상황별 딜레마의 경우, 피상적으로 접근할 경우, 마치 인터랙티브 스토리텔링 게임처럼 단지 주어진 상황에서 두 선택지 중 무엇을 선택할지 고르거나, 찬성과 반대 팀을 나누어서 주어진 입장을 뒷받침하는 증거를 찾아오는 정도에 그칠 수 있다. 그러나 중요한 것은 각 딜레마 상황에서 어떤 가치가 왜 충돌하고 있으며, 각 개인은 왜 특정한 가치를 다른 가치보다 더 중요하게 여기는 선택을 하게 되는지, 딜레마 상황이 근본적으로 어떻게 설정되어 있는지 메타적으로 묻는 것이다. 따라서 P4C 방법론을 적용한 철학 수업이라면, 딜레마 상황 자체가 토론의 촉발점인 내러티브 역할을 할 것이고, 단순히 딜레마 상황을 제시하고 어떤 선택을 할 것인지 고르고 그 이유를 대답하게 하는 것을 넘어서, 탐구공동체에 속한 개인은 스스로 질문을 형성할 수 있어야 할 것이다. 탐구공동체에서 만들어 낼 만한 질문은 다음과 같다.

- 이 딜레마가 우리에게 선택하도록 하는 가치들은 무엇인가?
- 그 가치들은 왜 중요한가?
- 하나의 가치를 선택함으로써 포기해야 하는 가치는 무엇인가?
- 가치 A를 가치 B보다 중요하게 여기는 사람은 왜 그렇게 생각하는가? 또 그 반대의 경우는 왜 그러한가?
- 실제 상황에서 이 딜레마의 선택지 중 하나를 택하였을 때 우리 삶에 미치는 영향은 무엇인가?
- 나와 다른 선택지를 고른 사람의 입장을 어떻게 이해할 수 있을까?
- 우리는 왜 이 딜레마에 대해서 고민해야 하는가?

이 질문 목록은 예시일 뿐이며, 탐구공동체는 얼마든지 다양한 질문을 만

들어 낼 수 있을 것이다. 다만 위 질문 목록에 대하여 두 가지만 언급하겠다. 우선, 예시의 질문은 모두 개별적이고 구체적인 딜레마 상황에 대한 질문이 아니라 딜레마 자체와 관련된 메타적인 질문(meta-questions)이다. 만약 구체적인 딜레마 하나를 놓고 그 상황에 대해서 물을 수 있는 질문의 목록을 제시한다면, 그러한 질문 목록은 기존의 디베이트(debate) 수업에서 찬반 토론을 할 때 각 팀의 주장 및 반론과 비슷한 형태를 띨 것이다. 예를 들어, 자율주행차와 관련된 트롤리 딜레마를 놓고 왜 초록색 신호에 길을 건너는 보행자 다섯 명 대신 운전자 한 명을 희생시켜야 하는지 물을 수 있다. 물론 이러한 질문도 탐구공동체에서 제기할 수 있는 질문이며, 탐구공동체에서 다루고 넘어가야 할 질문일 수 있다. 그러나 위 질문 목록에서 구체적인 딜레마 상황을 넘어서는 메타적인 질문만을 제시한 이유는, 구체적인 상황에 국한된 질문이 아닌 딜레마 전체를 바라봄으로써 딜레마 상황 자체에 대한 새로운 관점을 얻을 것을 기대할 수 있기 때문이다.

다음으로, 이미 딜레마를 다루던 철학 수업이라면 메타적인 질문들이 오히려 익숙한 내용이고, 각 딜레마에 대한 개별적이고 구체적인 질문 예시를 기대하였을 수 있다. 그러나 본 논문에서는 인공지능과 관련된 세 종류의 딜레마가 개별 맥락에서 따로 다루어지고 있긴 하여도 철학교육의 맥락에서 다루어지는 측면에 있어서는 부족했다고 생각하며, 인공지능 딜레마도 철학교육적 측면에서 다루어질 필요가 있다는 주장을 하는 것이기에 메타적인 질문만을 예시로 제시하였다. 그리고 이러한 메타적인 질문이 디베이트가 아니라 디스커션(discussion)을 목적으로 하는 철학 수업에 더욱 적절하다고 본다.

2) 〈인공지능 딜레마〉 교과목에 대한 제안

마지막으로, 인공지능 딜레마를 핵심 주제로 하면서 철학교육 방법론을 적용하여서 한 학기의 인공지능 철학 교과목을 구상하면 어떨지 한 가지 제안을 하고자 한다. 세 종류의 딜레마는 모두 탐구공동체의 맥락과 수준에 따라

서 조절될 수 있지만, 구체적으로 다음과 같은 구상이 가능할 것이다.[13]

〈표 4〉 〈인공지능 딜레마〉 교과목의 가능한 예시

차시	주제	주요 내용
1	오늘날의 인공지능, 인공지능 윤리의 필요성	인공지능 윤리 분야에 대한 전반적인 소개
2	인공지능이란 무엇인가?	인공지능의 역사 및 기술적 내용 이해
3	인공지능 윤리의 핵심 개념	투명성, 설명가능성, 블랙박스, 편향, 책임과 책무 등 인공지능 윤리의 핵심 개념 이해
4	[딜레마 1] 투명성의 딜레마(1) 정보 보호	투명성 vs. 정보 보호
5	[딜레마 2] 투명성의 딜레마(2) 인공지능의 발전	투명성 vs. 인공지능의 발전 방향
6	[딜레마 3] 자율주행차와 트롤리 딜레마	"무엇을 기준으로 하여 누구를 희생시킬 것인가?"
7	[딜레마 4] 인공지능의 감시와 빅브라더	사회적 공공선 vs. 사생활 보호/인권
8	[딜레마 5] 돌봄 노동과 제미노이드	사회적 공공선 vs. 인간의 본질
9	[딜레마 6] 킬러 로봇과 터미네이터	사회적 공공선/기술 합목적성 vs. 인간 존엄성
10	[딜레마 7] 인공지능과 지적재산권	사회적 공공선/기술 합목적성 vs. 인간 존엄성
11	[딜레마 8] 일만 하는 외계인의 공격	사회적 공공선 vs. 인간 존엄성/인간의 기본권
12	[딜레마 9] 다빈치 로봇 수술과 의료윤리	사회적 공공선/기술 합목적성 vs. 인간의 기본권
13	[딜레마 10] 디지털 천국과 Peter 2.0	사회적 공공선 vs. 기술 합목적성/인간의 본질/인간 존엄성
14	[딜레마 11] 군비경쟁: 죄수의 딜레마	개별 의사결정 주체 차원의 합리성 vs. 공동체 전체 차원의 합리성

13) 해당 커리큘럼은 중간고사와 기말고사를 포함한 16차시의 학부 교양 수업을 염두에 두고 14차시 내용으로 제안하였다. 그러나 상황에 따라서 수업 차시 및 난이도 조절이 가능할 것이다. 평가를 비롯한 구체적인 실행 방법은 필자의 다른 논문(홍예리, 2023)을 참조하기를 바란다.

제안된 커리큘럼은 한 학기에 걸쳐서 인공지능과 관련된 딜레마 포함하여 총 14차시에 걸쳐서 진행할 수 있도록 고안되었다. 딜레마와 함께 표의 우측에 실린 상충하는 두 가지 가치의 경우 단지 이해를 돕기 위한 예시일 뿐, 수강생들은 얼마든지 다른 가치를 생각하여 토론할 수 있다. 6차시의 경우 다양한 상황으로 응용되므로 상충하는 가치를 정확하게 표현하기 어렵기에 질문 형태로 표현하였지만, 이러한 질문은 교수자가 제시할 것이 아니라 수강생들의 탐구공동체가 스스로 형성해 내야 하는 질문이다. 즉, 표의 우측 부분은 이해를 돕기 위한 것일 뿐, 실제로 수강생들에게 미리 주어지는 내용이 아니다.

우선 1~3차시에는 기본적인 개념을 배운다. 인공지능의 경우, 어떻게 지금과 같은 형태로 발전하게 되었는지 그 발전의 역사 및 인공지능이 어떻게 작동하는지에 대한 기초적인 지식이 없으면 윤리적인 문제를 이해하기조차 힘든 경우가 있기 때문이다. 핵심 개념을 먼저 배웠기 때문에, 4~5차시에는 핵심 개념 중 하나인 투명성과 관련된 딜레마를 이어서 다룬다. 다음으로 6차시부터 13차시까지 아홉 차례에 걸쳐 상황별 딜레마를 다룬다. 최대한 구체적인 토론 상황을 제공하기 위해서 상황별 딜레마를 여러 가지로 쪼개 놓았지만, 수업의 여건 및 수강생의 수준에 따라서 상황별 딜레마를 줄이고 투명성의 딜레마나 죄수의 딜레마 비중을 늘릴 수도 있고, 인공지능과 관련된 다른 이슈를 추가할 수도 있을 것이다.

상황별 딜레마를 처음 다루기 시작하는 6차시에는 가장 잘 알려진 자율주행차와 트롤리 딜레마를 배치하였다. 7~11차시에는 김은경, 이영준(2023)이 개발한 딜레마를 각각 1~3(7차시), 4와 9(8차시), 7(9차시), 8(10차시), 5~6(11차시)으로 나누어 배치하였다. 1~3은 인공지능 스피커, 안면인식 CCTV, 경찰로봇을 주제로 하여, 사회적 공공선과 개인의 사생활 보호 및 인권 사이의 딜레마로 함께 다룰 수 있다. 4와 9의 경우, 김은경, 이영준(2023)은 서로 다른 범주에 속하는 것으로 보았지만, 고인을 재현한 인공지능 로봇이나 부모를 재현한 돌봄 로봇 모두 본질적으로는 실존 인물과 똑같이 만들어진 인공지능,

즉 ‘제미노이드’이기 때문에 같은 차시에서 다룰 수 있을 것이다. 킬러 로봇을 다루는 9차시와 생성형 인공지능의 작품을 둘러싼 저작권 문제를 다루는 10차시는 독자적인 논의가 가능한 차시이다. 11차시는 청소 로봇과 드론 배달의 경우가 모두 인공지능으로 인하여 변화하는 노동시장에 대한 것이므로 함께 다룰 수 있다.[14)]

12차시와 13차시는 앞의 장에서 제시된 적이 없는 딜레마다. 12차시에서 소개하는 다빈치 수술 로봇의 경우 오늘날 널리 이용되고 있지만, 넓게 보았을 때 인간의 의술이 단순히 인공지능 수술 기계를 이용하는 기술로 치환될지에 대한 문제를 제기한다. 이 주제는 11차시에서 다룬 노동시장 관련 딜레마에 이어서, 의료와 노동의 본질에 대한 질문으로 이어질 수 있다. 13차시는 실제로 필자가 철학 수업을 진행해 본 사례이다. 영국 출신의 공학 박사이자 루게릭병 환자였던 피터 스콧-모건은 세 차례의 큰 수술을 거쳐 장기 대부분을 기계로 교체하여 수명을 연장하는 것을 시도했던 실존 인물이다. 그는 장기를 교체하는 것을 넘어서서 두뇌까지 인공지능에 업로드하여 스스로 “Peter 2.0”이 되는 것을 희망하였으나, 의료진이 예측했던 2년의 시한부 인생보다 조금 더 살고 세상을 떠났다(Scott-Morgan, 2021). 실제 철학 수업에서 이 내러티브를 접한 수강생들은 Peter 2.0이 남아있는 배우자에게 과연 동일한 인물로 인식될지, 그러한 업로드 기술이 실현될 경우 과연 남아있는 사람들에게 정서적인 위로를 줄 수 있을지에 대하여 토론하였다.[15)] 이러한 모든 주제를 거친 후, 마지막 차시에는 인공지능 개발 전반과 우리가 나아갈 방향에 대한 딜레마인 군비경쟁 딜레마를 소개할 수 있을 것이다. 이 시간을 통해서 탐구

14) 이 차시에 ‘일만 하는 외계인의 공격’이라는 제목을 붙인 것은 마틴 포드(Martin Ford)의 비유 때문이다. 그는 인공지능을 쉬지 않고 일만 하며 오로지 일을 하는 데에서만 성취감을 느끼는 외계인에 비유하며, 이러한 외계인의 침공으로 지구인이 일자리를 잃는 시나리오를 제시한다. (Ford, 2016, 302-305.)

15) 이러한 질문들은 철학적 탐구공동체가 충분히 탐구해 볼 만한 주제이지만, 직접적으로 인공지능과 관련된 주제가 아닌 방향으로 토론이 흘러갈 가능성이 있는 주제이기도 하다. 상황에 따라서 보다 인공지능과 직접적으로 관련 있는 주제로 해당 차시 내용을 바꿀 수도 있다.

공동체는 지금까지 논의했던 다양한 주제들의 여러 측면을 정리하고 자신의 입장을 정할 수 있을 것이다.

위와 같은 커리큘럼이 가지는 또 한 가지 특징은, 딜레마 상황과 함께 실제 문학이나 영화에서의 내러티브를 제공할 수 있다는 점이다. 예를 들어, 자율주행차와 트롤리 딜레마 같은 경우 드라마 <굿 플레이스(The Good Place)>에서 트롤리 딜레마를 재현한 장면이 있으니 해당 장면을 소개할 수 있다.[16] 인공지능의 감시와 빅브라더의 경우 조지 오웰의 『1984』를 함께 소개할 수 있다. 디지털 천국과 Peter 2.0도 죽은 사람의 두뇌를 업로드한 디지털 천국을 묘사한 드라마 <업로드(Upload)>가 있고, 피터 스콧-모건은 실존 인물이며 그에 관한 책이 번역서도 출판되어 있다(Scott-Morgan, 2021). 이러한 내러티브 보충은 토론을 더욱 풍부하게 만들어 줄 수 있을 것이다.

Ⅳ. 나가며

지금까지 인공지능과 관련한 딜레마를 세 종류로 나누어 살펴보았고, 이러한 딜레마를 철학교육에서 토론의 출발점인 내러티브로 이용함으로써 탐구 주제인 메타적인 질문을 만드는 방향으로 발전시킬 수 있다는 것을 보였으며, 대략적이지만 인공지능 딜레마만을 이용하여 가능한 철학교육 커리큘럼도 하나 제안해 보았다.

마지막으로 당부할 점이 두 가지 있다. 첫째, 인공지능 딜레마에만 초점을 맞춘 철학교육 수업은 당연히 인공지능 윤리 전반을 다루기에 충분하지 않다. 이 커리큘럼에서는 인공지능 윤리의 핵심 개념을 심도 있게 다룰 시간이 부족하고, 인간과 인공지능의 차이 및 싱귤래러티 문제에 대해서 깊이 있는 토론

16) 미국 NBC의 드라마로, 등장인물 중 도덕 윤리 교수인 '치디 아나곤예'와 데카르트의 전능한 악마 같은 존재인 악마 '마이클'이 등장한다. 마치 철학교육을 염두에 두고 제작한 것 같은 드라마로 유용성이 높다.

을 할 수 있는 시간이 배제되었고, 알고리즘 편향에 대한 시간도 빠져있다. 현재 제안된 커리큘럼은 아마도 마이크로 디그리 과정을 위한 수업 정도로 응용할 수 있을 것이다. 그럼에도 불구하고 이러한 커리큘럼을 제안해 본 이유는, 인공지능과 관련된 딜레마만으로 한 학기 수업이 가능할 정도임을 보여주려는 의도에서였다. 인공지능과 관련하여 "딜레마"라는 표현이 언급되지만, 기존에는 그러한 딜레마를 종류별로 정리하였거나 딜레마만을 다룬 수업이 없었다. 그러나 이렇게 많은 주제가 나올 정도라면, 인공지능과 관련된 딜레마에 대하여 메타적으로 한 번쯤 짚고 넘어갈 필요가 있다고 본다. 또한 기존에는 자율주행차와 트롤리 딜레마 또는 상황별 딜레마 일부에만 주목하였지만, 인공지능이 초래하는 다양한 딜레마에 대해서 깊이 있게 고민해 보는 시간은 분명히 필요하다. 왜냐하면 이러한 종류의 딜레마는 더 이상 동화 속 골렘이 아니라 현실로 뛰쳐나온 골렘이기 때문이다. 따라서 인공지능과 관련된 딜레마에 대해서 다각도로 생각할 필요가 있고, 그러한 기회를 철학 수업이 제공할 수 있을 것이다.

둘째, 각 커리큘럼에서 다루는 주제와 수강생들의 토론 방향에 따라서, 토론 진행자인 교수자는 해당 주제가 전통적인 철학의 문제와 어떻게 맞닿아 있는지 보충하여 소개하고 설명할 필요가 있다. 배경지식이 토론보다 중요한 것은 아니지만, 인공지능이 초래하는 딜레마가 새로운 괴물이 아니라 오래된 동화 속 괴물이라는 것을 알면 탐구공동체가 철학에서 그 문제를 해결할 실마리를 찾는 데에 조금이나마 도움이 될 것이기 때문이다. 철학이 골렘을 단번에 무찌를 수 있는 엑스칼리버는 주지 못할 수도 있다. 하지만 호수의 여인은 만나게 해줄 수 있을 것이다.

【주제어】 **인공지능, 인공지능 딜레마, 철학교육, 투명성, 트롤리 딜레마, 어린이를 위한 철학교육(P4C)**

[참고문헌]

강용일 (2024. 1. 24). [CES 2024 인사이트] AI 혁신이 기존 산업 진화로 이어져... "AI·모빌리티·로봇·스마트홈·인프라" 주목. 아주경제, 출처: https://www.ajunews.com/view/20240123105639803

김건우 (2021). "인공지능 로봇의 윤리" 과목 강의 계획서. 광주과학기술원 MOOC 강의. 출처: https://gist.edwith.org/ethics-of-ai-and-robots

김보경 (2023). 사람들은 인공지능에게 어떤 윤리적 판단을 기대하는가?: 딜레마 판단을 중심으로. 사회과학연구, 62(3), 403-427.

김은경, 이영준 (2023). 인공지능 윤리 역량 신장을 위한 인공지능 윤리 딜레마 개발. 컴퓨터교육학회 논문지, 26(5), 31-42.

_____ (2022). Moral machine을 활용한 인공지능 윤리교육이 초등학생의 인공지능에 대한 인식에 미치는 영향. 컴퓨터교육학회 논문지, 25(3), 1-8.

김자미, 김용 (2022). 인공지능 기술 딜레마를 기반으로 한 인공지능 윤리 교육에 대한 소고. 창의정보문화연구, 8(2), 87-95.

김장현 (2024. 1. 25). [김장현 이제는 AI시대] 2024 CES가 보여주는 미래. 김장현 성균관대 글로벌 융합학부 교수(외부 칼럼). 파이낸셜뉴스, 출처: https://www.fnnews.com/news/202401251831220775

김효은 (2019). 인공지능과 윤리. 서울: 커뮤니케이션북스.

_____ (2020a). 의사결정 자동화에 대한 대응으로서의 인공지능 윤리 교육. 윤리교육연구, 55, 277~308.

_____ (2020b). 공학적 방법을 결합한 인공지능 윤리 학습. 윤리연구, 129, 133~153.

_____ (2023). "공정한 인공지능(Fair Artificial Intelligence)" 과목 강의계획서. DSC 공유대학, 출처: https://sugang.dscu.ac.kr/ko/module/sugang/@sugang_syllabus/143

변순용, 이연희 (2020). 인공지능 윤리하다. 서울: 도서출판 어문학사.

손혜숙 (2022). 인공지능 윤리 의식 함양을 위한 대학 교양교육 방안 연구. 교양학연구, 21, 33-61.

이상민, 김홍래 (2021). 인공지능 윤리 교육의 주요 이슈 분석. 2021년 한국컴퓨터교육학회 동계 학술발표논문집, 25(1), 49-54.

인공지능과 가치 연구회 (2021). 인공지능윤리: 다원적 접근. 서울: 박영사.

정유진, 김진수, 박남제 (2019). 초등학생 대상 블록체인 기술의 위변조 방지 핵심원리 이해와 교육방안 설계. 정보교육학회논문지, 23(6), 513-520.

홍예리 (2023). '질문하는 인간'을 길러내는 인공지능 윤리 교육에 대한 제안. 인간연구, 51,

39-75.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J., & Rahwan, I. (2018). The moral machine experiment. Nature 563(7729): 59-64.

Coeckelbergh, M. (2020). AI Ethics. 신상규, 석기용 옮김 (2023). AI 윤리에 대한 모든 것. 경기도 파주: 아카넷.

Daniels, G. (Creator). (2020-present). Upload. [TV series]. Amazon Prime Video.

Ende, M (1973). Momo. 한미희 옮김 (1999). 모모. 서울: 비룡소.

Ford, M. (2015). Rise of the Robots: Technology and the Threat of a Jobless Future. 이창희 옮김 (2016). 로봇의 부상. 서울: 세종서적.

_____ (2018). Architects of Intelligence. 김대영, 김태우, 서창원, 최종현, 한성일 옮김 (2019). AI 마인드. 서울: 터닝포인트.

Harris, T. & Raskin, A. (Center for Humane Technology). (2023, March 9). *T*he A.I. Dilemma [Video]. YouTube. https://youtu.be/xoVJKj8lcNQ?feature=shared

Lipman, M. (2003). Thinking in Education (2nd ed.). Cambridge: Cambridge University Press.

Orlowski, J. (Director). (2020). The Social Dilemma [Film, distributed by Netflix]. Exposure Labs; Argent Pictures; Agent Pictures; The Space Program.

Orwell, G (1949). 1984. 정회성 옮김 (2003). 서울: 민음사.

Schur, M. (Creator). (2016-2020). The Good Place. [TV series]. New York: NBC.

Scott-Morgan, P. (2021). Peter 2.0: The Human Cyborg. 김명주 옮김 (2022). 나는 사이보그가 되기로 했다. 파주: 김영사.

Shane, J. (2019). You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place. New York: Voracious.

[국문초록]

인공지능을 이야기할 때 “딜레마”라는 표현은 자주 눈에 띈다. 자율주행차의 의사결정과 관련하여 철학에서 오래된 트롤리 딜레마가 논의되기도 하고, 인공지능 윤리의 한 가지 주요 개념과 관련하여서도 몇 가지 딜레마가 언급되기 때문이다. 이외에도 전직 구글 디자인 윤리학자이자 현재 <인간적인 기술 센터(The Center for Humane Technology)>의 공동창립자인 트리스탄 해리스(Tristan Harris)는 2020년 ≪소셜 딜레마(The Social Dilemma)≫라는 다큐멘터리 영화에 참여하였고, 2023년에는 ≪인공지능 딜레마(The A.I. Dilemma)≫라는 제목의 강연을 펼쳤는데, 그는 ‘딜레마’라는 표현을 또 다른 맥락에서 사용한다.

그런데 인공지능과 관련하여 “딜레마”라는 표현을 사용할 때, 그 맥락이 모두 같은 것은 아니다. 특히 인공지능 윤리와 관련된 수업을 할 때 “딜레마”라는 표현은 여러 개념과 얽혀서 등장한다. 본 논문에서는 인공지능과 관련하여 등장하는 딜레마를 세 가지 종류로 정리한 후, 다양한 종류의 인공지능 딜레마를 철학교육에서 왜 다루어야 하며 어떻게 다룰 수 있을지 고민해 보고자 한다. 또한 이러한 고민과 노력의 일부로서 인공지능 딜레마만으로 구성된 한 학기 커리큘럼을 제안해 보고 그 가능성과 한계도 이야기할 것이다.

[Abstract]

Addressing Artificial Intelligence Dilemmas in Teaching Philosophy

Hong, Yeri (Ewha Institute of Philosophy)

In the realm of artificial intelligence ethics, the term *dilemma* often emerges prominently. The age-old philosophical trolley dilemma frequently surfaces in discussions regarding decision making in autonomous vehicles, while various dilemmas are associated with key concepts of artificial intelligence ethics. Moreover, Tristan Harris, a former Google design ethicist, has employed the term dilemma in a different context. His involvement in the 2020 documentary film *The Social Dilemma* and his lecture titled *The A.I. Dilemma* in 2023 shed light on unique dimensions of these dilemmas.

When employing the term *dilemma* in the context of artificial intelligence, the nuances vary significantly. Particularly within the realm of teaching artificial intelligence ethics, *dilemma* intertwines with various concepts. This paper aims to categorize three types of dilemmas inherent in artificial intelligence and elucidate why and how different types of artificial intelligence dilemmas should be addressed in philosophy education. Furthermore, as part of the endeavor to tackle artificial intelligence dilemmas in philosophy classes, a semester-long curriculum focused on artificial intelligence dilemmas is proposed, alongside an exploration of its possibilities and limitations.

【Keywords】 Artificial Intelligence(AI), Artificial Intelligence Dilemma(AI Dilemma), Transparency, Trolley Dilemma, Teaching Philosophy, Philosophy for Children(P4C)

논문투고일: 2024년 02월 24일 / 논문심사일: 2024년 04월 08일 / 게재확정일: 2024년 04월 25일

【저자연락처】 yeri88@ewha.ac.kr