유교적 인공지능 윤리를 위한 시론(試論)

: ‘인공적 도덕 행위자’를 중심으로

이재복*

【요약】

이 논문은 인공지능 윤리에 대한 유교적 접근 가능성을 ‘인공적 도덕 행위자’ 문제를 중심으로 살핀 논문이다. 실효성 있는 논의를 위하여 관련 쟁점을 ‘쉬운 문제’와 ‘어려운 문제’로 구분하고, 유교적 해결책을 마련할 수 있는 ‘쉬운 문제’에는 어떤 것이 있는지 고찰하였다. 먼저 지금까지 유교 연구 분야에서 인공지능을 이해하고 논의한 방식을 살피고, 그 과정에서 유교 연구자들은 대체로 인공지능을 인간에 대한 위협으로 상정하고 있으며 그에 대한 인간의 존재론적 지위를 확고하게 만드는 데 주력하고 있음을 드러냈다. 다음으로 유교적 성인 인공지능을 만들어 한국 사회의 문제를 해결하자는 제안을 검토함으로써 그러한 제안이 함의하는 바가 무엇인지 분석하였다. 마지막으로 유교적 접근이 가능한 ‘쉬운 문제’로서 도덕성의 편향 문제를 소개하고, 이 문제를 해결하기 위한 방안으로 인공지능을 도덕적으로 만드는 것과 인공지능의 행위를 도덕적으로 만드는 것에 대해 논하였다.

【주제어】 인공지능, 인공지능 윤리, 인공적 도덕 행위자, 유교, 유교적 인공지능 윤리

---

* 한양대학교 창의융합교육원 강사
https://doi.org/10.34162/hefins.2022..29.004

I. 들어가는 말

신기술의 등장은 그와 관련된 다양한 담론을 이끌어낸다. 존 매카시John McCarthy, 마빈 민스키Marvin Minsky 등이 주도한 1956년의 다트머스 모임에서 인공지능Artificial Intelligence(이하 'AI') 분야가 확립된 이후 법과 제도, 윤리 등의 영역에서는 AI가 인류에게 끼칠 영향에 대한 다양한 논의가 진행되고 있다. 그러나 AI 기술의 발전 속도는 전문가들의 예상을 넘어서고 있으며, 그에 대한 구체적이고 실효성 있는 담론이 요구되는 실정이다. 그러한 담론은 이전에는 물을 필요가 없었던 것들도 주제로 다루는 양상을 띤다.[1)]

AI 윤리에 대한 쟁점은 크게 ① 개념에 대한 이해에 근본적인 의견 차이가 있어서 합의된 해결책을 모색하기가 요원한 '어려운 문제'와 ② 기존에 정립된 논의틀framework을 그대로 유지하되 개념을 유연하게 적용함으로써 비교적 수월하게 해결책을 찾을 수 있는 '쉬운 문제'로 구분된다. 전자에는 AI의 위협에 직면한 인간의 존재론적 지위, '도덕적'의 의미는 무엇이고 그에 비춰보아 AI를 '도덕적 행위자'라고 인정할 수 있을지를 고민하는 문제 등이 포함되고, 후자에는 행위자가 될 수 있는 요건을 확장하고 AI가 그 요건에 부합할 경우 도덕적 행위자로 인정할 수 있을지 등을 논하는 문제 등이 포함된다. 이것은 이상욱이 '자율주행차autonomous car'와 관련된 쟁점을 분석하며 제안한 구분 방식이다.[2)] 그의 주장을 살펴보면, 자율주행차 또는 '운전자 없는 차driverless car'가 '객관적으로 바람직한 준칙에 따라 스스로를 규제하거나, 외부의 영향으로부터 자유롭게 주체적으로 행동한다'

---

1) 이상욱은 AI 기술의 급속한 발전이 우리 사회에 미치는 영향은 일시적 유행이 아니라 삶의 여러 영역에 파고 들 것이라는 점을 지적한다. 이에 그는 우리가 이전에는 물은 적 없는 '인공지능을 갖춘 기계의 도덕적 행위자로서의 성립 가능성'을 진지하게 검토해야 한다고 주장한다. 인공지능이 도덕적 행위자로 간주될 수 있는지에 대한 그의 연구는 이상욱 (2019), pp. 259-279 참조.

2) '쉬운 문제'와 '어려운 문제' 개념에 대한 설명은 이상욱 (2019), pp. 262-264 참조.

는 의미의 자율성autonomy을 지니고 있는지, 그와 관련하여 그렇다면 과연 인간은 얼마나 자율적인지를 묻는 것은 '어려운 문제'에 해당한다. 이 문제는 도덕적 행위자로 인정받기 위한 조건으로서의 자율성의 철학적 의미를 분석하고, 자율주행차의 자율성이 그 의미에 부합한다고 할 수 있는지 논하는 것으로서, 자율성 개념에 대한 연구자들의 의견 차이를 해소하기가 쉽지 않기 때문이다.

한편 '쉬운 문제'란 기존에 정립된 논의를 수용하며 이를 AI에 적용할 때 어떤 시사점이 있는지, 개념의 확장 가능성은 없는지를 쟁점으로 삼는다. 현재까지 개발된 운전자 없는 차는 스스로의 판단에 의거하여 진정한 의미의 자율주행을 하는 차가 아니다. 자의식을 갖추지 못했기 때문이다. 이상욱은 운전자 없는 차가 전통적 의미의 자율성을 지니지 못했다는 점에서 통상적 의미의 도덕적 행위자로 인정받지는 못할 것이라고 말한다. 하지만 그는 그럼에도 운전자 없는 차는 사고 등에 대한 책임의 주체가 될 수 있다고 본다. 자연인이 아닌 사단社團, 재단財團에 법적 인격legal person을 부여하는 것과 유사한 방식으로, 운전자 없는 차에도 인격을 부여함으로써 그것을 도덕적 행위자로 정의하는 방식의 해결이 가능하다는 것이다.[3)]

필자가 자율주행차, 더 넓게는 인공적 도덕 행위자Artificial Moral Agent (이하 'AMA')에 대한 '쉬운 문제'와 '어려운 문제' 구분에 주목하는 이유는, 이러한 구분과 문제 해결이 유교 연구자가 AI 윤리에 접근하는 데에 하나의 지침이 될 수 있기 때문이다. '쉬운 문제'의 상당수는 해결이 시급한 문제이다. 기술이 담론의 형성보다 먼저 발전하고 있는 상황에서 사고가 발생했을

3) 이와 관련하여 AI가 법인法人과 유사한 방식으로 행위자로서 인정받게 될 경우에는 어떤 처벌이 가능할지에 대해서도 함께 고민할 필요가 있다. 제리 캐플런 Jerry Kaplan은 AI가 법적 책임의 주체가 될 경우 처벌로서의 '갱생 rehabilitation'이 가능하다고 주장한다. '리프로그래밍reprogramming'으로써 교화에 해당하는 데이터 재학습을 강제하여 새로운 행위자로 거듭나게 하는 것이 처벌의 한 형태가 될 수 있다는 뜻이다. 제리 캐플런, 신동숙 역 (2017), pp. 159-197.

때 생산자, 소비자 그리고 AI 중에서 누구에게 책임을 물을 것인지와 같은 문제는 긴급을 요하는 물음이다. 예술 영역에서는 AI가 만든 작품이 대회에서 수상을 하는 일이 실제로 일어나기도 했다. 이 경우의 창작 주체는 누구인지, 저작권은 누구의 것인지 등의 문제는 더 이상 해결을 미룰 수 없는 문제가 되었다.[4)]

현재까지 진행된 AI 윤리 담론은 주로 서양 철학의 논의에 기반하고 있다. 가령 2017년 2월 유럽연합European Union은 AI에 '전자 인격electronic persons'이라는 법적 지위를 인정하는 결의안을 채택했다. AI가 자율성은 갖추지 못했더라도 법적 의무와 권리의 주체 또는 객체가 될 수 있다는 그들의 생각은 '인격'에 대한 서양 철학적 사유에 근거한다. 여기서 'person'은 생물학적 인간을 가리키는 개념이 아니라, 존 로크John Locke가 정의한 의미로서 '사고 능력을 갖춘 존재자'를 뜻한다는 것이다.[5)] 한편 AMA의 도덕성을 무엇에 기초하여 구현할 것인가의 쟁점에서 언급되는 입장은 의무론deontology, 덕 윤리학virtue ethics 그리고 결과주의consequentialism로서의 공리주의utilitarianism 등에 한정된다는 사실도 간과할 수 없다.[6)] 이러한 경향

4) 최근 미국에서는 이미지 생성 AI '미드저니Midjourney'가 만든 '시어터 오페라 스페이셜Théâtre D'opéra Spatial'이 지난 8월에 개최된 콜로라도 주립 박람회 미술경연(디지털아트 부문)에서 1위를 차지하며 사람들의 이목을 끌었다. 제이슨 앨런Jason Allen은 미드저니를 활용하여 여러 이미지를 생성하였고 그 중에서 이미지 세 개를 출력하여 자신의 이름으로 대회에 출품했던 것이다. 제이슨 앨런과 미드저니 중에서 누가 창작자인지, 1위 상금(이 경우 $300)은 누가 가져야 하는지 등의 논의가 부재한 상황에서 AI가 '창작자'로 등장한 것이다. 이것은 '쉬운 문제'에 해당하는 것으로서, '예술', '창작'의 본질을 재검토하는 '어려운 문제'와 구별된다. 상세한 내용은 Harwell, Drew. "He used AI to win a fine-arts competition. Was it cheating?", The Washington Post, September 2 2022. https://www.washingtonpost.com/technology/2022/09/02/midjourney-artificial-intelligence-state-fair-colorado/ 참조. (검색일: 2022.09.03.)

5) 김효은 (2019), pp. 15-25.

6) 이와 관련하여 스튜어트 러셀Stuart Russell은 한 가지 흥미로운 주장을 하고 있다. 그는 AI가 자의식을 가진다는 증거가 없이는 덕을 갖추거나 도덕 법칙에 따라 행동을 선택하는 AI를 만드는 것이 의미가 없다고 주장한다. 그는 결과를 수반하는 AI를 만들 때에는 그 결과가 인간이 선호하는 결과를 도출해내는 AI를

은 AI에 구현될 도덕성이 서양 철학의 도덕 개념, 윤리적 직관 등에 편향될 가능성을 내포한다는 점에서 비판적으로 살펴볼 필요가 있다.[7]

AI 윤리에 대한 논의가 복잡하게 진행되고 있는 상황에서 AI 윤리에 대한 유교적 접근은 상대적으로 부족한 실정이다. 여기서 유교는 '공맹孔孟의 가르침에 근거하여 동아시아 지역에 형성된 도덕 체계'를 지칭한다. 유교는 이상적 도덕 행위자인 '성인聖人'을 상정하고, 성인이 갖춰야 할 덕목, 성인이 되기 위한 수양법과 실천 방안 등을 제시하는 사상으로서 오늘날 우리의 가치 판단에 여전히 큰 영향을 끼치고 있다. 유교의 도덕 체계에 입각한 AI 연구가 부족할 경우, 우리는 AMA를 개발하는 과정에서 유교에서 중시하는 가치를 반영하거나 개발 지침을 제공하는 일을 심도 있게 진행할 수 없게 된다. 간혹 AI 혹은 그와 관련된 기술의 발전을 논의한 연구가 나오더라도 그 연구는 '새로운 시대를 위한 유교적 가치의 재확립'[8]이나 'AI가 변화시킬 가족 공동체의 의미와 유교의 역할'[9]처럼 유교적 가치나 개념과 관련된 '어려운 문제'를 다루는 경우가 대부분이다. 이러한 연구는 초지능 superintelligence의 등장 가능성 앞에서 실존적 위협을 느끼는 우리에게 성찰의 계기를 제공하므로 의미가 있을지는 모르나, AI 윤리 영역에서는 실효성 없는 논의가 될 가능성이 크다. AI의 발전에 대응하여 유교적 인간 본성을 강조하고 이 시대에 유교적 가족 공동체의 의미를 재고하는 일은 AI와 직접적 관련성이 적기 때문이다. AI 윤리에 대한 유교적 접근에서는, 연구자들이 개별적으로 '어려운 문제'에 천착하는 것보다, 유교적 접근이 가능한 '쉬운 문제'를 찾아서 담론을 형성해 나가는 것이 중요하다.

이상의 문제의식에 기초하여 이후에는 선행 연구에 대한 구체적인 검토를

만드는 것이 중요할 뿐이라고 말한다. 스튜어트 러셀, 이한음 역 (2021), pp. 309-357 참조.

7) AI 윤리 논의에 참여하는 주체의 다양성을 강조한 연구는 허유선, 이연희, 심지원 (2020), pp. 165-209 참조.

8) 김백희 (2018), pp. 391-411 참조.

9) 이현지 (2016), pp. 91-116 참조.

진행하고, AI 윤리에 대해 유교적으로 접근 가능한 문제는 무엇이 있는지, 그에 대해서는 어떤 주장이 가능한지 살펴볼 것이다.

Ⅱ. AI 윤리에 대한 유교적 이해와 대응

AI를 도덕적 행위자로 인정할 수 있는가의 문제를 포함하는 AI 윤리는 유교 연구에서 주요 관심 분야가 아니다. 물론 우리가 AI 윤리에 대해 관심을 가져야만 하는 당위는 없다. 하지만 알파고AlphaGo의 등장으로 많은 연구자가 적어도 바둑에서는 인간을 뛰어넘는 AI가 등장했음을 인지하고 있으며, 자율주행차에 대한 일반적 관심이 높아진 상황에서 AI에 대한 연구가 많지 않다는 것은 의외이다. 그 이유를 추정해보면, ① 기존 AI 윤리의 논의틀은 유교적 논의틀과 대응하지 않는다. 기존 AI 윤리에서 중요하게 다루는 '책무성accountability', '설명가능성explicability' 등은 AI 기술의 발달과 함께 대두된 개념으로서 그에 대한 문제의식이 없는 철학 체계에서는 쟁점이 되지 않았다. 유교 연구에서는 AI 윤리가 자신들이 다뤄야 할 주제로 인지되지 못했던 것이다. ② 기술에 대한 철학적 분석은 해당 기술에 대한 이해를 요구한다는 점에서 AI 윤리 연구는 접근성이 낮다. AI 윤리에 대한 연구는 유교 연구에 AI를 끼워 넣는 방식이 아니라, AI 윤리 연구에 유교적 접근을 시도하는 방식으로 이뤄진다. 기술로서의 AI에 대한 이해와 기존의 AI 윤리에 대한 비판적 검토가 선행되어야 한다는 점에서 진입 장벽이 존재한다.

그럼에도 최근 국내의 유교 연구 분야에서는 AI가 끼칠 사회적 영향, AI와 인간의 관계 설정, AI를 활용하여 우리 사회를 더 나은 방향으로 이끌고 갈 방안 등을 '유교적 도덕 행위자'와 관련지어 논의한 연구들이 나오고 있다. 여기서는 그러한 선행연구의 의미를 '쉬운 문제'와 '어려운 문제' 구분에 기초하여 확인함으로써 AI 윤리에 대한 유교 연구의 현재를 진단하고 보완해야 할 사항을 찾아보고자 한다.

1. 유교적 인간 본성의 견고화

AI를 다룬 최근의 유교 연구는 AI가 인간에게 가할 위협을 전제한 상태에서 인간과 AI의 존재론적 지위, 양자 사이의 관계 등을 어떻게 설정할 것인가와 같은 문제를 주로 다루는 경향을 보인다. 이 연구들은 유교적 인간 본성에 대한 이해를 기초로 AI의 지위를 규정하려 시도한다는 점에서 '어려운 문제'에 속한다. 우선, 양선진은 그의 연구에서 AI 기술 발전이 인간의 존엄성을 위협하고 인간의 본질이 무엇인지 묻게 만든다고 말한다.[10)]

양선진은 "인공지능이 등장한 시점에서 인간의 규정은 동물과의 차이점에서 정의내릴 것이 아니라 새로운 경쟁적 존재인 인공지능과의 차별성에서 정의해야 하는 것은 아닌가?"[11)]라고 물으며 AI 윤리를 논하기 전에 인간의 본질을 살펴본다. 그는 '존엄성'을 인간의 본질로 제시하고 인간이 어째서 존엄한지를 다양한 방식으로 설명한다. 그 중에서 유교적 가치관과 관련 있는 설명은 인간의 영성靈性을 존엄성의 근거로 제시한 것이다. 양선진은 왕수인王守仁이 심즉리心卽理를 주장한 것에 근거하여 인간이 존엄성을 획득할 수 있는 이유를 인간은 과도한 '생리적 욕망'과 '이기적 욕망'을 제거함으로써 자신 안에서 '천리天理'를 발견할 수 있기 때문이라고 주장한다. 그에 따르면 인간은 자신 안에서 천리를 발견함으로써 인간다움을 얻고 그것으로부터 존엄성을 획득한다.[12)]

인간의 본질이 무엇인지를 논한 이후 양선진은 'AI 기술 발전에 의한 기술 변화가 노동자에게는 큰 재앙이 될 것이고 대부분의 인간들은 일자리를 잃고 자본가들의 자본축적만이 이뤄지는 시대가 올 수도 있다'는 내용의

10) 양선진 (2016), pp. 479-507 참조.
11) 양선진 (2016), p. 483.
12) 양선진 (2016), pp. 491-492.

예측을 인용하며 AI가 인간에게 어떤 위협으로 다가올지 서술하였다.[13] 인간의 노동을 AI가 대체하는 산업구조의 변화가 불가피하다는 것, 많은 노동자가 일자리를 잃을 것이라는 등의 예측은 '4차 산업혁명'과 함께 AI에 의한 위협으로 많이 언급되는 사안이다. 양선진은 그러한 위협에 대비하여 우리는 인간만의 고유한 영역에 집중하고 그에 적합한 윤리를 개발해야 하며, 자본축적에 대한 욕망을 제거하기 위한 윤리교육이 강화되어야 한다고 주장한다.[14] 요컨대 그는 AI 기술이 인간 소외로 이어지지 않기 위해서는 인간 스스로 욕망을 제거하고 존엄성을 획득해야 한다고 강조하였다.

양선진은 인간의 본질을 유교적 수양론인 '존천리거인욕存天理去人欲'과 관련지어 검토하였으며, AI와 인간의 관계를 '일자리', '자본축적'과 같은 사회문제로부터 접근하여 설정하였다. 그의 주장에서 주목할 부분은 해당 논의가 '어려운 문제'를 해결하는 데 집중하고 있다는 사실이다. AI 기술의 발전이 야기할 산업구조의 변화, 인간의 존재론적 지위 확립 등은 'AI를 도덕적 행위자로 인정할 수 있는 조건은 무엇인가', 'AMA에게 법적 책임을 부과할 수 있는가' 등의 문제와 비교하면 의견 차이를 해소하기 어려운 문제에 해당한다.

임헌규의 연구도 양선진의 논조와 궤를 같이한다. 그 또한 AI가 단순 생산 분야뿐만 아니라 의료, 군사 등과 같은 분야에도 활용될 수 있음을 언급하며 인간의 대량 실업을 우려한다.[15] 그는 AI 기술의 본성과 파장에 대한 유교 연구의 필요성을 주장하며 다음과 같이 말하였다.

> 특히 공자 이래 儒敎는 그 학파의 명칭(儒=人+需: 人間에게 必需적인 것)이 시사하듯, 항상 "인간의 근본이 무엇인가"하는 물음을 제기하면서, 교학상장을 통해 인간다운 인간 혹은 인간의 이상을 정립·양성하려고 지난한 실천적 노력을 기울여 왔다. 인간에 의해 만들어진

13) 양선진 (2016), pp. 496-498.

14) 양선진 (2016), pp. 502-504.

15) 임헌규 (2018), pp. 123-143 참조.

> 인공지능의 시대에 인간 혹은 그 이상의 지적 작용을 수행할 수 있는 로봇이 만들어질 가능성과 함께, 이제 우리는 "무엇이 진정 인간을 인간답게 하는가?"라는 물음을 다시 제기하면서, 그에 대한 물음과 대답을 새롭게 정립할 필요가 있다.[16]

임헌규는 유교에서 인간의 본성을 다룬 논의인 심성론을 재검토하고 그것이 'AI 시대'에 어떤 의미가 있는지 논한다. 그에게 AI는 인간의 본성을 다시 한 번 되짚어볼 계기를 마련해주는 위험요소이다. 그의 관심사는 AI 자체에 있지 않고 AI에 대응하는 인간의 자기인식과 지위 확립에 있었다. 임헌규는 기술의 발전에 의하여 인간의 지능과 동일하거나 그것을 뛰어넘는 수준의 AI가 등장하더라도 다음과 같은 조건을 만족시키지 못하면 AI는 인간으로서의 지위를 획득할 수 없다고 주장한다.

> 1) 생물학적 몸을 주재하여 여타 동물과 구별되는 가치 있는 인간의 본성 혹은 심성을 지니고 있지 않다면, 2) 인간과 동일한 방식으로 자기의 본성을 자각하여(자기인식의 문제) 자율적 자기완성의 길(人道)을 가지 않는다면, 나아가 3) 자신의 본성의 덕을 밝히면서(明明德) 공동체적 존재로서 자신의 사명을 자각하여(親民) 타인과 만물의 본성의 덕을 실현시켜 주면서 함께 至善의 공동체를 건설하는데 能參하지 않는다면, 그것[AI]은 이념으로서 인간 혹은 인간의 이념에 위배된다는 것이다.[17]

유교적 인간 본성이 무엇인지를 재확인하고 그 본성에 근거하여 AI가 인간으로서의 이념에 부합하는지를 논하는 것은 '어려운 문제'이다. 인간의 본성을 반드시 유교적 심성론에 기초하여 규정해야만 하는 당위가 확보되지 않는다면 AI의 지위를 규정하는 데에 위와 같은 기준을 제시하는 것이 적절한가의 물음이 제기될 수밖에 없다. 또한 이 논의는 AI 윤리 연구 내에서 큰 주목을 받지 못할 가능성이 크다. AI 윤리 연구는 '인간과 AI를 어떻게

16) 임헌규 (2018), p. 126.
17) 임헌규 (2018), p. 141.

구별할 수 있을까' 보다는 'AI가 인간과 구별된다 하더라도 어떻게 그것에 인격을 부여할 수 있을까' 또는 'AI에 어떻게 책임을 부과할 수 있을까'[18] 등과 같은 문제에 더 비중을 두고 있기 때문이다.[19]

양선진과 임헌규는 각자의 연구에서 유교적 인간 본성에 대한 재검토로써 인간의 존재론적 지위를 견고하게 만들고자 했다. 이것은 AI가 인간에게 실질적인 위협으로 다가오는 상황에서 인간이 변화의 주도권을 잃지 않기 위한 노력의 일환으로 읽힌다. 그들은 도덕 행위자로서의 인간을 기준으로 AI의 지위를 규정하고 관계를 설정했다. 논의의 중점이 '유교적 인간 본성'을 강조하는 데 있었기에 AI 윤리에서 시급하게 해결해야 할 비교적 '쉬운 문제'를 다루지는 않았다. 다음 절에서는 AI를 위협으로서가 아니라 사회적 문제 해결을 위한 긍정적 도구로서 다룬 연구를 살펴보겠다.

2. 사회 문제 해결을 위한 도구화

유교 연구자라고 하여 모두가 AI를 인간에 대한 위협으로 상정하거나 AI는 도덕적 행위자가 될 수 없다는 입장을 취하는 것은 아니다. 최근 연구 중에는 AMA와 유교적 도덕 행위자인 성인聖人의 유사성에 주목하여 AI에

18) AI 윤리와 관련하여 가장 중요하게 다뤄지는 가치인 '책무성'에 대한 논의는 이중원 (2019), pp. 79-104 참조.

19) 황갑연의 연구도 AI의 존재론적 지위를 규정할 때 그 기준을 유교적인 도덕적 행위자에 두고 있다는 점에서 임헌규의 결론과 유사한 방식의 결론이 도출될 수밖에 없다. 황갑연은 비인간으로서의 AI가 도덕실천의 대상이 될 수 있는지, 권리를 가질 자격이 있는지, 도덕실천의 주체로서의 자격을 가질 수 있는지 등을 유교적 인간관에 기초하여 검토한다. 그는 AI를 도덕실천의 고려 대상에 포함시키고, 권리를 부여하는 것은 가능하지만, 지능, 도덕적 정감, 타인에 대한 배려와 의무의식 등을 지닌 인간과 AI를 동일한 수준의 도덕행위의 주체로 인정할 수는 없다고 결론 내린다. 의사 능력을 갖추지 못한 AI를 도덕적 행위자로 인정할 수는 없다는 것이다. 인정 여부를 판가름하는 기준이 '인간의 능력 또는 본성'에 있다는 점에서 그의 논의는 '어려운 문제'에 해당한다. 황갑연 (2019), pp. 243-262 참조.

대한 유교적 접근의 순기능을 조망한 연구도 있다. 정재현은 '한국철학'의 성립 가능성을 모색하는 연구의 일환으로 '유교 성인 AI 만들기'라는 가상의 프로젝트를 제안한다. 그는 유교적 성인을 본보기로 삼아 AMA를 만든다면, 우리는 그 과정에서 한국 사회에 내재한 여러 문제를 해결하는 단초를 얻을 수 있으며, 도덕 교육의 측면에서 많은 도움을 받을 수 있다고 주장한다.[20]

정재현은 유교 성인을 사회가 요구하는 도덕성과 공정성을 지닌 존재로 해석하며 "따라서 그런 [유교 성인] AI를 만들 수 있다면 그것은 현대 한국의 현실에 대한 파악은 물론이고, 이 한국 현실의 문제를 해결할 방안의 제시에 일정 부분 기여"[21]할 수 있다고 말한다. AI를 유교적 도덕 행위자로 만들기 위해서는 성인의 특성, 성인으로서 갖춰야 할 덕의 특징, 한국 사회의 문제 등에 대한 이해가 선행되어야 한다. 이러한 이해는 기존 유교 담론에 대한 분석으로 이뤄진다. 그가 보기에 '유교 성인 AI'를 만드는 과정은, 곧 우리 사회에 필요한 덕은 무엇인지, 그것을 어떤 방식으로 함양할 수 있는지(또는 구현할 수 있는지) 알아가는 과정이기에 프로젝트가 실패하더라도 얻는 바가 있다.

앞 절에서 다룬 두 연구에서 AI는 인간에 대한 실존적 위협으로 묘사된다. 그렇기에 두 연구자는 인간이란 무엇이고 우리는 AI와 어떤 관계를 맺어야 하는지 등의 문제를 제기한 후 유교적 인간 본성의 의미를 견고하게 만드는 데 집중하였다. 이에 비해 정재현은 AMA의 출현이 우리 사회의 문제를 해결하는 데 좋은 수단이 될 것이라고 기대한다. 이러한 기대는 그가 AI를 이해하는 방식에 기반을 둔다. 그가 생각하는 AI는 "인간을 그 한 부분으로 포함하며 이루어지는 하나의 집합 혹은 하나의 체계를 가리킨다. 그것은 인간과 별개로 존재하는 것이 아니라, 인간을 도와주면서 인간과 공존하는 체계"[22]이다. 그는 인간의 본질은 무엇이고 존재론적 지위를 어떻게 확립할

20) 정재현 (2021), pp. 91-119 참조.
21) 정재현 (2021), pp. 94-95.
22) 정재현 (2021), p. 96.

것인지를 논하기 보다는, AI 개념을 유연하게 적용하는 방식을 택하여 인간과 AI의 관계를 대립이 아니라 상보적인 관계로 설정한다. 시스템으로서의 AI에 직접 참여하여 AI를 우리 사회의 문제를 해결하는 수단으로 활용하자는 그의 주장은, AI가 우리 사회에서 도덕적 행위자로서 기능하기 위해 '인간 또는 AI의 존재론적 지위'라는 '어려운 문제'가 반드시 선결되어야 하는 것은 아니라는 점을 보여준다.

정재현은 AI의 기계학습machine learning을 "바람직한 처리 결과를 의식하면서 많은 주어진 데이터로부터 일종의 알고리즘 즉 문제처리 방법을 끄집어내는"[23] 것으로 이해한다. 그는 이와 같은 학습 방식이 유교에서 성인을 만드는 "구체적 사례를 통한 인문 교육의 방식"[24]과 유사하다고 생각한다. 그는 성인이란 어떤 도덕적 행위자인지를 다음과 같이 설명하였다.

> 성인은 추상적 도덕 원리를 논리적 추론을 통해 구체적 상황에 적용하여 도덕적 행위를 하는 것이 아니고, 구체적 상황을 통해 습득한 공감적 능력을 발전시키고 이런 발전된 공감적 능력을 구체적 상황에서 발휘하는 사람으로 보아야 한다.[25]

위의 인용문에 따르면 유교적 성인은 어떤 도덕 원리를 먼저 습득하고 그것을 구체적 상황에 적용하는 방식으로 행위를 하지 않는다. 달리 말하면 성인의 도덕성은 하향식Top-down으로 구현되지 않는다. 이에 대해 정재현은 성인의 도덕성이 구현되는 방식은 의무론, 공리주의처럼 도덕 규칙이 선재하는 방식과 다르다고 말한다.[26] 그는 성인을 개별적인 윤리적 문제들을 해결하는 과정에서 공감 능력을 발전시키고 덕을 실현하는 사람이라고 생각한다. 성인의 도덕성은 상향식Bottom-up으로 구현된다고 본 것이다. 그는 성인의

23) 정재현 (2021), p. 104.
24) 정재현 (2021), p. 102.
25) 정재현 (2021), pp. 102-103.
26) 정재현 (2021), p. 104.

도덕성이 형성되는 과정을 AI가 경험에 근거하여 도덕적 행동을 배우는 것과 동일한 과정이라고 이해하였다.[27]

정재현의 논의는 도덕성이 구현되는 방식의 유사성을 바탕으로 유교적 도덕 행위자 개념을 AMA에 적용한 것이다.[28] 성인에 대한 유비로써 AMA의 가능성을 인정하고 그것의 실용성을 논한 것이기도 하다. 이는 유교 연구 분야에서 AI 윤리에 접근하기 위해 어떤 방식으로 문제를 설정하고 해결하는 게 실효성이 있을지를 보여주는 사례에 해당한다. 그는 'AI에 어떻게 도덕성을 구현할 것인가'라는 문제에 대하여 성인을 하나의 표본으로 삼는 해결 방안을 제시하였다. 이러한 논의가 과연 얼마나 타당하고 받아들일 만한 것인가에 대해서는 더 비판적인 검토가 필요하다. 하지만 여기서 중요한 것은 유교 연구자들이 AI 윤리 담론에 대화 상대자로 참여하기 위해서는, 유교적 성인으로서의 AMA 개발을 제언하는 것처럼, '쉬운 문제'에 주목하고 그에 대한 해결 방안을 모색하는 게 도움이 된다는 사실이다.

'쉬운 문제'라고 해서 그것을 해결하는 게 단순하기만 한 것은 아니다. 정재현은 '유교 성인 AI'의 성립 가능성을 주장하기 위해 세 가지를 전제하고 있다. 덕의 상황주의적 해석 또는 행동주의적 해석, 감정의 기능주의적 해석, 깨달음에 대한 객관주의적 해석 등이다.[29] 쉬운 문제에 대한 해결은 인접한 문제에 대한 재검토를 요구하는 경우가 많으며, 그 과정에서 우리는 기존의 개념과 직관이 확장되는 것을 경험하게 된다. 이상욱은 그러한 '쉬운 문제' 해결이 점진적으로 이뤄지다 보면 '어려운 문제'도 해결될 수 있다고 주장한

27) AI의 도덕성을 구현하는 방식인 상향식, 하향식, 절충식에 대한 논의는 김효은 (2019), pp. 97-123 참조.

28) 성인이 도덕성을 갖추게 되는 방식을 AI에 적용하는 데에는 풀어야 할 과제들이 있다. 특정 문화에 맞춰서 상향식으로 도덕적 행위를 학습할 경우 그 AMA가 다른 문화권에서도 도덕적으로 행위를 할 수 있을 것인지, 개발자의 의도와 다르게 AI가 비도덕성을 구현하게 될 경우 어떻게 대응할지 등에 대한 논의가 병행되어야 한다.

29) 정재현 (2021), pp. 109-115 참조.

다.[30] 그의 주장처럼, 유교 연구에서 AI 윤리에 접근할 때에 우선적으로 해결해야 하는 문제는, 인간의 본성을 유교적으로 재정립하고 그것을 기준으로 AI를 도덕적 행위자로 인정할 수 있느냐가 아니라, 유교 내에 도덕적 행위자의 개념을 확장하여 AI를 포섭할 여지가 있느냐는 것이다. 이와 관련하여 다음 장에서는 유교적 AI 윤리 연구가 지향해야 할 바가 무엇인지를 논하겠다.

Ⅲ. AI 윤리의 다양성을 확보하기 위한 시도

AI 윤리 분야에서 시급하게 해결해야 할 문제로 편향bias과 다양성diversity 문제가 있다. AI 기술에는 크게 두 가지 요인에 의한 편향 문제가 잠재한다. 하나는 자료를 탐색하고 분석하여 규칙을 추론하는 '자료채굴data mining' 단계에서 발생하는 편향이고, 다른 하나는 알고리즘을 설계하는 전문가의 편향이 반영되어 발생한 편향이다.[31] AI는 결과를 추론하는 과정에서 '자료'와 문제를 해결하는 절차로서의 '알고리즘'에 의존할 수밖에 없다. AI가 수집하고 학습한 자료가 무엇인지, 알고리즘에 대한 인위적 조작 여부 등이 AI가 내놓는 결과에 영향을 끼친다. 이것은 기술적 문제에 해당하는 것으로서 자료의 다양성, 알고리즘의 투명성 등을 확보하면 개선 가능하다.[32]

문제는 두 번째 요인에 의한 편향이다. AI가 학습하는 자료가 편향되거나 알고리즘이 공정하지 않은 방식으로 설계되는 것은 결국 전문가 혹은 전문가가 속한 사회에 편향이 있기 때문이다. 이를 잘 보여주는 사례가 2018년에 폐기된 아마존의 채용 AI이다. 이 AI는 채용 과정에서 '여성'이라는 단어를

30) 이상욱 (2019), pp. 271-276 참조.

31) AI의 편향이 발생하는 이유를 기술적 특성에 근거하여 분석한 연구는 정원섭 (2020), pp. 165-209 참조.

32) AI의 편향성과 그에 대한 개선 방안을 논한 연구는 허유선 (2021), pp. 201-234 참조.

포함하거나 여자대학을 졸업한 지원자의 이력서에 감점을 줬다. 반면 남성 개발자들이 자주 사용하는 단어인 '실행execute', '수집capture' 등을 포함한 이력서는 긍정적으로 평가했다.[33] 이런 편향이 발생한 주요 원인으로 지목된 것이 IT 기업인 아마존의 개발직 성비 불균형이다. 기업 구성원의 대다수가 남성인 상황에서, AI가 지원자를 평가하기 위해 만든 기준에 활용된 자료는 남성 편향적일 가능성이 높다. 개발자 스스로 의식하지 못했더라도 AI가 개발되고 학습하는 환경에 이미 편향이 내재해 있기에 결과에 문제가 생긴 것이다.

편향은 공정한 채용을 위해 만들어진 AI에서만 문제되는 것이 아니다. 예측 가능한 더 심각한 문제는 AI의 도덕성이 편향될 때 일어난다. 자율주행차의 윤리적 곤경에 대한 논의에서 자주 언급되는 '광차문제trolley problem'처럼, AI의 '판단'은 사람의 목숨과 직결될 수 있다.[34] 만약 AI가 불가피한 사고 상황에서는 언제나 '특정한 연령, 성별, 인종 등의 조건을 만족하는 사람을 희생자로 결정하는 것이 옳다'는 편향된 결론을 내리게 된다면, 우리는 이전에는 없던 방식의 사회적 차별과 갈등을 경험하게 된다.[35] 또한 AI가 내린 결론이 반드시 인간이 납득할 수 있는 결론이 아닐 가능성도 있으므로 이로 인한 혼란도 해결해야 한다. 그렇기에 여러 영역에서 상용화를 목표로 기술이 개발되고 있는 지금 시점에 AI의 도덕성을 어떤 방식으로 구현해야 편향

33) 박소정 (2018), "이력서에 '여성' 들어가면 감점...아마존 AI 채용, 도입 취소", https://www.chosun.com/site/data/html_dir/2018/10/11/2018101101250.html. (검색일: 2022.09.08.)

34) 광차문제는 제동장치가 고장 난 탄광 수레가 소수와 다수의 사람 중 어느 한 쪽을 희생시킬 수밖에 없을 때 어떤 선택을 해야 하는가의 문제이다.

35) 인간 운전자도 동일한 상황에서는 윤리적 곤경을 겪는다. 하지만 인간이 운전자인 경우에는 사고에 대한 책임의 주체가 분명하다. 또한 운전자의 진술을 통해 그가 그러한 판단을 내리게 된 절차를 확인할 수도 있다. 하지만 AMA에 대해 광차문제가 끊임없이 제기되는 이유는 그것이 '인공적' 도덕 행위자이기 때문이다. 여기에는 AMA를 책임의 주체로 놓는 것이 정당한지에 대한 쟁점, AMA가 판단을 내린 과정에 윤리적 결함은 없었는지 확인하기 어렵다는 쟁점 등이 포함된다는 점에서 인간의 경우와 다르다.

문제를 해결 할 수 있을지 고민하는 것은 중요하다.

현재 논의되는 AI의 도덕성을 구현하는 방식에는 앞서 언급했던 것처럼 의무론, 덕 윤리학 그리고 공리주의 등이 있다. AI를 연구하는 학자들은 세 가지의 접근법 중에서 어떤 방식이 AMA를 만드는 데 더 적합한지를 두고 논쟁을 하고 있으며, 여기에는 어느 한 방식만 채택하기 보다는 각자의 방식이 지니고 있는 한계를 보완하기 위해 절충적 방식이 적용될 것이라고 전망하는 입장도 있다.[36] 우리는 이미 'AI는 도덕적 행위자가 될 수 있는가'를 묻는 단계를 벗어나 '도덕적 행위가 가능한 AI를 어떻게 만들 것인가'를 고민하는 단계에 있다.

문제는 이러한 담론이 철저하게 서양 철학을 중심으로 형성되고 있다는 사실이다. 칸트적 의무론에서 제시하는 도덕 법칙이나 공리주의에서 우선적으로 고려하는 가치를 구현한 AMA가 한국과 같은 유교 문화권의 도덕 체계 안에서도 도덕적 행위자로 기능할 수 있는지에 대한 고려가 충분히 이뤄지지 않았다는 의미이다. 또한 어떤 유교적 가치가 도덕 법칙으로 정립되기에 적절한지 또는 사회적 유용성을 측정하는 지표가 될 수 있는지 등을 논한 연구도 많지 않다. 의무론이든 공리주의이든 AI에 구현된 도덕성은 해당 체계를 구성하는 도덕 개념과 직관에 의존한다. 그렇기에 AI 윤리 분야에서 지금까지 논의된 의무론, 공리주의 등을 그대로 AI에 적용하면 유교를 포함한 동양 철학적 도덕 개념과 직관은 설자리를 잃을 것이며 AMA는 도덕적 편향을 갖게 될 것이다. 그런 점에서 정재현이 '유교적 성인'의 도덕성을 구현하는 AMA에 주목한 것은 도덕적 다양성의 확보라는 측면에서 유의미하다. 다만 '유교 성인 AI'는 온전히 유교적인 도덕 행위자를 구현한다는 특수성에 초점을 맞추고 있기에 보편성을 확보하기 위해서는 또 다른 차원의 연구가 수행되어야 한다.

예를 들면, 공자가 말한 성인은 '자신을 희생하여 인을 이루는[殺身成仁]'

36) 이와 관련해서는 자율주행차를 의무론적 자율주행차와 공리주의적 자율주행차로 구분하여 논한 변순용·이연희 (2020), pp. 196-202 참조.

인물이다.[37] 만약 자율주행차가 무단횡단을 하는 사람을 살리기 위해 탑승자를 태운 채 방향을 꺾어 벽과 충돌하는 '살신'을 선택한다면, 이 선택은 모두를 만족시키는 도덕적 행위가 될 수 있을까를 고민해야 한다. 의무론의 관점에서는 '자신을 희생하여 인을 이뤄야 한다'는 명제가 보편적으로 타당한 도덕 규칙인지, 공리주의의 관점에서는 자신을 희생하는 것, 인을 이루는 것, 무단횡단을 하는 자를 살리는 것 등의 사회적 효용은 얼마나 되는지 등을 평가해야 한다. AMA의 도덕적 편향을 방지하고 다양성을 확보하기 위해서는 종합적인 검토가 이뤄져야 한다는 것이다.

AMA의 도덕적 편향을 방지하기 위해서는 AI의 도덕성 담론에 더 다양한 논의 주체가 참여해야 한다. 이것은 유교 성인 AI, 기독교 성인 AI처럼 AMA 종류의 다양성 확보를 도모하자는 주장과 다르다. 이것은 AI가 구현하게 될 도덕성에 유교적 가치를 포함한 다양한 도덕 개념과 직관이 반영되게 해야 한다는 것을 의미한다. 특정한 도덕 체계에 기초하여 만든 AMA는 그 체계 안에서만큼은 잘 작동하는 행위자가 되겠지만, 그 체계를 벗어나 다른 도덕 체계에 기초하여 제작된 AMA를 만났을 때에는 윤리적 충돌을 일으킬 수 있다. 우리는 그러한 충돌이 우리에게 어떤 영향을 끼칠지 아직 정확히 판단하지 못한다. 판단을 위한 경험적 근거가 부족하기 때문이다. 그렇기에 AI 윤리는 특수성 보다는 보편성을 지향하며 그 안에서 지속적인 학습을 거쳐 특수성을 확보하는 방향으로 연구되는 것이 바람직하다.

그렇다면 AI 윤리 담론에 대한 유교적 접근은 어떤 방식으로 가능할까? 먼저 생각해 볼 수 있는 방식은 AMA가 유교적 덕목을 체득하게 만드는 것이다. 대표적 덕목으로 공자가 증삼曾參에게 '오직 한 가지로 관통하였다[一以貫之]'고 말했을 때 그 '한 가지'에 해당하는 '서恕'가 있다.[38] 자공子貢은 공자에게 종신토록 행할만한 한 글자를 물었고 공자는 다음과 같이 대답한다.

---

37) 『論語』, 「衛靈公篇」 "志士仁人, 無求生以害仁, 有殺身以成仁."

38) 『論語』, 「里仁篇」 "子曰, 參乎, 吾道一以貫之. 曾子曰, 唯. 子出, 門人問曰, 何謂也. 曾子曰, 夫子之道, 忠恕而已矣."

> 선생께서 말하시길, “그것은 ‘서’이다. [‘서’는] 내가 원치 않는 바를 남에게 베풀지 말라[는 것이다].”[39]

윤리학에서 ‘서’는 ‘소극적 황금률negative golden rule’로 이해된다.[40] 다양한 종교와 사상에서 제시된 황금률은 ‘다른 사람들이 당신에게 행하기를 원하는 것을 그들에게 행하라’라는 형식을 띤다. ‘서’는 그것을 금지문의 형태로 내세운 것이다. ‘서’는 공동체 내에서 자신과 타인이 어떻게 관계를 맺어야 하는지를 제시한 실천적 덕목이라 할 수 있다. 만약 AMA가 이 황금률을 체화한다면 그것은 유교적 도덕 행위자로 인정받기에 적절한 요건 중 하나를 충족한 것이다.[41] 하지만 여기서 한 가지 고려해야 하는 것은 AMA는 인간과 달리 욕망[欲]을 결여하고 있다는 사실이다. AMA는 스스로 자신이 ‘원하는 바’가 무엇이고 ‘원치 않는 바’가 무엇인지 판단하지 못한다. 그러므로 AMA가 ‘서’를 행하기 위해서는 두 가지가 함께 해결될 필요가 있다. 하나는 AMA가 ‘인간이 원하지 않는 바’라는 개념을 이해하고 실제로 원하지 않는 바에 어떤 것들이 있는지 학습해야 한다는 것이다. 다른 하나는 ‘서’를 AMA에 적용할 수 있는 법칙으로 변환하는 것이다. AMA에 적용될 ‘서’는 다음과 같이 변환 가능하다.

> 어떤 행위를 했을 때 그 결과가 인간에게 해를 끼칠 것으로 예측되는 경우, AMA는 그 행위를 해서는 안 된다.

한 가지 문제 상황을 가정할 수 있다. 자율주행차를 다시 소환해 보면,

---

39) 『論語』, 「衛靈公篇」 “子曰, 其恕乎, 己所不欲, 勿施於人.”

40) ‘서’를 황금률의 관점에서 다각적으로 분석한 논문은 강진석 (2017), pp. 7-33 참조.

41) 윤리학 내에는 황금률이 도덕 법칙으로 적절한 원리인지에 대한 논쟁이 있다. 그러나 ‘서’ 자체를 분석하고, 황금률의 도덕 법칙으로서의 적절성을 논하는 것은 이 글의 주제를 벗어나는 것이기에 여기서는 다루지 않는다. 황금률에 대한 분석은 박종준 (2016), pp. 227-255 참조.

어떤 자율주행차가 무단횡단을 하는 사람과 부딪힐 위험에 처한 경우 이 자율주행차는 아무런 판단도 하지 못할 가능성이 있다. 자율주행차가 무단횡단을 하는 사람을 치고 계속 달려가는 일이 발생할 수도 있다는 뜻이다. 무단횡단을 하는 사람을 치는 행위와 그 사람을 피해 탑승자를 태운 채 벽으로 방향을 꺾는 행위는 어느 쪽이든 인간에게 해를 끼치기 때문이다. 그래서 이 법칙을 AMA에 적용할 때에는 이를테면 '살신성인'과 같은 법칙이 함께 적용될 필요가 있다. '자율주행차는 도로 위를 횡단하는 사람을 칠 것 같은 경우에는 탑승자의 희생을 감수하더라도 벽으로 돌진하여 주행을 멈춰야 한다'와 같은 법칙을 따르게 만들어야 한다는 것이다. 물론 유교가 아닌 도덕적 직관을 가진 체계에서는 '살신성인'을 정당하지 않은 것으로 여길 수도 있다. 필자의 주장은, 그것이 정당한지 정당하지 않은지 등을 AI 윤리 담론 내에서 다양하게 다뤄야 한다는 것이다.

또 다른 방식은 AMA가 유교적 덕목을 실천하는 것을 선호하게 만드는 방식이다. AI가 도덕적인 것과 AI의 행위가 도덕적인 것은 다르다. AI가 도덕법칙을 따르게 하거나, 덕을 갖추게 만든다고 해서 그 AI의 행위가 언제나 인간에게도 도덕적일 것이라고 판단할 근거는 없다. 그런 맥락에서 스튜어트 러셀Stuart Russell은 애초에 AI는 인간에게 이로워야하며, AI의 목적은 인간이 선호하는 것을 실현하는 것이어야 하고, 인간이 선호하는 바가 무엇인지는 인간의 행동으로부터 학습해야 한다고 말한다.[42] 이에 비춰보면 우리가 고민해야 할 문제는 '어떻게 AI에 도덕성을 구현할 것인가'가 아니라, '어떻게 AI가 도덕적으로 행위를 하게 만들 것인가'가 된다. AI를 도덕적 행위자로 인정할 수 있는지의 문제에서 AI의 도덕성이 아니라 AI의 행위의 도덕성이 쟁점이 될 때 고려해야 하는 것은, 행위의 도덕성을 판가름할 때 다양한 가치를 고려해야 한다는 것이다. 이와 관련해서는 정약용丁若鏞의 주장을 하나의 모델로 삼아 논의해볼 수 있다.

---

42) 스튜어트 러셀, 이한음 역 (2021), pp. 251-268 참조.

주지하듯, 정약용은 성性을 기호嗜好로, 인의예지仁義禮智를 인간에게 내재한 본성이 아니라 선한 행위에 대한 평가적 명칭으로 이해하였다. 그는 사덕四德을 실천으로 구현되는 것, 사회적 평가로 구성되는 것이라고 생각했던 것이다. 정약용의 주장이 유교를 대표하는 것은 아니지만, 우리가 그의 주장을 택하여 AMA에 적용한다면, 우리는 AI가 도덕적 본성을 갖추고 있는지를 묻는 대신에 그것의 행위를 도덕적이라고 평가할 수 있는지에 주목할 수 있게 된다. 정약용이 언급한 '기호'는 스튜어트 러셀이 강조한 '선호'와 맞닿아 있으며, 행위의 도덕성이 실천으로 구현되는 것이라는 정약용의 생각은 AMA의 행위가 인간의 행동을 학습하여 수행되는 과정에서 도덕적인 것으로 평가될 수 있다는 관점으로 해석할 여지가 있다. 그렇기에 정약용의 주장은 유교적 덕목을 실천하는 AMA를 설계하는 데 적절한 모델로 기능한다.

자세히 살펴보면, 정약용은 인간의 본성에 대해 "이제 인간의 성을 논하면, 인간은 선을 좋아하고 악을 부끄러워하지 않음이 없다."라고 말하였다.[43] 이에 따르면 인간의 본성은 선을 좋아하고 악을 부끄러워하는 기호이다. AI를 만들 때, 우리는 이와 같은 인간 본성의 특징을 AI에 적용할 수 있다. AI가 선을 행하는 것을 선호하고 악을 행하는 것을 꺼리는 경향성을 갖게 설계하는 것이다. 이때 선과 악이 무엇인지에 대한 판단은 인간 행위에 대한 학습을 거쳐 이뤄진다. 정약용은 유교적 덕목으로서의 인의예지를 일이 행해진 뒤에 성립하는 것으로 규정하며 다음과 같이 주장하였다.

> 인의예지라는 명칭은 일을 행한 뒤에 이뤄지는 것이다. 그러므로 사람을 아껴준 후에 그것을 인이라고 하니 사람을 아껴주기 전에는 인이라는 명칭이 성립하지 않는다. 나를 선하게 한 뒤에 이것을 의라고 하니, 나를 선하게 하기 전에는 의라는 명칭이 성립하지 않는다. 손님과 주인이 절하고 읍한 뒤에 예의 명칭이 성립하는 것이다. 사물이 명료하게 분별된 뒤에 지의 명칭이 성립하는 것이다.[44]

---

43) 丁若鏞, 『心經密驗』, 「心性總義」 "今論人性, 人莫不樂善而恥惡."

44) 丁若鏞, 『孟子要義』, 「公孫丑」, <人皆有不忍人之心章> "仁義禮智之名, 成於

정약용에게 사덕은 인간이 태어나면서 품수한 덕목이 아니라 행위에 대한 평가로서 성립하는 명칭이다. 도덕이라는 것은 실천으로써 구현되는 것이지 처음부터 내 안에 갖춰져 있는 것이 아니라는 뜻이다. 이를 참고하면, 우리는 AI가 처음부터 덕을 갖고 있거나 도덕법칙을 지니고 있어야 한다고 생각할 필요가 없다. 선을 행하는 것을 선호하고 악을 행하는 것을 꺼리는 경향성을 가진 AI에게 인간의 행위 중에 어떤 것이 사덕을 실현한 행위인지 학습시키고, 그것을 행하는 것을 선호하게 만듦으로써 우리는 유교적 도덕을 구현하는 도덕적 행위자를 만들 수 있기 때문이다.

이상에서 살펴본 두 가지 방식 ① 유교적 덕목을 갖춘 AMA 설계와 ② 유교적 덕목을 실천하는 AMA 설계는 AI의 도덕성에 대한 담론에 유교는 어떤 방식으로 참여할 수 있는지를 고민하는 과정에서 실험적으로 제안한 것이다. AMA가 편향된 도덕성을 구현하는 문제는 '쉬운 문제'에 해당한다. 도덕성에 대한 여러 담론을 고려하고 반영하여 다양성을 추구하는 과정에서 충분히 해결할 수 있는 문제이기 때문이다. 이때 유교 연구는 AMA가 갖춰야 할 덕목은 무엇이고 그것을 어떻게 AI에 적합하게 변형시킬지, AI가 실천해야 할 덕목에는 어떤 것이 있는지 등을 고찰함으로써 해결책 마련에 한 역할을 담당할 수 있다.

Ⅳ. 나가는 말

유교 연구 분야의 AI에 대한 논의는 대체로 '어려운 문제'를 해결하는 데 주력하고 있는 것처럼 보인다. 그러한 연구 경향이 생긴 이유는 연구자들이 인간만 가능하다고 생각했던 영역을 AI에게 침범 당함으로써 인간의 존재론

行事之後. 故愛人而後謂之仁, 愛人之先, 仁之名未立也. 善我而後謂之義, 善我之先, 義之名未立也. 賓主拜揖而後, 禮之名立焉. 事物辨明而後, 智之名立焉."

적 지위가 위태로워졌다고 생각했기 때문이다. 그들은 AI의 위협에도 흔들리지 않는 인간의 지위를 확보하기 위해 유교적 인간 본성을 강조하는 방향의 연구를 수행하였다. AI와 비교하여 인간이 존엄한 이유, AI가 도덕적 행위자가 될 수 없는 이유 등을 유교적 개념에 근거하여 밝히려 했던 것이다. 이러한 문제의식에서 출발한 연구는 결과적으로 AI 윤리 담론 자체에는 영향을 끼치지 못할 가능성이 높다. 현재 진행되고 있는 AI 윤리 연구는 '어려운 문제'보다는 '쉬운 문제'를 해결하는 데 관심을 갖고 있기 때문이다. 인간과 AI의 존재론적 차이를 논하거나, AI를 도덕적 행위자로 인정할 수 없는 유교적 인간관 등은 '어떻게 하면 AMA에 책임을 물을 수 있을까'와 같은 '쉬운 문제'에 대한 실효성 있는 해결책을 제시하기 어렵다.

한편, 연구 중에는 유교적 성인이 되기 위해 수양하는 과정과 AI가 학습을 하는 과정의 유사성에 주목하여 유교적 AMA의 효용성을 논한 연구도 있다. 한국 사회에 내재한 문제를 해결하기 위해서는 한국 사회에 최적화된 도덕적 행위자가 있어야 하며, AI를 유교적 성인으로 만들면 문제가 해결된다고 생각한 것이다. 성인을 AMA의 모형으로 삼은 이 연구는 유교적 인간 본성을 검토하는 것과 같은 '어려운 문제'를 직접 해결하지 않아도 AI 윤리에 대한 유교적 접근이 가능하다는 사실을 보여줬다는 점에서 의미가 있다. 다만, AMA의 도덕성을 유교적 도덕 체계에 기초하여 구현하는 것은 도덕성의 편향을 야기할 가능성이 있으며, 다른 체계의 도덕성을 구현하는 AMA와 충돌을 일으킬 수 있다. 그렇기에 AMA를 설계할 때에는 도덕의 특수성보다는 보편성을 지향하는 AMA를 만드는 것이 적절하다.

AI가 도덕적 행위자가 되기 위해서는 AI 스스로 도덕적거나 AI의 행위가 도덕적이어야 한다. 이에 따라 AMA에 유교적 도덕성을 구현하는 방식으로는 두 가지를 고려할 수 있다. 하나는 유교적 덕목을 갖춘 AMA를 설계하는 것이고, 다른 하나는 AMA의 행위가 유교적 덕목에 부합하게 만드는 것이다. 이 문제에서 염두에 둬야 할 것은 AI가 인공적으로 도덕성을 품수한다고 해서 그것의 행위가 언제나 인간에게도 도덕적이리라는 것을 보장할 수는

없다는 것이다. 또한 AI가 인간이 선하다고 판단하는 행위를 학습하는 과정에서 정보의 편향을 통제하는 확실한 방안을 마련하는 것도 중요하다.

이 연구는 AI 윤리에 대한 유교적 접근 방식을 모색하는 연구이기에 구체적인 논의를 진행하는 데에는 제한이 있었다. 필자가 무엇보다 강조하려고 하는 바는 AI 윤리에서 '쉬운 문제'는 '어려운 문제'를 거쳐야만 해결할 수 있는 문제가 아니라는 사실이다. 초지능의 위협이 하나의 계기가 될 수는 있겠지만, 인간의 존재론적 지위를 유교적 인간 본성에 대한 담론에 근거하여 확립하는 일은 AI 윤리와 별개의 논의이다. 유교 연구가 AI 윤리 담론에 참여하기 위해서는 담론 내에서 유교적 접근이 가능한 '쉬운 문제'를 검토하고 적극적으로 해결 방안을 찾아야 한다. AMA의 책임에 대한 문제를 유교에서는 어떻게 설명할 수 있는지, 법에 위배되는 행위를 명령 받았을 때 AMA는 어떤 선택을 해야 하는지를 '충忠'과 연관 지어 고찰하는 등의 연구는 AI 윤리 담론 내에서도 충분히 유의미하다. '쉬운 문제'에 대한 관심과 해결이 점진적으로 확대되다 보면 유교적 인공지능 윤리의 정립도 가능해질 것이다.

참고문헌

공자, 『論語』

정약용, 『孟子要義』

정약용, 『心經密驗』

강진석 (2017), 「주자 충서론의 다층적 해석에 관한 논의」, 『인문학연구』, 53: 7-33.

김백희 (2018), 「신과학기술혁명시대의 유교윤리주체」, 『동서철학연구』, 89: 391-411.

김효은 (2019), 『인공지능과 윤리』, 서울: 커뮤니케이션북스.

박종준 (2016), 「현대 황금률의 도덕철학적 문제」, 『철학사상』, 60: 227-255.

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

양선진 (2016), 「양명학을 통해 본 인공지능(AI)시대의 과학기술윤리」, 『양명학』, 45: 479-507.

이상욱 (2019), 「인공지능의 도덕적 행위자로서의 가능성: 쉬운 문제와 어려운 문제」, 『철학연구』, 125: 259-279.

이중원 (2019), 「인공지능에게 책임을 부과할 수 있는가?: 책무성 중심의 인공지능 윤리 모색」, 『과학철학』, 22(2): 79-104.

이현지 (2016), 「유교사상과 인공지능시대의 가족에 대한 시론」, 『사회사상과 문화』, 19(3): 91-116.

임헌규 (2018), 「인공지능시대 유교 심성론의 의미: 공맹과 퇴계를 중심으로」, 『대동철학』, 84: 123-143

정원섭 (2020), 「인공지능 알고리즘의 편향성과 공정성」, 『인간·환경·미래』, 25: 165-209.

정재현 (2021), 「인공지능으로 유교성인 만들기: 한국철학의 정초를 위한 실험철학적 시론」, 『동양문화연구』, 35: 91-119.

허유선, 이연희, 심지원 (2020), 「왜 윤리인가: 현대 인공지능 윤리 논의의 조망, 그 특징과 한계」, 『인간·환경·미래』, 24: 165-209.

허유선 (2021), 「인공지능 시스템의 다양성 논의, 그 의미와 확장: 인공지능의 편향성에서 다양성까지」, 『철학·사상·문화』, 35: 201-234.

황갑연 (2019), 「유가철학에서 인공지능로봇 지위 설정에 관한 試論」, 『중국학보』, 88: 243-262.

스튜어트 러셀 (2019), 이한음 역 (2021), 『어떻게 인간과 공존하는 인공지능을 만들 것인가』, 파주: 김영사.

제리 카플란 (2016), 신동숙 역 (2017), 『제리 카플란 인공지능의 미래』, 서울: 한스미디어.

박소정 (2018), "이력서에 '여성' 들어가면 감점...아마존 AI 채용, 도입 취소", https://www.chosun.com/site/data/html_dir/2018/10/11/2018101101250.html. (검색일: 2022.09.08.)

Harwell, Drew (2022), "He used AI to win a fine-arts competition. Was it cheating?", The Washington Post, September 2 2022. https://www.washingtonpost.com/technology/2022/09/02/midjourney-artificial-intelligence-state-fair-colorado/. (검색일: 2022.09.03.)

【Abstract】

Preliminary Study on Confucian Artificial Intelligence Ethics : Focusing on ‘Artificial Moral Agent’

Lee, Jae-Bok

This paper envisions potential role of Confucianism in the context of artificial intelligence ethics by focusing on the issue of artificial moral agent, and it suggests what researches in the field should aim for. To make the discussion productive, AI-related issues have been categorized into easy and difficult problems and this essay moves on to highlight the former to find the niche for Confucianism. First, this study analyzes how Confucian researchers have understood and formed discourse regarding AI so far. Next, this study examines the implications of a research that claims that we should solve social issues in South Korea by making an AI that plays the role of Confucian Sage. Finally, the issue of morality bias is presented as an example of easy problem and Confucian solutions are suggested.

[Key Words] artificial intelligence, artificial intelligence ethics, artificial moral agent, Confucianism, Confucian artificial intelligence ethics

논문 투고일: 2022. 09. 15
심사 완료일: 2022. 10. 19
게재 확정일: 2022. 10. 19