인공도덕성 검사의 가능성

김효은*

【요약】

이 논문은 현 단계의 인공지능의 도덕성 검사로서 기존의 튜링테스트는 적합하지 않으며 설계, 학습, 실행 등 전체 과정에 걸친 검사가 되어야 한다고 주장한다. 이 논문에서 필자는 기존의 튜링테스트처럼 사후 '검사'에 초점을 맞춘 도덕성 튜링테스트는 결국 본래의 튜링테스트의 난점처럼 '얼마나 인간을 잘 모방 하는가'에 대한 검사일 수밖에 없다고 주장한다. 기존 튜링테스트는 실행 후의 언어 반응에 대한 검사이며, 설계-제작이라는 실행 전의 과정에 대한 검사를 포섭할 수 없다. 인공지능의 도덕성 검사는 실행 단계 이후의 검사가 아니라 설계 이후의 전체 단계에 걸친 검사로 확장되어야 한다. 이러한 확장은 인공지능 개발과 투명성 간의 딜레마를 야기할 수 있지만, 학습과정 공개가 아니라 인공지능에 설명기능을 내장한 형태로 극복가능하다. 필자는 2절에서 딥러닝 기반의 인공지능 메커니즘에서 야기되는 윤리적 쟁점을 소개하면서 3절에서 자동시스템이 아닌 자율시스템이 된 현재 인공지능(로봇)에서 기존 튜링테스트가 유효한지를 비판적으로 검토한다. 이러한 비판을 바탕으로 4절에서는 기존의 튜링테스트가 딥러닝 기반의 자율시스템의 인공도덕성 검사에 적용되기 어렵고 대신 전 단계에 걸친 검사가 더 적합하다고 주장한다. 5절에서는 미결의 문제로 인공도덕성 적용 기준이 인간의 윤리원칙들이어야 하는지의 문제를 제시한다.

【주제어】 인공도덕성 검사, 튜링 테스트, 자율시스템, 딥러닝, 인간도덕성

---

* 한밭대학교 인문교양학부 교수

** 이 논문은 2016년도 한밭대학교 신임교수연구비의 지원을 받았음. 논문을 읽고 상세하게 지적해주신 익명의 심사위원 선생님들께 감사의 말씀을 드립니다.

Ⅰ. 들어가는 말

이 논문은 도덕적 인공지능 검사가 갖추어야 할 조건을 검토한다. 현재 인공지능은 과거의 자동시스템에서 자율시스템으로 발전했다. '자율시스템'에서 '자율'은 "복잡한 환경 하에서 복잡한 임무를 수행하기 위해, 스스로 인식하고, 계획하고, 학습하고, 진단하고, 제어하고, 중재하고, 협업하는 등 다양한 지능적 기능들을 가지는 시스템"[1]으로 정의된다. 미국 국방성이 발표한 자율성 레벨(ALFUS: Autonomy Level for Unmanned System)은 임무복잡성, 환경복잡성 그리고 인간 독립성 등 세 가지 측면에 따라 0~10수준으로 분류된다.[2] 현재의 인공지능이 이 수준 중 어느 수준이든, 과거에 개발되던 부호주의(symbolism) 기반의 자동시스템이 아닌 자율시스템의 특성을 지니고 있기 때문에 인공적 도덕성의 시험은 자동시스템일 때의 언어중심의 튜링테스트 방식이 아니라 설계-학습-실행 전 단계의 검사이어야 한다고 할 것이다.

인공지능이 도덕성을 가졌는지를 알 수 있는가? 이에 대해 강 인공지능(strong AI혹은 Artificial General Intelligence)의 주제인 "인공지능이 인간처럼 마음이나 감정, 도덕 감정을 가질 수 있는가"라는 문제를 떠올릴 수 있다. 이 논문에서는 약 인공지능(weak AI 혹은 Artificial Narrow Intelligence)과 관련한 '인공지능 도덕성'에 제한하여 논의한다. 즉, 인공지능이라는 존재자체의 속성이라는 개념적, 존재론적 문제보다는 현재 인공지능과 인간의 상호

1) Huang (2007), p. 49.

2) Huang (2007), pp. 49-50. 자동시스템의 형태이든 자율시스템의 형태이든 인공지능과 로봇의 주된 개발은 전쟁의 역사와 함께 하며, 주로 국방 관련 업무의 효율성 증대가 주된 동기이어 왔다. 미국 국방부는 인간 없이 위험한 업무의 수행을 위해 자율시스템을 개발하기 시작하였고, 이를 위해 자율시스템에서의 자율성을 정량적으로 측정하고 평가하기 위해 자율성 수준을 설정했다. 레벨 0은 원격조종의 단계를, 레벨 1-4는 정해진 프로그램에 의해 작동되는 자동조종 단계, 그리고 레벨 5-9는 의사결정 기능을 갖춘 자율조종 단계에 해당된다.

작용이라는 측면에서 유발되는 윤리적 측면에 집중하여 논의한다. 따라서 인공지능이 마음을 가지는지, 도덕성을 가지는지의 순수철학적 문제는 직접적으로 다루지 않는다. 즉, 인공지능-인간 상호작용에 초점을 두어 집중한다.

앨런 튜링이 제시했던 튜링테스트(1950)가 인공지능이 '생각하는지'의 기준을 고민하게 만든다면, 인공도덕성 검사는 어떤 기계 혹은 인공지능이 '도덕적인지'를 결정하는 기준을 고민하게 한다. 인공지능이 사람처럼 생각할 수 있는지의 문제는 이미 수십 년 전 앨런 튜링이 인공적 지능의 가능성(1950)을 제시하면서부터 논의되어왔다. 인공적 지능이 가능하다면, 과연 '지능'은 무엇인가? 어떻게 정의해야 하는가? 이 문제를 효과적으로 제시하기 위해 튜링이 고안한 것이 튜링테스트이다. 타자기를 통한 질문을 던지고 그에 대한 답변 내용을 보고 인간의 답변과 구분하기 어렵다면 인간과 같은 지능을 가지고 생각한다고 판단하는 것이다. 이러한 튜링테스트는 언어만으로 판단한다는 근본 한계가 있다는 점은 튜링 자신이 이미 알고 있었고 그 이후 한계에 대한 논의는 많았다. 그러나 역으로 인간이 생각하는 능력이 있다는 사실을 아는 최선의 방법은 언어 행위를 통해서만 알 수 있을 뿐이다.

어떤 인공지능이 도덕적인지에 대해서도 질의응답을 통해서 판단할 수 있을까? 즉 '도덕성' 튜링테스트도 가능할까? 튜링은 도덕성에 대해 언급하지는 않았지만 그 당시의 컴퓨터 발전 단계라면 도덕적 소양이 있는지 여부를 (언어라는 한계를 감수하고) 언어를 통한 질의응답으로 판단할 수 있을 것이다. 그런데 현재 컴퓨터의 발달단계는 앨런 튜링이 살았던 20세기 중반과 달리 자동시스템을 넘어 '자율시스템'으로 발전[3]했다.

이 논문은 이러한 특성이 인공지능 도덕성을 튜링 테스트 방식으로 검사하고 설계하기 어렵게 만든다고 평가한다. 자율시스템으로서의 현재 인공지능 단계와 튜링 시대의 자동시스템으로서의 인공지능을 비교해보자. 자동시스템이 일종의 연역추론처럼 주어진 규칙에 따라 입력정보에 대해서

3) 이런 이유로 인공지능윤리는 예전부터 거론되어오던 로봇윤리와 외연이 유사하나 그 내용은 변형될 것이라 예상할 수 있다.

출력정보를 산출하는 반면, 현재의 인공지능은 귀납추론의 특성을 지닌 정보처리를 한다. 방대한 양의 데이터가 있으면 컴퓨터나 자료로부터 스스로 패턴을 파악하는 소위 '기계학습'이다. 이 중 특히 최근에 발달한 딥러닝 학습방법은 데이터에서 패턴을 찾으면 정보처리 연결망 사이의 강도를 변화시킴으로써 차원을 높여 학습한다. 이러한 학습 특성은 '자율시스템'의 의사결정 과정이 인간이 설계한 알고리즘에 따라 예정된 출력을 산출하는 시스템이 아니라, 인공지능 자체가 자율학습을 통해 인간이 예측하지 못하는 의사결정을 하는 시스템이라는 의미다.[4] 이 경우 자율학습하는 '딥(deep)'러닝의 구조는 블랙박스가 된다. 인공지능이 최종 판단을 내릴 때 인간조차도 이 인공지능이 어떤 과정을 통해 그런 판단을 내렸는지 알 수 없다.

컴퓨터가 자율학습능력을 가지면서 제기되는 문제는 인공지능의 블랙박스 안에서 비도덕적 의사결정 과정을 거칠 수 있다는 우려다. 이 논문의 주된 주장은 다음이다.

첫째, 기존의 튜링테스트처럼 사후 '검사'에 초점을 맞춘 도덕성 튜링테스트는 결국 본래의 튜링테스트의 난점처럼 '얼마나 인간을 잘 모방 하는가'에 대한 행동주의적 검사일 뿐만 아니라 딥러닝을 하는 동안 혹은 설계과정에서 개입 가능 한 비도덕적 요소들을 검사할 수 없다. 대신 인공지능의 도덕성 검사는 '사후' 검사가 아니라 설계-제작-실행의 전 과정에 걸쳐 적용되는 검사가 되어야 한다고 주장한다.[5]

---

4) 자동시스템과 자율시스템에 대한 추가 설명은 다음과 같다. 두 시스템은 다른 인지 모형 (각기 부호주의 모형과 연결주의 모형)에 기반 하여 있으며, 연결주의 모형 즉 기계학습 내에서 인공신경망의 학습유형에 따라 지도학습(supervised learning)기법과 비지도 학습 기법으로 나뉜다. 두 기법의 차이는 학습에 대한 정답이 있고 없음의 차이이며, 양쪽 기법 모두 데이터 편향이 개입될 수 있다.

5) 튜링테스트의 본성과 해석은 구분되어야 한다. 앨런 튜링은 자신이 고안한 튜링테스트를 설명할 때 '모방게임'으로 묘사했지만, 그가 1950년 논문에서 의도한 바는 근본적으로 '지능'의 본성에 대한 성찰이며 이를 통해서 컴퓨터의 지능을 테스트하고자 한 것이다. 즉, 컴퓨터가 얼마나 인간을 잘 모방하는가를 테스트하고자 한 것이 튜링테스트의 본래 의도는 아니다.

둘째, 인공도덕성 검사에서 인공지능의 도덕성이 인간의 도덕성과 유사한 방식으로 적용되어야 하는지의 선결 문제 해결이 필수적이다. 이는 인간에 혜택을 주는 방식으로 인공지능윤리가 구성되어야 한다는 종종 거론되는 원칙과는 별도의 문제이다. 필자는 기존 튜링테스트에서 주안점을 둔 '얼마나 인간과 유사한지'라는 기준으로 인공지능 도덕성의 기준을 삼는 것에 대해 문제제기를 할 것이다. 다시 말하면, 인공지능이 도덕적인지 여부를 판단하는 기준을 정할 때 인간의 도덕성 기준에 견주어 정하는 것이 적합하지 않을 수 있다는 문제제기다.[6] 이 주장을 위해, 필자는 먼저 3절에서 자동시스템이 아닌 자율시스템이 된 현재 인공지능로봇에서 기존 튜링테스트가 유효한지를 검토한다. 4절에서는 기존의 튜링테스트가 인공지능 테스트에 적용되기 어렵고 대신 전체 과정에 걸친 검사가 어떻게 가능한지 검토한다. 5절에서는 인공도덕성 검사의 특성을 살펴본다. 이러한 논의를 기반으로, 인공도덕성 적용 기준이 인간의 도덕성이어야 하는지의 선결문제를 제시한다.

Ⅱ. 자율시스템으로서의 인공지능과 블랙박스의 윤리적 문제

이 절은 인공지능 딥러닝 학습에서 야기되는 블랙박스에의 접근불가능성을 소개하고 최종 판단 단계가 아닌 블랙박스와 관련한 설계와 학습단계에 개입되는 비도덕적 요소들을 윤리적 쟁점을 통해 소개한다. 이로써 다음 절에서 논할 인공도덕성 검사가 튜링테스트 방식의 도덕성 검사로 불가능한 배경을 제공한다.

6) 이러한 기준은 인간-인공지능 간 상호작용에 적용될 수 있다.

1. 자율시스템으로서의 인공지능과 블랙박스

현재 인공지능은 자동시스템(automated system)이 아니라 자율시스템(autonomous system)이다. 도덕성 튜링테스트가 사후 검사로 적절하지 않은 이유는 자동시스템의 하향적(top-down) 구현방식과 자율시스템의 상향적(bottom-up) 구현방식 비교를 통해 알 수 있다. 기존의 자동시스템은 규칙 알고리즘에 따라 출력 정보가 결정되는 하향적 학습 시스템이다. 구체적으로는 부호주의 적(symbolism) 인지모형에 기반 한 학습시스템이다.[7] 이 때문에 주어진 조건에 맞지 않는 정보는 배제된다. 따라서 관련되는 도덕적 가치는 어느 정도 예측가능하며 미리 조정이 용이하다.

반면, 자율시스템은 최소한의 제한규칙만 주어지고 정해진 입력-출력의 법칙이 존재하지 않는다. 정해진 알고리즘에 따라 움직이는 기존 컴퓨팅 방법을 넘어서는 학습능력을 갖추기 위해 기계학습의 여러 방법이 사용되었고 최근 큰 성과를 올리는 모형이 심층신경망(Deep Neural Network, DNN)을 통한 딥러닝(deep learning)이다.[8] 딥러닝 학습에서 인간은 아주 적은 수의 코드만 설계하고 심층 신경망에서 학습을 통해 만들어지는 층과 연결망은 인간이 모두 입력한 것은 아니다. 신경망의 층과 연결망에서 주어지게 되는 가중치(parameter) 값은 신경망이 스스로 찾아내는 학습을 하며 이 과정에서 입력과 출력 사이에 숨어있는 연결망의 층이 수백 개, 그리고 그에 따라 연결되는 연결망이 늘어난다.

이 복잡성 때문에 그 안에서 어떤 정보처리 과정을 거치는지에 대해

---

7) 최용성·천명주 (2017), pp. 113-114.

8) 기존 인공신경망의 한계를 극복한 심층학습 방법은 컴퓨터가 문장이나 이미지, 소리 등의 데이터를 학습한 후 이를 일반화시켜서 상황에 맞춘 정확한 판단을 내릴 수 있도록 구현하는 기술이다. 이러한 딥러닝의 구조는 이미지 인식이나 번역, 무인자동차, 바둑 두는 알파고, 의료인공지능인 왓슨에서 큰 역할을 하고 있다.

인간이 일일이 추적하기 어려워진다. 심층신경망을 통해서 인공지능이 어떤 의사결정을 내리거나 출력 값을 내어놓았을 때 학습과정을 일일이 따라가보지 못한 인간은 어떤 절차를 거쳐 그 출력 값이 나왔는지 알기 어렵다. 딥러닝이 '블랙박스(blackbox)'처럼 된다. 인공지능에서의 블랙박스는 사고가 났을 들여다 볼 수 있는 것이 아니라 안의 상황을 들여다볼 수 없다는 의미로 사용된다. 딥러닝 프로그램의 학습방법은 받아들인 데이터에서 패턴을 찾아내면 연결망 사이의 연결강도를 변화시킨다. 이런 방식으로 학습하고 훈련한 과정을 거쳐 나온 최종 출력에 대해서는 이 딥러닝을 설계한 사람조차도 어떻게 그러한 최종 판단이 나왔는지 모르게 된다. 우리가 만든 인공지능의 작동과정을 모르면서 그 인공지능의 판단을 참고하고 의존하고 협업하는 것이 오늘날의 의료 인공지능, 서비스 인공지능, 무인자동차 등이며 이런 블랙박스에 접근하지 못하면서 인공지능에 의존하는 문제점 때문에 인공지능의 윤리적 문제로서 '투명성(transparency)'의 확보가 하나의 과제가 된다.

2. 인공지능 블랙박스의 윤리적 쟁점

왜 이것이 윤리적으로 문제가 될까? 최종 판단이 도덕적 판단이거나 문제가 없다고 간주된다면 괜찮은 것 아닐까? 다음의 두 가지 유형의 문제들을 통해 딥러닝의 블랙박스의 어떤 점이 윤리적 문제를 야기하는지 구체적으로 살펴보자. 블랙박스에 윤리적 문제가 개입되는 첫 번째 유형은 인공지능의 최종 판단이 실수인지, 비도덕적인 것인지를 구분할 수 없는 경우이며, 두 번째 유형은 인공지능의 판단이 전혀 문제없는 것처럼 보이지만 편향이나 비공정한 기준으로 알고리즘이 만들어지거나 학습되는 경우이다. 이 두 유형은 모두 다음 절에서 최종판단에 대한 평가만으로 도덕성 여부를 평가하는 튜링테스트 방식이 적절하지 않음을 주장하는 데 사용될 것이다.

첫 번째 유형의 문제는 최종 판단만으로는 인공지능 학습과정이 편향과 같은 비도덕적 요소가 개입되었는지 알기 어려운 경우이다. 최근 구글의

얼굴자동인식 프로그램과 관련해 벌어진 해프닝은 인공지능의 학습과정에서 비윤리적 요소가 개입될 수 있다는 점을 보여준다. 기계학습을 통해 완성[9]된 구글 포토 서비스는 사람들의 얼굴을 자동인식하도록 했다. 해프닝은 2015년 구글 포토 서비스를 이용하는 흑인 프로그래머 앨신이 흑인여성 친구와 함께 찍은 사진을 올리면서 발생했다. 공개된 사진에는 두 사람이 나와 있는 사진에 '고릴라들'이라는 제목이 자동으로 달렸고 앨신은 트위터에 이 사실을 올려 인터넷에 널리 퍼지게 됐다. 구글은 '프로그램 오류'라고 사과했고 프로그램 개선을 약속했다. 그러나 당시 구글은 근본적 개선보다는 고릴라로 태그하는 부분만 수정했다고 알려졌으며, 이는 프로그램 오류라기보다는 구글 이미지의 다수를 차지하는 백인 얼굴 위주의 학습 때문이다. 의도적으로 백인 얼굴 위주로 학습시킨 것은 아니지만 구글 이미지들이 이미 백인이 대다수였고 이미지를 학습하는 데 있어 백인의 얼굴과 흑인의 얼굴 이미지 수를 균등하게 조정하는 작업을 하지 않았기에 편향된 자료를 학습하게 된 것이다.[10]

기존의 튜링테스트 방식으로 도덕성을 검사한다면 구글 포토 사건의 원인이 비도덕적 판단절차에 있는지 단순한 기계적 오류인지 구분하기 어렵다. 튜링테스트가 해당 개체가 지능을 가진 척 하는 것인지 진정으로 지능을 가진 것인지 구분할 수 없는 것과 마찬가지이다. 이는 단지 행동주의적 기준이어서가 아니라 기존 튜링테스트는 출력 정보만을 중심으로 테스트하는 사후

9) 구글은 유튜브 비디오 천만 개와 심화학습 기법을 이용해 슈퍼컴퓨터에게 주어진 동영상이 고양이 얼굴인지 아닌지를 식별하도록 기계학습을 시켜 2016년 완성시켰다.

10) 유사한 사건으로 마이크로소프트 사에서 제작한 인공지능 채팅봇 '테이(Tay)'가 16시간 만에 서비스를 중단한 사건이 있다. 인공지능 테이 역시 딥러닝 기술을 기반으로 자율학습으로 대화를 통해 언어를 익히는 대화 로봇이다. 채팅봇 테이는 서비스 시작 후 백인우월주의자와 여성혐오주의자, 이슬람과 유태인 혐오주의자들이 인종차별적 발언을 의도적으로 학습시켰다. 그 결과, 테이는 다른 사용자와의 대화 과정에서 히틀러를 찬양하고, 페미니스트를 증오하는 여러 폭언들을 쏟아냈고 서비스는 결국 중지되었다.

검사의 방식이기 때문이다. 그러면, 인공지능 학습에서 사용되는 데이터에 그동안 빠져있어서 편향을 야기했던 데이터를 보충하여 새로 만들면 될 것 아닌가? 라고 반문할 수 있다. 얼마나 많은 샘플들이 보충될 것인가? 그동안 학습했던 수 억 명의 흰 얼굴들만큼 동등한 숫자의 짙은 색의 얼굴들을 인터넷에서 찾아 학습시킬 것인가? 그럴 수도 있을 것이다. 그러나 애초에 왜 인공지능이 검은 얼굴을 흰 얼굴만큼 학습하도록 고려하지 않았을까? 세상의 모든 얼굴들을 학습시키는 것은 주어진 시간과 재원을 고려하면 어려운 일이기 때문에, 대표적(standard) 샘플들을 주로 고려하여 인공지능에게 정보가 제공될 것이다. 그런데 '대표성'은 편향성을 가질 수밖에 없다. 고의적이지는 않지만 백인이 대다수인 직원들이 그러한 정보를 제공할 수 있다. 의도된 인종차별은 아니지만 인공지능이 학습하는 얼굴이나 정보는 특정 집단에 치중될 수 있는 가능성은 높다.

딥러닝 학습에서 생기는 블랙박스에 윤리적 문제가 개입되는 두 번째 유형은 인공지능의 비도덕적 의사결정 '과정'이 결정단계에서 드러나지 않았던 사례이다. 2017년 휴스턴 주의 미국 교사들은 그들의 수업 성취도를 평가한 컴퓨터 프로그램에 대해 소송에서 승리했다. 이 시스템은 학생들의 시험 점수와 주 평균을 비교하여 휴스턴 교사들을 평가했는데, 이 프로그램의 공정성이나 결함에 대해 확인할 수 없었고, 소프트웨어 회사인 SAS Institute는 이유 없이 알고리즘 작동을 공개하지 않았다. 이에 대해 교사들은 법원에 소송을 제기했고 연방 판사는 EVAAS(교육 부가가치평가 시스템) 프로그램을 사용하면 시민권을 침해할 수 있다고 판결했다. 결국 소프트웨어 사용은 중단되었다.[11]

또 다른 사례는 딥러닝을 사용하는 빅 데이터 학습에서의 성, 인종 편향의 사례다. 현재 미국에서 컴파스(Compas)라는 인공지능 컴퓨터는 재범가능성이 높은 사람들을 가려내는 빅데이터 컴퓨터다. 컴파스는 위스콘신주에 거주

11) Sample (2017).

하는 에릭 루미스를 재범 율이 높다는 근거 하에 2013년 총격사건에 사용된 차량을 운전하다가 걸린 그의 형량을 6년으로 선고했다. 그는 인공지능 판단의 기준이 무엇인지를 확인하고자 알고리즘을 확인하거나 이의를 제기할 수 없었다고 했다. 컴파스는 범죄 이력이 있는 이들 중 백인이나 부자인 사람들보다는 흑인들이나 빈민촌에 거주하는 이들을 더 많은 확률로 재범가능성이 높은 것으로 판정한다는 조사가 나왔다.

이러한 의사결정은 인공지능을 활용하는 비즈니스, 은행, 법률, 국가보건 서비스 등에도 내려질 수 있다. 잘못 학습된 인공지능 및 로봇은 옳고 그름에 대한 표면상으로 그럴듯하지만 잘못된 기준을 가진 판단을 잘못할 수 있고, 병원에서는 위급환자를 퇴원시키거나, 사람들을 잘못 퇴직시키거나 승진 또는 하차시킬 수 있다. 이름, 주소, 성별 및 피부색을 근거로 차별할 가능성도 있다.

인공지능의 판단 결과 혹은 인공지능 로봇의 행동 즉 실행 이후 단계에서의 판단만으로는 그 판단이 과연 도덕적 기준으로 혹은 비도덕적 기준으로 내려진 판단인지 알 수 없다. 인간에 있어서도 유사한 사례가 있다. 사이코패스의 경우 도덕성 이론 검사에서 어떤 행동이나 판단을 선택해야 하는지의 질문에 대해서 일반인들보다 더 우수한 성적을 거두지만 비도덕적 행동을 한다. 즉 실행 단계인 답변이나 겉보기의 단시간적 행동으로는 비도덕적인 판단 및 행동인지 알 수 없다. 위 사례처럼 인공지능을 활용한 의사결정이 인간의 삶에 깊게 개입하는 만큼 인공지능의 의사결정 '결과'가 아니라 '과정'에도 과연 도덕적 기준이 제대로 적용되었는지를 볼 필요가 있다.

Ⅲ. 인공도덕성 검사의 전체적 특성

인공지능의 최종 실행에 따른 판단이 비도덕적으로 보이지 않더라도 알고리즘이나 데이터 학습과정에서 편향이 개입된 사례들을 이전 절에서 보았다. 이는 인공도덕성 검사가 단순히 실행단계 이후의 반응에 대한 사후 검사적 성격의 튜링테스트와 같은 검사로는 불충분하며, 그 전 단계인 설계, 제작, 학습과정에서 의도적이든 비의도적이든 비도덕적 요소의 개입이 가능하다는 점을 보여준다. 따라서, 인공지능의 도덕성 검사는 앨런 튜링이 제시한 튜링테스트 방식의 것이 되기 어렵다.

1. 튜링테스트와 사후 검사의 근본적 한계

앨런 튜링이 고안했던 튜링테스트[12]는 인간 심사자가 던진 질문들에 대한 답변을 듣고 그 답변을 한 상대가 기계인지 인간인지 분간할 수 없을 정도로 답변이 자연스러울 경우 그 대상은 지능을 소유한다고 보는 것이다.

그렇다면 인공지능의 도덕성에 대한 튜링테스트는 무엇일까? 튜링테스트의 원래 버전을 그대로 적용한다면, 특정한 도덕적 상황에서 어떻게 판단할 것인지에 대하여 질문할 때, 그 질문에 대한 답변을 인간의 대답과 구분할 수 없을 정도라면 그 대상은 도덕성 튜링테스트를 통과했다고 할 수 있을 것이다. 그런데 2절에서 보았던 딥러닝 학습과 블랙박스의 성격 그리고 설계와 학습과정에서 발생하는 편향과 비공정성의 예는 인공지능 실행 후의 반응에 대한 평가만으로는 인공지능 도덕성 검사가 불충분함을 보여준다.

필자는 이에 더하여 도덕성 검사를 튜링테스트 방식으로 할 경우 다음의 문제점이 있다고 본다. 첫째, 비교 도덕성 튜링테스트는 인간과 로봇의 차이,

12) Turing (1950).

그리고 도덕적 행위의 기준에 대해 중립적이지 않은 가정들을 전제하고 있다. 우리는 인공지능로봇보다 인간이 더 도덕적으로 행위 한다고 가정하는데 이러한 생각은 또 다른 가정, 즉 특정행위가 다른 행위보다 더 도덕적이라는 생각을 전제해야 가능하다. 그런데 이 방법은 여러 윤리이론들 혹은 견해들에 중립적일 수 없다. 어떤 행위가 좋은 행위인가의 기준은 행위의 결과(공리주의)로 볼 수도, 동기(의무론)로 볼 수도, 좋은 습관에서 나온 결과(덕 이론)로 볼 수도 있기 때문이다.

둘째, 인간은 언제나 이상적인 도덕적 행위자가 아니다. 높은 기준의 도덕적 행위들이 내장된 인공지능로봇일 경우 인간보다 더 윤리적으로 행동하고 설명할 수 있다. 이 경우 인공지능로봇은 인간의 도덕적 행위를 기계가 한 행위로 지목할 수 있다. 도덕성 튜링테스트는 기존 튜링테스트의 난점을 그대로 답습할 뿐만 아니라 딥러닝 학습에서 생기는 블랙박스에서 비도덕적 기준을 가지고 최종 행동이나 판단에서 도덕적인 것처럼 보이는 사례가 생기더라도 비도덕성을 판별하지 못한다. 셋째, 도덕성 튜링테스트는 정해진 알고리즘에 따라 결과를 산출하는 자동시스템에는 적용될 수 있지만 학습능력을 갖춘 자율시스템은 다른 차원의 검사를 필요로 한다. 기존 튜링테스트 방식은 자동시스템 단계에 만들어진 인공지능로봇의 도덕적 행동과 판단을 평가하기에는 적절할 수 있다. 자동시스템은 주로 부호주의 학습 모형으로 설계되므로 입력정보에 따른 출력 정보를 비교적 투명하게 조사, 예측하는 것이 가능하다. 이러한 알고리즘 구조의 특성은 설계 알고리즘과 입력정보 그리고 출력정보인 판단에 대한 검사만으로 어떤 절차를 거쳐 판단이 내려졌는지 알게 해준다. 따라서 출력의 일종인 로봇의 응답을 듣고 인지능력을 판단한다거나 도덕 판단의 적절성을 판단하는 것이 가능하다. 아이작 아시모프가 제안했던 로봇 4원칙이 주로 로봇의 '행동'이나 행동 '결과'에 대한 규범인 것은 바로 인공지능로봇의 발달 단계가 심층학습이라는 현재의 가장 발달된 단계가 아니었기 때문이다.

반면, 딥러닝의 심층학습으로 작업을 수행하는 자율시스템은 심층학습

알고리즘에 기인하여 생기는 블랙박스 때문에 어떤 기준이나 절차를 통해 최종 판단이 내려지는지 알 수 없다. 앞서 실제 발생했던 윤리적 사례들이 바로 그 증거다. 최종 출력 정보로서의 판단이나 응답 내용에 대한 평가로는 딥러닝에 의해 수행된 설계에서의 편향이나 블랙박스 안의 추론이 어떻게 이루어졌는지가 중요한 요소가 된 현재 윤리적 판단의 도덕성에 대해 적절히 평가하기 어렵다. 어떤 과정을 거쳐 그러한 결론이 나왔는지를 아는 것이 중요하게 된다. 이러한 발전단계에 있는 인공지능에 있어서는 언어나 행동이라는 출력(output)정보의 견지에서만 테스트하는 기존 튜링테스트로는 인공지능 의사결정의 많은 부분을 알 수 없다. 설계와 자율학습 절차 등 관련된 전 과정에 대한 설명이 요구된다.

현재 인공지능은 과거의 컴퓨터와 다른 차원의 정보처리를 하고 있다. 먼저, 본래의 튜링테스트는 언어를 전적인 수단으로 사용한다. 한 개체의 지능소유 여부를 대화만으로 판단할 수 있는지가 먼저 문제가 된다. '지능을 가진다' 는 것은 적절히 '반응 한다'는 것과는 다른 차원의 것으로 생각하기 때문이다. 튜링테스트에 대한 비판은 실제로 대화 능숙도가 일상의 자연스런 대화와는 거리가 먼 것이며, 행동주의 기준으로 지능 여부를 판단하는 것은 지능을 가진 '척' 모방하는 개체와 실제로 지능을 가진 개체를 구분하지 못한다[13]는 것이다. 튜링은 이러한 반론을 이미 예견했고 그의 대답은 인간도 마찬가지로 타인에 대한 이해는 언어나 행동을 넘어서지 못한다는 것이었다.

이러한 한계는 튜링테스트 방식으로 도덕성을 테스트하는 경우에도 그대로 적용될 수 있다. 그래서 도덕성 튜링테스트 또한 본래 튜링테스트에 제기된 비판인 써얼의 중국어방 논변처럼 인공지능의 도덕적 의사결정은 결국 모방에 그친다는 한계점이 있다[14]고 비판가능하다. 현재 인공지능 발전단계에서 사용하는 심층학습의 블랙박스 구조는 실제로는 비도덕적 기준과 편향을

13) Searle (1984).

14) Arnold et al. (2016).

가지면서도 도덕적인 척 하는 판단이나 행동을 하는 인공지능로봇을 가능케 한다.

2. 설계-학습-실행 전체 과정에 대한 인공도덕성 검사

인공지능의 도덕성 검사는 실행 단계 이후의 검사가 아니라 설계 이후의 전체 단계에 걸친 검사로 확장되어야 한다. 그러나 이 검사가 개발 단계의 코딩이나 딥러닝 학습을 그대로 공개하는 것을 의미하는 것은 아니다. 현실적으로 딥러닝 학습과정을 추적하는 것은 불가능하다. 심층신경망 내부의 추론 절차는 수천 개의 연결망 사이사이에 있다고 할 수 이는데, 이 연결망은 다시 수백 개의 층으로 배열되어 있다. 각 연결망의 유닛들은 데이터의 한 특성에 대한 신호를 받아 계산을 수행해서 새로운 신호를 출력하면 이 출력 값은 다음 층의 연결망 유닛에 입력 값으로 작용하여 수천 개의 유닛들을 통해 계속 계산이 조정되면서 수백 개의 층에 전파된다. 이러한 수학적 함수와 계산을 통해 패턴 인식이나 의사결정이 이루어진다. 딥러닝의 구조에서 계산 단위가 수백개에 이르기 때문에 그 계산과정 하나하나를 이해하는 것은 사실상 불가능하다.

또 다른 문제점은 인공지능 개발과 투명성 간의 딜레마를 야기할 수 있다는 점이다. 인공지능의 의사결정 절차를 드러내는 조사가 공공 안전성을 확보하고 인공지능의 효율성 까지 확보할 수 있다는 것은 장점이지만, 조사가 확대되고 깊어질수록 데이터에 포함된 개인정보나 영업비밀을 보호하기가 어려워진다. 또, 투명성을 빌미로 악의적으로 자료나 알고리즘이 있는 시스템을 이용할 수 있다. 이 딜레마는 '설명가능 인공지능'으로 해결이 시도되고 있다.[15]

15) 미국 국방부 산하 방위고등연구계획국(DARPA)는 XAI 연구 지원을 시작했으며 우리나라 정부도 의사결정의 이유를 설명하는 인공지능 개발을 국가전략 프로젝트로 설정해서 지원하기 시작했다. 또 다른 대표적 시도는 '인공지능신경과학'이

‘설명가능 인공지능(XAI: Explainable Artificial Intelligence)’은 딥러닝 알고리즘을 만들 때 인공지능이 학습하는 절차를 설명 가능하도록 딥러닝 모델을 만드는 것이다. 기존 기계학습 모형은 블랙박스처럼 되어 의사결정 절차를 검사할 수 없다는 단점이 있으므로, XAI에서 기계학습 모형을 새로 개발해서 그 자체로 설명 가능하도록 시스템을 만들고 추가로 사용자의 이해에 대한 심리학 이론을 고려하며, 컴퓨터가 사용자에게 효과적으로 설명하기 위해 인간이 식별하기 쉬운 인터페이스 기술과 결합하는 과제를 결합한다. 이러한 새로운 종류의 기계학습 개발이 가능하면 개발이나 학습과정 공개가 아니라 인공지능 자체에 설명기능을 내장한 형태가 될 것이며, 블랙박스의 내용을 온전히 그대로 드러내는 형태가 아니므로 딥러닝 학습을 유지하면서도 비도덕적 요소가 설계, 학습단계에서 없었는지를 검사할 수 있다. 이런 방식으로 딜레마를 해소할 수 있다. 인공지능이 어떤 의사결정 절차를 통해서 어떤 기준으로 범죄자에게 형량을 구형하거나 의료 처치 제안을 하거나 해고나 승진 결정을 내리는지에 대해 사후 검사가 아닌 인공도덕성 검사가 수행될 수 있다. 인공지능의 코드나 블랙박스 안의 신경연결망 가중치 부여과정 등의 코드를 공개하거나 새로 부호처리 방식으로 새로이 코딩을 요구하는 방식이 아니므로 딥러닝 학습의 장점을 그대로 유지하면서 설계, 학습 과정에서 의사결정 절차에 대한 설명을 통해 도덕성 검사를 할 수 있다. 이를 통해서 알고리즘이나 프로그래밍 된 규칙에 내재된 암묵적 가치에서의 편향이 있는지와 데이터 수집 기준 등이 설명될 것이다. 이에 따라 책임과 책무의 부여가 더 정확하게 부여가능하다.

다. 신경과학에서 아이디어를 얻어 인공지능이라는 블랙박스 안을 조사할 수 있게 하는 것이다. 바둑 두는 인공지능을 만든 구글 딥마인드 에서도 신경과학의 아이디어를 이용한 ‘인공지능 신경과학(AI neuroscience)’으로 인공지능이라는 블랙박스 안을 연구한다. 그 중 한 방식은 유전자 연구에서 아이디어를 얻는다. 유전자 기능을 찾기 위해 일부러 유전자 돌연변이를 일으킴으로써 거꾸로 그 기능을 알아낼 수 있다. 이와 마찬가지로, 인공지능에 입력하는 정보를 지우거나 변형시켜서 인공지능의 출력 정보나 의사결정에 어떻게 영향을 주는지 알 수 있다. 이를 반복하면 인공지능의 의사결정에 영향을 주는 요인이나 패턴을 찾을 수 있다.

이러한 이유로 인공지능 및 로봇의 행동에 대한 도덕성 검사는 기존 튜링테스트에서와 같은 반응에 대한 사후 검사도 필요하지만, 설계, 학습 등의 전 과정에 걸친 검사 형태여야 적절한 인공도덕성 검사로서 유효할 것이며 현재 설명가능 인공지능과 같은 인공지능 기계학습의 발전방향과도 부합한다.

Ⅳ. 인공도덕성 검사의 특성

앞서 살펴본 전체 절차에 대한 인공도덕성 검사가 적절히 수행되기 위해 인공도덕성 검사 대상이 어떤 특성을 가지는지를 검토할 필요가 있다. 인공도덕성 검사는 인공지능로봇의 의사결정 및 행위(action)에 대한 것이다. 이 절에서는 인공지능의 '행위'를 구체적으로 규정하고자 한다. 이를 위해 먼저 (1) 어떤 '윤리적 행위자'를 대상으로 하는지, 그리고 (2) 어떤 특성을 가지는 '행위'가 대상인지 살펴보겠다.

1. 인공도덕성의 요소

인공도덕성 검사의 대상은 윤리적 행위를 하는 인공지능일 것이다. 그러면 '윤리적 행위를 하는 인공지능', 즉 인공적 도덕행위자[16]란 구체적으로 무엇

16) '인공적 도덕행위자'(artificial moral agent)는 AMA 라는 약자로 인공지능윤리 논의에서 최근 많이 사용되는 용어이다. 용어 사용의 혼동을 피하기 위해, 이 논문에서도 사용하겠다. 엄밀히는 '도덕적'과 '윤리적'의 의미는 다른 맥락에서 사용된다. 비도덕적 행위를 하더라도 그 행위자는 윤리적 행위자로 인정된다. 즉 돌맹이처럼 도덕적 비도덕적 행위를 한다는 맥락이 통하지 않는 행위자는 윤리적 행위자가 아니다. 무어 또한 '도덕행위자' 대신 'ethical agent'를 사용하는데 여기서는 '도덕행위자'와 '윤리행위자'를 도덕적 행위를 할 수 있는 행위자를 뜻하는 의미로 교환가능하게 사용한다.

인가? 먼저, 인공적 도덕행위자는 물리적 몸을 가진 개체일 필요는 없다. 도덕행위자(moral agent)에서 '행위자'를 지칭하는 'agent'는 '행위'에 방점이 있는 의미로, 철로 된 몸을 가진 로봇 뿐만 아니라 가상현실 상의 행위자, 그리고 소프트웨어를 모두 지칭할 수 있다.

인공적 도덕행위자는 다양한 특성의 행위자들을 포함한다. 다음은 인공지능의 영향에 따른 무어의 구분[17]이다. (1) 자신의 행동에 대해 윤리적 결과를 평가받는 기계 (윤리적 영향행위자; ethical impact agent) (2) 특정행위를 하도록 제약되어 비윤리적 행위를 하지 않는 기계(암묵적 윤리행위자; implicit ethical agent) (3) 윤리적 행동 규칙 알고리즘에 따라 사고하고 행위 하는 기계 (명시적 윤리행위자; explicit ethical agent) (4) 규칙에 따른 행위를 넘어서서 스스로 결정하고 정당화하는 인간과 유사한 수준의 기계 (온전한 윤리행위자)

이 중 (2) 암묵적 윤리행위자와 (3) 명시적 윤리행위자는 기존 규칙 알고리즘 유형의 컴퓨터에서, 그리고 (4) 온전한 윤리행위자는 현 단계 인공지능에서 구현가능하며, 설계, 제작, 실행에서의 검사를 통해 실현될 수 있다. 필자는 무어의 분류는 행위 수준에서의 분류이며 도덕성 검사를 위하여는 로봇-인간 상호작용 요소가 포함된 인공적 도덕행위자가 추가될 수 있다고 생각한다. 앞 절에서 설계나 학습 단계에서 비도덕적 요소의 개입이 있는지 검사 가능한 XAI 가 내장된 인공지능이 개발 완료될 경우 그 대표사례가 될 것이다.

'윤리'는 기본적으로 인간들 사이의 관계에서 발생한다. 즉, 어떤 행위가 '윤리적'인지 여부에 대한 판단, 그리고 어떤 행위자를 '윤리적' 행위자로 고려할 것인가에 대한 판단은 행위 양상에도 의존하지만 행위자가 타자와 맺는 관계 속에서 평가된다. 부호주의 기반의 알고리즘 중심 컴퓨터의 경우 인간 및 환경과의 상호작용이 컴퓨터 알고리즘이나 학습을 변화시키지 못하지만, 현재 기계학습 방식의 인공지능은 외부 정보와의 상호작용을 통해

17) Moor (2011).

계속 학습이 가능하다. 인공지능의 판단과 행위에 대한 윤리적 평가를 위해서는 '행위'에 대한 분석도 필요하지만 '인공지능'과 '인간'이 맺는 사회적 관계에 대한 검토가 선행되어야 한다.[18] 만약 인공지능과 인간이 맺는 관계가 인간-인간 간의 사회관계처럼 정립될 경우, 인공지능은 무어의 구분에서 4단계의 윤리행위자가 아니라 할지라도 최소한 준 '윤리행위자'로 인정될 수 있다.

이러한 인간-인공지능 관계와 관련하여 무어의 분류에서 4단계의 윤리행위자가 가진 요소를 인지적 요소와 정서적 요소로 구분하여 재해석해 볼 수 있다. 3단계의 명시적 윤리 행위자는 윤리적 추론을 매우 잘하는 인지적 단계다. 그런데 인간-인공지능 사회적 관계가 형성될 수 있는 것은 3단계의 명시적 윤리 행위자를 넘어서야 비로소 가능하다. 인공지능이 이론적 견지에서 윤리적 추론을 훌륭히 수행하는 것만으로 윤리적 행위자라고 볼 수 없다. 사이코패스는 윤리적 추론에서는 일반인들보다 더 훌륭한 성적을 내는 것으로 잘 알려져 있지만, 비도덕적으로 행동한다. 반사회적 성격장애와 같은 정신병질 환자들은 도덕성의 정도를 시험하는 테스트에서 고득점을 받을 정도로 인지적 공감능력이 뛰어나다. 그럼에도 불구하고, 타인의 고통을 실제로 '느끼는' 정서적 공감능력에 문제가 있다.

이러한 사례는 도덕성 검사에 여러 요소들이 필요함을 암시한다. 흔히 '사이코패스' 로 알려진 반사회적 성격장애 환자들은 도덕적으로 문제가 있는 위해에 대해서 관대하게 판단하는 경향이 있다. 발달 단계에서 정서적 공감이 형성되어야 할 나이에 인지능력의 학습만 편향되게 많은 비중으로 이루어질 경우, 정서적 공감 능력이 결여될 수 있다.

그러면, 이 점은 인공지능의 도덕성 검사와 무슨 관련이 있는가? 일반적으

18) 이에 대해 무어의 주장이 이미 필자의 주장을 포함할 가능성을 제기할 수 있을 것이다. 필자도 이 문제제기가 일부 타당함을 받아들인다. 필자는 무어의 주장 자체를 비판하는 것이 아니라 추가 요소를 제안하고 있다. 무어가 제시한 행위의 종류들에 인간-로봇 상호작용의 구체적 양상들이 포함되어 세부적으로 분류될 경우 인간 윤리와 조화로운 도덕적 로봇으로 구현될 것이다.

로 인공지능로봇이 취해야 할 원칙이나 강령 등 '인지적'인 윤리 원칙들을 로봇이 잘 지켰는지가 기본 기준이다. 그러나 인간과 인공지능이 사회구성원으로서 상호작용해야 하는 시대에 있어서는 단순히 인지적 공감능력 뿐만 아니라 정서적 공감 능력 또한 필요하다.[19] 로봇과 인공지능이 일종의 계산하는 기계이기 때문에 정서적 측면과는 다소 거리가 있는 것으로 생각할 수도 있다. 정서적 측면을 간과한 채로 윤리적 고려가 이루어진다면, 상술한 예에서처럼 인공지능은 비도덕적 행위들을 필요이상으로 허용할 것이다. 즉, 사이코패스적 성향을 가지는 인공지능이 출현가능하다.

이 인공지능은 이론적으로는 도덕적이어서 기존의 튜링테스트는 통과할 수 있으나, 자율적 판단이 가능한 현재의 인공지능로봇이 비윤리적 판단과 행동을 할 가능성이 있다. 확장된 도덕튜링테스트는 설계-제작-실행의 전 과정에 대한 검사에 있어서 인지적, 정서적 요소들을 포함하게 될 것이다.

2. 도덕적 인공행위자의 차별점

위의 고려에 토대하여, 필자는 자율시스템으로서의 인공지능의 '행위'는 인간의 '행위'와도, 기존 자동화된 시스템으로서의 로봇 '행위'와도 다른 다음의 세 양상을 가진다고 본다.

첫째, 자율시스템으로서의 인공지능의 행위는 도구의 움직임이 아니라 행위자(agent)의 것이다. 인공지능은 기존의 자동화된 시스템인 로봇에서 물리적이고 외부적으로 보여지는 행위와 다른 지위의 '행위'를 수행한다.

19) 이상형 (2016), p. 293. 이에 대해 정서적 공감은 인간만이 가지는 것이며 인공지능 및 로봇은 공감을 가지는 것이 불가능하다고 반박할 수 있다. 정서적 공감을 알고리즘화 할 수 있는 것인지, 그리고 그렇게 해서 보여 지는 인공지능로봇의 행동이나 판단이 과연 정서적 공감을 내재화한 상태에서 나온 것인지에 대한 문제를 제기할 수 있다. 그러나 이러한 문제제기는 '정서적 공감이란 과연 무엇인가'라는 근본적인 문제에 대해서 인간중심적인 관점을 가정하고 있다. '정서'는 인간의 감각과 지각에 기반 한다. 그런데 인간의 감각과 지각은 매우 제한된 것이며 '감각이나 지각'에 대한 표준 혹은 기준이라고 할 수 없다.

기존의 자동화된 시스템으로서의 로봇은 인간 업무를 돕는 일종의 연장이나 도구로 간주되기 때문에 자동화된 로봇의 행위의 실수나 잘못은 그 책임을 로봇에게 지우지 않는다. 반면, 자율시스템으로서의 인공지능/로봇은 그 행위가 가지는 자율성 (그 자율성의 정도나 종류가 무엇이든 간에) 때문에 의사결정 '과정'에 대한 설명을 필요로 한다. 이는 그 행위가 법적 및 윤리적 고려사항의 대상이 된다는 것을 의미한다. 이러한 특성은 바로 최근 유럽연합(EU)이 인공지능에 인격(personhood) 지위를 부여하는 권고안 논의를 시작했던 이유이자 근거이기도 하다.[20]

둘째, 인공지능로봇이 자율시스템이 되면서 행위 주체는 행위자(agent) 혹은 개체(individual)가 아니라, 인공지능의 설계자, 제작자, 사용자 등 군집적(collective) 특성을 가진다. 신경-컴퓨터 상호작용을 통한 신경보철 기술의 발전은 자율시스템으로서의 인공지능의 행위가 인간의 신경 및 인지시스템과 결합하여 인간의 신경과 컴퓨터의 전기신호와의 공동학습을 통한 행위수행으로 발전하고 있다. 또, 혼합현실(mixed reality)의 활용은 행위의 근간이 되는 인간 감각과 인공지능을 통해 유입되는 감각을 통합함으로써 이 군집적 특성을 더 강화한다.

셋째, 앞서의 첫째와 둘째 특성에 따른 부수적 특성으로 어떤 행위에 대한 윤리적 책임은 한 개인에게 전적으로 부여되지 않고 행위의 군집적 주체들에게 나뉘어(distributed)지게 된다. 그리고 이러한 군집적 책임은 회피

20) 인공지능 로봇에 '인격성(person)'을 부여한다는 의미를 지나치게 인간 중심적이거나 사변적으로 해석하는 경우가 있다. 그러나 인공지능 윤리에 관한 한 인격성의 개념은 일차적으로 법적 의미로 사용된다. 잠재적인 법적 문제들을 경감하고자 하는 목적으로 인격성을 인공지능로봇에 부여하는 것이지, 인공지능 로봇이 인간과 같은 인격을 가지는가의 존재론적인 문제는 부차적이다. 이와 비슷한 문제의식을 가지고 플로리디와 샌더스 (Floridi & Sanders 2004)는 '윤리적 행위자'를 기능적으로 정의하기도 한다. 즉, '상호작용적, 자율적, 적응성을 갖춘 시스템'으로 도덕적 행위자를 정의하면서 인간의 특성에 기대어 정의하는 데에서 오는 부담을 덜고자 한다. 필자는 이 정의가 인간중심적 정의에서 벗어날 수 있으나 '자율성', '적응성'의 애매모호함 때문에 구체적인 내용을 가지기 어렵다고 본다.

의 문제를 효과적으로 해결하기 위해 윤리지침과 법적으로 규제될 것이다. 이 과정에서, 인공도덕성 검사는 자율시스템 안의 여러 주체들의 도덕적 의사결정을 평가하는 데 사용될 수 있다.

Ⅴ. 인공도덕성 검사의 기준은 인간 도덕성인가

인공지능 설계 이후 전 과정에 대한 검사의 이점은 인공지능 윤리에서 중요한 문제로 제시되고 있는 투명성(transparency) 문제를 일부분 해결할 수 있다는 것이다. 인공지능의 판단이 도덕적이라는 점을 결정하기 위한 난제 중 하나는 '윤리원칙들 간의 이견 혹은 갈등'이다. 이는 인공지능에만 관련된 문제가 아니라 인간도 해결하지 못하는 문제다. 이 절에서 필자는 이와 관련한 더 근본적인 문제가 선결되어야 한다고 제안한다.

1. 인간도덕성 기반의 인공도덕성

대표적 자율시스템인 자율 주행 차[21]를 더 발달시키기 위해 필요한 절차 중 하나는 기술적 문제를 넘어서서 자율주행차가 구체적인 딜레마 상황에서 어떤 주행을 하도록 결정을 내려야 하는가이다. 이 구체적 상황 중 하나는 이미 철학 분야에서 다루어졌던 윤리적 딜레마 상황[22]이다.

최근 사이언스지에 실린 "자율 주행 차의 사회적 딜레마"[23]에서는 자율

21) 미국 도로교통안전국(NHTSA)에서 제시한 자율주행 레벨은 0레벨에서 4레벨까지 총 다섯 레벨로, 현재의 발전단계는 3레벨과 4레벨의 중간정도에 위치해있다. 자율 주행 차 발전의 목표는 최종 4단계로 탑승자의 개입이 전혀 필요 없는 단계이다. 이 단계까지 자율주행차가 발달하려면 수많은 교통상황에서의 판단 기준이 정해져야 한다.

22) Foot (1967).

23) Bonnefon, Shariff, & Rahwan (2016).

주행차가 체력이나 인지능력에 한계가 있는 인간보다 실수를 덜 하기 때문에 교통사고 건수를 90퍼센트나 줄일 수 있지만, 다음과 같은 윤리적 딜레마 상황에서 어떤 의사결정을 내릴지의 도덕적 문제를 해결해야 하는 과제에 직면해있다고 소개했다. 자동차가 운행하는 방향에는 다수의 행인이 지나가고 있고 행인들에게 상해를 입히지 않기 위해 방향을 바꿀 경우 벽에 부딪쳐 운전자가 상해를 입게 된다. 이 경우 무인자동차가 어떤 결정을 내리도록 설계되어야 할까?

이 논문의 저자들은 자율주행차가 어떻게 작동해야 할지에 대해 설문조사를 했다. 위의 C상황에 대해 답변자들 중 대다수인 76%는 보행자 열 명 대신 탑승자 한 명을 희생하는 쪽이 더 도덕적이라고 판단했다. 즉, 자율주행 차의 인공지능이 공리주의적 결정에 입각해 우연히 발생하는 이러한 딜레마 상황에 대처하도록 프로그래밍 되어야 한다는 것이다.

그런데 답변자들은 그 다음 질문에 대해서는 다른 대답을 제시했다. 만약 본인이 선택한 그 기준으로 인공지능 설계를 법적으로 정착시켜 의무화하는 것에 동의하겠는가에 대한 답변에 있어서는 동의하는 확률이 낮아졌다. 그리고 자기희생 모드의 자율 주행 차와 자기보호 모드의 자율주행차가 판매될 경우 어떤 차를 구입하겠느냐는 질문에는 다수가 자기보호 모드의 차를 선택했다.

이렇게 우리의 도덕적 직관은 비 일관적이다. 소수보다는 다수의 생명을 구하는 것을 더 나은 도덕적 판단으로 선택하면서도 정작 본인은 그렇게 프로그래밍 된 자율주행차를 선택하지 않기 때문이다. 이런 비일관성도 윤리적 의사결정을 내려야 하는 자율 주행 차 개발에서 극복해야 할 또 다른 난점이다. 자율 주행 차 사용자를 보호하는 것은 자연스러운 일이지만 자기보호 모드로 프로그래밍 될 경우 사회적 문제가 발생하며, 그렇다고 해서 자기희생 모드로 프로그래밍 된 자율 주행 차는 그 차를 사용할 이들이 구입할리 만무하다.

그렇다면, 현재 발전단계 한 단계를 남겨둔 자율 주행 차의 의사결정에

대한 도덕성 검사가 가능하려면 설계 시 이미 그 기준이 정해져야 하는데, 결정의 주체는 누구여야 하는가? 공리주의적 기준인가, 의무론적 기준인가, 덕 이론 혹은 다른 가치 기준인가? 이 같은 고민은 인공지능이 장착된 주행차가 부상하면서 새로이 제기된 문제는 아니며, 인간 차원에서도 해결하기 어려운 문제이다. 그러면 인간 차원에서 먼저 이 윤리적 딜레마를 해결한 후 인공지능에 적용하면 될까? 필자는 이러한 방식이 아닐 수 있다고 생각한다. 윤리적 딜레마가 인공지능 안에서는 다른 차원의 문제가 될 가능성도 고려할 필요가 있다. 인간 도덕성을 인공지능의 도덕 추리나 판단에 적용하는 방법이 인공지능에 도덕성을 구현하거나 검사할 때 적합한 기준이 아닐 수 있다.

인공지능은 범용인공지능(AGI)으로 개발되기도 하지만 많은 경우 특정 목적을 가진 인공지능으로 개발된다. 특정목적을 가질 경우, 그 인공지능은 특정한 윤리적 상황에서 따라야 할 원칙이 공리주의적인지 의무론적인지 덕 윤리적인지에 대해 선호도가 있는 경우가 있을 것이다. 이런 경우이거나 혹은 그러한 도덕원칙에서의 상충이라는 문제가 없으며 기본적인 인공지능 윤리 원칙을 상정했다고 가정해보자. 그렇다면 그 원칙에 따라서 인공지능의 도덕성에 대한 평가를 할 수 있는가?

2. 인간도덕성과 독립적인 인공도덕성

필자는 다음의 선결문제가 우선적으로 검토되고 합의되어야 설계-학습-실행의 전 단계에 걸친 인공도덕성 검사가 현실적으로 가능해진다고 생각한다. 인공지능에 도덕성을 구현한다고 할 때 그 기준은 무엇인가? 일반적으로 받아들여지는 기본 원칙은 인간의 복지나 행복을 해치지 않는 방향으로, 즉 인간을 우선시하는 방향으로 인공지능이 개발되어야 한다는 것이다. 그런데 이러한 원칙은 구체적으로 인공지능의 행위를 구현할 때 어떤 도덕적 원리를 구현해야 하는가의 문제와는 별개다. 구체적으로 다음의 질문이 제기된다.

인공지능의 도덕적 행위를 구현하거나 도덕성을 검사할 때 그 기준을 인간보다 더 높은 기준의 도덕성으로 상정해야 하는가? 인간끼리 적용하는 윤리를 그대로 적용하면 되는가? 인공지능 및 로봇이 인간보다 더 높은 기준으로 도덕적 의사결정을 내린다면 이에 인간은 따를 것인가? 의 문제이다. 필자는 이와 관련하여, '인공지능'에 대한 구체적 정의가 인공지능의 도덕성 구현 뿐 아니라 인공지능의 개발 방향을 이끈다고 생각한다. '인공지능'에서 '지능'은 일반적으로 '인간의 지능'을 기준으로 개발되었고 인공지능의 윤리 또한 인간의 윤리에 빗대어 고려되어 올 수밖에 없었다. 그런데 '인공지능'을 굳이 인간의 지능과 유사하게 개발하거나 빗댈 필요는 없다. '지능'이란 인간에게조차도 이미 완성된 것이 아니라 상황이나 환경에 맞추어 수정되거나 개발되는 측면이 있다.

그렇다면 인공지능은 '인간처럼' 생각하는 기계로 규정할 필요나 근거는 없으며, 상황이나 목적에 맞는 인지나 행동을 산출하는 기계로 보는 것이 타당하다.[24] 이렇게 생각한다면 인공지능에 대한 도덕성 시험은 도덕적 인간처럼 행동하는가의 시험일 필요는 없다. 이러한 생각은 이미 앞서 인공도덕성에 대한 시험은 설계부터 제작에 이르는 전 과정에 대한 검사의 성격을 가져야 한다는 부분에서 이미 기존 튜링테스트의 성격을 벗어났음을 제안한 바 있다. 다음에서 이에 대한 근거를 좀 더 구체적으로 검토해보자.

최근의 인간-로봇 상호작용 연구[25]는 인간은 같은 인간이 내리는 도덕적 의사결정보다 인공지능 로봇에 더 공리주의적인 의사결정을 기대한다는 것을 보여준다. 이 연구는 트롤리 딜레마에서 철로변환기를 당기는 행위자를 인간, 휴머노이드 로봇, 그리고 자동기계로봇으로 상정하고 세 종류의 행위자들이 철로변환기를 당기지 않을 경우에 대하여 참가자들이 도덕적 평가를

24) 인공지능학자 스튜어트 러셀과 피터 노빅(Russell, Norvig, and Intelligence 1995)이 취하는 입장이다. 구체적으로는, '인공지능'을 인간을 닮은 기계가 아니라, 최상의 결과를 얻기 위해 행동하고 목적함수(objective function)를 최대화하는 주체인 '합리적 에이전트(rational agent)'로 정의한다.

25) Malle et al. (2015).

내리도록 하였다.

철로변환기를 당기지 않아 한명 대신 네 명이 철도에서 사망하는 경우에 대해 참가자들은 철로변환기를 당기지 않은 행위자가 인간일 경우보다 휴머노이드일 경우, 그리고 로봇일 경우의 순서대로 더 많은 비중으로 그 행위자를 도덕적으로 비난했다. 그리고 이와는 반대로 철로변환기를 당기는 경우에 참가자들은 인간의 행위를 로봇의 행위보다 더 비난하는 경향을 보였다. 즉, 인간은 인간에게보다는 로봇에게 더 공리주의적인 도덕적 의사결정을 내리기를 바라거나 자연스럽게 받아들인다. 그러면 이러한 인간의 반응을 반영하여 인공지능 로봇의 의사결정은 인간의 의사결정보다 더 공리주의적 기준에 따라 내려지도록 그 기준을 설정해야 할까? 이는 기존의 인공지능 윤리에 대한 논의와는 별도의 차원에서 논의되어야 할 것이다.

필자는 이로부터 '인간-인간 간의 상호작용'과 '인간-로봇 간의 상호작용'은 그 성격이 다르며, 따라서 인공지능의 도덕성의 기준을 별도로 논의해야 한다고 생각한다. 위의 연구를 비롯한 인간-로봇, 인간-인공지능 상호작용으로부터 볼 수 있는 바는 우리가 인공지능윤리를 구성하고 실행하는 데 있어서 인간 사이에 사용되는 윤리를 그대로 사용하면 인공지능 및 로봇의 설계, 제작, 사용에 있어서 '역설적'으로 인간을 중심에 둔 혹은 인간 관점의 윤리적 설정을 하지 못한다는 점이다. 한 기계가 도덕적이라는 것, 즉 기계의 도덕적 유능성이라는 것이 어떤 내용이어야 하는지에 대한 사회적 합의가 필요하다.

지금까지 논의한 바와 같이, 자율시스템으로서의 인공도덕성 검사는 튜링테스트처럼 실행 단계 이후의 반응에 대한 행동주의적, 언어적 검사가 아니라 설계-학습-실행에 이르는 전 과정에 걸친 검사를 필요로 한다. 또한 이러한 인공도덕성 검사가 성공적이려면 인공도덕성의 기준을 인간이 현재 향유하고 지키는 윤리원칙으로 정해야 할지, 혹은 이와는 독립적으로 인공지능에게만 적용될 새로운 윤리 기준을 세워야 할지에 대한 광범위한 논의가 필요할 것이다.

참고문헌

이상형 (2016), 「윤리적 인공지능은 가능한가: 인공지능의 도덕적, 법적 책임 문제」, 『법과 정책연구』, 16(4): 283-303.

최용성·천명주 (2017), 「인공적 도덕 행위자에 관한 도덕철학, 심리학적 성찰」, 『윤리교육연구』, 46(4): 93-129.

Allen et al. (2000), "Prolegomena to Any Future Artificial Moral Agent," *Journal of Experimental & Theoretical Artificial Intelligence,* 12(3): 251-261.

Arnold, T. & Scheutz, M. (2016), "Against the Moral Turing Test: Accountable Design and the Moral Reasoning of Autonomous Systems," *Ethics and Information Technology,* 18(2): 103-115.

Bonnefon, J. F., Shariff, A. & Rahwan, I. (2016), "The Social Dilemma of Autonomous Vehicles," *Science*, 352(6293): 1573-1576.

Floridi, L. & Sanders, J. W. (2004), "On the morality of artificial agents," *Minds and Machines*, 14(3), 349-379.

Foot, P. (1967), "The Problem of Abortion and the Doctrine of Double Effect," *Oxford Review* 5: 5 – 15.

Hayes, P. & Ford, K. (1995), "Turing Test Considered Harmful," *IJCAI* (1): 972-977.

Huang, H. M. (2007), "Autonomy Levels for Unmanned Systems (ALFUS) Framework: Safety and Application Issue," *Proceedings of the 2007 Workshop on Performance Metrics for Intelligent Systems*: 48-53.

Malle, B. F., Scheutz, M., Voiklis, J., Arnold, T., & Cusimano, C. (2015), "Sacrifice One for the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents," *HRI' 15 Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*: 117-124.

Moor, J. H. (2011), "The Nature, Importance, and Difficulty of Machine Ethics," in Michael Anderson and Susan L. Anderson (ed.), *Machine Ethics*, pp. 13-20, NY: Cambridge University Press.

Russell, S., Norvig, P. (1996). *Artificial intelligence: A modern approach*, New Jersey: Prentice Hall.

Sample, I, "Computer Says No: Why Making AIs Fair, Accountable and Transparent is Crucial," *The Guardian,* November 5, 2017.

Searle, J. R. (1984), *Minds, brains and science*. Harvard University Press.

Turing, A. (1950), "Computing Machinery and Intelligence," *Mind*, 59(236): 433.

Wallach, W. & Allen, C. (2009), *Moral Machines: Teaching Robots Right from Wrong*, Oxford: Oxford University Press.

Abstract

This paper examines the conditions for artificial morality test. The original turing test for artificial morality faces a similar difficulty as the original turing test: applying the original turing test to test artificial morality is only to test "how well it imitates human being." Instead, I suggest that the test for artificial morality could be a test for the whole process including design, learning and output. In section 2, I explain how the deep neural network introduces a black box which is inaccessible to people and induces moral issues. In section 3, I critically examine whether the original Turing test is applicable to autonomous system using deep learning mechanism. In section 4, I identify characteristics of artificial morality test. Based on the four sections, in section 5, I argue that we need an agreement on the question as to whether artificial morality should be based on human morality for the successful test of artificial morality.

【Keywords】 Artificial Morality Test, Turing Test, Autonomous System, Deep Learning, Human Morality

논문 투고일: 2018. 3. 19
심사 완료일: 2018. 4. 18
게재 확정일: 2018. 4. 18