인공지능 발전에 따른 법적 쟁점과 대응

(2019. 10. 31.)

영남대학교 법학전문대학원
교수 양 천 수

목 차

I. 인공지능의 법적 문제

II. 법적 문제에 대한 법철학적 기초

III. 법적 인격 개념

IV. 법적 인격 개념의 변화

V. 탈인간중심적 인격 개념의 가능성

VI. 인공지능의 법적 문제 검토

I. 인공지능의 법적 문제

1. 알파고 충격

- 이세돌 9단과 알파고의 바둑 대결 (2016)
- 인공지능에 대한 법적 연구의 전환점
- 다양한 영역에서 인공지능의 법적 문제 연구

I. 인공지능의 법적 문제

2. 인공지능의 법적 문제

- 인공지능과 민사책임 가능성 (자율주행자동차)
- 인공지능의 거래행위 주체성 (투자 AI)
- 인공지능과 저작권
- 인공지능의 형사책임
- 인공지능의 사법적 활용 가능성

II. 법적 문제에 대한 법철학적 기초

1. 탈인간중심적 법학의 가능성

- 탈인간중심적 법체계 및 법학의 가능성
- 탈인간중심적 법체계의 가능성
- 탈인간중심적 법학(학문체계)의 가능성
- 탈인간중심적 사고가 가능한가?

II. 법적 문제에 대한 법철학적 기초

2. 탈인간중심적 사고의 가능성

- 인간중심적 사고 vs. 탈인간중심적 사고
- 탈인간중심적 사고의 전통
- 구조주의
- 체계이론 (Niklas Luhmann)
- 인간중심적 사고와 작별하기

III. 법적 인격 개념

1. 법에서 인격의 의미

- 법적 문제 해결의 출발점
- 법적 권리와 의무의 주체가 될 수 있는 자격
- 인간(human)과 인격(person)의 분리

III. 법적 인격 개념

2. 인격의 기능

- 주체 보호
- 책임귀속
- 인격체의 상대방 보호
- 법적 관계의 명확화
- 법체계의 안정화

IV. 법적 인격 개념의 변화

1. 인간중심적 인격 개념

- 인간과 인격의 개념적 분리
- 인간에 바탕을 둔 인격 개념

- 자연적 인격 개념 (Hans Kelsen)

- 자율적인 이성적 인간으로서 인격

- but, 역사적으로 노예, 여성, 아동 배제

IV. 법적 인격 개념의 변화

2. 인격 개념의 확장

- 자연적 인격 개념의 확장

- 노예, 여성, 아동의 포함

- 법인의 등장

- 사회적 체계로서 법인

- 동물권 운동

V. 탈인간중심적 인격 개념의 가능성

1. 법적 인격의 인정 요건

- 도덕적 인격과 법적 인격의 구별
- 법적 인격의 요건

- 인간?

- 자율성?

- 소통가능성?

V. 탈인간중심적 인격 개념의 가능성

2. 법적 인격의 모델

- 인간중심적 모델
- 불완전한 탈인간중심적 모델
- 완전한 탈인간중심적 모델

V. 탈인간중심적 인격 개념의 가능성

3. 탈인간중심적 법적 인격의 요건

- 독립된 소통의 귀속지점
- 자율성
- 소통참여 가능성

VI. 인공지능의 법적 문제 검토

1. 인공지능의 법적 인격 주체성

- 원칙

- 탈인간중심적 법적 인격 주체성 가능

- 강한 인공지능

- 가능

- 아직은 현실적 문제가 아님!

VI. 인공지능의 법적 문제 검토

- 약한 인공지능

- 소통의 귀속 가능성 (◯)

- 소통의 참여 가능성 (◯)

- 자율성(Δ)

▶ 자율성을 어떻게 판단할 것인가?

▶ 획일적 판단이 아닌 유형적 판단 필요성

VI. 인공지능의 법적 문제 검토

2. 인공지능의 법적 문제 검토

- 가능성의 차원

- 필요성의 차원

- 무엇 때문에 권리와 의무를 인정?

VI. 인공지능의 법적 문제 검토

※ 현행 법체계의 법리

- 소유자 책임
- 미성년자의 감독자 책임
- 사용자 책임
- 위험책임

VI. 인공지능의 법적 문제 검토

3. 인공지능의 권리주체성 인정 문제

● 가능한가?

● 필요한가?

- 인공지능 자체의 보호
- 인공지능의 저작권 보호?

VI. 인공지능의 법적 문제 검토

4. 인공지능의 거래 주체성

- 필요한가?
- 거래관계의 명확화?
- 무엇이 효율적인 수단인가?

VI. 인공지능의 법적 문제 검토

5. 인공지능의 민사책임 문제

- 자율주행자동차의 경우
- 필요한가?
- 책임법과 보험법의 결합?

VI. 인공지능의 법적 문제 검토

6. 인공지능의 형사책임 문제

- 가능한가?
- 필요한가?

- 인공지능의 폐기처분 or

- 인공지능에 대한 형벌 부과?

경청해 주셔서 고맙습니다.

고등과학원 인공지능과 포스트휴머니즘 연구단
워크숍 '인공지능의 발전에 따른 윤리적·법적 쟁점과 대응'

의사결정 자동화에 대한 대응으로서의 AI윤리 교육

김효은 (한밭대 인문교양학부)

1. 인공지능윤리는 컴퓨터 윤리나 과학기술윤리와 무엇이 다른가
2. 컴퓨터 안에는 없고 인공지능 안에는 있는 '윤리'
3. 의사결정 자동화에 필요한 윤리원칙은 무엇인가
4. 자율주행 등급의 어디에서 윤리가 관련되는가
5. 윤리적 문제를 공학적으로 해결할 수는 없을까
6. 사회적 대응으로서의 인공지능윤리 교육

1. 인공지능윤리는 컴퓨터 윤리나 과학기술윤리와 무엇이 다른가

인공지능윤리는 과학기술윤리나 공학윤리에 속하는 하나의 예가 아닐까? 왜 인공지능윤리를 중요하게 이야기할까? 기존의 과학기술들과 인공지능은 어떤 점이 다르길래 인공지능 회사의 중역들이 나서서 인공지능윤리 프로젝트에 투자를 하고 목소리를 높이는 것일까. 사업 수익이나 기술 발전에 윤리가 어떤 역할을 하는가?

현재 컴퓨터의 발전 단계는 컴퓨터 시스템이 '지능'을 가지는 인공지능 단계에 와 있다. 이러한 발전은 그와 관련한 윤리 차원 또한 달라지게 한다. 현재 인공지능 시스템을 사용하는 로봇은 더 이상 단순한 기계가 아니라 '의사 결정'을 하는 '자율 시스템'이다. 자율 시스템이란 무엇이며, 단순한 기계와 어떤 차이가 있는가? 인공지능을 장착하지 않은 단순 기계는 주어진 규칙에 따라 작동한다. 반면 인공지능이 장착된 로봇이나 시스템은 데이터로부터 규칙이나 패턴을 파악해낸다. 데이터를 스스로 학습해 패턴을 인식하고 결정을 내린다.

인공지능의 윤리적 차원을 이야기하려면 먼저 '인공지능'을 어떤 의미에서 사용하는지 정확히 밝힐 필요가 있다. '인공지능'이라는 단어를 사람마다 다른 의미로 사용하기 때문이다.
기본적으로 '인공지능'은 지각, 인지, 행동을 하는 개체를 의미한다. 로봇처럼 몸을 가질 수도 있고 소프트웨어처럼 몸 없이 존재할 수도 있다. 기술과 산업의 관점에서는 일의 종류에 따라 인공지능이 여러 하위 분야로 나뉜다. 반면 수행 능력 차원에서 인공지능은 크게 다음 세 가지로 나뉜다.

- 좁은 인공지능 (ANI, Artificial Narrow Intelligence): 특정 영역의 문제를 풀어내는 인공지능
- 범용 인공지능 (AGI, Artificial General Intelligence): 인간 수준의 인공지능
- 초인공지능 (ASI. Artificial Super Intelligence): 모든 분야에서 사람보다 뛰어난 인공지능

좁은 인공지능의 대표적 예는 2016년 이세돌 구단과 바둑 경기를 치룬 바둑 인공지능 알파고(AlphaGo)다. 이 알파고 시스템의 기본 원리를 바탕으로 바둑뿐 아니라 다른 분야에서도 패턴이나 규칙 인식을 학습할 수 있는 범용 인공지능이 현재 발전하고 있다. 한편, 초인공지능에 대해서는 최근 한국의 대중매체에서 종종 언급되었다. 주로 '인공지능이 사악한 의도를 가지고 인간을 지배할 위험'이 다루어지는데, 이는 범용 인공지능이나 초인공지능과 관련된 이야기로 좁은 인공지능과는 무관하다. 이 책은 좁은 인공지능과 관련한 윤리적 이슈들을 주로 다룬다. 범용 인공지능이나 초인공지능과 관련한 윤리적 이슈는 별도로 다루어져야 할 것이다.

좁은 인공지능의 윤리적 쟁점에 초점을 맞추는 이유는 다음과 같다. 첫째는 시의성이다. 범용 인공지능, 초인공지능의 윤리 문제들보다는 좁은 인공지능의 윤리 문제들이 현재 우리가 먼저 해결해야 할 문제들이다. 범용 인공지능과 초인공지능과 관련한 윤리 문제들 ("인공지능은 감정이나 자아를 가질 수 있는가" 등)은 이후에 생각해야 할 쟁점들로, 시급성이 상대적으로 적다. 물론 이와 연관된 주제가 등장할 때 언급은 할 것이다.

둘째, 좁은 인공지능과 범용 인공지능을 분명히 구분해야 이론적, 정책적 등의 측면에서 혼란이 없기 때문이다. 예컨대 "로봇에 '인격'을 부여할지"에 대한 최근 논쟁은 사실 두 종류의 논의로 구분해야 한다. 이는 좁은 인공지능과 관련해서는 법적 논의가, 범용 인공지능과 관련해서는 철학적 논의가 된다. 당면한 문제는 전자인 법적 논의다. 좁은 인공지능 관련 논의에서 '인격'은 인공지능을 활용했을 때 생기는 사고의 법적 처리를 위한 법적 지위를 의미한다. 인간에게 부여되는 심리적, 도덕적 인격과는 의미가 다르다. 반면 범용 인공지능과 관련해 '로봇도 인간처럼 심리적, 도덕적 인격을 가지는가'라는 주제가 논의되곤 하는데, 여기서 '인격'은 법에서 사용되는 인격의 의미와는 다른 맥락을 띤다.

인공지능에 적용되는 윤리는 '착하게 살아야 한다'는 좁은 의미의 윤리를 넘어선다. 여기서 윤리는 복잡한 상황에 적합한 의사 결정으로 관련 의사 결정 주체나 대상의 지속가능성을 확보한다는 넓은 의미를 띤다. 이렇게 윤리적 고려가 산업, 경영, 과학기술 등 거의 모든 분야를 이끄는 주요한 요소가 되는 것은 윤리 패러다임의 변화라고 할 수 있다.

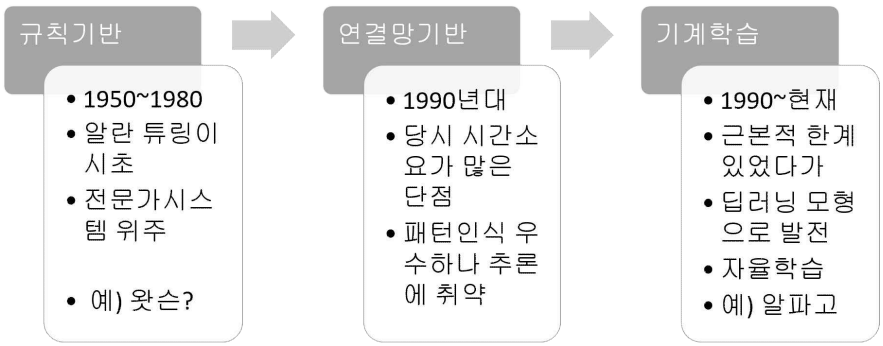
2. 컴퓨터 안에는 없고 인공지능 안에는 있는 '윤리'

인공지능윤리를 이야기하기 전에 인공지능의 간략한 역사를 볼 필요가 있다. 인공지능이라는 아이디어와 컴퓨터가 인간과 같은 지능을 가질 수 있다는 가능성은 영국의 논리학자이자, 수학자인 앨런 튜링(Alan Turing)이 제시했다. 튜링은 1945년 복잡한 계산과 논리 문제를 처리하는 초기 형태의 컴퓨터를 고안했고 이로부터 오늘날의 딥러닝까지 발전해왔다. 그리고 튜링이 1950년 철학 저널 Mind지에 실은 "계산기계와 지능(Computing Machinery and Intelligence)"이라는 논문에서 계산하는 기계가 지능이 있는 것으로 간주될 수 있는 가능성을 제안한다.

'인공지능'이라는 용어는 튜링의 초기 컴퓨터와 계산기계의 지능이라는 아이디어가 제시된 후 1956년 여름 미국 다트머스 (Dartmouth) 대학에서 열린 학술회의에서 존 매카시(John McCarthy), 마빈 민스키(Marvin Minsky), 클로드 섀넌(Claude Shannon) 등의 학자들이 모여 컴퓨터의 추론과 탐색에 대해 대화를 나누었고 여기서 '인공지능'이란 용어가 탄생했다. '인공지능'이란 단어는 인간과 똑같은 수준으로 기계 혹은 컴퓨터가 지능을 가진다는 의미보

다는 형식화 가능한 지능의 요소들을 개발하고자 하는 방향을 의미한 것이다. 지능을 기계적 계산과정으로 설명할 수 있다는 아이디어는 계산을 수행하기 위한 '기호'와 이 기호들을 계산할 '규칙'을 통해 가능하기 때문에 초기 인공지능은 규칙기반(Rule-based AI)로 발전하였다.

현재 주로 사용되고 있는 인공지능은 '규칙기반'이 아니라 뇌신경망을 모형으로 하여 발전된 '신경망 기반'의 인공지능으로 과거에 해결할 수 없었던 컴퓨터 파워와 충분히 많은 양의 데이터 덕분에 발전하고 있다. 기존의 규칙기반 인공지능은 규칙이 주어지면 그에 해당하는 답들을 내어놓는 일종의 자동시스템이다. 이와 반대로 신경망 기반 인공지능은 자료들이 주어지면 그로부터 패턴이나 규칙을 찾아내는 일종의 자율시스템이다. '자율시스템'이라고 해서 '기계가 자율성이나 의식을 가진다'는 의미는 아니며 그와는 구분되어야 한다. 알고리즘 방식의 변화에 따라 기계 및 컴퓨터에 자기 학습 기능이 추가되었다는 의미다. 자율시스템으로서의 인공지능은 규칙기반 인공지능에서는 성공적으로 하기 어려웠던 패턴 인식 등이 가능하게 되어 심층학습을 통해 사물의 특징을 스스로 파악할 수 있게 되었다.



3. 의사결정 자동화에 필요한 윤리원칙은 무엇인가

도덕적으로 행동하는 인공지능 로봇을 만들고 인공지능의 판단이 윤리적이길 바라는 이유는 무엇인가? '도덕적으로 유능한 인공지능'이란 무엇이며 그 기준은 무엇인가? 인공지능 로봇이 인간을 보조하는 영역이 확대되고 우리의 의사 결정이 인공지능에 의존하는 비중이 커지고 있다. 반면 인공지능의 학습 절차를 인간이 알기는 점점 어려워지고 있다. 이때 우려되는 점은 인공지능이 인간에 피해를 줄 가능성이다. 이 때문에 우리나라를 포함한 여러 나라에서 최근 발표한 인공지능윤리 원칙들 중 우선순위로 꼽는 원칙이 바로 "인간의 복지나 행복을 해치면 안 된다"는 원칙이다.

'원칙'이란 기본 방향을 알려 주는 역할을 한다. 원칙 자체는 추상적이며 일상생활의 실제 상황에 그대로 적용하기 어렵다. 소설가 아시모프의 로봇 삼 원칙, 여기에 '0원칙'이 추가된 사 원칙, 그리고 일반 윤리 이론의 원칙들도 그 성격상 추상적일 수밖에 없다. 이런 원칙들은 우리가 실제 일상생활에서 부딪치는 복잡한 상황에 그대로 적용하거나 알고리즘으로 만들기 어렵다. 로봇 삼 원칙이나 공리주의, 의무론 중 어느 것도 인공지능이 탑재된 무인자동차가 구체적 교통 상황에서 어떤 결정을 내려야 하는지 답을 주지 못한다.

"인간의 복지나 행복을 우선으로 한다"는 원칙을 인공지능 로봇 설계에 이용하거나 인공지

능 소프트웨어의 판단 기준으로 삼는다고 해 보자. 사회에서 상호 경쟁 혹은 갈등 관계에 있는 이익 집단들은 '복지'나 '행복'에 대해 다른 해석을 내릴 수밖에 없다. '타인을 해치지 말라'는 원칙 또한 누구나 동의할 만한 윤리 원칙인 것 같지만 정작 구체적 상황에는 합의가 어렵다. 예컨대 정당방위로 살인한 경우를 윤리적, 법적으로 허용할 수 있을 것인지는 인간조차도 쉽사리 합의하기 어려운 문제다. 만약 정당방위로서 살인을 지나치게 관대한 수준으로 허용하면 살인을 정당화하는 데 이용될 것이고, 다른 한편으로 아주 좁게 허용할 경우 범죄에 매우 취약하게 되고 피해가 늘어날 것이다.

인공지능의 '지능'은 처음에는 인간의 지능을 목표로 개발되기 시작했다. 하지만 기억이나 검색, 패턴 인식 등의 일부 분야에서는 인간 지능보다 더 높은 수행 능력을 보이며 발전 중이다. 무인자동차를 비롯한 여러 인공지능의 윤리적 판단 및 행동 기준은 인간이 보통 따르는 윤리적 판단 및 행동 기준에 맞춰 유사하게 잡아야 할까, 더 높게 혹은 더 낮게 혹은 아예 다른 기준으로 잡아야 할까. 인간보다 더 높은 인지 수행 능력을 갖추고 있으니 더 높은 도덕 기준을 적용해야 할까. 이에 대한 결정은 사회적 합의를 거쳐야 하며, 일부만이 임의로 합의할 경우 사회적 갈등이 생길 수 있다. 인공지능 로봇과 인간이 원활하게 상호작용하는 데에 인공지능윤리가 인간 윤리와 비슷해야 하는지 여부를 미리 실험해 보면 유용할 것이다. 그 방법 중 하나가 가상으로 인공지능 시스템의 상황을 연출하고 피험자의 반응을 보는 것이다.

4. 자율주행 등급의 어디에서 윤리가 관련되는가

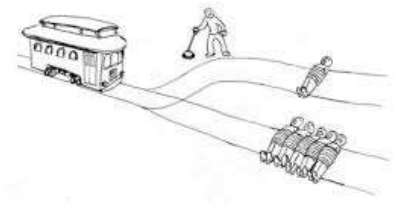
무인자동차는 체력이나 인지 능력에서 한계가 있는 인간보다 실수를 덜 한다. 졸음운전이나 좁은 시야, 늦은 판단력의 문제가 없다. 그래서 무인자동차는 교통 사고 건수를 기존보다 90퍼센트나 줄일 수 있으므로 안전도를 높이는 데 기여한다고 한다. 그러나 무인자동차가 구체적인 딜레마 상황들에서 어떤 의사 결정을 내릴지 정해져야 가장 높은 수준의 '자율주행' 자동차가 완성 가능하다.

인공지능을 탑재한 무인자동차의 자율 주행 기술 등급은 미국 도로교통안전국(NHTSA)에서 정한 바에 따르면 레벨 0부터 레벨 4까지 총 다섯 단계다. 레벨 0은 자동차 주행 보조 장치가 없는 일반 자동차 상태다. 레벨 1은 현재 일부 자동차에 장착돼 있는, 일정 속도로 계속 달릴 수 있는 크루즈 컨트롤이나 일부 고속버스에 장착된 차선 이탈 금지 기능과 같이 운전 보조 기능이 더해진 상태다. 레벨 2는 레벨 1 기능에 더하여 앞차가 정지하면 자동으로 정지하고 앞차가 움직이면 자동으로 출발하는 적응형 크루즈 컨트롤 기능이 장착된 상태다. 레벨 3은 차의 운행을 무인자율차에 온전히 맡길 수 있고, 비상시에 운전자가 차를 제어할 수 있는 상태다. 레벨 4는 모든 환경에서 자율주행이 가능하며 탑승자의 개입이 필요하지 않은 상태다. 현재는 세 번째 단계인 레벨 2를 개발 중이며, 레벨 3을 거쳐 레벨 4가 완성되어야 자율 기능이 온전히 완성된 '무인자동차'라 할 수 있다. 이 단계에서야 비로소 인간의 개입이 전혀 없이도 운행이 가능한 상태가 되며 보통 '무인자동차의 윤리'가 문제되는 단계이다.

여기까지 도달하기 위해 선결되어야 할 사항 중 하나는 특정 교통 상황에서 무인자동차가 어떤 결정을 내리게 할지 정하는 일이다. 일어날 수 있는 상황을 미리 모두 예측할 수는 없지만 어떻게 무인자동차가 운전을 결정할 것인지를 합의하고 이를 알고리즘으로 만들어야 사고 발생의 손해를 최소화하고 법적, 도덕적 처리를 용이하게 할 수 있다.

도덕판단의 기준을 알아보기 위한 사고실험: 트롤리 딜레마(Trolley's dilemma)

특정 교통 상황에서 직면하는 딜레마는 이미 철학 분야에서 '트롤리 딜레마(Trolley Dilemma)'로 다루어진 바 있다. 트롤리는 기차의 일종으로, 기차 대신 큰 상해나 죽음을 야기할만한 동물이나 위험물로 대체해서 생각해도 가능하다. 트롤리 딜레마는 일종의 사고실험으로 먼저 제시되었다. 사람들이 실제로 구체적 상황에서 어떤 결정을 내릴지에 대해서는 알기 어렵다. 어떤 구체적 상황에 많은 사람들을 동시에 모아 그 반응을 살펴보기란 현실적으로 불가능하기 때문이다. 그래서 실제 상황에 사람들이 어떻게 반응할 것이라고 생각하는지를 알아보기 위해 가상 시나리오를 사용하는 '사고실험(thought experiment)'를 한다.
이렇게 고안된 사고실험 중 도덕판단에 관한 것이 영국 윤리철학자가 1967년 제시한 트롤리 딜레마(trolley's dilemma)[1] 이다. 이 딜레마의 문제상황은 위급한 상황에서 누구를 살릴 것인가이다. 구체적 상황은 "다섯 사람을 구하기 위해 한 사람을 죽이는 것이 도덕적으로 허용가능한가?"이다.



<트롤리 딜레마> 정차 기능이 고장난 트롤리는 선로를 따라 달려오고 있고, 선로에는 다섯 사람이 있다. 당신은 선로 밖에 서 있고 다섯 사람을 구하기 위해서는 선로변환기를 당기면 되지만 그렇게 되면 다른 선로에 있는 다른 한 사람이 죽게 된다. 선로변환기를 당기는 행위는 도덕적으로 허용가능한가?

이와 함께 많이 소개되는 윤리적 딜레마의 또다른 대표적 사례는 트롤리 딜레마의 변형 시나리오인 인도교 사례이다.

<인도교 사례> (오른쪽 그림) 트롤리는 선로를 따라 달려오고 있고, 선로에는 다섯 사람이 있다. 당신은 선로를 사이에 둔 인도교 위에 서 있고, 바로 옆에는 상당히 무거운 사람이 한 명 서 있다. 다섯 사람을 구하는 유일한 방법은 옆에 서 있는 사람을 아래로 밀쳐서 그 무게로 트롤리를 멈추게 하는 것인데, 이는 도덕적으로 허용가능한가?

이 두 시나리오로부터 수많은 변형된 시나리오가 개발되었고, 여러 가지 변형된 시나리오를 통칭하여 '트롤리 딜레마'로 부른다. 2000년대에는 심리학자들이 트롤리 딜레마에 관심을 가지게 되어 많은 수의 피험자를 대상으로 서베이를 통해 사람들의 직관을 알아보기 시작했다. 또 특정 상황에 대한 사람들이 생각을 할 때 뇌의 반응을 보는 인지신경과학에서는 이 심리학

1) 트롤리 문제는 Philippa Foot(1967)이 처음 제시하였고 Thompson(1976). Killing, Letting Die, and the Trolley Problem, *The Monist* 59, pp.204-217이 변형하여 제시하였다. Foot이 제시한 트롤리 문제에서는 의사결정을 내리는 주체가 트롤리 밖에 서 있는 제 3자가 아니라 트롤리를 운전하는 사람으로 기술된다.

실험에서 사람들이 반응할 때 뇌의 어떤 영역이 반응하는지를 보면서 사람들의 도덕판단이 어떤 기능과 관련되어 있는지를 알아보고 있다.

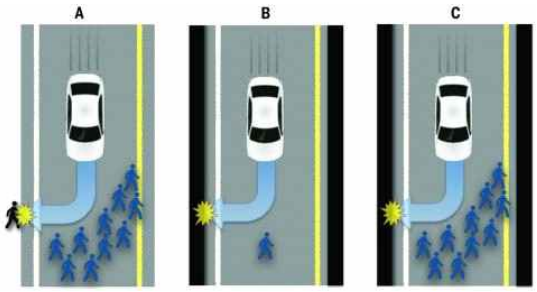
심리학자들의 서베이에서 밝혀진 사실은 사람들이 동일한 유형의 딜레마 사례에 다르게 반응한다는 것이다. 트롤리 사례에서는 피실험자들의 85%가, 인도교 사례에서는 12%만이 그런 행위가 도덕적으로 허용가능하다고 대답했다. 이러한 반응은 성별의 차이, 교육의 차이, 문화의 차이와 상관없이 공통적이었다. 여기서 제기되는 문제는 두 사례가 '다섯 사람을 구하기 위해 한 사람을 죽이는' 내용을 공통적으로 가짐에도 불구하고 피험자들이 왜 두 사례에 다른 반응을 보이는가 의 질문이다. 피험자들은 트롤리 사례는 도덕적으로 허용 가능한 것으로, 인도교 사례는 허용 불가능한 것이라 보는 도덕적 반응의 차이는 어떻게 설명될 수 있는가.

이렇게 트롤리 딜레마는 철학을 벗어나 심리학의 서베이를 통해 새롭게 조명되면서 사람들이 동일한 상황이라 할지라도 구체적 맥락에 따라 비일관적인 윤리적 선택을 한다는 것을 드러냈다. 선로가 어떤 구조로 되어 있는지, 선로에서 일하는 사람들이 나와 관련있는 사람들인지, 선로 변환기가 아니라 다른 방식으로 선택하는 경우인지 등 구체적 상황들에 따라 사람들은 윤리이론에 해당하는 공리주의적 선택을 하기도 하고 의무론적 선택을 하기도 했다.

트롤리 사고실험은 무인자동차와 무슨 관련이 있는가

이런 선택의 문제는 스스로 학습해서 상황을 인식하는 인공지능이 장착된 무인자동차에서도 나타난다. 무인자동차가 운전자의 희생과 타인의 희생 중 하나를 선택할 수밖에 없는 상황에 처할 경우의 선택 문제 이다. 2016년 ≪사이언스≫에 실린 "무인자동차의 사회적 딜레마(The social dilemma of autonomous vehicles)"라는 논문에서는 무인자동차가 유사 상황에서 어떤 선택을 해야 하는지 딜레마 상황과 사람들의 반응을 소개한다.

소개한 여러 딜레마 상황들 중 'C 딜레마' 상황을 중심으로 이야기해 보자. 무인자동차의 운행 방향에는 다수의 행인이 지나가고 있고 행인들에게 상해를 입히지 않기 위해 방향을 바꾸며 벽에 부딪쳐 운전자가 상해를 입게 된다. 이때 무인자동차는 어떤 결정을 내리도록 설계되어야 할까?



이 논문의 저자들은 무인자동차가 어떻게 작동해야 할지에 대해 설문조사를 했다. 이 C 상황에 대해 전체 답변자들 중 3분의 2가 넘는 76퍼센트의 답변자들은 보행자 열 명 대신 탑승자 한 명을 희생하는 쪽이 더 도덕적이라고 판단했다. 즉, 무인자동차의 인공지능이 공리주의적 결정에 입각해 딜레마 상황에 대처하도록 프로그래밍되어야 한다는 것이다. 이 대답은 앞서 트롤리 딜레마에 대한 사람들의 응답과 유사하다.

그런데 이런 대답은 제3자의 입장을 반영한다. 실제 본인이 그 무인자동차에 타고 있다면

어떨까? 연구자들은 두 번째 질문을 던져 보았고, 사람들은 입장을 바꿨다. 두 번째 질문은 "만약 본인이 선택한 기준으로 인공지능 설계를 법적으로 의무화한다면 동의하겠는가"였다. 이에 대해서 답변자의 대다수가 동의하지 않았다. 첫 번째 질문에 대해서는 제3자의 입장에서 공리주의적 답변을 지지했다면, 자신이 운전자가 되는 두 번째 질문에서는 공리주의적 결정을 더 이상 지지하지 않았다.

연구자들이 던진 세 번째 질문은 "자기 희생 모드의 무인자동차와 자기 보호 모드의 무인자동차가 판매될 경우 어떤 차를 구입하겠는가"였다. 이에 대해 사람들은 다수가 자기 보호 모드의 차를 선택했다. 이렇게 우리의 도덕적 직관은 비일관적이다. 소수보다는 다수의 생명을 구하는 것을 더 좋은 도덕 판단으로 선택하면서도 정작 본인은 그렇게 프로그래밍된 무인자동차를 선택하지 않는다. 사용자 보호 모드를 선호하는 것은 사용자에게는 자연스런 반응이다. 그런데 이렇게 프로그래밍된 무인자동차가 잘 팔릴까? 자기 보호 모드로 프로그래밍되면 사용자의 생명은 구하겠지만 타인에게 상해를 입히므로 이 역시 바람직한 상황은 아니다. 그렇다면 무인자동차 개발이 마지막 완성 단계에 다다를 때 그 의사 결정 기준은 어떻게 정해져야 하는가? 결정의 주체는 누구여야 하는가? 윤리 이론들 중 공리주의, 의무론, 덕 이론 혹은 다른 가치 기준으로 정해야 하는가? 이 같은 고민은 인공지능이 장착된 무인자동차가 부상하면서 새로이 제기된 문제는 아니며, 인간 차원에서도 해결하기 어려운 문제다.

그렇다면 다른 방식으로 문제해결할 방법은 없을까? 윤리이론이나 특정한 하나의 가치를 정해서 무인자동차에 입력하는 방식이 유일한 대안은 아니다. 윤리이론이나 인간의 가치는 인간이 같은 인간과 사회적으로 상호작용할 때 필요해서 제시된 것이기 때문에 단순한 사고실험을 넘어선 다양한 해결방식을 생각해볼 수 있다.

5. 윤리적 문제를 공학적으로 해결할 수는 없을까

무인자동차는 윤리적 문제를 이미 안에 포함하고 태어난 기술이다. 그래서 무인자동차의 윤리적 선택의 문제를 윤리이론의 선택이나 윤리적 분석으로 해결해야 한다고 생각하는 것이 자연스러울 수 있다. 앞서 소개한 여러 연구들은 우리의 그런 자연스런 반응을 기반으로 한 것이다.

그러나 무인자동차의 윤리적 문제를 꼭 윤리적으로 해결해야 하는 것은 아니다. 무인자동차의 운행은 교통신호나 보행자 뿐만 아니라 도시 설계 등과 관련하여 넓게는 사회체계, 행정, 등과 연결되어 있기 때문에 다양한 관점에서 접근할 필요가 있다. 중요한 것은 무인자동차가 사회에서 안전하게 문제없이 운행되고 인간 삶의 질을 높이는 것이다.

무인자동차의 문제는 무인자동차라는 인공지능 운영체계와만 관련되는 것이 아니라 넓게는 교통체계의 문제이므로, 한 예로 도시공학적으로 접근해볼 수 있다. 무인자동차와 다른 자동차 그리고 다른 보행자와의 관계가 문제가 되기 때문에 트롤리의 문제가 무인자동차에 해결해야 할 과제가 된다면, 어느 보행자를 선택할 것인가의 프레임으로 해결하려고 하지 말고 도로 형태를 기존과 다르게 완전히 바꾸면 된다. 그 한 예로, 무인자동차가 다니는 도로는 다른 자동차나 보행자가 보행할 수 없도록 하고 무인자동차 운행 도로, 일반 차 운행 도로, 보행자 도로 등을 완전히 분리해서 트롤리 딜레마와 같은 상황 자체를 발생하지 않도록 할 수 있다.[2)]

---

2) 한밭대 2017학년도 2학기 인공지능윤리 교과 수강학생들의 토론으로부터 발췌했음.

이런 해결책은 무인자동차의 윤리적 선택의 문제를 윤리적으로가 아니라 도로 구조를 새롭게 함으로써 해결하는 공학적 해결 사례이다. 한국의 세종시처럼 스마트 시티를 새로 구축하는 데에도 이런 통합적인 해결책을 적용할 수 있다. 인공지능이나 사물인터넷 기술을 전체 도시 운영의 많은 부분에 응용하므로 스마트 시티를 구성할 때 처음부터 인공지능과 관련한 윤리적 문제가 발생하지 않도록 설계할 수 있다.

윤리 딜레마가 무인자동차 설계에 이야기해주지 않는 점은 무엇인가

자율주행 자동차가 완성된다는 것의 의미는 단순히 0-4단계까지 있는 무인자동차기술의 완성이나 윤리적 선택의 문제의 해결 이상의 의미를 지닌다. 기술의 완성이나 윤리적 선택 문제의 해결은 기술이나 윤리 자체로는 온전히 해결하기 어렵다. 기술이나 윤리는 '사회'라는 구체적 맥락 속에서 사용되고 의미를 부여받기 때문에 사회를 구성하는 정책, 문화, 교육 등 거의 모든 요소들의 영향을 받고 영향을 준다. 따라서 이들 요소들 중 하나가 변화하면 기술이나 윤리 또한 영향을 받을 수 있다. 윤리적 선택의 문제는 무인자동차 설계에 중요한 요소이기는 하지만 전부는 아니며 통합적인 시각을 필요로 한다.

독일의 무인자동차 윤리 가이드라인을 떠올려보자. 왜 윤리적 선택의 문제를 차후의 문제로 미루었을까? 왜 무인자동차에 사용되는 인공지능 알고리즘이 누구를 살릴지를 선택하게 하면 안된다고 했을까? 무인자동차에 사용되는 알고리즘은 부호 기반의 규칙 계산 시스템이 아니라 실시간으로 업데이트되는 교통정보를 기계학습을 통해서 학습하고 그를 통해 의사결정 또한 새로이 업데이트하는 방식으로 정보처리를 한다. 따라서 애초에 여러 구체적 상황에서 누구를 살릴지를 입력한다고 해도 그러한 규칙들이 적용되지 않은 상황에서는 인간이 입력한 규칙과는 별도로 의사결정을 내릴 수 있다. 심층학습이라는 모델을 통해 인공지능이 작동하는 한 인간의 의도를 벗어난 의사결정을 내릴 가능성을 배제할 수 없다. 독일의 가이드라인은 이러한 위험성을 미리 배제하고자 한 것이다.

트롤리 딜레마, 그리고 트롤리 딜레마를 응용한 무인자동차에서의 선택문제는 인공지능으로 작동하는 여러 기술들이 윤리적 문제를 애초부터 가지고 있음을 잘 보여주었다. 그리고 이 문제를 어떻게 해결해야 하는가를 각 국가 교통국에서 열심히 고민하게 된 계기를 제공했다. 그런데 윤리적 문제는 트롤리 딜레마의 선택 문제에만 국한되어 있지 않다. 트롤리 딜레마에서 어떤 기준으로 선택을 하든 사람들의 선택은 하나가 아니라 다양하며 다수의 견해라는 이유만으로 그 선택을 알고리즘에 반영하는 것은 합리적 근거가 되기 어렵다.

그러면, 무인자동차의 선택 문제는 어떻게 생각해야 할까? 무인자동차는 단순히 이동수단의 문제일 뿐만 아니라 교통도로, 정책, 문화가 모두 포함된 복합적 문제로 접근하면 여러 가지로 접근이 가능할 것이다. 여기서는 그 중 하나의 시각과 접근만 소개하고 나머지는 각자의 근로 영역이나 전공에 따라 해결을 시도해볼 수 있고 이런 여러 시각들을 모아 여러 모델을 만들어 볼 수 있다.

6. 사회적 대응으로서의 인공지능윤리 교육

현재 한밭대에서 4학기 째 운영 중인 [인공지능윤리] 교과는 도덕, 윤리 개념보다는 '지속가능성'의 관점에서 출발하는 것이 특징이며, 딥러닝 + 인지과학/AI윤리 이슈 + 법적 변화 + 경영적 변화 로 구성되어 있다. 구체적 교과 내용은 다음과 같다.(한밭대학교 인공지능윤리 교과

2019년 2학기 수업계획서)

주	강의내용
1	강의소개
2	왜 인공지능윤리인가
3	인공지능의 딥러닝 구조 AI윤리 주요개념들
4	전자인격 논쟁 알고리즘 편향 설명가능성
5	책무성과 책임
6	뇌-컴퓨터 인터페이스와 정신의 경계
7	특강: 성로봇은 윤리적인가
8	중간고사
9	설계자, 제작자, 사용자 윤리
10	무인자동차 윤리
11	상향식, 하향식 인공도덕성 구현
12	인공지능의 법인격
13	인공지능의 법적 책임
14	인공지능과 지적재산권
15	기말고사

또한 알고리즘과 데이터에 내재되어 있는 인공지능윤리의 특성을 경험하고 훈련하기 위해, 다음을 프로젝트로 논의하고 구성해보도록 하고 있다.

학생들의 프로젝트: 선택한 인공지능기반 로봇 및 봇에 대해

1. 어떤 딥러닝 알고리즘으로 구현되는지
2. 어떤 데이터를 선택하여 인공지능에 학습시키는지, 그에 따라 결과가 어떻게 달라지는지
3. 데이터 선택에 편향이 필수적이라면 방안은?
4. 알고리즘 변수(요인)을 구성해보고 변수에 가중치를 매겨보기
5. 가중치를 달리 할 때 AI가 어떤 의사결정을 내리게 될지 추측해보기
6. 바람직한 가중치를 주는 것에서 개인차가 있는 이유를 생각해보기

참고문헌

김효은 (2019) [인공지능과 윤리] 커뮤니케이션 북스.
Foot, P. (1967). The problem of abortion and the doctrine of double effect. Oxford Review, 5, 5~15.
Greene, J. D. (2014). Moral tribes: Emotion, reason, and the gap between us and them. Penguin.
Huang, H. (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.
NHTSA (2017). Automated driving system: vision for safety, U.S. Department of Transportation, https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf
Federal Ministry of Transport and Digital Infrastructure (2017). Ethics commission: Automated and connected Driving, https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html?nn=187598