Ethics of Artificial intelligence

2014103473 수학과 석진우

2014103486 수학과 이종률

발표 목차

1. 도덕 개념의 역사

2. 도덕성 학습의 가능 여부

3. 인공지능에게 도덕성이란?

4. 도덕성의 의미 & 결론

도덕 개념의 역사

- 도덕성의 정의
  - 사회적 규범과 관습에 의해 판단되는 인간의 내면적 가치

- 신분제도와 노예제도의 발생 & 붕괴
  - 사회 구성원 다수의 합리성보다는 소수의 이익을 위해 실현된 제도
  - 인권이 확립되기 이전, 계급 사회에서 현대 사회로 넘어오며 성립
  - 사회 구성원 다수의 합리성 결여로 인해 근현대에 들어 붕괴

- 인권 확립과 법치주의의 확대
  - 사회 구성원 각자의 권리와 책임을 인지한 인권 개념의 확립
  - 국가 기초법인 헌법을 통해 인권의 법치적 보장
  - 사회 구성원 다수의 합리성이 보장된 '법'의 발생

도덕성 학습의 가능 여부

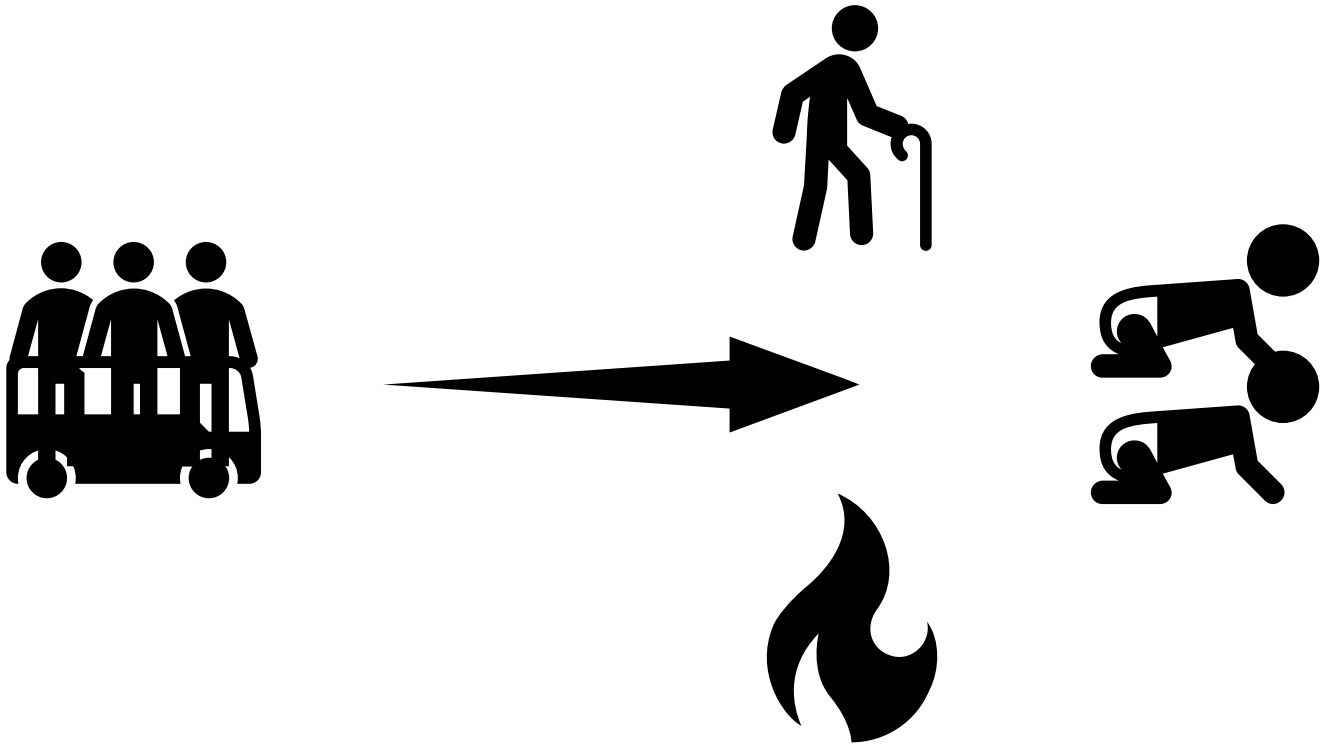
- 인공지능 학습 알고리즘
  - 현상과 지표 중심의 데이터 학습에 기반을 두고 있음
  - 학습의 옳고 그름(효율성)을 판단하는 기준 역시 인공지능 학습의 필요충분조건

- 인공지능 학습 데이터 = 법 + 판례
  - 법 = 개인의 내면적 가치를 넘어선 다수의 의견이 반영된 사회적 합의
  - 판례 = 사회적 인식에 기반을 둔 도덕성에 대한 다양한 사례를 접할 수 있는 사례

- 법과 문화의 주관성
  - 미국의 보복성 범죄와 한국의 보복성 범죄의 차이
  - 데이터 학습 기준에 따라 도덕성의 판단 기준이 매번 달라질 수 있음

인공지능에게 도덕성이란?

- 인공지능 자율주행 자동차의 사고 시뮬레이션 中



도덕성의 의미 & 결론

- 도덕성의 두 가지 구성 요소
  - 행동 주체의 자율성
  - 행동 주체가 소속되어 있는 사회의 성질
  - 도덕성 = 자율성과 사회성의 복합적 작용

- 법과 판례 데이터에 근거한 인공지능 학습
  - 불법인 데이터들에 대한 능동적인 학습 가능
  - 인공지능이 소속되어 있는 사회의 법적 문화적 데이터 습득 → 사회의 구성 요소로서 인정 가능

- 인공지능의 자율성
  - 데이터의 습득과 예측에 초점이 맞춰져 있는 인공지능에게 자율성을 기대하기는 어려움.
  - 자율성을 습득한다 해도 랜덤변수로 작동되는 이상 무의미
  - 사회 규범화 과정은 가능하지만 도덕적 판단의 자율성이 결여 → 진정한 의미의 도덕성 부여는 어렵다

EOD