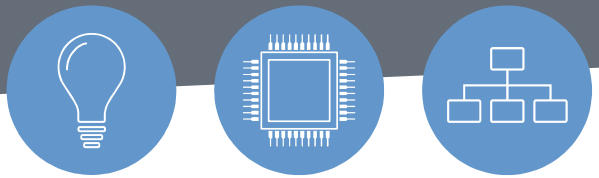
인공지능 사회적 영향, 윤리적 접근

2023년 3월 7일

목차

인공지능 사회적 영향 (2) – 역기능



인공지능 신뢰성

인공지능 취약점

인공지능 위협요소

인공지능 신뢰성

▪ 인공지능이 내린 결과에 대한 신뢰성

- 인공지능이 내놓는 결과는 과연 신뢰할 만한가?

- 신뢰성 이슈 사례
  - 우버, 테슬라의 자율주행차 사고
  - IBM Watson의 폐암 진단율 17.8%
  - 교통 표지판 오인식 사례

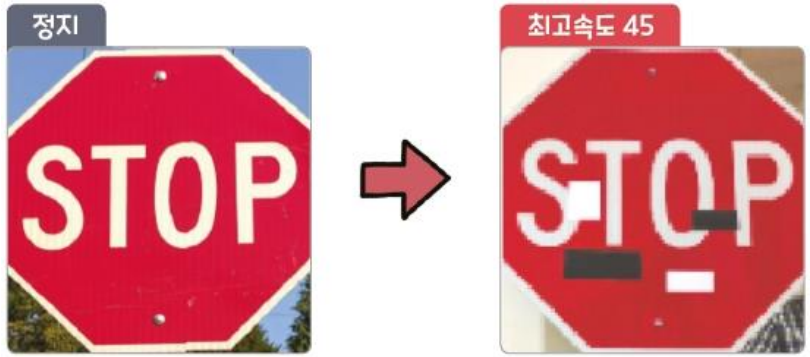


그림 4-19 '정지'를 '최고속도 45'로 잘못 인식한 인공지능

▪ 인공지능의 학습모델의 정확도 이슈

- 주어진 데이터의 불완전성
  - 학습모델의 정확도 100% 보장 어려움
  - 편향 발생 가능성도 있음

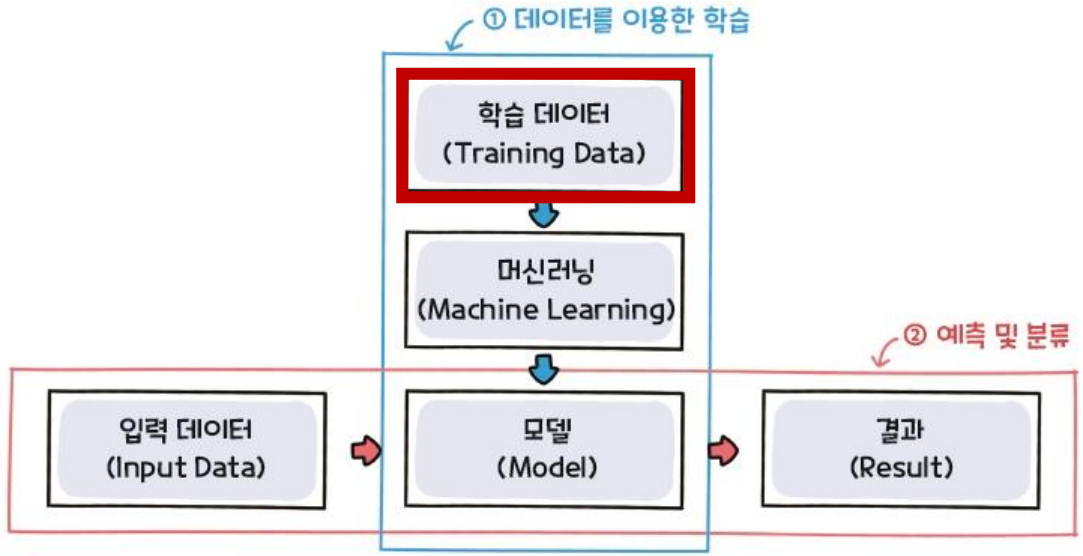


그림 4-8 인공지능 학습 방법

인공지능 신뢰성 대응 방안

▪ 인공지능에서의 블랙박스 문제

- 불확실한 데이터 학습 과정
  - 인공지능의 학습과정을 인간이 이해하기 어려움

- 인공지능 시스템의 판단 근거 및 결정 과정에 대해 설명할 수 없는 상태

그림 4-11 인공지능 블랙박스 문제

▪ 설명가능 AI (XAI, eXplanable AI)

- 데이터, 학습과정과 예측 과정을 시각화

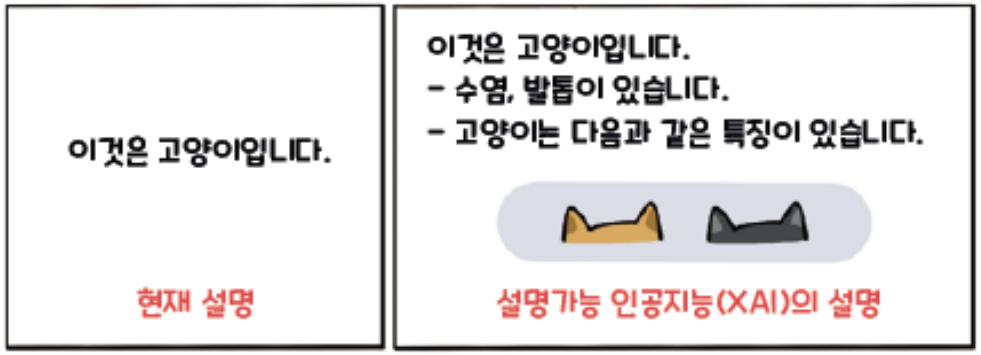


그림 4-20 설명가능 인공지능의 설명 방법

- 현실적으로 정확성은 떨어지더라도 그 과정을 설명하여 대응가능하도록 함

인공지능 취약점 및 대응 방안

- 데이터 변조, 악의적 데이터 주입
  → 잘못된 판단/행동 유발

- 데이터 탈취
  → Privacy, 기업 기밀 노출 이슈

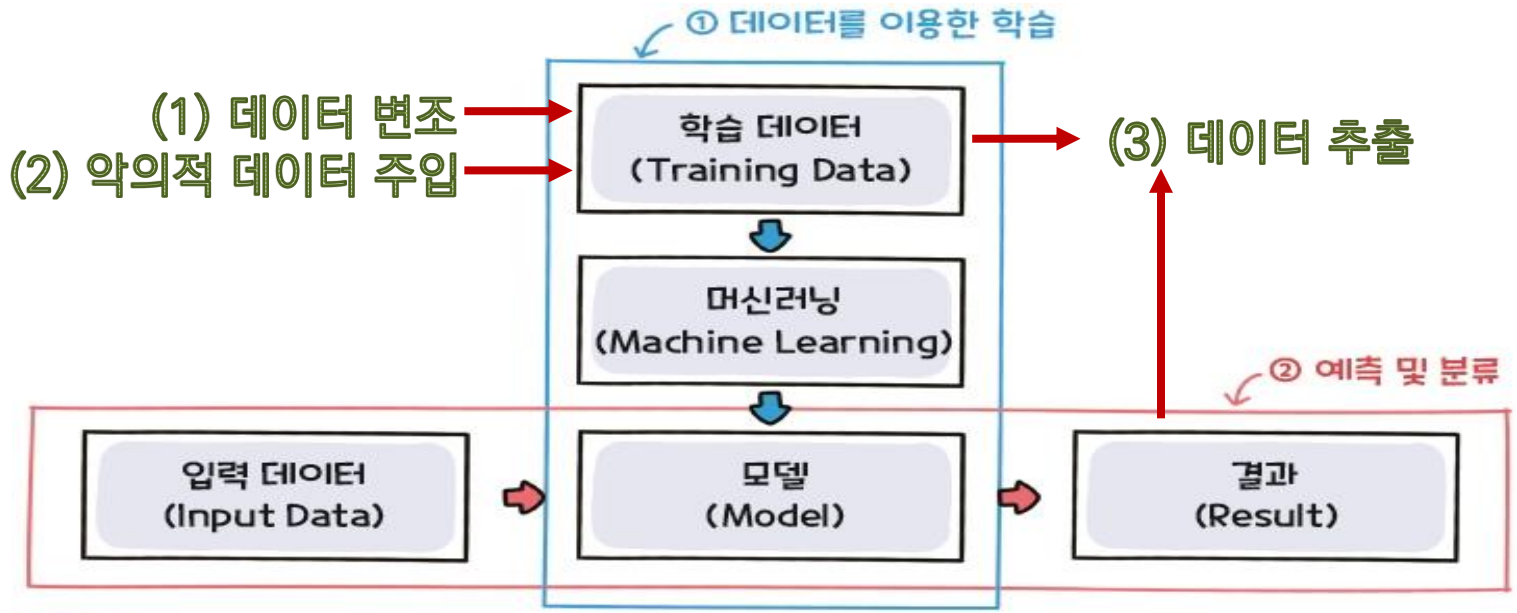


그림 4-8 인공지능 학습 방법

- 대응안
  - 변조된 데이터, 악의적 데이터 학습하여 식별
  - 모델 학습 후 데이터 파기

- 대응안
  - 데이터변환, 개인정보 비식별화, 연합학습 활용
  - 질의 횟수 조정

인공지능 위협요소

▪ 프라이버시 침해

- 개인정보를 대량 수집하여 인공지능 학습에 활용
  - 예) 수억 개의 CCTV로부터 특정 개인의 위치와 상태를 감시하는 등 프라이버시 침해 가능

그림 5-6 CCTV 데이터로 인한 프라이버시 침해

▪ 인간의 일자리/소득에 대한 위협

- 인간의 노동/업무를 인공지능이 대체
  → 일반 근로자의 소득 감소
- 인공지능의 창의적 활동 : 미술, 문학, 음악
- 인공지능 기술 독점 → 부익부 빈익빈 현상

▪ 인간의 존엄성 파괴

- 인공지능 특이점(Singularity)
  - 인공지능이 인간의 지능을 초월하는 시점으로, 사람이 기술의 발전을 따라잡을 수 없는 시기

그림 3-1 인공지능 특이점

- 인공지능이 더 이상 인간의 명령을 따르는 않고, 위험한 인격성을 가진 책임 주체가 될 가능성이 있음
  - 영화 '2001 스페이스 오디세이', 'AI'

인공지능 위협요소 대응 방안

▪ 프라이버시 보호

- 정책과 규범 수립
  - 인공지능이 보유한 개인정보는 다른 인공지능에서도 활용될 수 있으므로 정보에 대한 파기 및 이동에 대한 규율이 필요
  - 국내 데이터3법, 마이데이터법 등
- 기술적인 개인정보보호
  - 인공지능을 위한 서비스 개발 시 설계 단계부터 개인정보를 보호할 수 있는 방안(Privacy By Design)이 적용되어야 함.
- 데이터의 투명성
  - 완벽한 데이터 보안은 불가능하므로 정보의 주체자인 개인이 정보의 흐름을 확인할 수 있어야 하며, 언제든지 삭제할 수 있는 기술적·제도적 방안이 마련되어야 함

▪ 명확한 책임 소재

- 인공지능 시스템의 법적 지위 부여 이슈
  - 유럽의회 결의안 통과 → AI 로봇, AI 법학, AI 윤리전문가 162명 반대 공개 서한
- 인공지능 제조사 및 시스템 개발자, 사용자가 책임을 부담해야하는 상황
- → 인공지능으로 인간이 희생되기 전, 대응이 가능해야 함

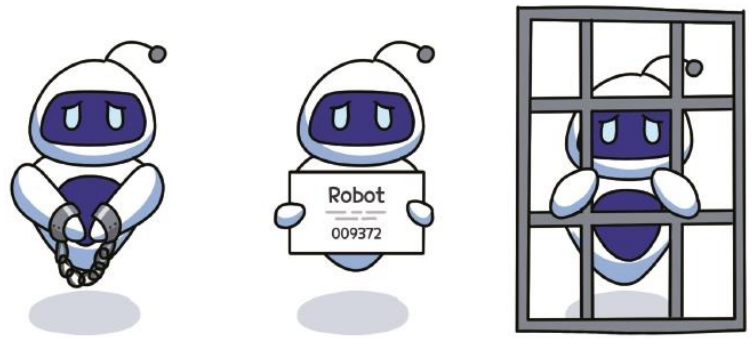


그림 5-14 인공지능 관련 법의 필요성

인공지능 윤리적 접근

인공지능 윤리적 이슈

인공지능 윤리규범 제정 현황

인공지능 윤리적 이슈

▪ 윤리적 편향

- 사례 1 : 인공지능 챗봇 '이루다'
  - 일부 사용자들이 '이루다'의 학습 능력을 악용해 부적절한 단어들을 주입
  - '이루다'가 혐오 발언을 가감 없이 내놓는 사태 발생

- 사례 2 : 재범률을 예측하는 프로퍼블리카(ProPublica)
  - 흑인의 재범률을 백인에 비해 실제보다 더 높게 추론

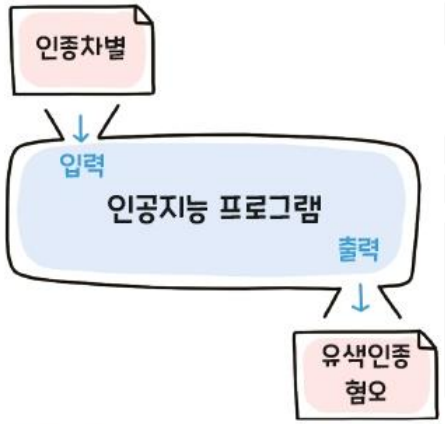


그림 3-15 인공지능의 윤리 문제

- 사례 3 : 아마존(Amazon)의 채용 인공지능
  - 여성차별 문제가 불거지면서 프로그램을 자체 폐기

▪ 윤리적 딜레마 – 트롤리 딜레마

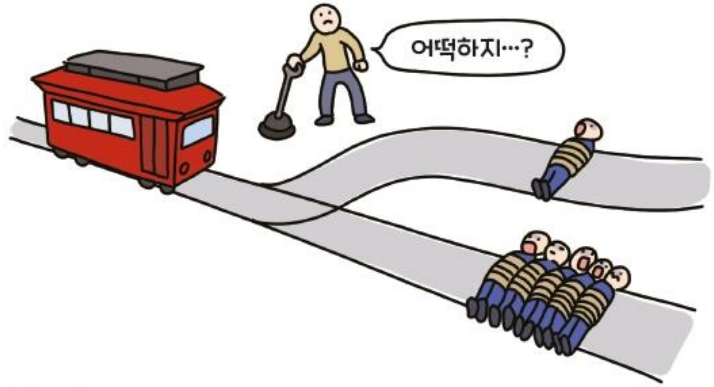


그림 3-16 트롤리 딜레마

- 자율주행에서의 트롤리 사례

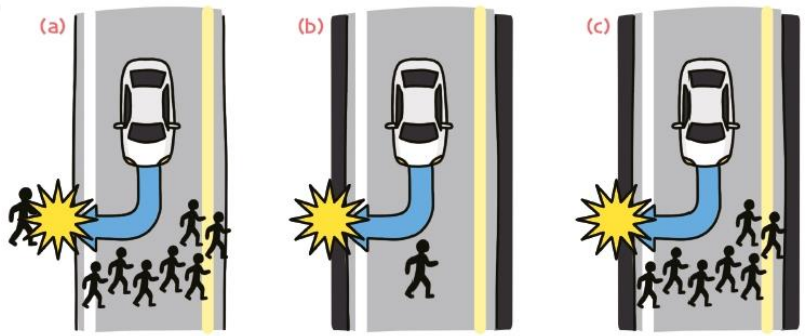


그림 3-18 자율주행차의 트롤리 사례

인공지능 윤리규범 제정 현황

▪ 인공지능 윤리 주요 사례

- 국가적 차원
  - 국내 '인공지능 윤리 기준'
  - 미국 알실로마 AI 3분류 측면 23가지 원칙
  - EU 신뢰할 수 있는 인공지능 윤리 가이드라인 (적법성, 윤리성, 견고성)
- 기업적 차원
  - 카카오 인공지능에 관한 윤리 규범 제정
  - 구글 인공지능 개발 및 활용에 지켜야할 권고사항 6가지 항목
  - MS 책임 AI 원칙

▪ 국내 '인공지능 윤리 기준'

- 최고가치 : 인간성 (Humanity)
  - 인간성을 위한 인공지능(AI for Humanity)을 위한 3대 원칙·10대 요건 제시 (인공지능 전 과정에서 고려/만족 필요)
- 3대 기본원칙: ❶ 인간의 존엄성 원칙, ❷ 사회의 공공선 원칙, ❸ 기술의 합목적성 원칙
- 10대 핵심요건 : ① 인권 보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성

[참고] 인공지능 관련 원칙

아실로마 인공지능 원칙 (Asilomar AI Principles)

- 2017년, 캘리포이나 아실로마에서 열린 AI Conference에서 채택
- 인공지능 연구의 목적은
  인간에게 유용하고 혜택을 주어야 하며,
  인간의 존엄성/권리/자유/이상 등과 양립할 수 있어야 하며,
  장기적으로 위험에 대응하고 공동의 이익을 위해 활용되어야 한다는 원칙
  - 연구이슈(5), 윤리가치(13), 장기이슈(5)

로봇 3원칙 (The Three Laws of Robotics)

- 아시모프 〈라이어!〉 (1941년)에서 제시

그림 3-22 로봇 3원칙

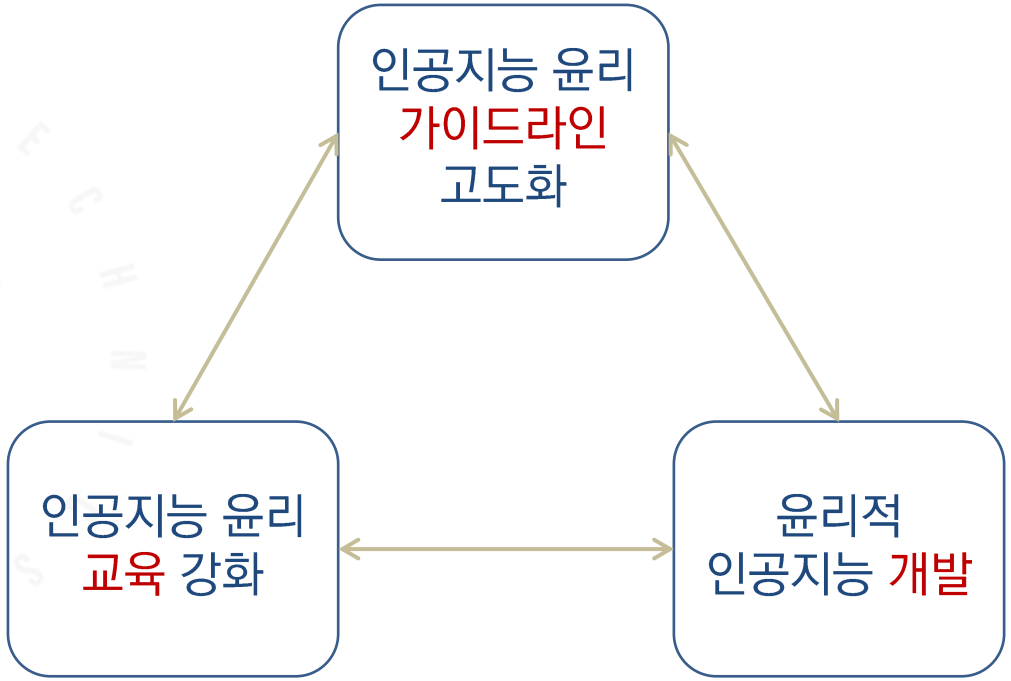
- 〈로봇과 제국〉 (1985년) 제0원칙 추가
  - 로봇은 인류에게 해를 가할 만한 명령을 받거나 행동을 하지 않음으로써 '인류'에게 해가 가해지는 것을 방치해서도 안 된다 (제1원칙의 확장)

인공지능 안전성 및 윤리 확보

인공지능 안전성 및 윤리 확보

- 인공지능의 기술·구조적 한계 등으로 발생가능한 의도치 않은 각종 위험들에 대비

- 인공지능 윤리 추진 방향



수고하셨습니다~~~ ^^