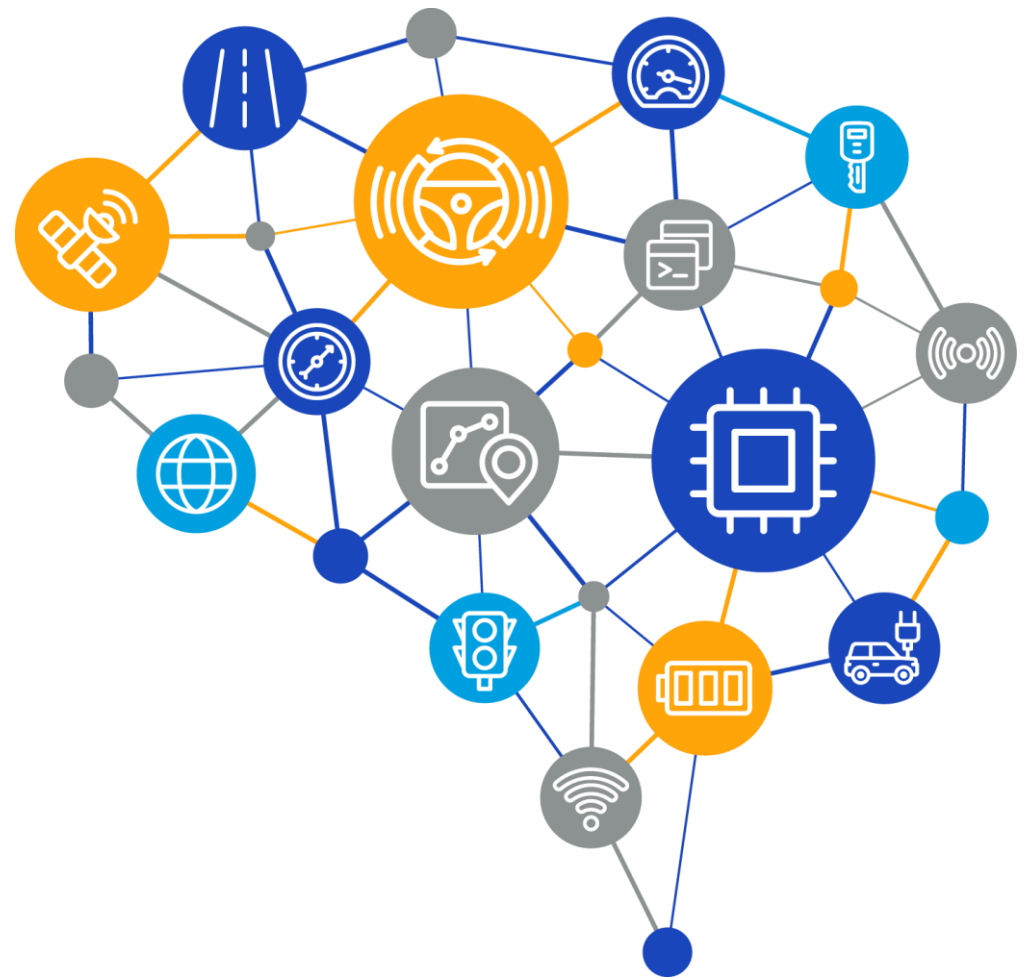


인공지능 윤리

AI Ethics

2024.11

강환수 교수



AI Experts Who Lead The Future

CONTENTS

- 01 | 인공지능 윤리의 필요성
- 02 | 인공지능 윤리적 딜레마
- 03 | 인공지능 윤리안

AI Experts
Who Lead
The Future

01

인공지능 윤리의 필요성

• 주목받는 인공지능 윤리 (1/2)

사례1: 인공지능 챗봇 '이루다'

- ✓ 일부 사용자들이 '이루다'의 학습 능력을 악용해 부적절한 단어들을 주입하였습니다.
- ✓ '이루다'가 혐오 발언을 가감 없이 내놓는 사태 발생하였습니다.
- ✓ 서비스 시작 20일 만에 서비스를 중단하게 되었습니다.

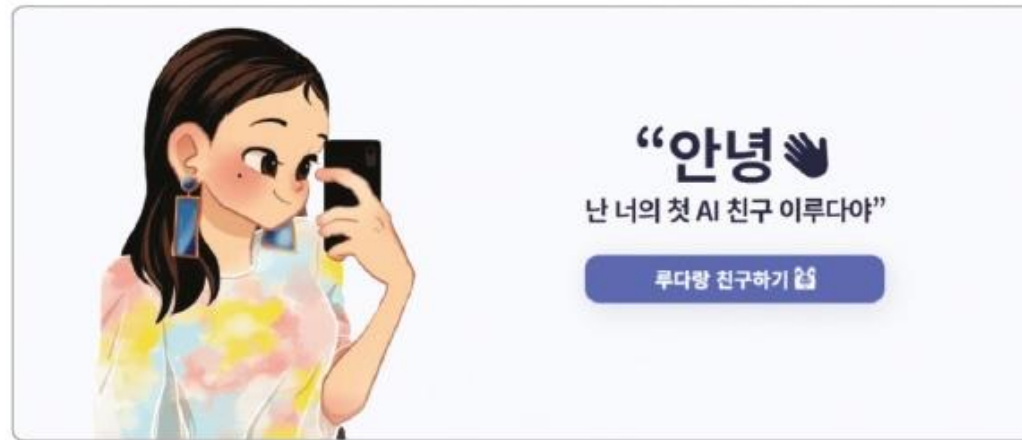


그림 3-11 인공지능 챗봇 '이루다'

[사진출처] 난생처음 인공지능 입문 (출판사: 한빛아카데미)

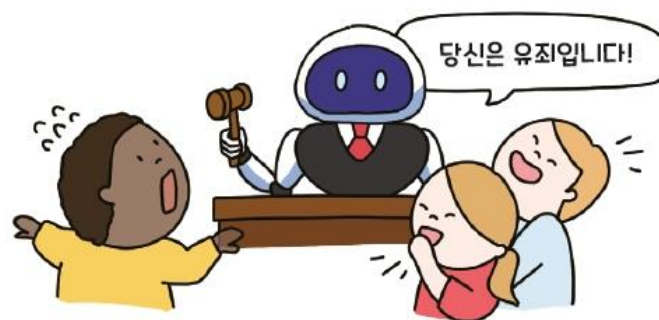
• 주목받는 인공지능 윤리 (2/2)

사례2: 재범률을 예측하는 프로퍼블리카 (ProPublica)

- ✓ 인공지능이 흑인의 재범률을 백인에 비해 실제보다 더 높게 추론하였습니다.

사례3: 아마존 (Amazon)의 채용 인공지능

- ✓ 인공지능을 활용한 채용 프로그램의 여성차별 문제가 불거지면서 프로그램을 자체 폐기하였습니다.



(a) 인공지능 판결



(b) 인공지능 면접

그림 3-12 인공지능을 테스트하면서 발생한 문제

[사진출처] 난생처음 인공지능 입문 (출판사: 한빛아카데미)

- 인공지능 윤리의 필요성 (1/3)

- Microsoft 최고경영자인 사티아 나델라 (Satya Nadella)는 인공지능 윤리에 대한 화두를 던졌습니다.

“인공지능 활용에 앞서 윤리가 우선시되어야 한다.”



그림 3-13 마이크로소프트의 최고경영자 사티아 나델라

- 인공지능 윤리의 필요성 (2/3)
 - 구글의 이미지 인식 (Image Recognition) 사례
 - 이미지 인식 중 흑인 여성을 고릴라로 인식
 - 위챗 (WeChat) 번역 과정 사례
 - 니그로 (Negro)라는 단어를 사용

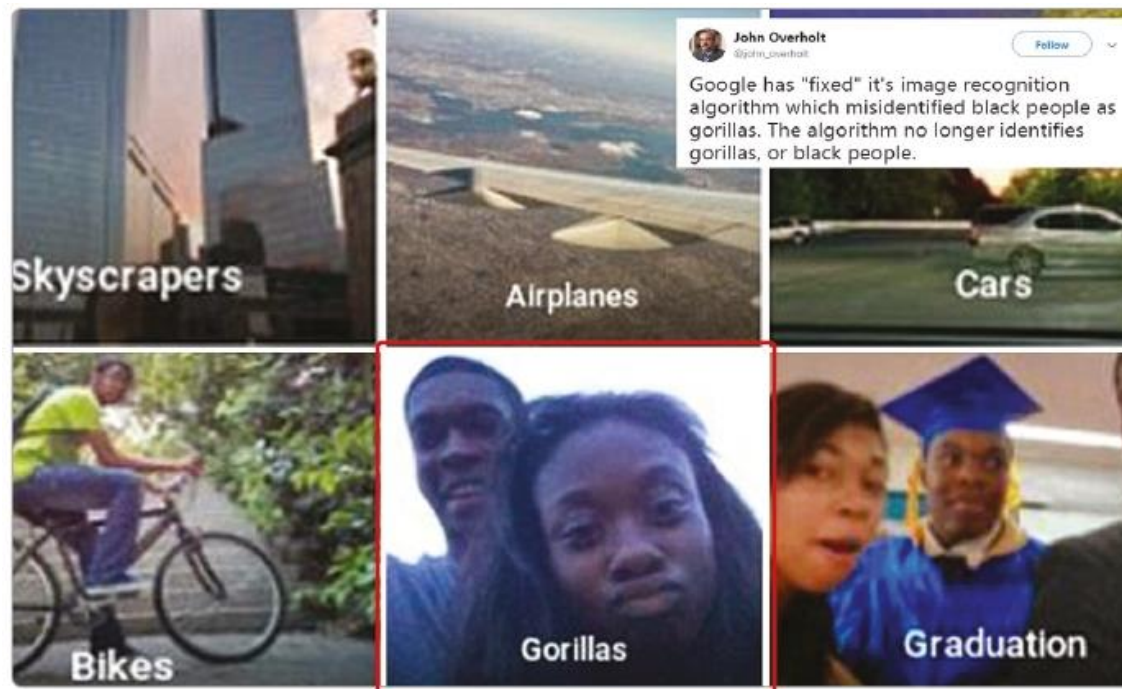


그림 3-14 구글의 이미지 인식 오류 사례

[사진출처] 난생처음 인공지능 입문 (출판사: 한빛아카데미)

• 인공지능 윤리의 필요성 (3/3)

- MIT 미디어랩의 발표 자료

- 인공지능이 백인 남성 얼굴을 인식하는 과정에서 오류를 일으킬 확률은 1%
- 흑인 여성의 경우 오류 발생 확률이 35%까지 상승

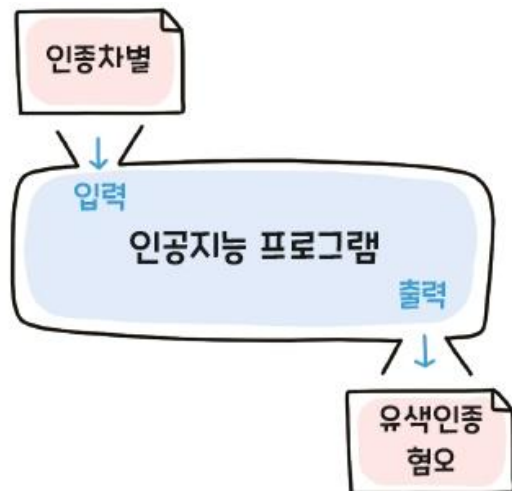


그림 3-15 인공지능의 윤리 문제

인공지능 알고리즘은 어떤 데이터를 입력하는지에 따라 결과가 달라집니다.

인공지능에게 어떠한 데이터를 주입할 것인지는 바로 인간의 몫입니다.

AI Experts
Who Lead
The Future

02

인공지능 윤리적 딜레마

• 트롤리 딜레마 (Trolley Dilemma)

- 윤리학 분야의 사고실험 중 하나입니다.
 - 다섯 사람을 구하기 위해 한 사람을 죽이는 것이 도덕적으로 허용 가능한지에 대한 질문
- 트롤리 딜레마와 관련된 대표 사례
 - 트롤리 사례
 - 육교 사례

“ 트롤리버스 ” (Trolleybus)

차량 내부에 탑재된 연료를 이용하는 것이 아닌,
외부의 전기를 직접 받아 이것을 연료로
이용하여 운행하는 버스를 뜻합니다.



[사진출처] https://en.wikipedia.org/wiki/Trolleybus#/media/File:Tr%C3%A5dbuss_Landskrona.JPG

• 트롤리 딜레마 (Trolley Dilemma): ① 트롤리 사례

가정하는 상황

- ✓ 트롤리 전차가 철길 위의 5명의 인부들을 향해 빠른 속도로 돌진하고 있습니다.
- ✓ 당신 옆에 트롤리의 방향을 바꿀 수 있는 레일 변환기가 있습니다.
- ✓ 트롤리의 방향을 왼쪽으로 바꾼다면 왼쪽 철로에서 일하는 1명의 인부가 사망합니다.
- ✓ 트롤리의 방향을 바꾸지 않는다면 5명의 인부들이 사망합니다.

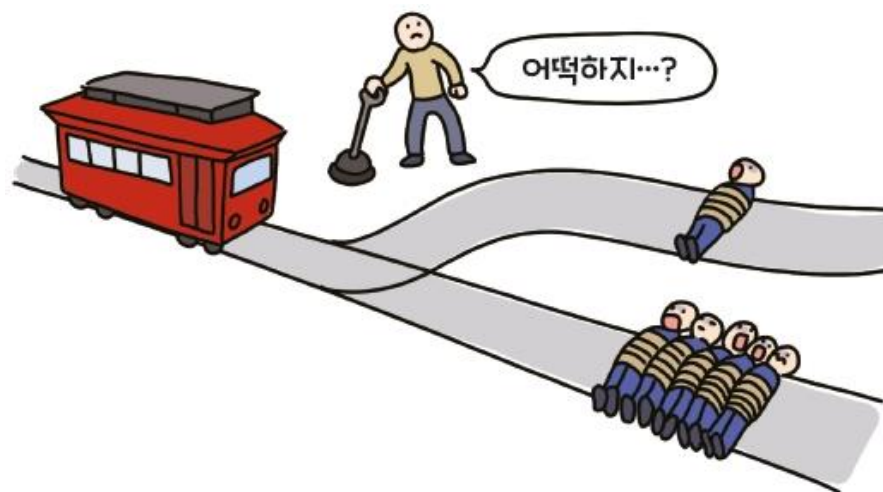


그림 3-16 트롤리 딜레마

[사진출처] 난생처음 인공지능 입문 (출판사: 한빛아카데미)

• 트롤리 딜레마 (Trolley Dilemma): ② 육교 사례

가정하는 상황

- ✓ 철길 위 5명의 인부들을 향해 돌진하고 있는 트롤리를 육교에서 보고 있습니다.
- ✓ 당신 옆에 몸집이 큰 사람이 있는데, 전차를 세우려면 이 사람을 육교 아래로 떨어뜨려야 합니다.
- ✓ 떨어진 1명의 사람은 죽겠지만, 철길 위의 5명의 인부들의 목숨은 구할 수 있습니다.



그림 3-17 육교 사례

[사진출처] 난생처음 인공지능 입문 (출판사: 한빛아카데미)

• 인공지능이 처할 수 있는 윤리적 딜레마 (1/2)

자율주행 자동차의 트롤리 사례

- ✓ (a) 여러 사람이 희생되는 것과 한 사람이 희생되는 것 중에 어떤 판단을 내려야 할까요?
- ✓ (b) 자율주행 자동차는 그냥 보행자를 치고 지나가야 할까요?
아니면 운전자가 다치게끔 방향을 꺾어야 할까요?
- ✓ (c) 여러 사람의 목숨과 운전자의 목숨 중, 자율주행 자동차는 어느 쪽에 더 비중을 두고 판단을 내려야 할까요?

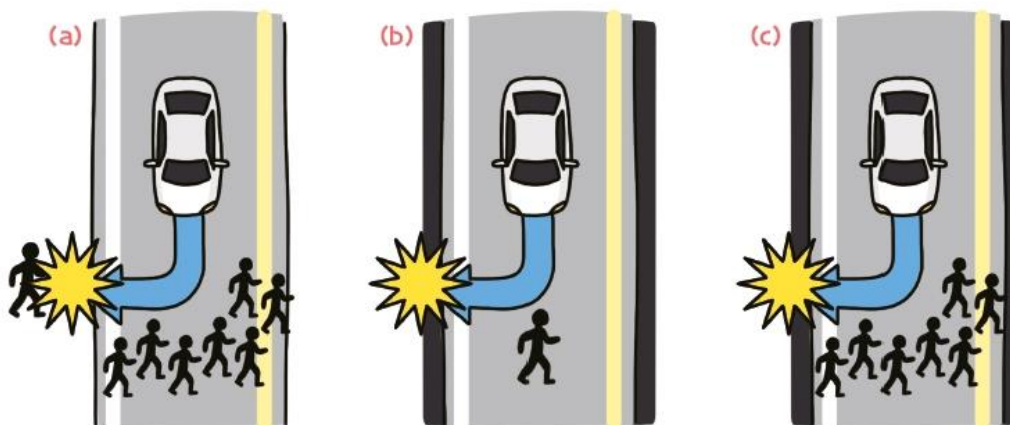


그림 3-18 자율주행차의 트롤리 사례

[사진출처] 난생처음 인공지능 입문 (출판사: 한빛아카데미)

• 인공지능이 처할 수 있는 윤리적 딜레마 (2/2)

자율주행 자동차 시대에 맞닥뜨리게 될 가장 기본적인 윤리적 이슈
더 늦기 전에 알고리즘의 윤리성에 대해 고민해 볼 필요가 있습니다.



그림 3-19 사람과 자율주행차에 대한 이중 잣대

AI Experts
Who Lead
The Future

03

인공지능 윤리안

아실로마 인공지능 원칙(Asilomar AI Principles) (1/2)

인공지능 활용 Python language

- 인공지능 개발의 목적, 윤리, 가치 등에 대해 개발자가 지켜야 할 23가지 준칙입니다.
- 이 원칙은 총 3가지 부분으로 구성되었습니다.
 - ① 연구 관련 쟁점, ② 윤리와 가치, ③ 장기적 이슈



그림 3-20 아실로마 인공지능 원칙

아실로마 인공지능 원칙(Asilomar AI Principles) (2/2)

인공지능 활용 Python language

① 연구 관련 쟁점

- ✓ 연구 목표, 연구비 지원, 과학정책 연계, 연구 문화, 경쟁 회피 등

② 윤리와 가치

- ✓ 안전, 실패의 투명성, 사법적 투명성, 책임성, 가치 일치, 인간의 가치, 개인정보보호, 자유와 프라이버시, 이익의 공유, 번영의 공유, 인간 통제, 사회전복 방지, 인공지능 무기 경쟁 지양 등

③ 장기적 이슈

- ✓ 역량 경고, 중요성, 위험성, 자기개선 순환, 공동의 선 등

로봇 3원칙(The Three Laws of Robotics) (1/3)

인공지능 활용 Python language

- 로봇이 반드시 따라야 할 3가지 원칙입니다.
- 로봇 3원칙을 제시한 작가 아이작 아시모프 (Issac Asimov)는 이 원칙들만 잘 지킨다면 로봇이 인간에게 위협이 될 일은 없을 것이라고 생각하였습니다.

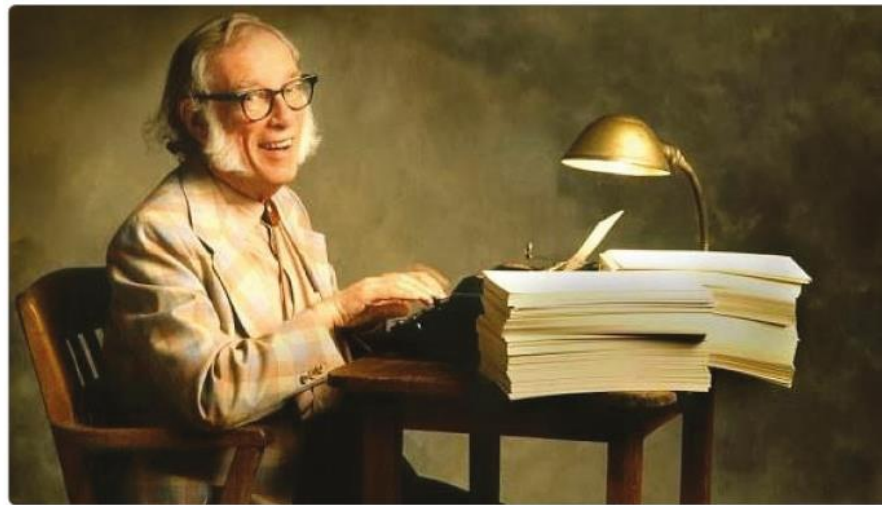


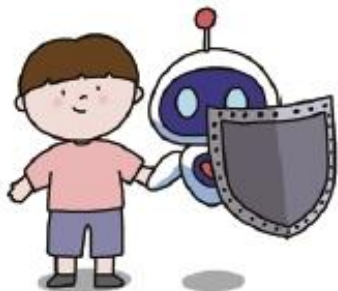
그림 3-21 로봇 3원칙을 제시한 아이작 아시모프 작가

로봇 3원칙(The Three Laws of Robotics) (2/3)

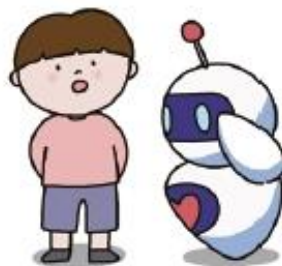
인공지능 활용 Python language

로봇 3원칙의 내용

- 제1원칙 : 로봇은 인간에게 해를 입혀서는 안 되고, 위험에 처한 인간을 방치해서도 안 된다.
- 제2원칙 : 제1원칙을 어기지 않는 한, 로봇은 인간의 명령에 복종해야 한다.
- 제3원칙 : 제1원칙과 제2원칙을 어기지 않는 한, 로봇은 로봇 자신을 지켜야 한다.



제1원칙 로봇은 인간에게 해를 입혀서는 안 되고, 위험에 처한 인간을 방치해서도 안 된다.



제2원칙 제1원칙을 어기지 않는 한, 로봇은 인간의 명령에 복종해야 한다.



제3원칙 제1원칙과 제2원칙을 어기지 않는 한, 로봇은 로봇 자신을 지켜야 한다.

그림 3-22 로봇 3원칙

로봇 3원칙(The Three Laws of Robotics) (3/3)

인공지능 활용 Python language

- 제0원칙을 추가로 제안
 - 이후 아이작 아시모프는 단편소설인 『로봇과 제국(Robots and Empire)』

제0원칙

- ✓ 로봇은 인류에게 해를 가할 만한 명령을 받거나 행동을 하지 않음으로써 '인류'에게 해가 가해지는 것을 방지해서도 안 된다 (제1원칙의 확장).