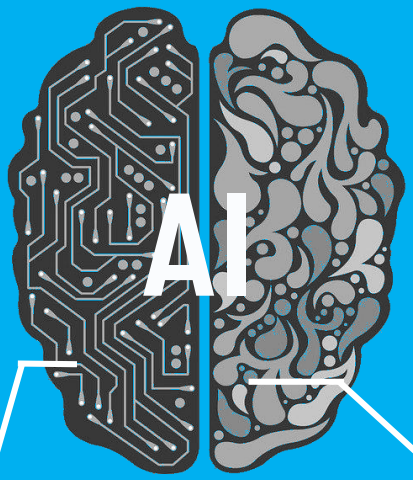
인공지능과 풀어야 할 사회/윤리 문제

서창호
KAIST 전기 및 전자공학부
KAIST AI Institute 부원장

2021년 10월 27일
On세상 지식탐구생활


AI

비서

튜터

채용 심사

형량 결정

대출 심사

두 가지 질문

1. AI 과연 무엇이길래?

2. AI 시대, 우리가 풀어야 할
   사회/윤리 문제?

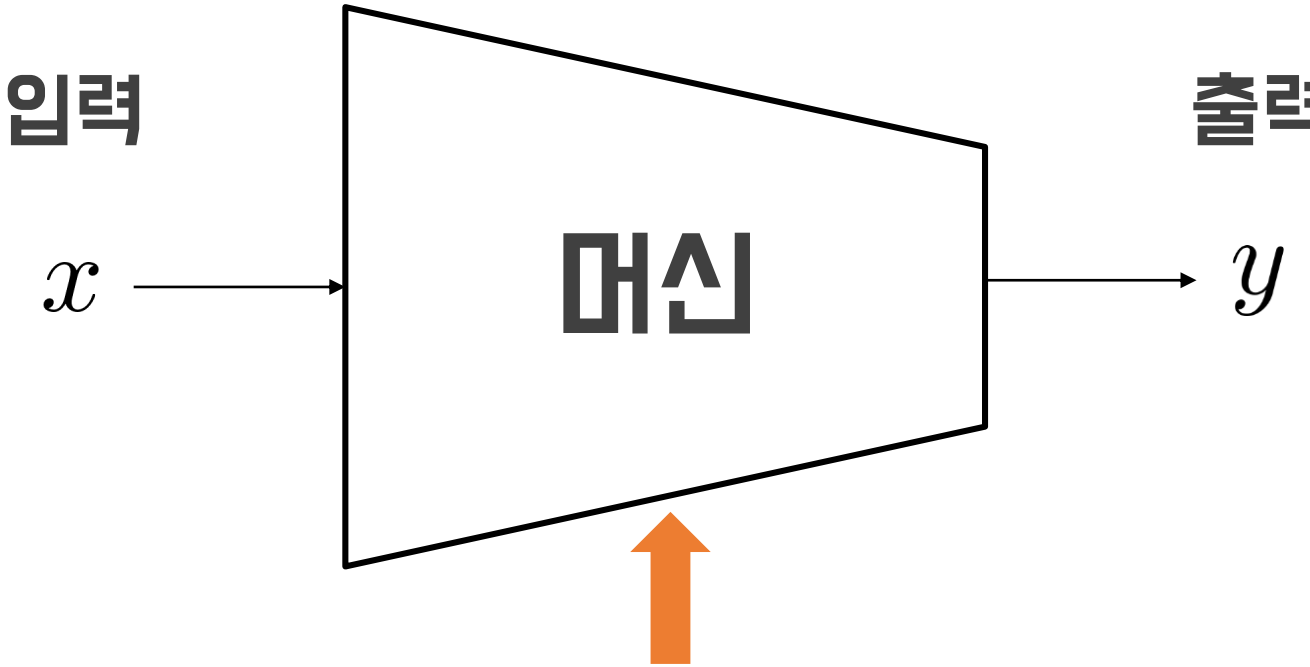
본 강연에서는 ...

1. AI의 기본 개념
   두 가지 키워드: 머신러닝, 딥러닝
2. 사회/윤리 문제 #1
   사례: 자율주행 AI
3. 사회/윤리 문제 #2
   사례: 신뢰할 수 있는 AI

AI의 기본 개념

1. 머신러닝
2. 딥러닝

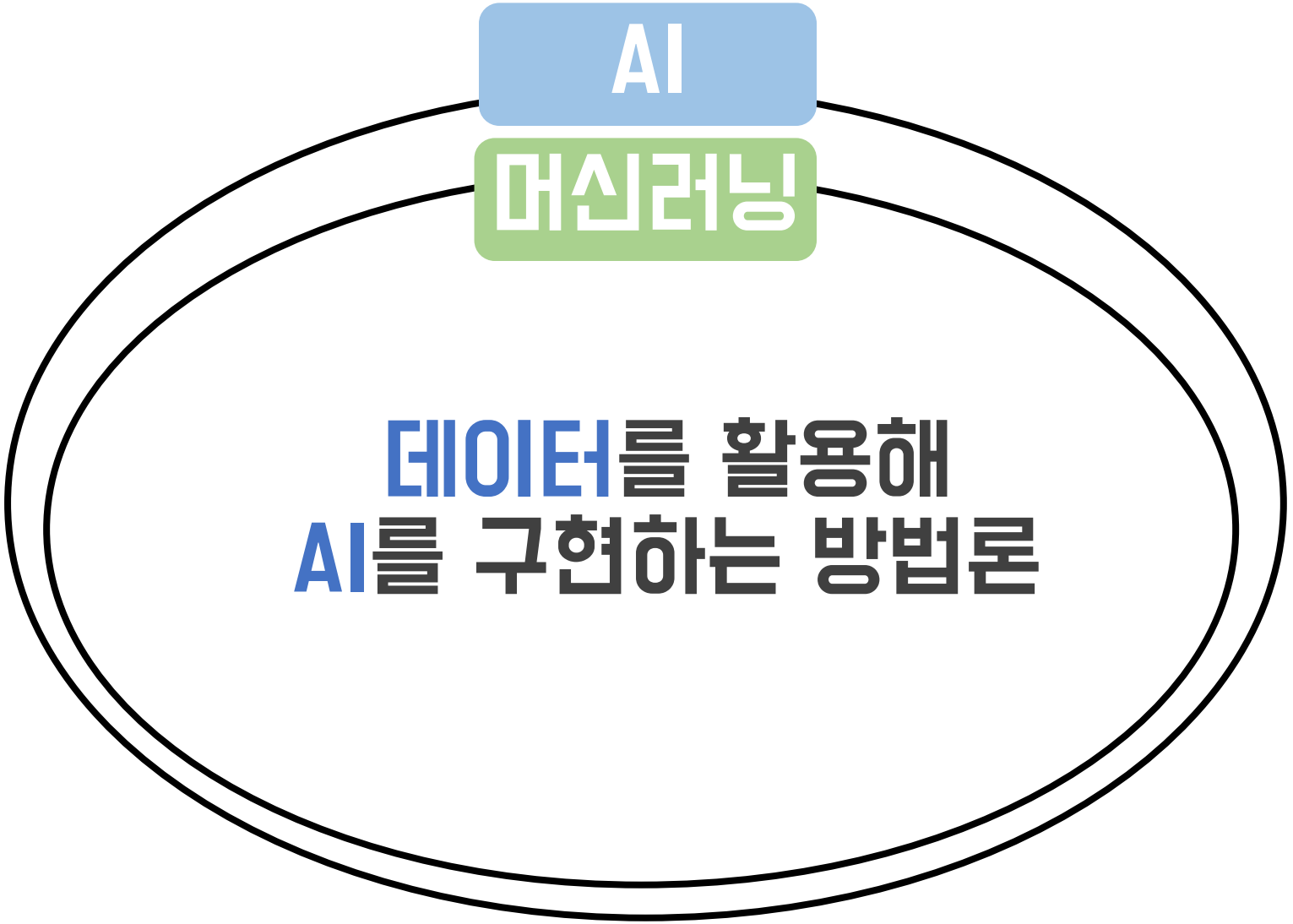
머신러닝



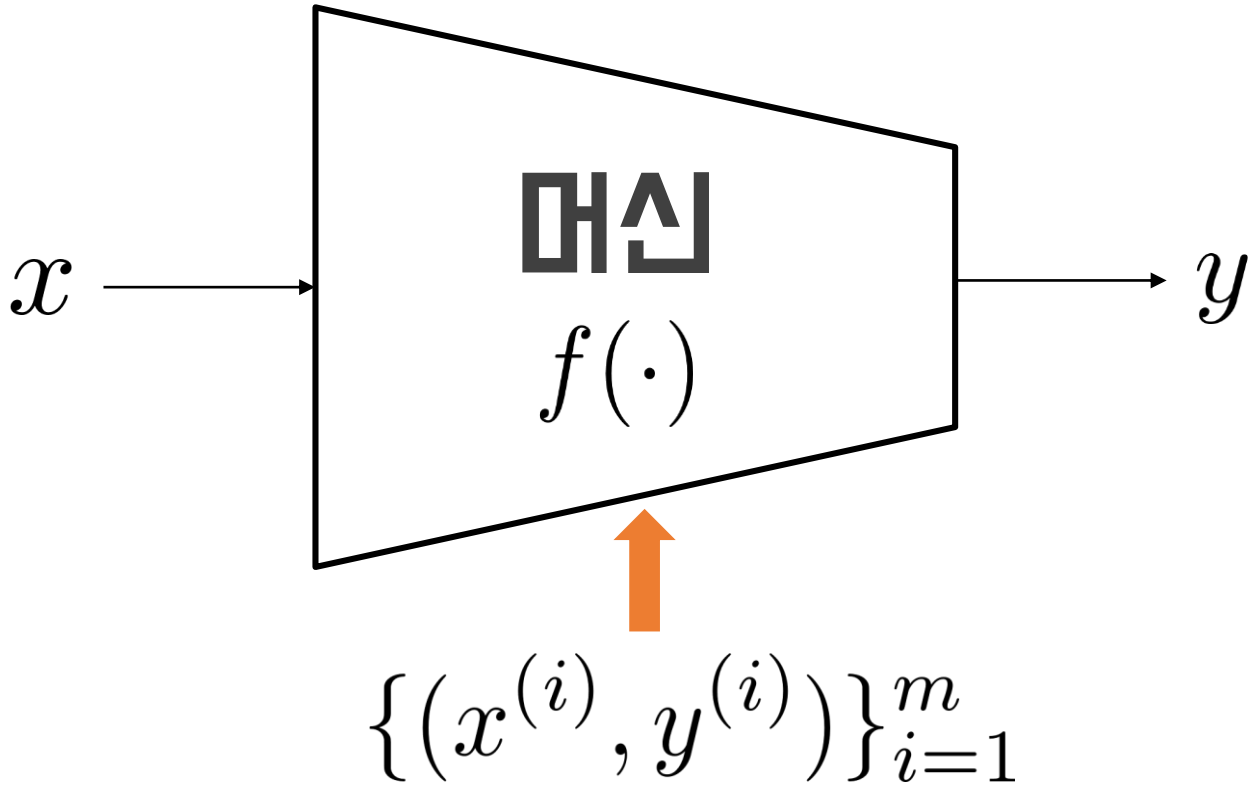
$\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$ (입출력 예시)

"데이터"

머신러닝 vs AI



머신러닝: 학습방법



최적화를 통해!

최적화란?

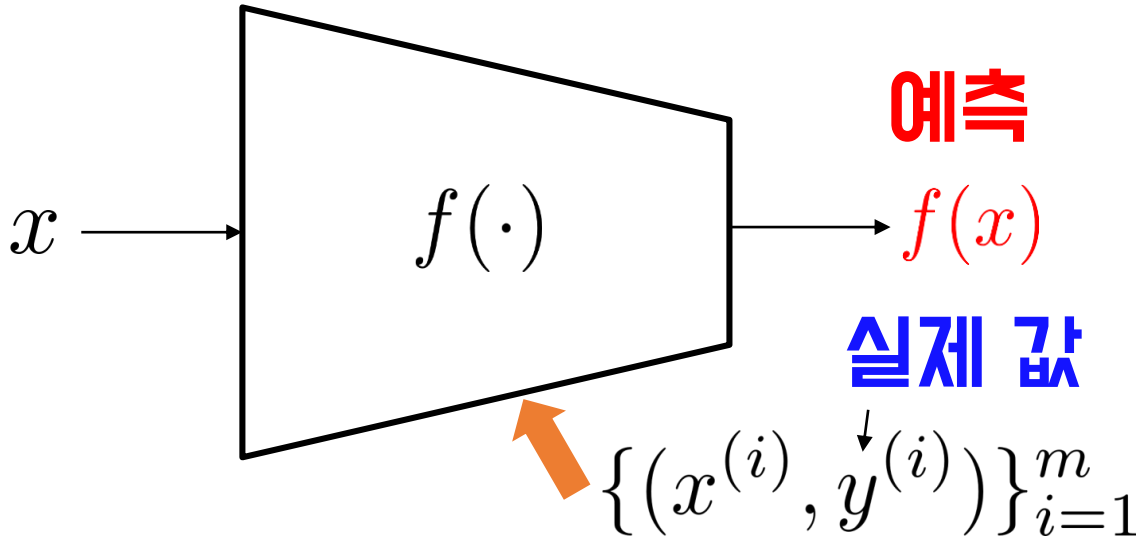
다음 두 가지 컨셉에 기반함:

1. 목적함수

2. 최적화 변수

개념: "목적함수"를 최소화(혹은 최대화)시키는
"최적화 변수"를 찾는 행위

목적함수?



목적: $f(x^{(i)}) \approx y^{(i)}$

Q. "예측"과 "실제 값"이 가까운 정도 어떻게 수치화?

A. **손실함수** 활용: $\ell(y^{(i)}, f(x^{(i)}))$

예) $\ell(y, \hat{y}) = (y - \hat{y})^2$

최적화 변수?

$$\underset{f(\cdot)}{\text{minimize}} \sum_{i=1}^{m} \ell(y^{(i)}, f(x^{(i)}))$$

무엇이 "목적함수"에 영향을 주는가?

예측함수 $f(\cdot)$

Challenge: **예측함수** 후보군이 무한히 많음

(1차함수, 2차함수, 3차 함수 ...)

"함수 최적화" 다루는 방법

$$\underset{f_w(\cdot)}{\text{minimize}} \sum_{i=1}^{m} \ell(y^{(i)}, f_w(x^{(i)}))$$

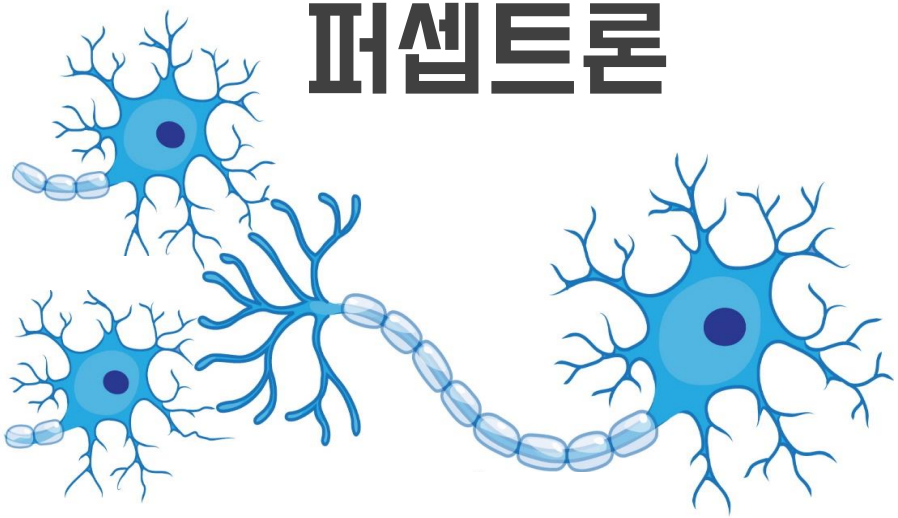
대표적인 방법:

1. 함수클래스 지정
   (예: 1차함수, 2차함수, …)
2. 예측함수 $f(\cdot)$ 를 **파라미터** w 로 표현
   (1차함수 경우: 기울기, y절편)
3. 파라미터를 "최적화 변수"로 고려

함수클래스?

1957년 인간의 뇌를 모사한 인공신경망:

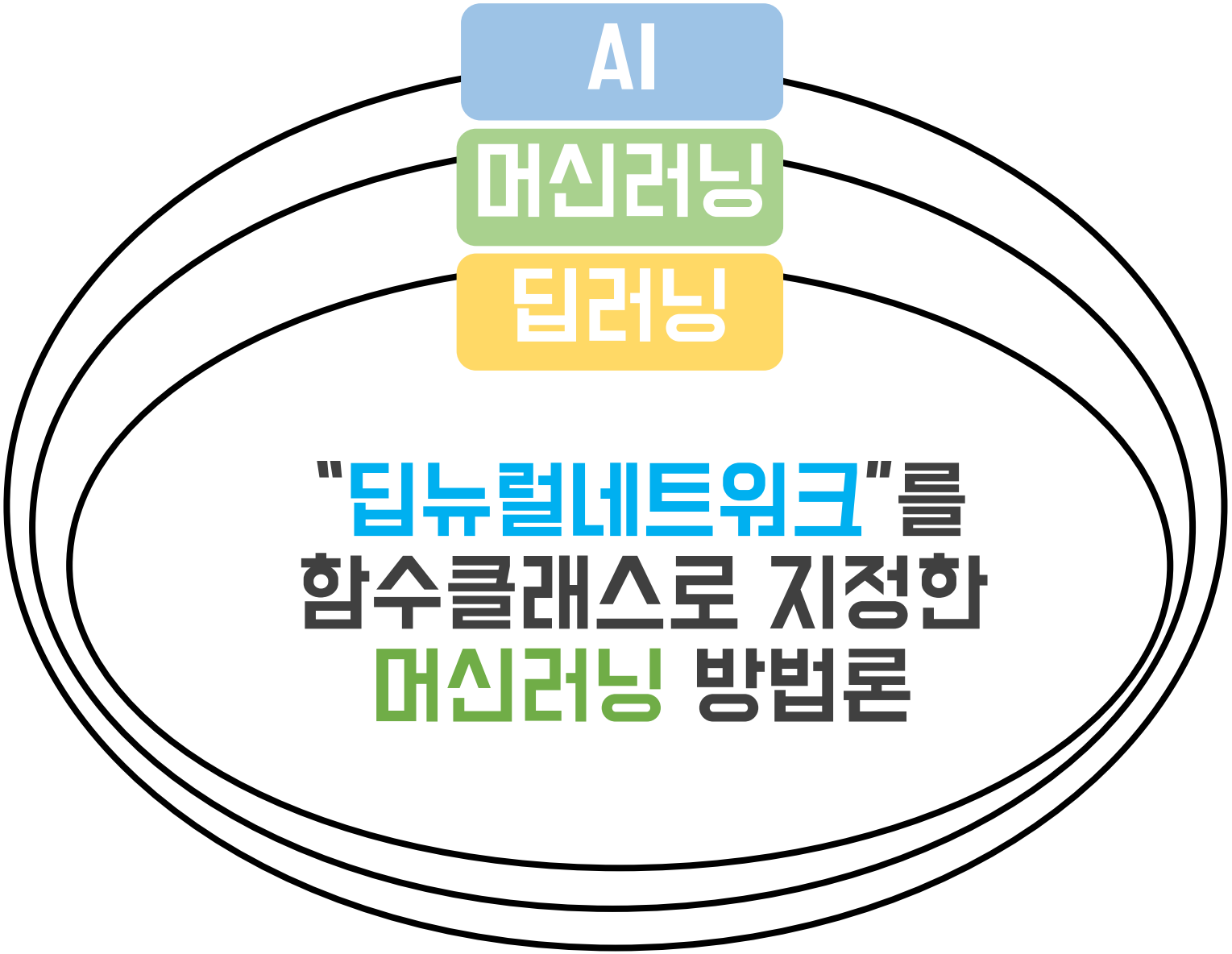
퍼셉트론



Frank Rosenblatt

퍼셉트론의 확장: 딥뉴럴네트워크

딥러닝



AI 성능 vs 데이터

딥뉴럴네트워크의 특징:

학습할 파라미터의 개수 매우 많음

딥러닝기반 AI의 좋은 성능을
위해 **빅데이터** 필요

사회/윤리 문제 #1

사례: 자율주행 AI

자율 주행

Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
운전자 보조 기능			자율주행 기능		
無 자율주행 (No Automation)	운전자 지원 (Driver Assistance)	부분 자동화 (Partial Automation)	조건부 자동화 (Conditional Automation)	고도 자동화 (High Automation)	완전 자동화 (Full Automation)
운전자는 상황을 파악하고 운전함	운전자는 상황을 파악하고 운전함	운전자는 상황을 파악하고 운전함	운전자가 시스템의 요청 시 운전함	운전자가 시스템에 개입하지 않음	시스템의 역할
운전자의 역할	시스템이 운전자의 가/감속 또는 조향을 보조함	시스템이 운전자의 가/감속과 조향을 보조함	시스템이 상황을 파악하고 운전함		

자율 주행

Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
운전자 보조 기능			자율주행 기능		
無 자율주행 (No Automation)	운전자 지원 (Driver Assistance)	부분 자동화 (Partial Automation)	조건부 자동화 (Conditional Automation)	고도 자동화 (High Automation)	완전 자동화 (Full Automation)
운전자는 상황을 파악하고 운전함	운전자는 상황을 파악하고 운전함	운전자는 상황을 파악하고 운전함	운전자가 시스템의 요청 시 운전함	운전자가 시스템에 개입하지 않음	시스템의 역할
운전자의 역할	시스템이 운전자의 가/감속 또는 조향을 보조함	시스템이 운전자의 가/감속과 조향을 보조함	시스템이 상황을 파악하고 운전함		

현 기술 상황

"완전자동화"를 위한 필수 기술

인지

예측

제어

부족한 부분

돌발상황 대처

돌발상황 대처기술 부족한 이유

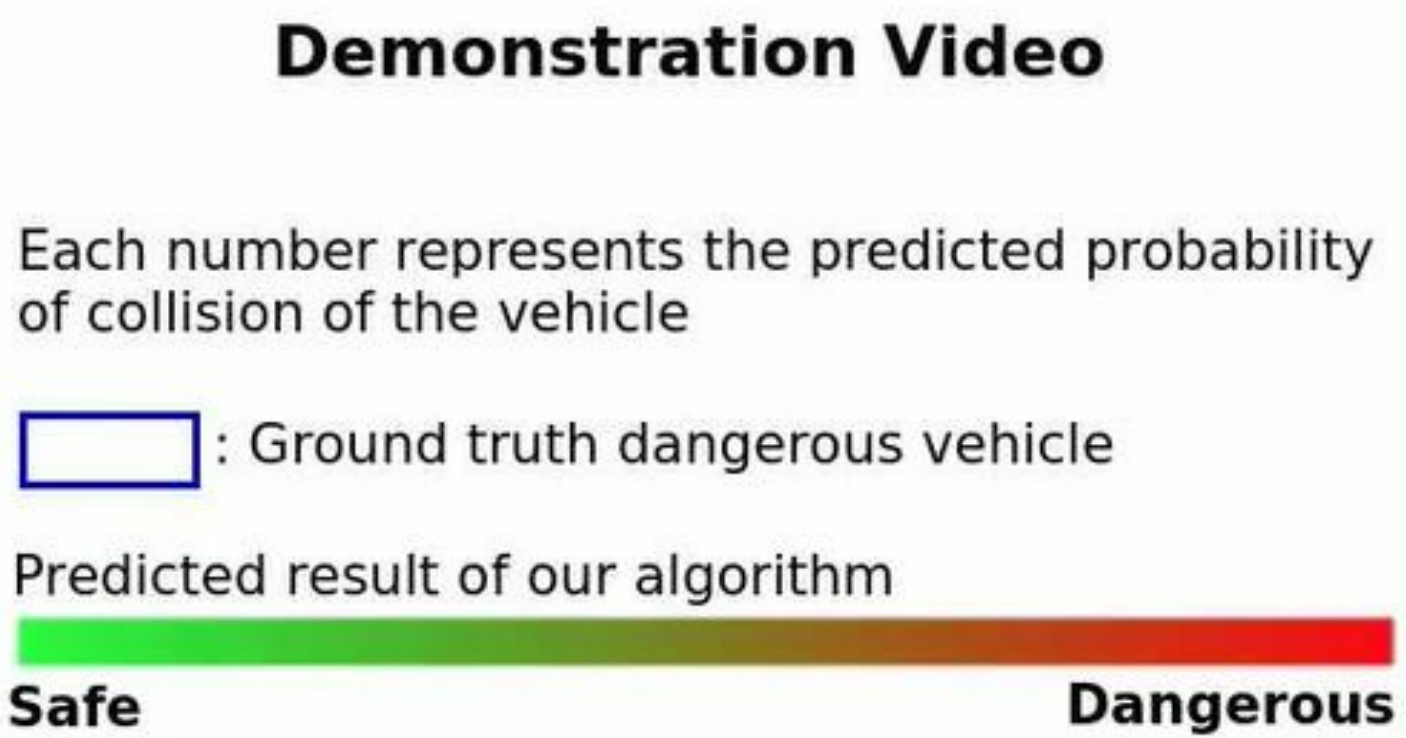
돌발상황 관련 "데이터" **부족**

돌발상황 대처기술 개발 노력

컴퓨터 시뮬레이터 활용: 사고장면 연출

서창호 교수 연구진, AAAI 2019 (국제인공지능학회)

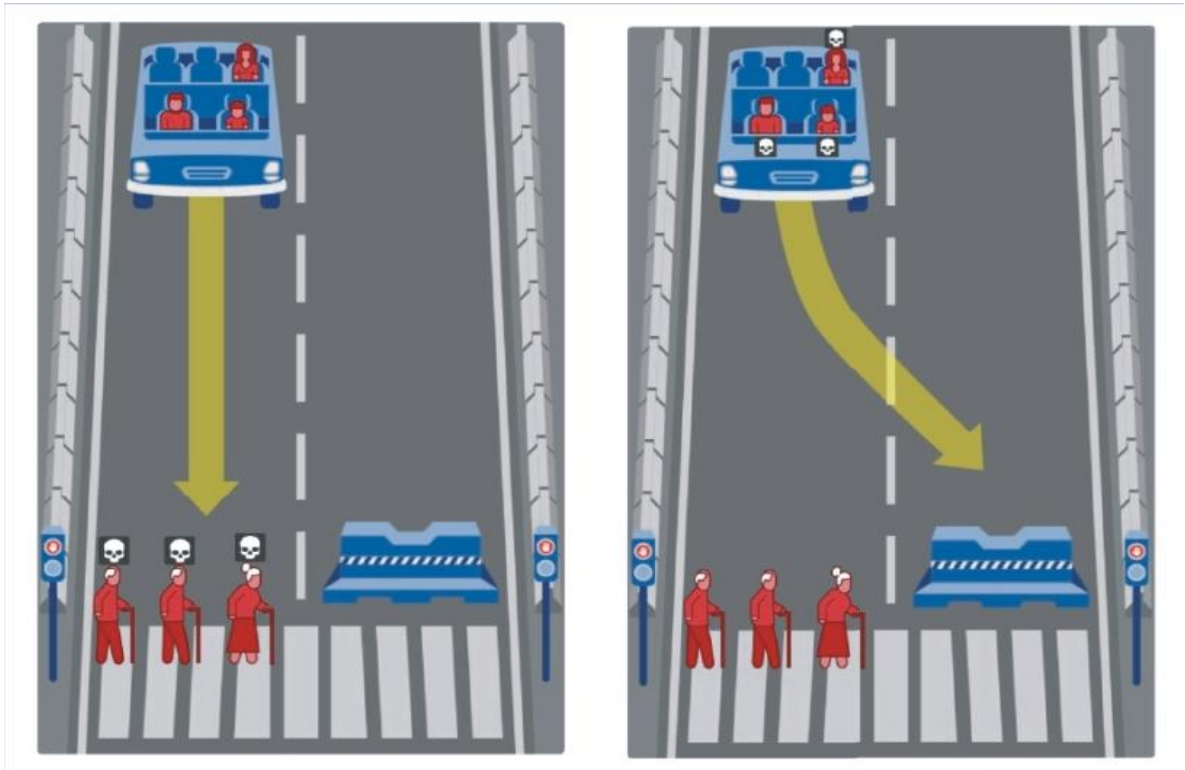
사고 위험도 예측 성능



서창호 교수 연구진, AAAI 2019 (국제인공지능학회)

사회/윤리 문제

예시: 트롤리 전차 딜레마 상황



정의로운 주행이란?

한 가지 해결방안

"정의를 잘 반영하는 데이터" 수집!

이슈: "정의" 개념에 대한 서로 다른 의견

→ 공론화된 윤리기준 도출을 위해

"빅데이터" 수집 필요

정의로운 빅데이터 수집을 위한 노력

Massachusetts Institute of Technology

VE RI TAS HARVARD

UBC

Toulouse School of Economics

"모럴머신" 데이터수집 플랫폼 (네이처 2018)

- 4천만개 이상의 딜레마 상황 연출
- 233개 국가 230만 명을 대상 데이터 수집

한국의 현 정책 현황

1. 자율주행기술개발혁신사업단 (2021년 출범)

   목표: 2027년까지 자율주행 4단계 상용화
   완전자율주행 5단계를 위한 생태계 구축

2. 자율주행차 데이터 표준화 위원회

   목표: 한국의 도로/교통 상황을 반영한
   차량/인프라 데이터의 국가표준화

추가해야 할 노력

공론화된 윤리기준을 도출할 수 있는
빅데이터 수집을 위해:

데이터수집 온라인 플랫폼 개발

사회/윤리 문제 #2

사례: 신뢰할 수 있는 AI

신뢰할 수 있는 AI

필요성:

1. 인권이나 공정성 관련된 영역까지 AI가 인간을 대체하고 있음

2. 기계가 지배하는 세상을 막기 위해 AI는 의도된대로 설계되어야 함

현 기술 상황

많이 부족한 상황

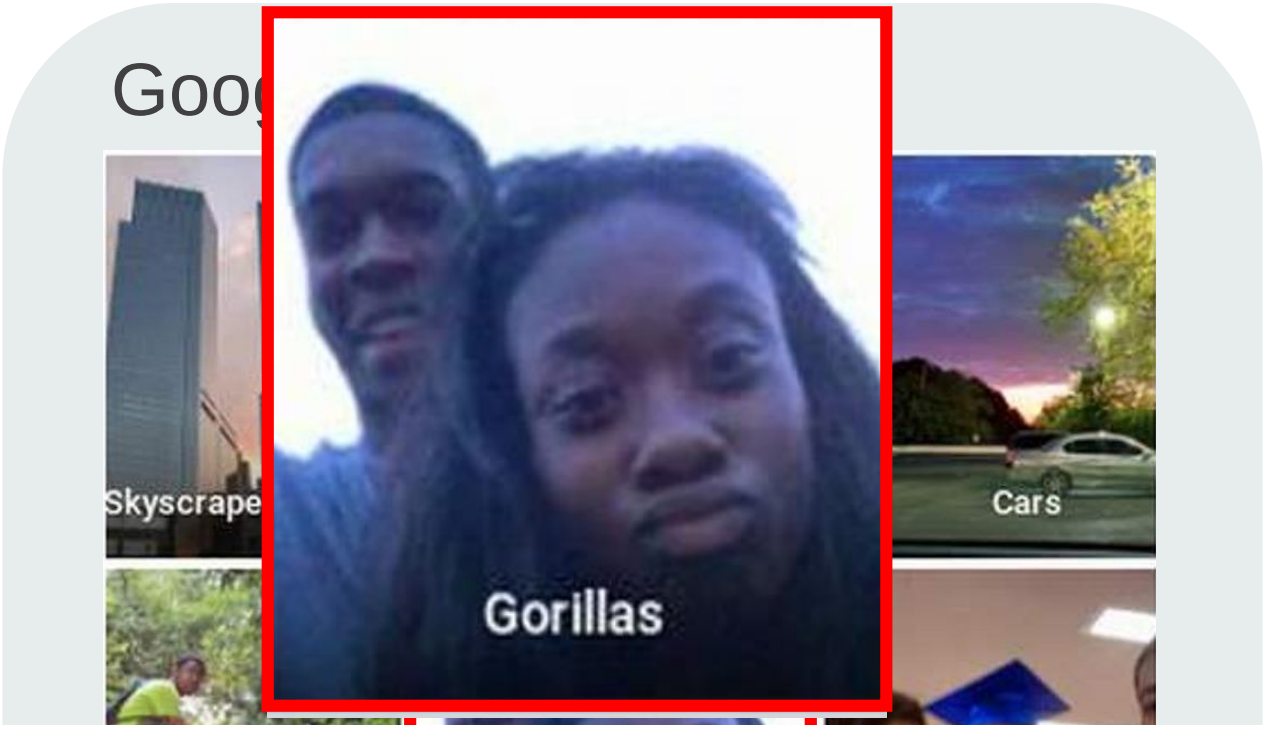
공정성 논란 사건 #1:

Amazon scraps secret AI recruiting tool that showed bias against women

출처: Reuters

현 기술 상황

공정성 논란 사건 #2:



Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

출처: The Wall Street Journal

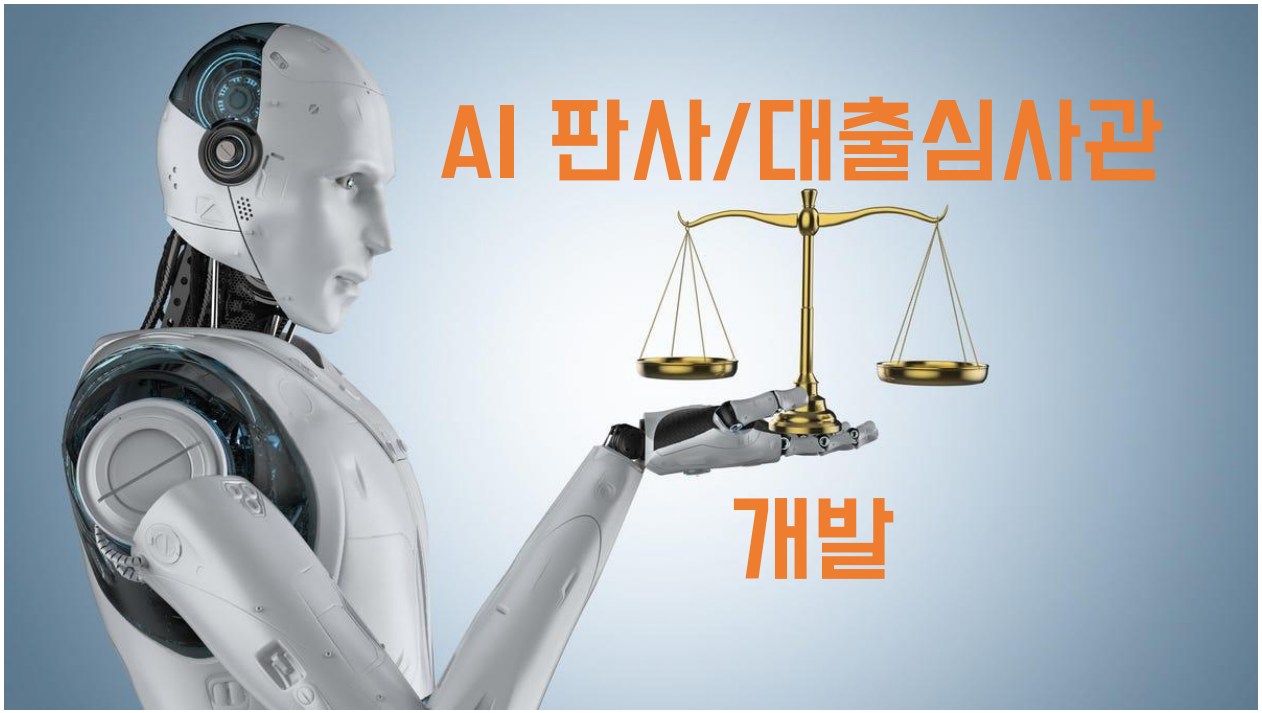
공정성 논란의 원인

데이터 **편향성**!

데이터 편향성 극복을 위한 노력

AI Ethics:

편향된 데이터로도 공정성을 보장하는 기술



서창호 교수 연구진, NeurIPS 2020 (국제신경망학회)

사례: AI 대출심사관

개인정보

대출 심사

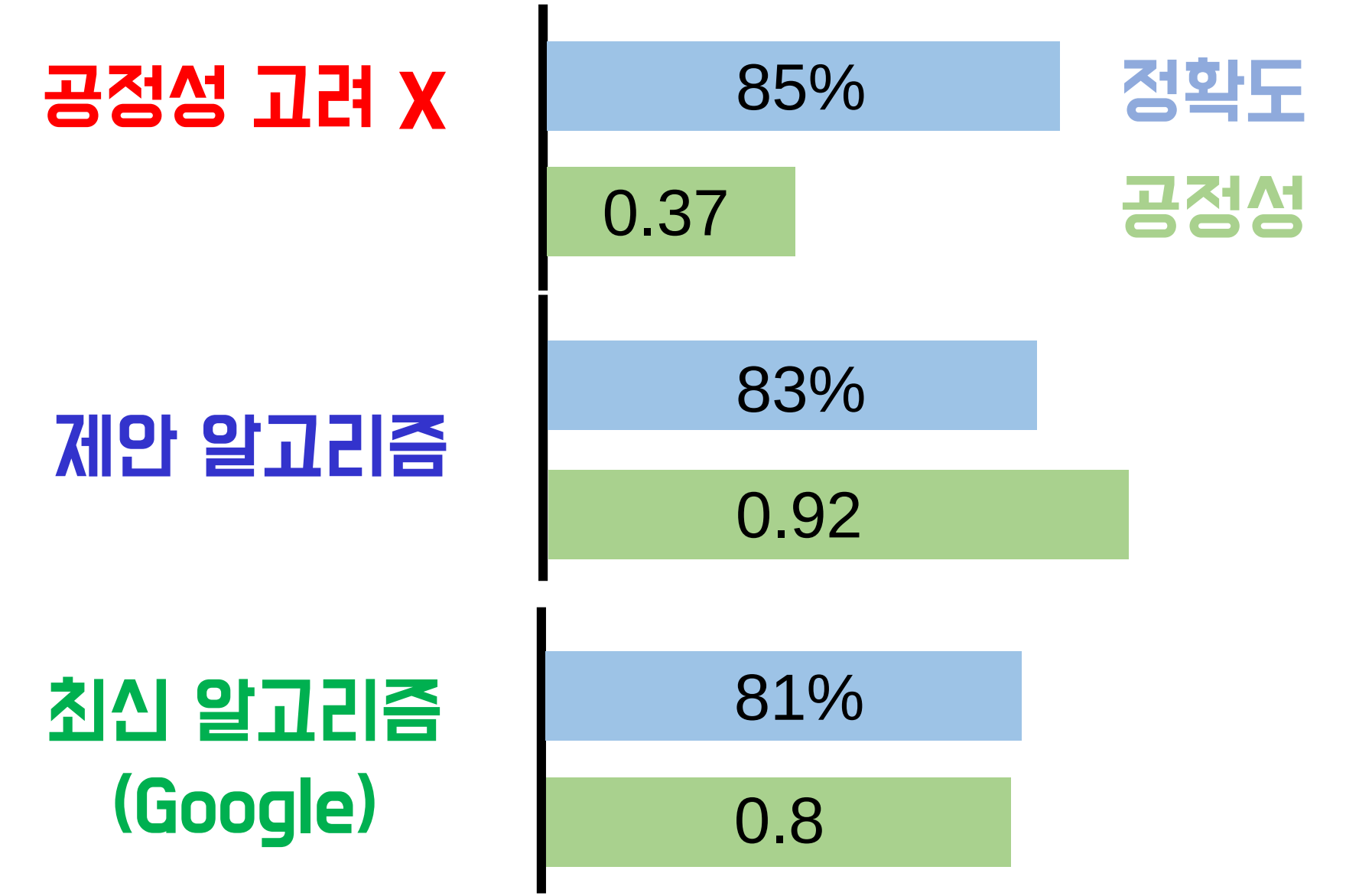
합격 or 불합격

(인종에 따라 편향된)
UCI Adult 데이터셋

VS

성능

서창호 교수 연구진, NeurIPS 2020 (국제신경망학회)



사회/윤리 문제

AI Ethics:

공정성/윤리에 대한 명확한 개념을 필요로 함

도전적 과제

1. 공정성/윤리에 대한 올바른 개념 확립

2. 이를 AI 기술에 접목시키는 능력 배양

장기적인 노력방안

윤리성을 겸비한 AI 인재양성

단기적인 노력방안

1. 공정성/윤리성 관련 기술 표준화

예) 한국정보통신기술협회 (TTA):
"신뢰할 수 있는 AI 기술" 표준화

2. 윤리성 보장을 위한 법제화

예) 비윤리적인 데이터 수집/활용관련
개인정보보호법 강화

마무리

1. AI의 기본 개념
2. 사회/윤리 문제 #1
   자율주행 AI: 정의로운 주행?
3. 사회/윤리 문제 #2
   신뢰할 수 있는 AI: 올바른 윤리 개념?