AI와 공정: 공학자의 시각

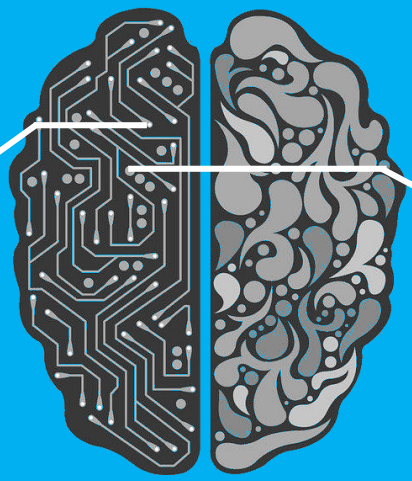
KAIST AI Institute: AI & Society 세미나

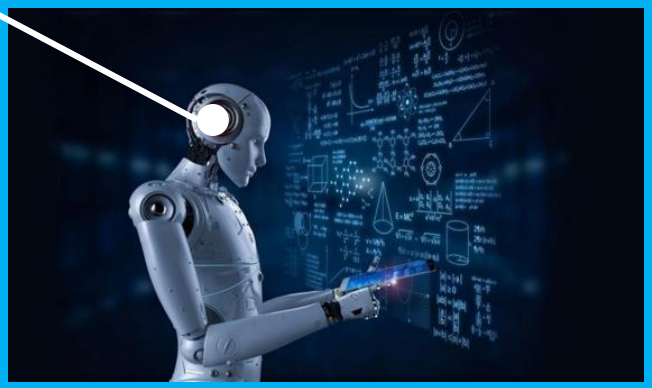
2022년 2월 8일

서창호

KAIST 전기 및 전자공학부

AI 킬러 앱:



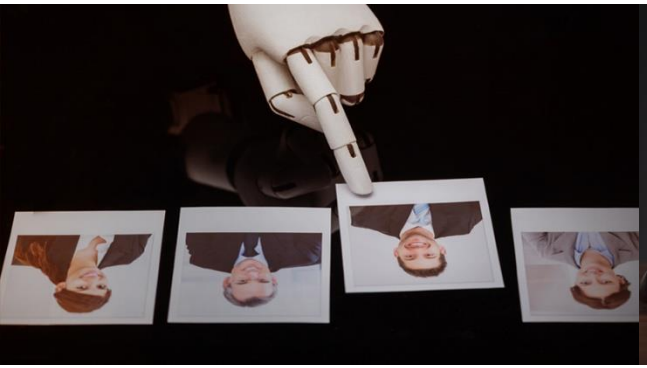


AI 비서

AI 튜터

인권 관련된 영역까지 AI가 대체:

채용 심사



형량 결정

대출 심사

상기 애플리케이션의 공통점:

공정성을 요구함!

공정성 관련 AI 기술 현황

많이 부족한 상황

공정성 논란 사건 #1:

Amazon scraps secret AI recruiting tool that showed bias against women

출처: Reuters

공정성 관련 AI 기술 현황

공정성 논란 사건 #2:

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

출처: The Wall Street Journal

공정성 논란의 원인

AI 성능을 좌우하는 것:

데이터!

데이터 **편향성**!

한 가지 해결책

"공정성을 잘 반영하는 데이터" 수집!

이슈: "공정" 개념에 대한 서로 다른 의견

→ 공론화된 공정성 기준 도출을 위해

"빅데이터" 수집 필요

빅데이터 수집을 위한 노력

Massachusetts Institute of Technology

VE RI TAS HARVARD

UBC

Toulouse School of Economics

"모럴머신" 데이터수집 플랫폼 (네이처 2018)

- 233개 국가 230만 명을 대상 데이터 수집

데이터수집 플랫폼의 한계점

1. 빅데이터 **수집**의 어려움

   예) 모럴머신: 18개월의 시간

   230만명의 인력 동원

2. 여전히 데이터 **편향성** 문제

   예) 샘플링 바이어스 발생할 경우

   → 특정 집단의 지엽적인 의견 반영

고도화된 공정성보장 AI 방법론

편향된 데이터로도 공정성을 알고리즘으로 보장하는 방법론

구글 연구진, AIES 2018 (국제 AI/윤리/Society 학회)

황의종/서창호 교수 연구진, ICML 2020 (국제머신러닝학회)

서창호 교수 연구진, NeurIPS 2020 (국제신경망학회)

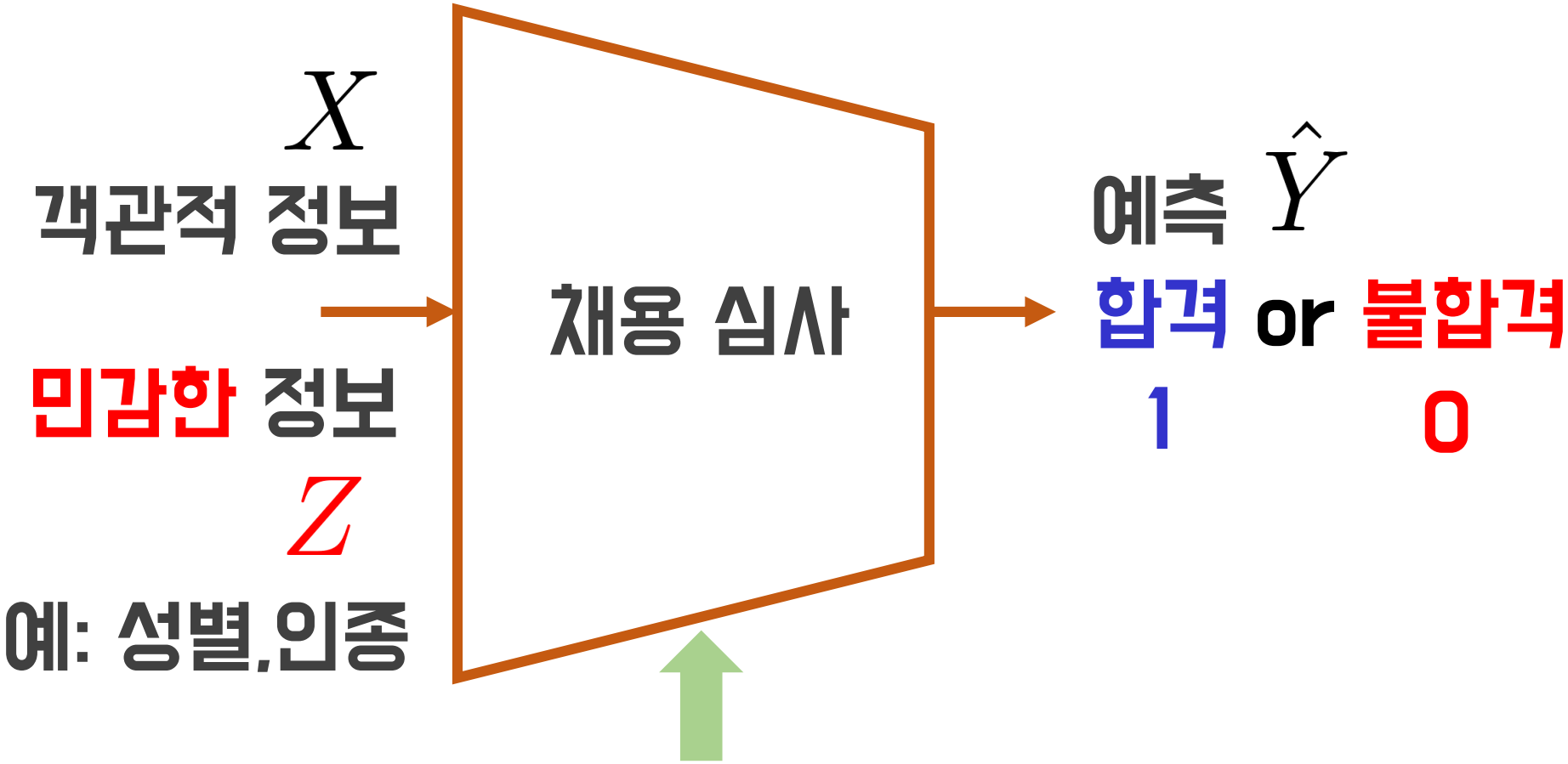
:

학술적 명칭: 알고리즘 공정성

특징: 공정성 개념을 수학적인 언어로 바꾸는 과정 포함

채용심사를 예시로 이 과정을 설명하겠음

채용 심사 사례



$$\{(X^{(i)}, Z^{(i)}, Y^{(i)})\}_{i=1}^{m}$$

대표적인 공정성 개념

민감한 정보: Z 예측: $\hat{Y}$

Demographic Parity(그룹간 평등)

$(Z, \hat{Y})$: 서로 상관 없음

수학적인 언어로 독립

독립의 정의

민감한 정보: Z 예측: $\hat{Y}$

Demographic Parity(그룹간 평등)

$(Z, \hat{Y})$ 독립:

$$\mathbb{P}(\tilde{Y} = 1 | Z = z) = \mathbb{P}(\tilde{Y} = 1)$$

모든 z에 대해

독립의 정도 수치화 과정

Demographic Parity(그룹간 평등)

$(Z, \hat{Y})$ 독립:

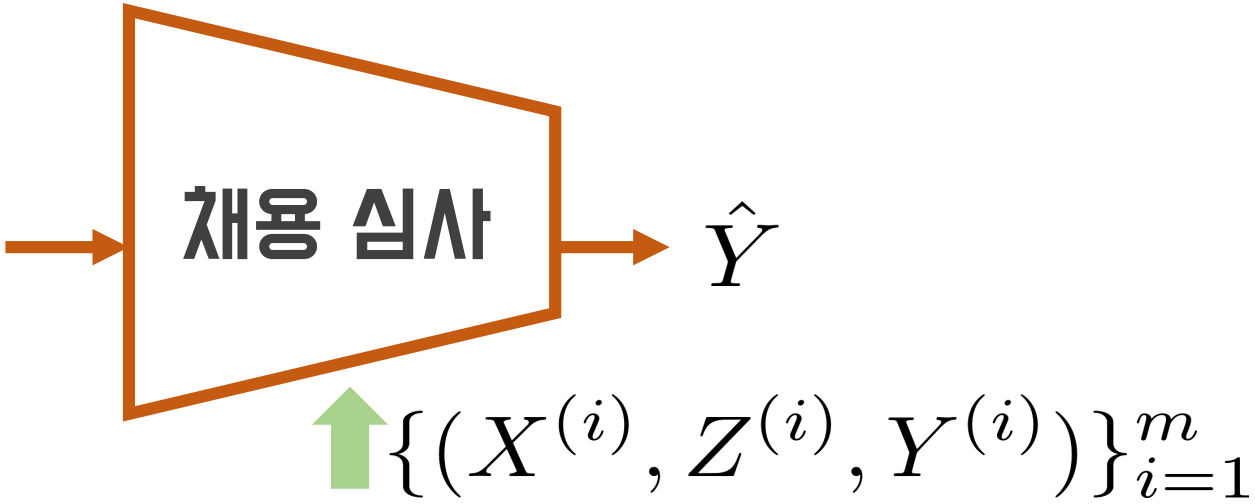
모든 z에 대해

$$\mathbb{P}(\tilde{Y} = 1 | Z = z) = \mathbb{P}(\tilde{Y} = 1)$$

$$\sum_{z} |\mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1)|$$

작을수록 독립에 가까움

공정한 AI 설계 방법론



1. 공정성을 위해:

$$\sum_{z} |\mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1)|$$ 작도록

2. 높은 예측 정확도를 위해:

$$\hat{Y}^{(i)} \approx Y^{(i)}$$ 모든 예시에 대해

생각해 볼 문제

1. Demographic Parity 적절한가?

   채용심사 데이터: 남녀 합격률이 현저히 차이나는 경우?

2. 공정성 정도 수치화 하는 방법 적절한가?

   수치화된 값이 어느 값 이하여야 공정하다 판단할 수 있을까?