한국실천공학교육학회, 2025 종합학술발표대회

인공지능의 편향과 공정성의 문제

실증적 사례를 통한 인공지능 학습 알고리즘의 문제 탐구

홍영준, 김광진

지도 교수 : 이효재

한국폴리텍대학 성남캠퍼스 인공지능소프트웨어과

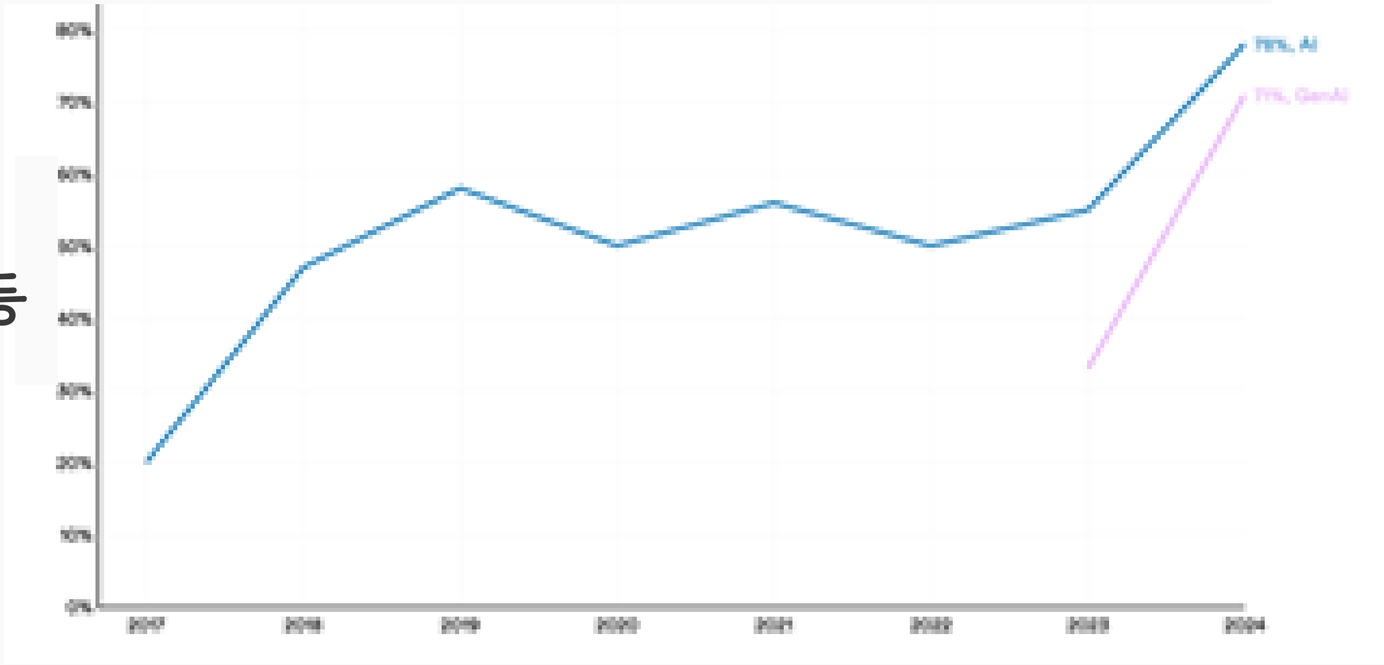
Contents

연구 배경

인공지능 기술은 **의료**, **금융**, **채용**, **법률** 등 다양한 분야에서 의사결정의 효율성과 정확성을 높이며 의사결정 자동화의 핵심 기술.

그러나, 인공지능은 "공정할 것"이라는 기대와 달리 다양한 원인으로 부터 인간의 편향을 그대로 학습함.

(2017-2024) 기업 AI 도입률 및 생성형 AI 활용 현황 추이 그래프



기업 AI 도입률 : 2017 20% → 2024 78%

생성형 AI 활용 : 2023 30% → 2024 70%

연구 목적

"공정성과 책임성 확보의 중요성 제시" 하는 것.

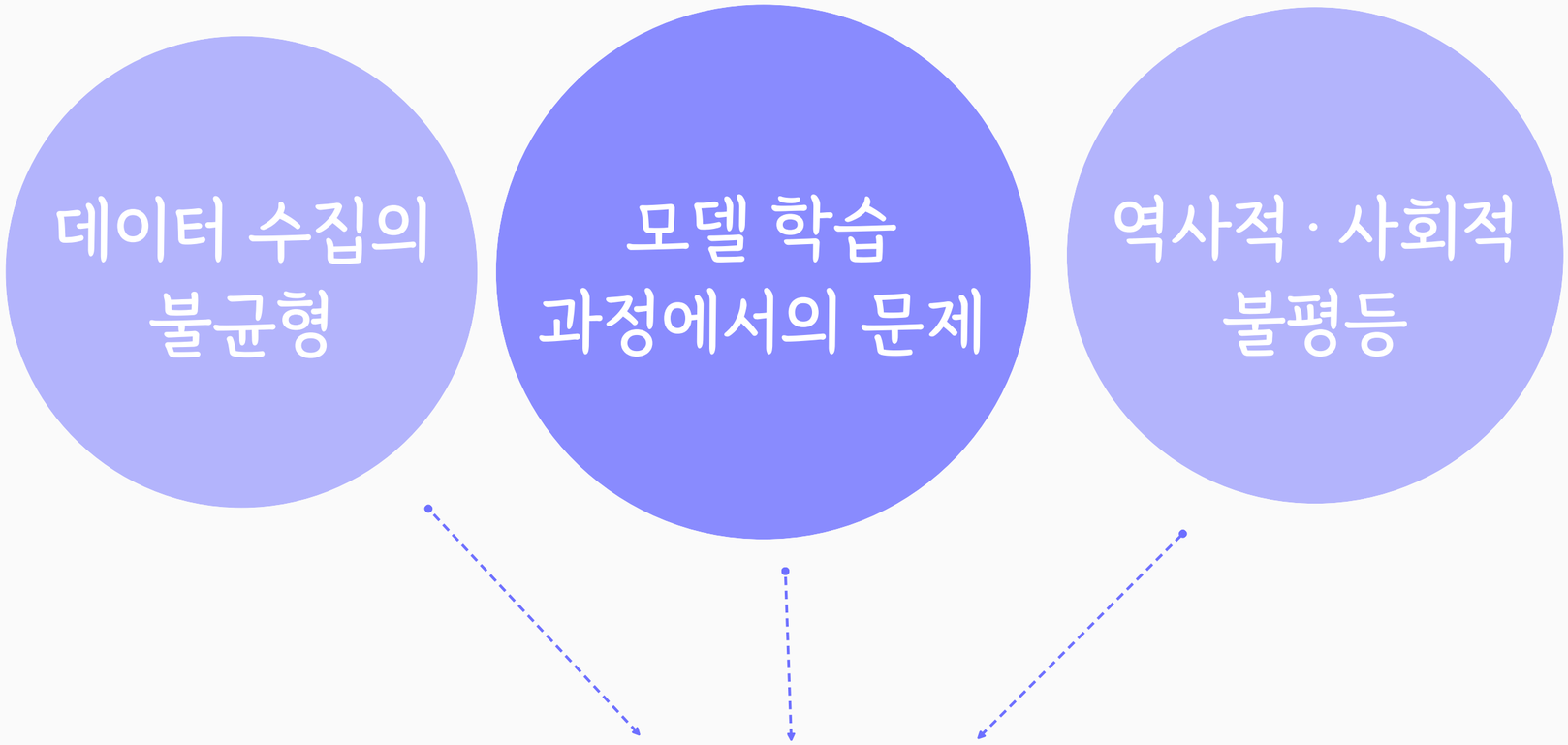
인공지능이 실제 사회 속에서 어떤 방식으로 편향을 나타내는지 실증적으로 살펴보고, 그 원인과 공정성 확보 방안을 탐구하는 데 있음.

인공지능 편향의 개념

데이터 수집, 학습, 알고리즘 설계, 과정에서 발생하는 왜곡된 결과나 불공정한 판단을 의미.

즉, 인공지능이 "사람 대신 판단할 때", 그 판단이 특정 그룹에게 유리하거나 불리하게 작용하는 상황.

인공지능 편향의 원인



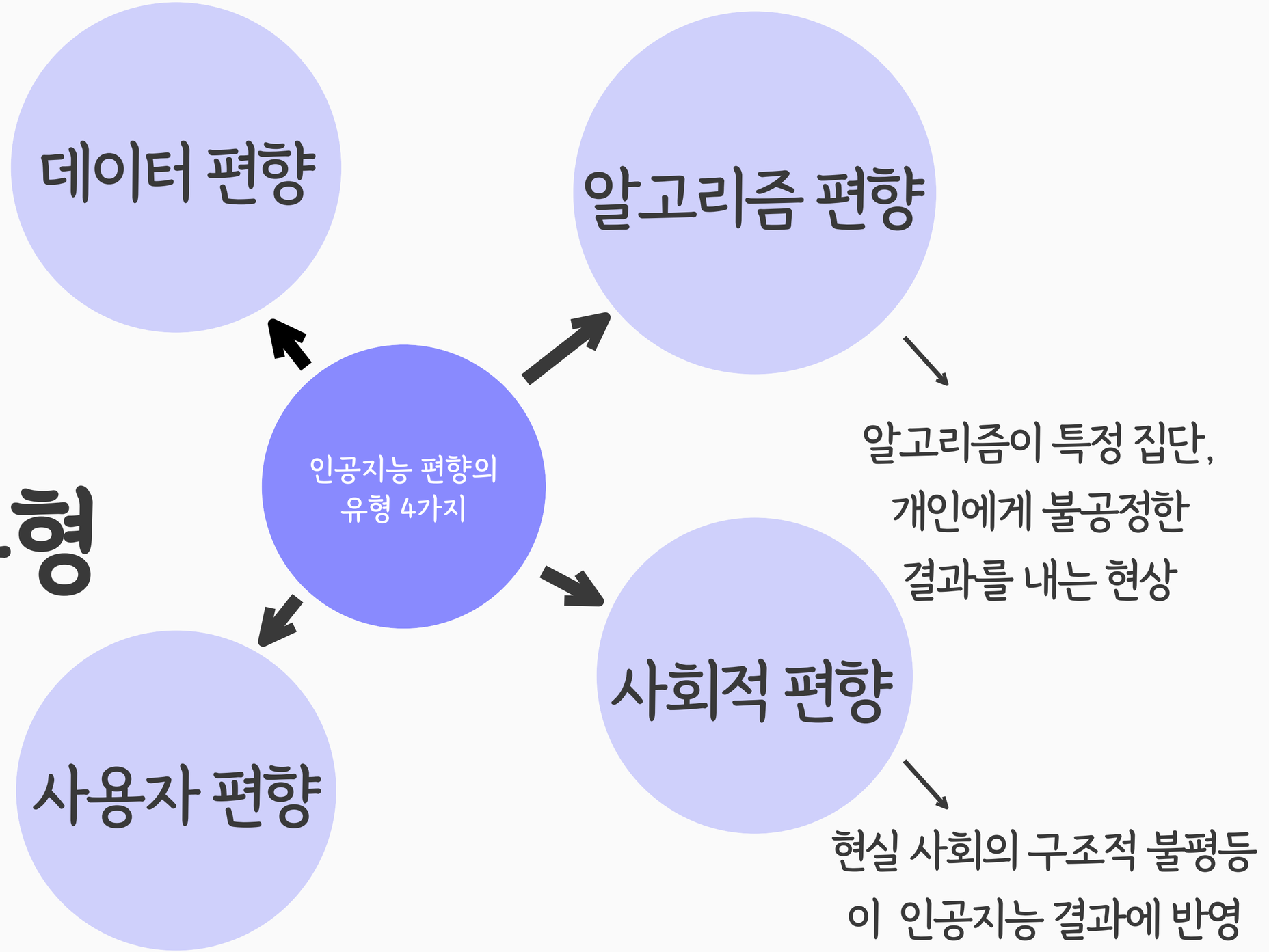
왜 생기는가? (근본적 원인)

인공지능은 "스스로 생각"하지 않고, 주어진 데이터를 학습.

그 데이터가 이미 사회적 편견을 담고 있다면,

인공지능은 그걸 그대로 '정답'이라고 믿고 복제.

인공지능 편향의 유형

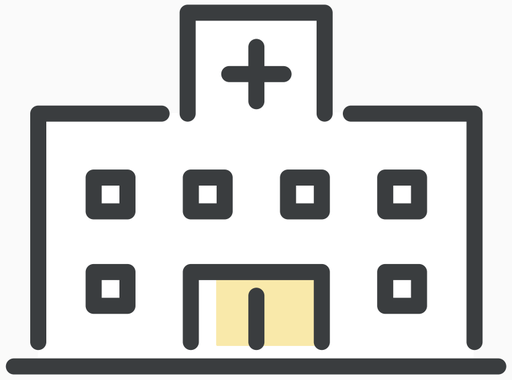

인공지능 편향의 유형 4가지
데이터 편향
학습 데이터가 특정 집단을 과소, 과대 대표할 때 발생
알고리즘 편향
알고리즘이 특정 집단, 개인에게 불공정한 결과를 내는 현상
사용자 편향
인공지능을 사용하는 사람이 편향적 결정을 내릴 때 발생
사회적 편향
현실 사회의 구조적 불평등이 인공지능 결과에 반영

실증 사례

amazon

AMAZON 채용 알고리즘

과거 10년간 남성 중심 채용 데이터를 학습. 여성 지원자의 점수를 자동으로 낮게 평가.



미국 의료 알고리즘

의료 데이터 기반의 알고리즘이 흑인 환자 집단의 치료 필요성을 과소평가

COMPAS 범죄 예측 모델

미국 범죄자 재범 위험 점수 예측 모델. 흑인 범죄자에게 높은 위험 점수를 부여,백인에게 낮은 점수를 부여

자동화된 진로 추천 시스템

여성에게는 간호사, 교사와 같은 직업을, 남성에게는 엔지니어, 의사와 같은 직업을 추천

결론 및 시사점

AI 편향은 기술적 오류가 아니라 사회적 구조의 반영이다.

데이터 · 모델링 · 운용 전 단계에서 편향을 인식하고 관리해야 한다.

공정성 · 투명성 · 책임성 확보가 핵심이다.

윤리적 AI 구현은 선택이 아니라 필수이다.

한국실천공학교육학회, 2025 종합학술발표대회

감사합니다.