# 알고리즘 공정성 평가 지표:

# 불가능성 정리 넘어서기

2024. 1. 17.

김병필(KAIST 기술경영학부 / 변호사)

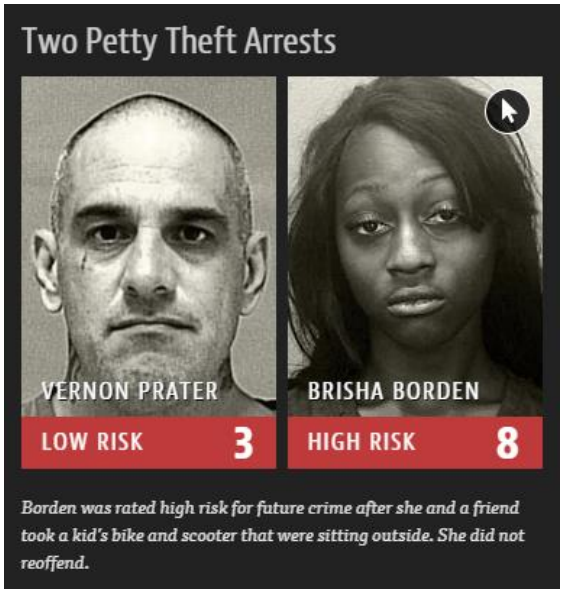
# 문제제기 – COMPAS 재범 예측 시스템

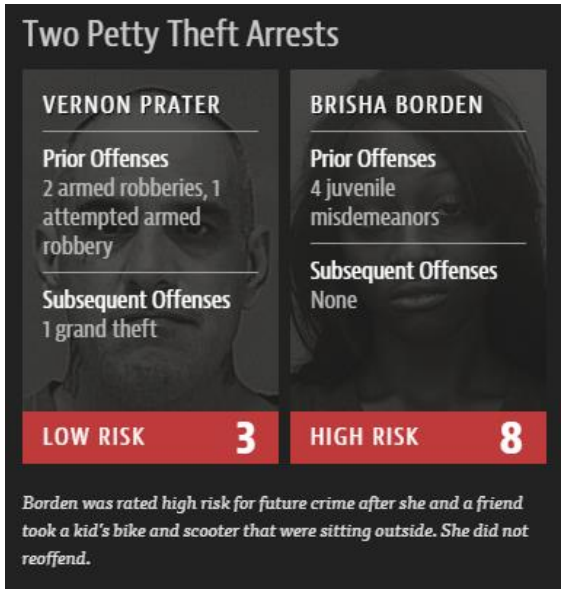
## Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*

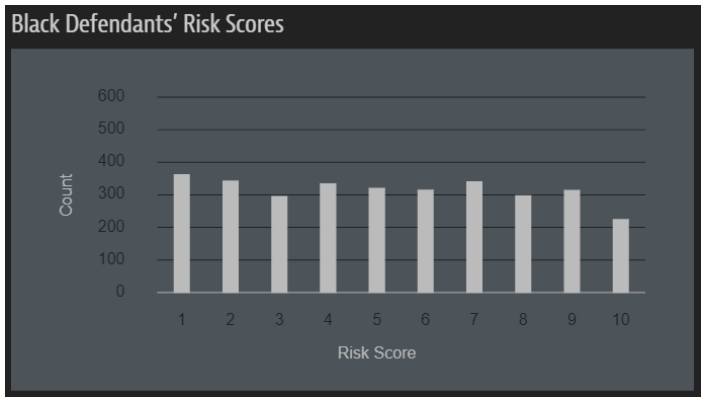
May 23, 2016

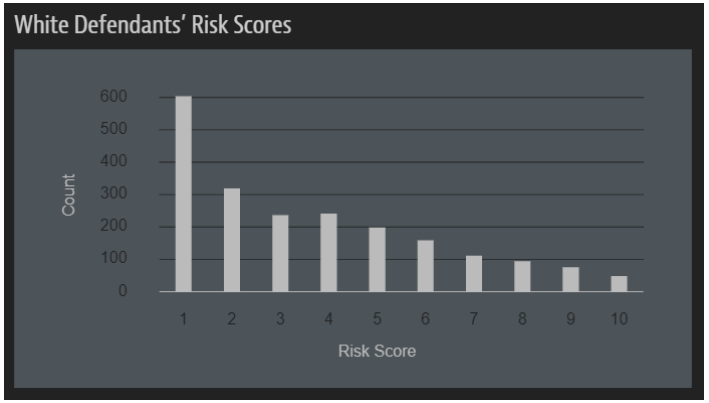


# 제기된 문제점

- 인종별 위험도 분포 차이




- 인종별 오류율 차이

**Prediction Fails Differently for Black Defendants**

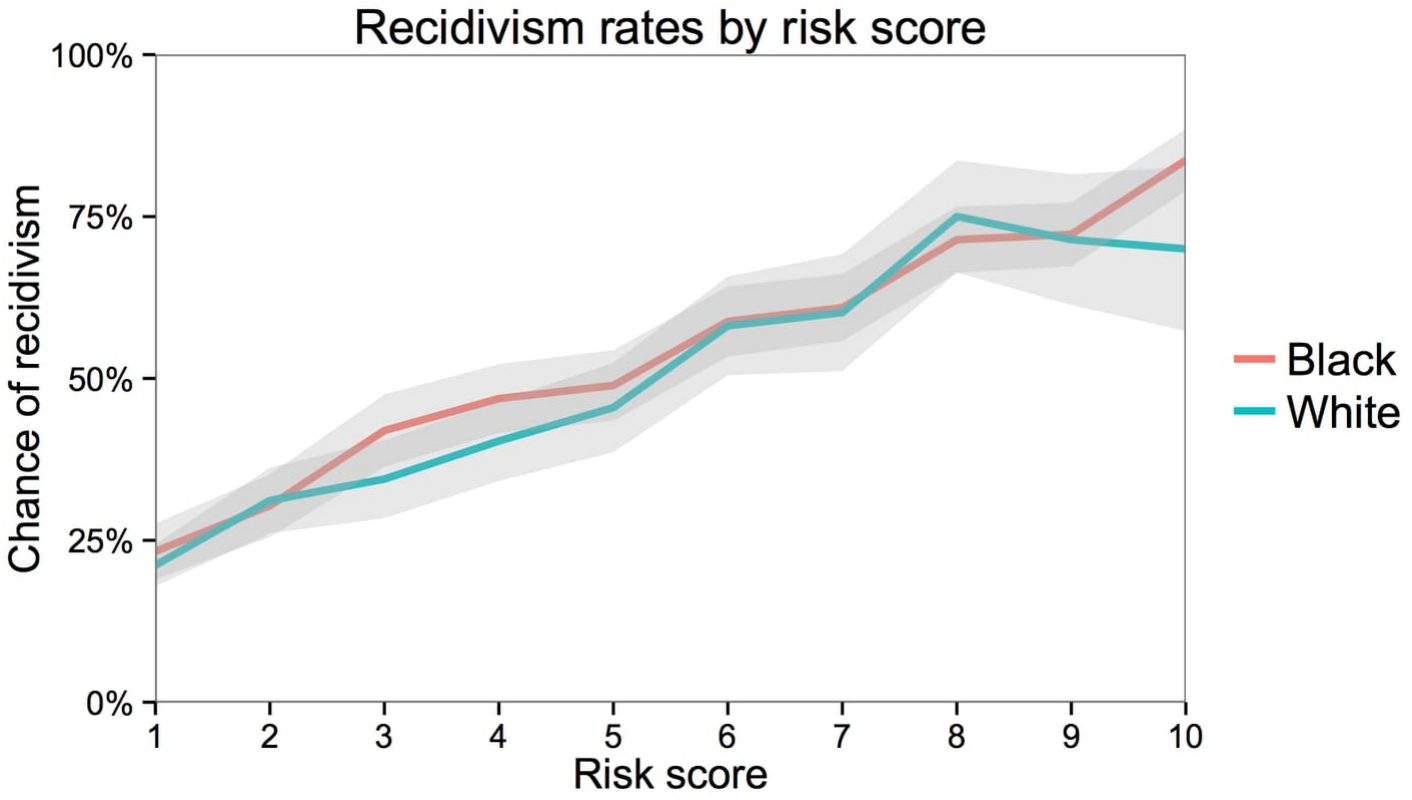
| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

***Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.*** *(Source: ProPublica analysis of data from Broward County, Fla.)*

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# COMPAS 개발사의 반론

- **COMPAS는 예측도 동등성 기준을 달성하고 있음**
  - 예측도 동등성 기준 = Risk Score가 재범율을 정확하게 예측하는 것 (= “Calibration” 기준)




https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

# COMPAS 사례 – 오류율 동등성과 예측도 동등성의 충돌

- **COMPAS 개발사: 예측도 동등성을 기준으로 감사 수행**

False Discovery Rate (FDR) = 1 – PPV

$$= \frac{\text{Labeled Higher Risk But Didn't Re-Offend}}{\text{Labeled Higher Risk}}$$

False Omission Rate (FOR) = 1 – NPV

$$= \frac{\text{Labeled Lower Risk Yet Did Re-Offend}}{\text{Labeled Lower Risk}}$$

| | White | African American |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 41% | 37% |
| Labeled Lower Risk, Yet Did Re-Offend | 29% | 35% |

Table 3.1: Propublica's table with correct target population errors at the study cut point (Low vs. Not Low) for the General Recidivism Risk Scale.

- **ProPublica: 오류율 동등성 기준으로 감사 수행**

False Positive Rate (FPR) = 1 – TNR

$$= \frac{\text{Labeled Higher Risk But Didn't Re-Offend}}{\text{Didn't Re-Offend}}$$

False Negative Rate (FNR) = 1 – TPR

$$= \frac{\text{Labeled Lower Risk Yet Did Re-Offend}}{\text{Did Re-Offend}}$$

**Prediction Fails Differently for Black Defendants**

| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

*Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.* (Source: ProPublica analysis of data from Broward County, Fla.)

# 대표적 알고리즘 공정성 평가 지표

- **입력 변수에 대한 기준**
  - Unawareness 기준

- **출력 변수에 대한 기준**
  - 결과 동등성 기준 (= Demographic Parity)
  - 오류율 동등성 기준 (= Equal Opportunity, Equalized Odds)
  - 예측도 동등성 기준 (= Calibration)

# 입력 변수 기준 – Unawareness 기준

- **Unawareness 기준**
  - 보호 속성을 입력변수로 사용하지 않는 기준
  - 어떤 분류모형 $f_{classfier} : X \mapsto \hat{Y}$ 에 대해, $f$가 $\hat{Y}$을 예측하기 위하여 보호 속성 $A$를 사용하지 않는 경우 $f_{classifier}$는 Unawareness 기준을 충족함

- **보호 속성의 결정**
  - 법에 명시된 차별금지 사유 또는 개인정보 보호법제의 민감 속성을 기준으로 원용
  - 법에 명시되지 않았더라도 일반인에 대한 의견 조사 등을 통해 결정

| Feature | Q. 1 (a priori) | Q. 2 (if more accurate) | Q. 3 (if increases disparity) |
|---|---|---|---|
| # prior offenses | 95% | 93% | 83% |
| arrest charge description | 86% | 92% | 71% |
| arrest charge degree | 85% | 91% | 69% |
| # juvenile felony offenses | 74% | 80% | 61% |
| # juvenile misdemeanor offenses | 65% | 71% | 53% |
| # juvenile other offenses | 63% | 69% | 52% |
| age | 44% | 61% | 32% |
| gender | 26% | 55% | 24% |
| race | 21% | 42% | 17% |

Table 1: Comparing user judgment of fairness of each feature, when the user has different knowledge about the impact of incorporating that feature in the decision making process. We show the percentage of users who categorized each feature as fair according to the 3 questions described in Section 2.1.

Grgic-Hlaca et al. (2018)

# Unawareness 기준의 한계

- **실효성에 대한 비판**
  - 보호 속성을 입력에서 배제하더라도, 인공지능은 다른 입력변수를 통해 보호 속성을 추론할 수 있음
  - 결국 보호 속성을 입력변수로 넣은 경우와 차이가 없는 결정을 내리게 됨(Gillis & Spiess, 2020)

- **대용변수(proxy)에 의한 차별 가능성(Cofone, 2019)**
  - 전이(transfer) 대용변수 – 전과 정보를 차단하면, 흑인에 대한 차별이 심화됨
  - 축소(reducing) 대용변수 – 민족성 정보를 차단하면, 이름에서 민족성이 드러나는 경우만 차별
  - 확장(expanding) 대용변수 – 임신 계획 여부 정보를 차단하면, 여성 전체에 대한 차별 심화
  - ➔ 정보 차단이 의사결정에 미치는 영향을 전반적으로 고려해야 함

- **적극적 평등실현 조치 또는 차별 감사의 제한**
  - 보호 속성이 해당 보호 집단에게 오히려 긍정적인 영향을 끼치는 경우도 존재
  - 인공지능 차별 감사를 위해서는 보호 속성 정보를 수집할 필요도 있음
    - EU AI 법안 제10조 제5항은 편향 모니터링, 감지 및 교정 목적의 민감정보 처리를 허용
    - 다만, 적절한 보호 조치(재사용에 대한 기술적 제한, 최선의 보안 및 프라이버시 보호 조치) 필요

# 출력변수 기준 (1) 인구통계적 동등성

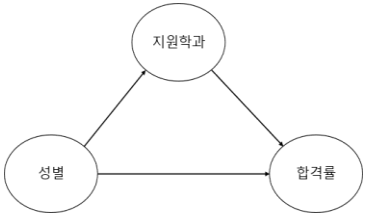
- **인구통계적 동등성(demographic parity, DP) 지표**
  - 어떤 분류모형 $f_{classifer} : X \mapsto \hat{Y}$ 에 대해, 보호속성 $A$가 $\hat{Y}$와 독립이어야 함
    - $A$에 속한 모든 $a, a'$에 대해 $P\{\hat{Y} \mid A = a\} = P\{\hat{Y} \mid A = a'\}$ 만족
  - 인구통계적 동등성, 통계적 동등성, 독립성 등으로도 불림
  - 완화된 기준 - $\epsilon$ 까지의 차이를 허용 (EEOC 4/5 원칙)
    - $\frac{P\{\hat{Y}|A=a\}}{P\{\hat{Y}|A=a\prime\}} \geq 1 - \epsilon$

- **함의**
  - 이진 분류시 특정 결과에 해당할 확률(즉, 선발률, 대출 승인율)이 집단별로 동등할 것을 요구
  - 비교 – 집단별 인구비례 선발을 요구하는 것은 아님
    - (예시) 25명을 선발하는데 남성이 200명, 여성이 50명 지원, 그 중 남성 20명, 여성 5명을 선발 → 남성 합격률과 여성 합격율은 모두 10%로 동일하므로, '인구통계적 동등성'은 충족

# 인구통계적 동등성 기준의 한계

- **법적 제한**
  - 인구통계적 동등성 기준을 충족하는 것이 오히려 역차별에 해당하여 직접차별 금지 위반 가능

- **심슨 패러독스 문제**
  - 1973년 버클리대학 대학원 성별 차별 사례
    - 남성 합격률 44%, 여성 합격률 35% 불과 → 체계적인 성별 차별을 암시하는가?
  - 여성 지원자들은 경쟁률이 높은 학과에 지원하였기 때문인 것으로 드러남




➔ 단순히 외견상 드러나 성별에 따른 합격률 격차만으로 차별적 관행을 추론할 수 없음

- **부정적 데이터의 역사적 누적 가능성**
  - 충분한 자격을 갖추지 못한 지원자들이 선발되면 해당 집단에 대해 부정적 데이터가 누적될 우려

- **역사적 차별 시정 비용 부담의 문제**
  - 결과 동등성 충족 위해 발생하는 비효율성(추가적 교육 필요성 등)의 비용 부담
  - 결과 동등성 기준을 적용함으로써 불합격하게 되는 다수 집단 구성원의 피해 가능성

# 조건부 인구통계적 동등성

- **조건부 인구통계적 동등성 기준**
  - 비차별적인 설명 요인에 의한 차등적 결과는 허용
  - 해당 요인을 통제한 상태에서 조건부로 결과 동등성이 달성될 것을 요구

- **예시**
  - 버클리 대학원 입학률 성별 격차 – 비차별적 매개변수인 '지원학과'를 구분하여 성별 합격률 비교
  - 보험료에서의 성별 격차
    - 운전경력 연수를 비차별적인 매개변수로 고려
    - 동일한 운전경력 연수를 가진 사람들 간에 비교하였을 때 성별간 보험료 격차 여부를 비교
    - 만약 여성이 평균적으로 운전경력 연수가 적다면 전체 집단에 대해서는 여성의 보험료가 더 높을 수도 있으나, 조건부 인구통계적 동등성 기준에서는 이러한 격차를 허용함

- **한계 – 비차별적 설명요인 설정의 어려움**
  - 어떠한 요인은 비차별적이고, 어떠한 요인은 차별적인지 개별 변수마다 판단하기 어려움

# 출력변수 기준 (2) 오류율 동등성

- **예시 사례**
  - 회사의 직원 출입구에 얼굴인식 서비스를 이용하여 직원인지 여부를 인식
  - 인공지능은 다른 피부색의 직원에 대해 서로 다른 오류율을 보임
  - 파란 피부를 가진 직원 – 평균적으로 100번 통과시 90번 통과, 10번 거절(위음성율 10%)
  - 나머지 피부색을 가진 직원 – 100번 통과시 99번 통과, 1번 거절(위음성율 1%)

- **의의**
  - 어떤 분류모형 $f_{classifer} : X \mapsto \hat{Y}$ 에 대해, 보호속성 $A$가 진정한 목표변수 값 $Y$를 조건으로 $\hat{Y}$과 독립인 경우 $f_{classfier}$는 오류율 동등성을 충족함
  - 위음성율(FNR)의 동등성과 위양성율(FPR)의 동등성으로 구분
  - 위음성율(FNR)의 동등성을 **‘기회 균등(Equal Opportunity)’** 조건이라고도 함
  - 양자 모두를 요구하는 기준을 ‘분리성’ 조건 또는 Equalized Odds 조건이라고도 함

# 오류율 동등성의 한계

- **법적 정당화 근거의 불명확성**
  - 기존 차별금지법제에 비추어 해당 기준이 어떻게 정당화될 수 있는지 명확하지 않음
    - 데보라 헬먼 – "오류율 동등성 지표만으로 알고리즘이 공정하거나, 불공정한 지 결정되지 않음"

- **평가 데이터 구축의 문제**
  - 오류율 동등성 기준은 평가 데이터를 기준으로 오류율을 측정하여 집단간 비교
  - 정확한 평가 데이터 구축이 어려운 경우가 많음
    - 예 – 선별적 라벨 문제

- **실무 구현상의 한계**
  - 다양한 보호속성 간의 충돌 문제
    - 싱가포르 통화청 감사 사례 - 성별에 대한 오류율 동등성 충족시 교육 변수에 대한 오류율 격차 증가
  - 단일 문턱값 원칙 위배 가능성
    - Hardt, Price & Srebro (2016)은 오류율 동등성 준수를 위해 집단간 상이한 문턱 값을 설정
    - 법적으로 집단간 상이한 문턱 값을 결정하는 것이 명시적으로 금지되는 사례
    - 그렇지 않다고 하더라도 직접차별 금지원칙 위반 가능성

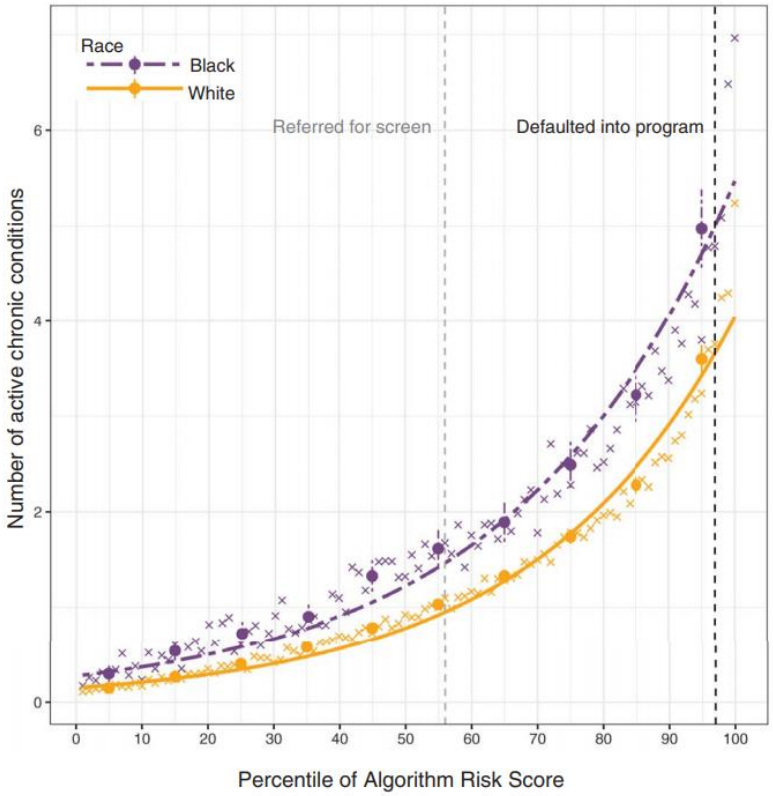
# 출력변수 기준 (3) 예측도 동등성

- **예시 사례**
  - 의료보험회사가 환자의 건강위험도를 평가하여 예방적 조치를 취하고자 함
    - 고위험군 환자를 식별하여 더 자주 검진을 받도록 함으로써 질병을 조기에 발견하도록 함
  - 고위험으로 예측된 흑인은 백인에 비해 더 건강상태가 나쁜 경우가 더 잦다는 문제 발견
  - "고위험" 흑인 환자는 1년 내 입원이 필요한 질병이 발견될 확률은 20%
  - "고위험" 백인 환자는 1년 내 입원이 필요한 질병이 발견될 확률이 10%

- **의의**
  - 어떤 분류모형 $f_{classifer} : X \mapsto \hat{Y}$ 에 대해, 보호속성 $A$가 $\hat{Y}$을 조건으로 진정한 목표변수 값 $Y$와 독립인 경우 $f_{classfier}$는 예측도 동등성을 충족
  - 양성 예측도 동등성과 음성 예측도 동등성으로 구분
  - 집단 캘리브레이션 조건과 동등함
    - 평가 대상에 대해 어떤 확률을 부여하는 인공지능에 있어, 인공지능이 예측한 확률이 실제 예측된 집단에 대해 결과가 발생한 비율과 일치할 것을 요구
    - 예시 – 재범 위험성 예측 알고리즘이 500명에 재범확률 0.2를 부여함. 만약 500명 중 약 100명이 실제로 재범을 저질렀다면, 인공지능이 산출한 재범확률 0.2는 신뢰할 수 있는 예측에 해당함.

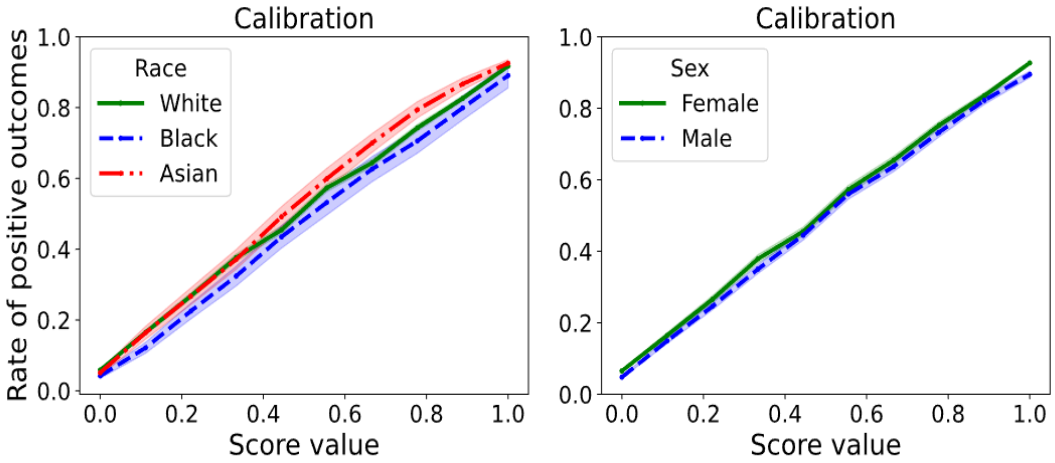
# 예측도 동등성 달성 실패 사례

- **의료 위험 점수 – Obermeyer et al (2019)**
  - 의료 위험 점수 – 위험 점수가 높은 의료보험 가입자에게 예방적 조치를 취하는 제도
  - 동일한 의료 위험 점수에 대해 흑인의 건강 상태가 나쁨
  - 지출 의료비를 기준으로 하여 의료 위험 점수를 산정함으로써 예측도 동등성 실패 발생



# 예측도 동등성의 한계

- **차별 보존적 특성**
  - 인공지능을 개발하거나 도입하는 주체가 차별 시정을 위한 추가적 개입을 하지 않더라도 예측도 동등성 기준은 자동적으로 충족될 수 있음
    - 다만, 적절한 목표 변수를 설정한 것이 전제되어야 함
  - 단순히 예측도 동등성을 보장하는 것만으로는 인공지능의 공정성이 보장되지 않음
    - 예시 - 예측된 소득 수준을 이용하여 공공 서비스나 복지 혜택 대상자를 선정하는 경우
    - 만약 인종이나 성별에 따른 소득 분포에 심각한 차이가 있고, 그 결과 해당 인공지능은 결과 동등성이나 오류율 동등성 조건을 중대하게 위배하고 있을 수 있음
    - 이 경우, 인공지능이 예측도 동등성을 준수한다고 하더라도, 그 활용이 규범적으로 정당화되지 못할 수 있음




Barocas, Hardt & Narayanan (2022)

# 불가능성 정리와 공정성 지표 선택 논쟁

- **불가능성 정리**
  - Barocas, Hardt, & Narayanan (2022) - 출력변수에 대한 공정성 기준 중 (i) 결과 동등성, (ii) 오류율 동등성, (iii) 예측도 동등성은 원칙적으로 동시에 충족하는 것이 불가능함
    - 예외 – (1) 각 집단간 기본 비율이 동일한 경우(즉, 원래 집단간 아무런 차이가 없는 경우) (2) 알고리즘이 완벽한 예측을 하는 경우는 위 불가능성 정리가 성립하지 않고 여러 공정성 기준을 동시에 충족 가능

- **공정성 지표 선택 논쟁**
  - **결과 동등성 지지 견해** – 공정성 지표는 차별시정적 역할을 해야 함
    - Wachter et al. (2021), Hertweck et al. (2021)
  - **예측도 동등성 지지 견해** – 사회적 불평등을 시정하는 것도 중요하지만, 이는 다른 정책수단을 활용해 수행해야 함
    - Hedden (2021) – 인공지능에 대해서는 확률을 정확히 예측하는 것만을 요구해야 함
  - **오류율 동등성 지지 견해** – 오류율 동등성을 다른 지표의 한계를 극복한 절충안으로 고려
    - Zafar et al. (2019) – 오류율 동등성을 “차등적 부당대우”를 막는 것으로 개념화
    - 결과 동등성으로 인한 역차별 문제를 피할 수 있는 장점을 지적

# 기존 공정성 지표의 한계

- **비교 대상의 결여**
  - "감사 대상 인공지능이 공정한가" → "감사 대상 인공지능이 대안과 비교하여 더 공정한가?"
  - 인간 심사자 또는 종전 알고리즘과의 비교 평가가 필요

- **절대적 평가 기준의 부재**
  - 기존 지표는 통계적 정확성 지표의 상대적 동등성만을 고려하고, 그 절대적 수준을 고려하지 못함
    - 위음성율 A 집단 30%, B 집단 25%, 차이 5%
    - 위음성율 A 집단 8%, B 집단 3%, 차이 5%
  - 인공지능의 성능 개선을 통해 공정성 지표가 개선될 수 있음을 염두에 두어야 함

- **불균등 원인의 미고려**
  - 예측 대상 집단 간 노이즈의 차이가 집단간 성능 격차의 원인으로 지적되고 있음
  - 품질이 더욱 우수한 데이터를 수집함으로써 예측 정확도를 개선하는 것이 해법으로 고려되어야 함

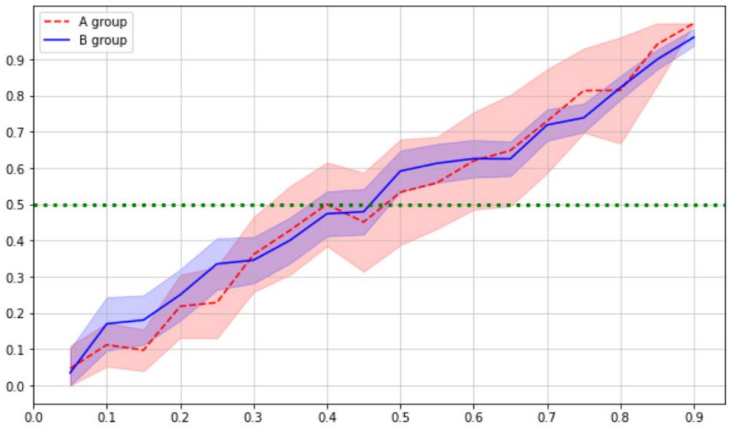
- **인공지능 활용 개인·조직에 대한 미고려**
  - 이용자가 해당 모형의 예측을 어떻게 활용하는지에 대한 고려가 필요
  - 모형 예측에 대한 과의존이 문제되는 경우와 이를 전혀 신뢰하지 못하는 경우는 모두 문제임

# 대안적 공정성 심사 방안

- **인간 평가자 또는 종전 인공지능과의 비교 검증**
  - 인공지능 공정성 감사시 비교 대상인 인간 평가자 또는 대안적 모형을 선정하여 비교

- **예측 값의 불확실성과 신뢰 구간의 제시**
  - 예측 값이 불확실하다면 인간 의사결정자가 해당 사실을 반영할 수 있어야 함(아래 그림 참조)
  - 통계학적 가설 검정 방법론을 적용하는 방법도 고려 가능

- **이해관계자의 참여 및 숙의**
  - 불공정한 인공지능이라도 피해를 받을 개인에 적정한 보상을 제안하고 도입할 가능성이 있음
  - 인공지능 감사 시 관련 이해관계자의 참여와 숙의 과정을 포함하는 것이 바람직함



# 감사합니다!