광고 알고리즘과 차별에 대한 감사(auditing)

2022. 6. 8.
디지털 플랫폼 연속포럼

장여경 (정보인권연구소)

알고리즘 감사(auditing)

美 FTC PrivacyCon

- 연구자들의 다양한 알고리즘 감사 방법론 발표 및 제안
- 광고 알고리즘과 차별 관련 논문
  - 2020년
    Muhammad Ali, Northeastern University 외,
    "최적화를 통한 차별: 어떻게 페이스북 광고 전달 메커니즘이 편향적인 결과로 이어지는가(Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes)"
  - 2021년
    Basileal Imana, University of Southern California 외,
    "구직 광고 전달 알고리즘의 차별에 대한 감사(Auditing for Discrimination in Algorithms Delivering Job Ads)"

英 4개 규제기관 공동

❖2020년 규제기관 공동 포럼 구성

- 알고리즘 감사와 규제기관 역할 모색
  - 경쟁 규제(CMA), 방송통신 규제 (Ofcom), 개인정보보호 규제 (ICO), 금융규제(FCA) 기관 공동
- 2022. 4. <알고리즘 감사> 보고서
  - 알고리즘 감사 방법론을 3종으로 제안 : 거버넌스[관리] 감사, 경험적 감사, 기술적 감사
  - 경험적 감사(Empirical audit): 입력 또는 출력을 사용하여 알고리즘의 효과를 측정하는 방법론. 언론과 학계가 알고리즘 블랙박스에 대하여 미스터리 쇼핑객 방식으로 감사할 때 사용하여 옴

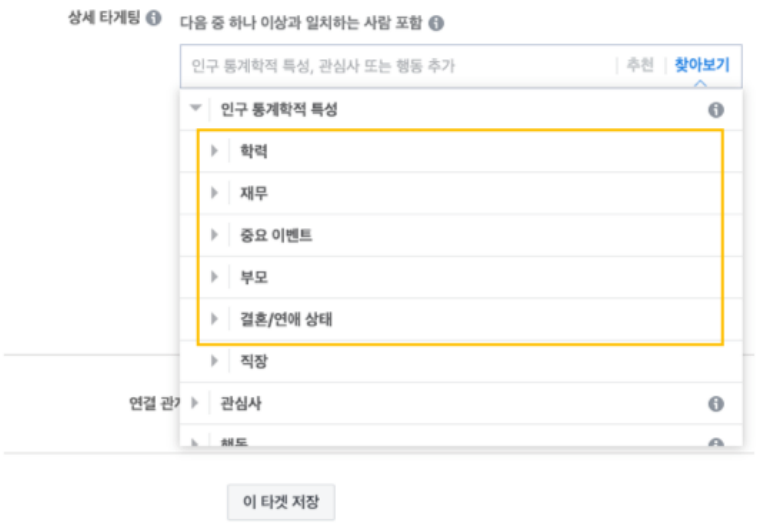
페이스북 광고 소송

- 2016년, 탐사언론 <프로퍼블리카> 폭로
  - 페이스북 구인/주택 광고가 광고주 선택으로 특정 성별, 인종 배제
- 2018년, 미국시민자유연합(ACLU) 고용평등위원회 고발
  - 페이스북 광고가 여성에게는 지붕 수리, 트럭 운전, 기계엔지니어 모집 광고를 노출하지 않음
- 2019년 3월, 페이스북 일부 타겟 광고 중단
  - 국민공정주택연맹(NFHA), ACLU 등이 페이스북을 상대로 한 5건의 차별 소송에 대한 합의
- 그럼에도 2019년 3월 28일, 美 정부 주택도시개발부(HUD) 소제기
  - 페이스북 주택광고가 외국인이나 비(非)기독교인, 장애인, 히스패닉, 이슬람교도 등 이용자 차별 주장

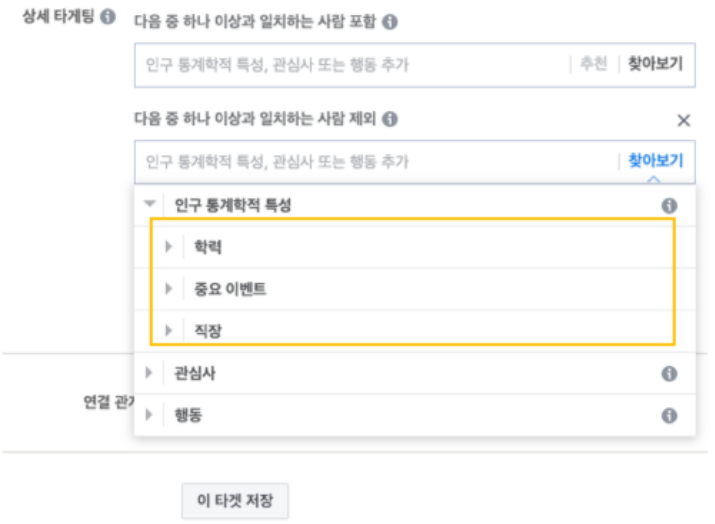
소송 합의 전후 페이스북 정책 변경

- 주택, 고용, 신용 광고에 특정 범주를 사용 못하도록 자동 감지 도구 구축
- 페이스북 광고 정책을 위반하지 않았음을 광고주가 자체 인증하도록 함
- (소송 합의) 주택, 고용, 신용 광고에서 연령, 성별 등에 기반한 타겟팅을 더 이상 허용하지 않으며 "차별 특성이 묘사되거나 관련된" 속성 차단

일치타게팅



제외타게팅



최적화를 통한 차별: 어떻게 페이스북 광고 전달 메커니즘이 편향적인 결과로 이어지는가 (2020)

연구 요약

- 연구자와 언론인들은 광고주가 광고를 보는 이용자 집단을 타겟팅하거나 제외할 수 있다는 사실을 다양하게 폭로해 옴
  - 그러나 광고 전달 프로세스의 역할에 대해서 상대적으로 관심이 적어
- 광고 프로세스: 광고주가 의도하지 않은 편향적 전달 가능성
  - 특정 인구집단 이용자가 광고를 볼 가능성이 다른 이용자보다 적어
- 연구 결과,
  - 페이스북의 시장 및 재정적 최적화와 이용자 집단별 '관련성'에 대한 자체 예측으로 광고 전달 편향이 발생하였음
  - 광고주 예산과 광고 콘텐츠가 페이스북 광고 전달의 편향에 기여함
  - 중립적인 타겟팅 매개변수에도 불구하고 고용 및 주택에 대한 실제 광고에서 성별 및 인종에 따라 전달 편향이 발생함

디지털 광고 플랫폼

- 광고 작성 단계: 광고주가 광고 콘텐츠를 구성하는 텍스트와 이미지를 제출하고 타겟팅 매개변수를 선택
- 광고 전달 단계: 플랫폼에서 광고 전달
  - 이때 광고주 예산, 광고 실적, 이용자의 광고 관련성 예상을 비롯한 여러 요인을 기반으로 특정 이용자 지정
- 정확한 타겟팅 기능으로 높은 재정적 성공
  - 타겟팅 기반: 인구통계학적 속성, 행동 정보, 이용자 PII, 모바일 장치 ID, 웹추적 픽셀 등

광고 전달 알고리즘과 차별

- 광고 '관련성'에 따른 편향
  - 플랫폼은 광범위한 이용자 관심 프로필 + 광고 실적 추적으로 이용자가 광고와 상호작용하는 방식 파악
    → 광고에 관심이 있을 가능성이 가장 높은 이용자에 광고 전달
  - 광고주가 의도 않았거나 인지 못한 집단별 광고 전달 편향 발생 가능
    - 신용, 주택, 고용 편향은 법적으로 금지된 차별
- 시장 효과와 재정적 최적화 - 이용자 집단별 선호도와 가용성에 따른 편향
  - 낮은 예산의 광고주는 '가치 있는' 이용자에 대한 경매에서 패배
  - '가치'가 차별금지 대상 속성인 경우 예산만으로 차별적 효과
- 페이스북은 HUD 소송에서 광고 전달 시스템의 차별 역할 부인

연구 대상 청중

- 청중 선정
  - 광고 자체 경쟁을 방지하기 위해 서로 다른 청중 집단 구성
  - 몇몇 실험은 맞춤 청중 < 미국 이용자 일반에 대하여 반복시행
- 맞춤 청중
  - 지역 전화번호 기반으로 1백만명x20개 리스트 무작위 샘플링
    - 페이스북에 20개 리스트 업로드: 각 22만명 이용자 일치 보고
  - 페이스북은 광고 전달 보고서에서 성별을 분류하지만 인종 분류X
    - DMA 지역을 대리변수로 사용하여 인종별 전달 결과를 분류
    - 유권자 명부(이름, 주소, 인종 등)의 DMA 지역별로 인종 샘플링 후 전달 결과를 DMA별 = 인종별 분석
      - DMA: 미국 카운티별로 구분되어 있는 표준 시장 지역. 페이스북은 광고 전달 위치 기준으로 DMA를 사용함

인종별 광고 청중 구성

DMA(s) [58]	# Records (*A*)		# Records (*B*)		# Records (*C*)	
	White	Black	White	Black	White	Black
Wilmington, Raleigh–Durham	400,000	0	0	400,000	900,002	0
Greenville-Spartanburg, Greenville-New Bern, Charlotte, Greensboro	0	400,000	400,000	0	0	892,097

- Records (A): 특정 카운티 백인 + 비슷한 인구대 다른 카운티 흑인으로 구성
- Records (B): (A) 그룹의 역으로 구성
- Records (C): (A) 그룹의 확대 구성

결과 수집

- 페이스북 마케팅 API를 사용
  - 2분마다 광고 전달 통계를 구함
  - 광고 전달 결과를 연구 속성별로(연령, 성별, 지역) 분류
  - 페이스북은 요청된 각 인구 속성별로 다음과 같은 결과 제공
    - 노출(impressions): 광고 노출 횟수
    - 도달(reach): 광고가 노출된 고유 이용자수
      ← 연구는 이 지표를 사용
    - 클릭(clicks): 광고 수신후 클릭수
    - 고유한 클릭(unique_clicks): 광고 클릭한 고유 이용자수
- 연구는 성별/인종에서 이진값으로 편향을 연구하고자 하였으나, 성별이나 인종이 이 값으로 양분되어 있다고 주장하는 것은 아님

광고 시행

- 모든 광고를 동일한 시간대에 시행함
  - 하루 중 서로 다른 시간대에 페이스북을 이용하는 서로 다른 이용자 집단으로 인한 시간대 효과를 통제하기 위함
- 광고주 옵션 설정
  - 목표: 고려 / 트래픽
  - 최적화 목표: 링크 클릭
  - 트래픽 목적지: 외부 웹사이트
  - 광고창작물: 모든 광고는 관련된 단일 이미지와 텍스트로 구성
  - 청중 선택: 맞춤 청중 + 미국 거주 성인(18세 이상) 추가 제한
  - 예산: 하루 $20 예산으로 시행, 통상 6시간 후에는 중단

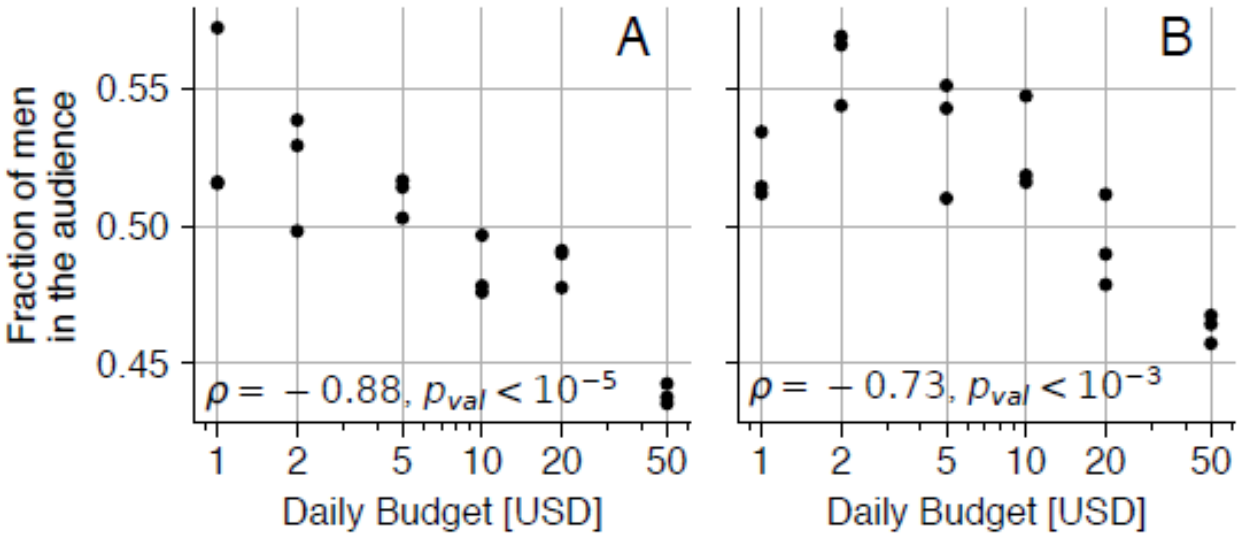
예산 효과

- Lambrecht et al. 가설 검증
  - 여성의 높은 클릭율(CTR)이 여성 대상 광고비를 더 비싸게
  = 같은 가격에 남성용보다 여성용 광고 입찰에 질 가능성이 높다
- 실험 설계
  - 광고창작물과 맞춤 청중을 일관되게 유지
  - 입찰 전략 / 일일 한도에만 변화: $1, $2, $5, $10, $20, $50
    - 통계적 확실성을 위해 각 예산 한도를 다양한 사례에서 운영
  - 2차례 실험 실시: (1) 임의의 지역 전화번호 맞춤 청중 대상
    (2) 모든 미국 이용자 대상 (같은 효과 확인)

예산 효과

❖실험 결과: 광고 시행 일일 예산에 따라 청중 성별이 분할됨
= 높은 예산은 더 높은 비율로 여성으로 할당됨

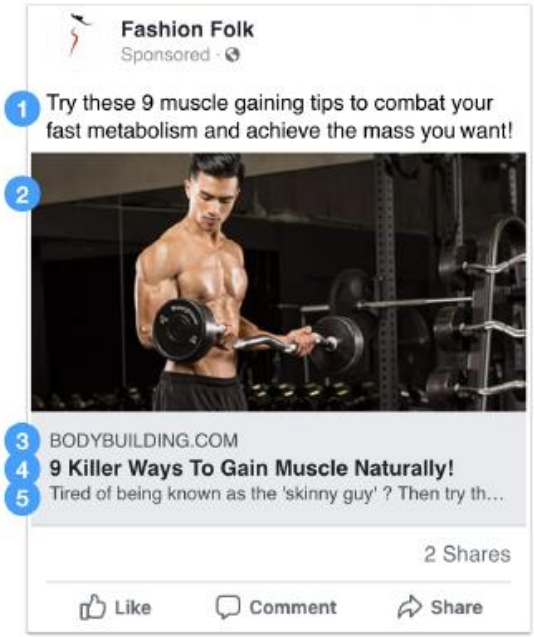
(좌) 모든 미국 이용자 (우) 임의 전화번호 맞춤 청중

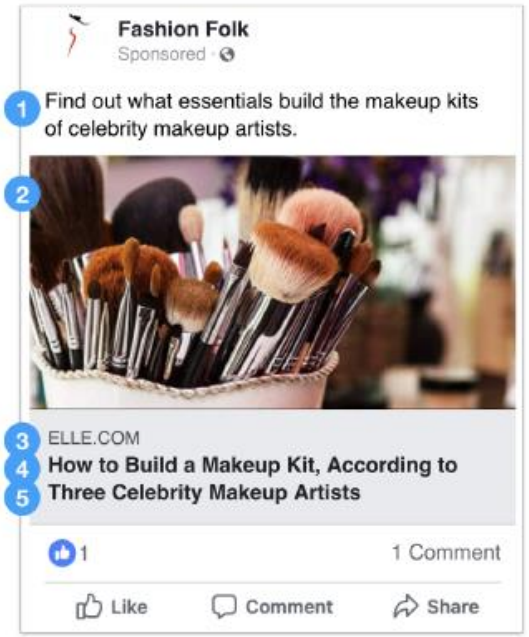


광고창작물 효과

❖남성 vs 여성 성별 전형적 광고 제작

(좌) 보디빌딩 광고 (우) 화장품 광고



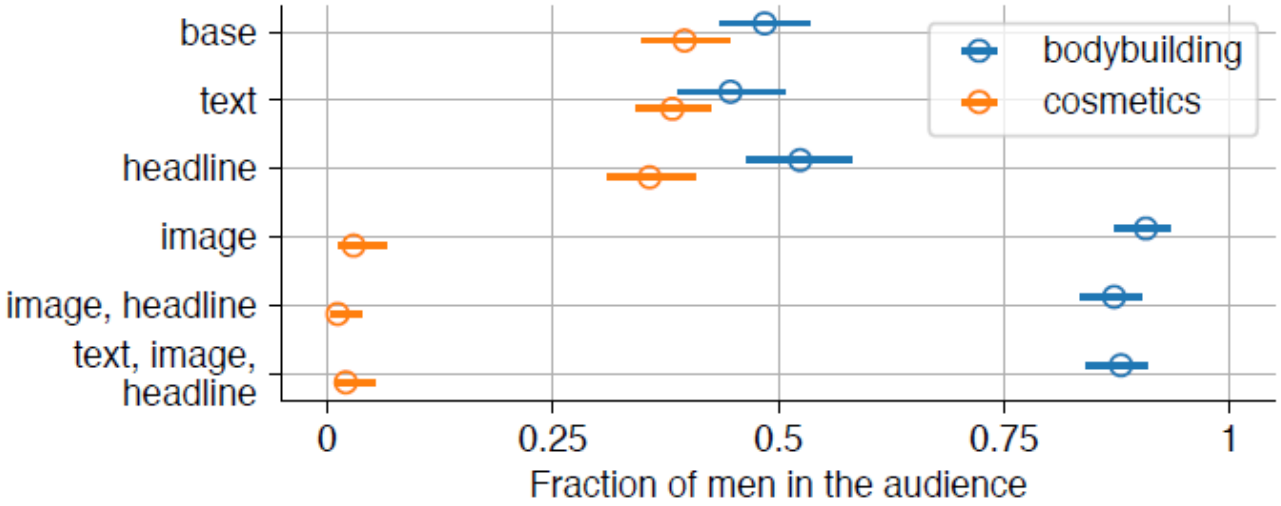


광고창작물 효과

- 실험 설계
  - 동일한 입찰 전략과 예산으로 시행
  - 성별을 명시적으로 타겟팅하지 않음
  - 기준 vs 요소별: 다른 청중 대상 시행 (미국 18세 이상 이용자)
  - 기준(base): 빈 헤드라인+빈 텍스트+빈(흰색) 이미지 +링크 (보디빌딩 bodybuilding.com / 화장품 elle.com)
  - 성별 전형적 광고에서 편향 요인 파악 위해 요소별로 빈 문자열이나 빈 이미지로 대체하여 끔(turn off)

광고창작물 효과 (개별 요소)

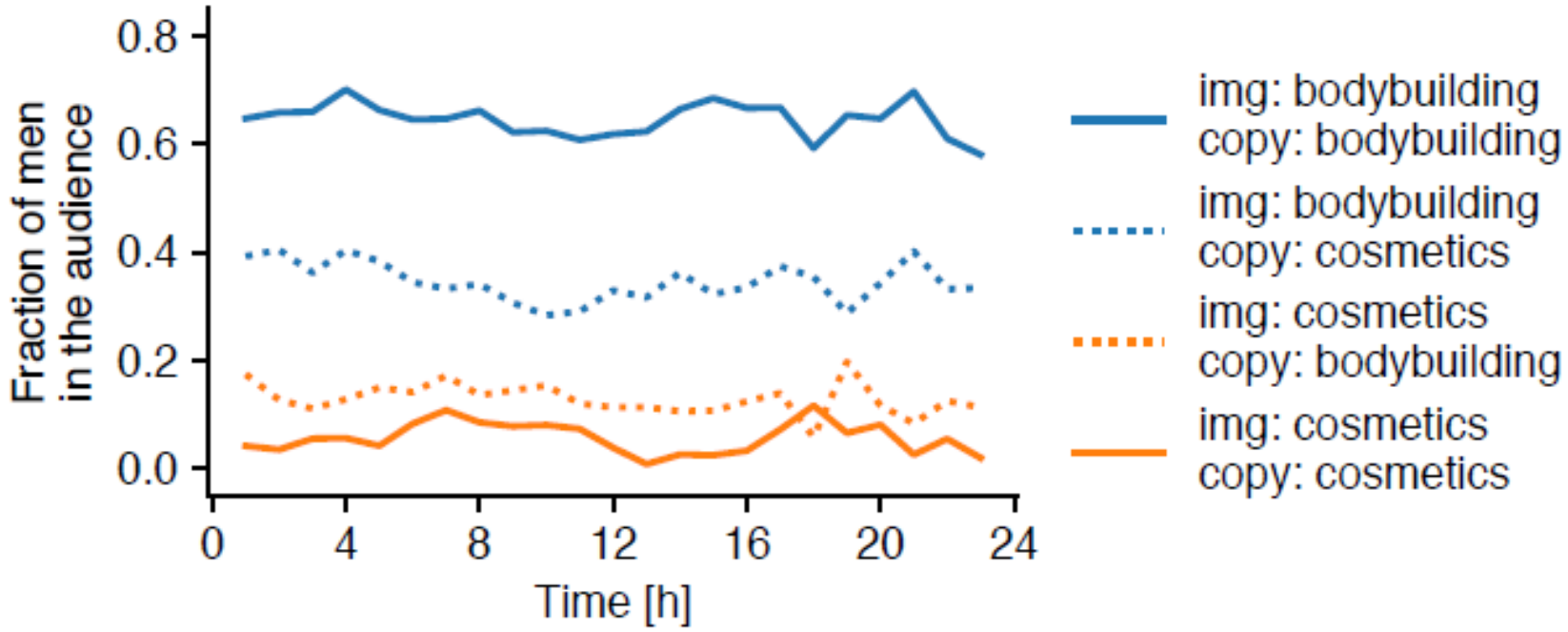
- 실험 결과: 광고 '이미지' 설정에서 극적인 변화가 나타남
  - 보디빌딩 기준(빈 이미지)은 48% 남성 vs 화장품 기준은 40% 남성에 도달
  - 동일한 맞춤 청중 집단에서 보디빌딩 이미지는 평균 75% 이상 남성 도달 vs 화장품 이미지는 평균 90% 이상 여성 도달



광고창작물 효과 (이미지 교체)

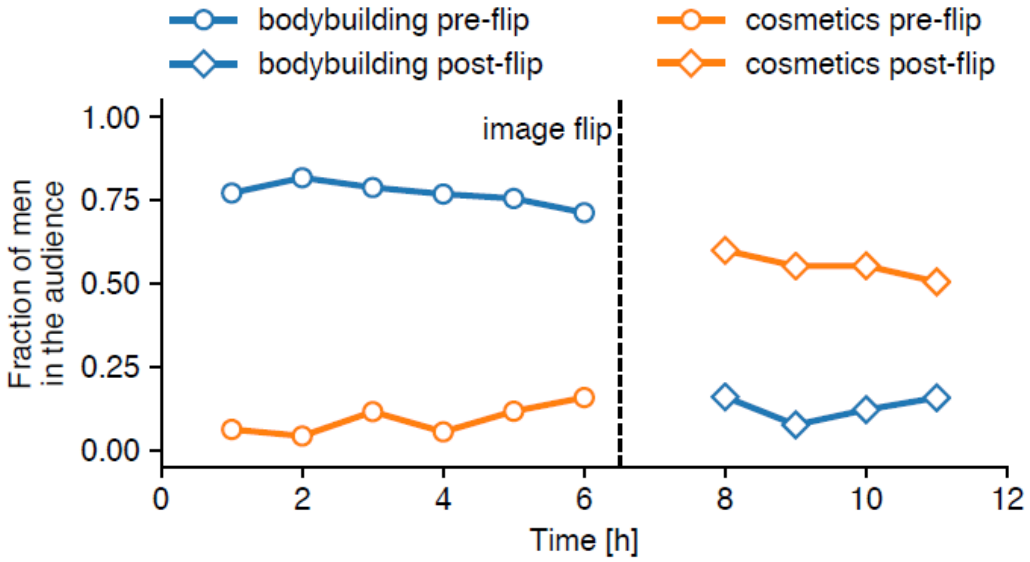
- 실험 설계: 불일치 '이미지'로 교체 후 시간대별 전달 비교
  - 불일치 이미지 광고
    : 보디빌딩 헤드라인, 텍스트, 목적지 링크 + 화장품 이미지
  - 그 반대 불일치 이미지 광고
  - 모두 일치 광고: 헤드라인, 텍스트, 목적지 링크, 이미지
- 실험 결과: 시작 시점부터 편향 발생
  - 단, 일치보다 불일치 이미지에서 편향 정도가 낮음

광고창작물 효과 (이미지 교체)



광고창작물 효과 (이미지 중간 교체)

- 실험 설계: 광고 시행 후 불일치 '이미지'로 교체
  - 일치 광고를 6시간 동안 시행 후 불일치 이미지로 교체
- 실험 결과: 매우 짧은 시간에 광고 전달이 극적으로 반전됨



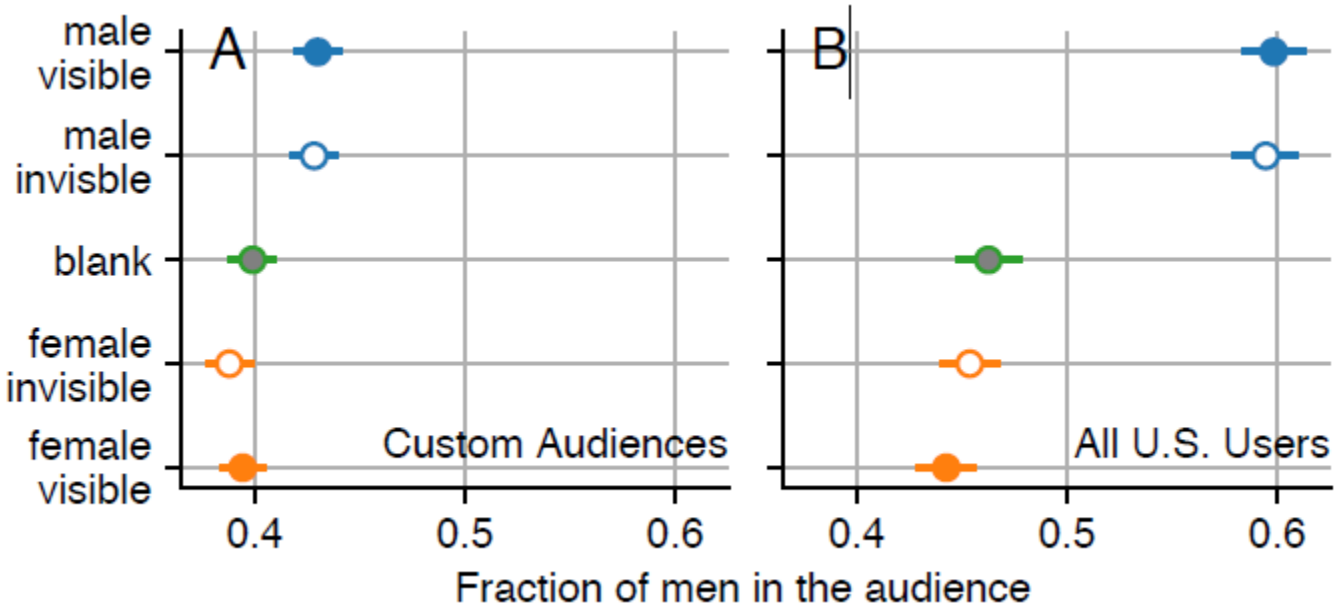
편향 요인 (자동화)

- 가설: 페이스북이 자동으로 광고를 분류하고 이용자별 예상 관련성 점수를 계산한다
- 사람에게 보이는 이미지 vs 기계만 인식하는 이미지
  - 이미지에 98% 불투명도의 알파 채널 사용 (인간에게 비가시적)
- 실험 설계
  - 남/녀 전형적 이미지 10개 x 2종(각 가시/비가시)
    + 비교를 위해 실제 비어있는 공백 이미지 5개 준비
  - 청중별로 (헤드라인, 텍스트, 링크 동일하지만)
    이미지가 다른 3개 광고 전달(남성/여성/공백)

No.	Masculine		Feminine	
	Visible	Invisible	Visible	Invisible
1				
2				
3				
4				
5				

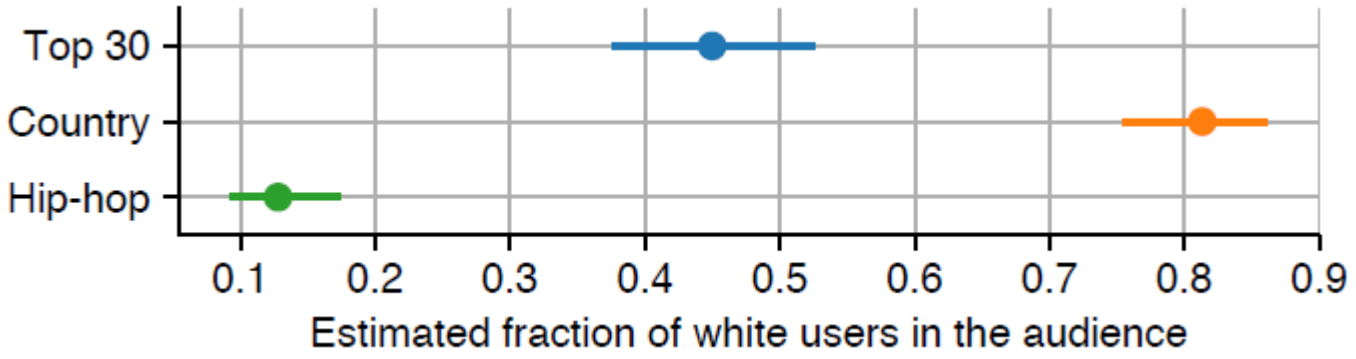
편향 요인 (자동화)

- 실험 결과: 남성vs여성 이미지 전달 차이
  - 남성 전형 이미지는 43% 남성 전달 vs 여성 이미지는 39% 남성에 전달
    - 모든 미국 이용자 대상에서는 각각 58%와 44%
    - 가시/비가시 이미지 결과는 유사함
    - 자동화된 이미지 분류를 시사함



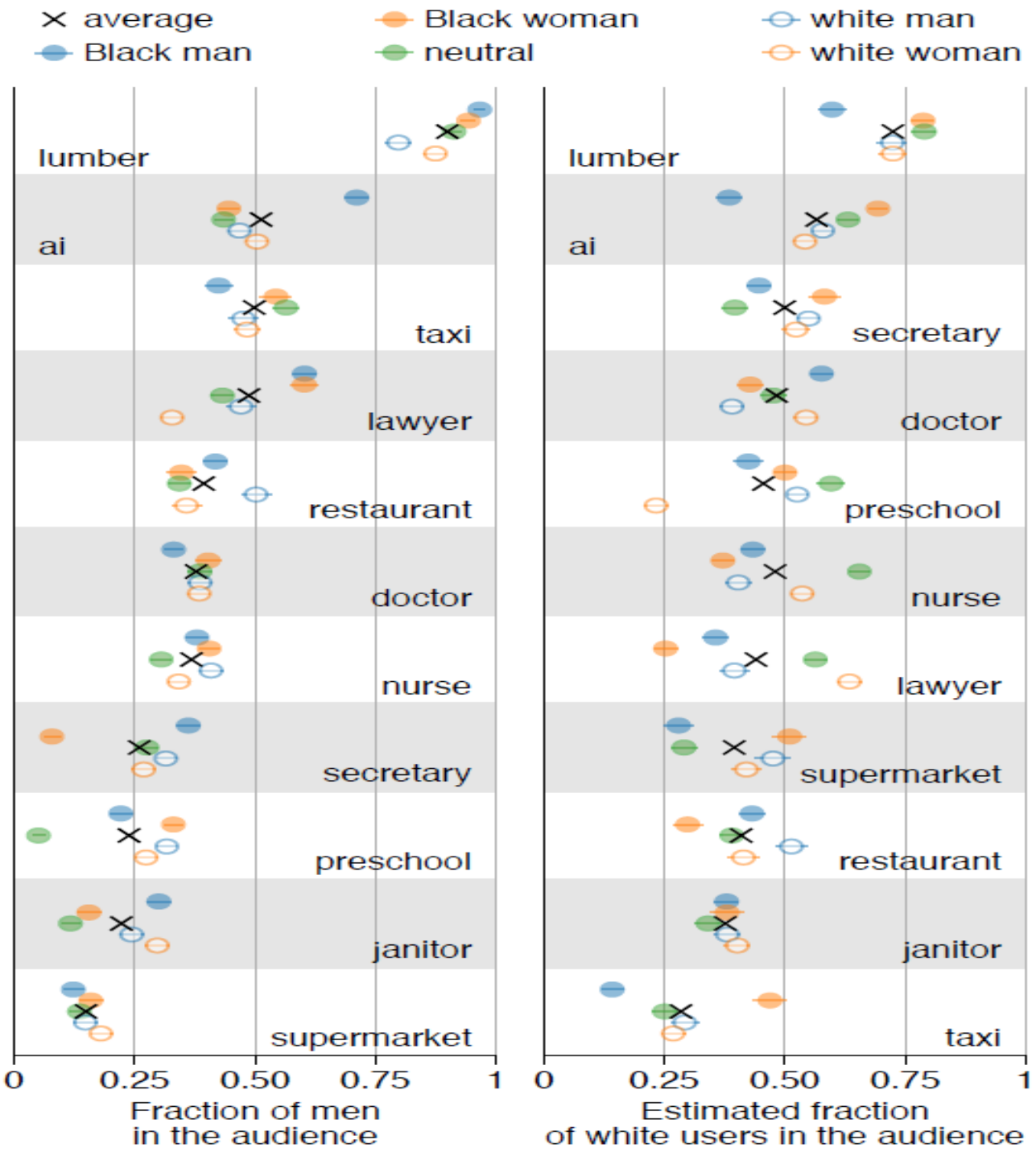
대중문화 광고

- 광고 설계
  - 흑백 인종에게 전형적인 대중문화 광고를 맞춤 청중 인종별로 시행
  - 전년도 베스트셀러 음반 목록으로 이어지는 광고 3종
    - 탑30(중립) / 컨트리뮤직(백인 전형적 취향) / 힙합(흑인 전형적 취향)
    - 모든 전달 방식과 입찰 전략 동일
- 시행 결과
  - 중립적인 광고는 45%의 백인 청중 vs 컨트리뮤직(80%) 힙합(13%)



고용 광고

- 고용 광고는 차별금지 대상
- 광고 설계
  - 인공지능 개발자, 의사, 환경미화원, 변호사, 목재업, 간호사, 유치원 교사, 식당 계산원, 비서, 슈퍼마켓 점원, 택시 운전사 등 11가지 직업 광고
  - 실제 취업 사이트에 있는 적절한 직업 목록 카테고리로 링크
  - 각 직업별 광고 이미지를 5종 준비:
    백인 남성, 백인 여성, 흑인 남성, 흑인 여성, 사람이 없는 이미지
  - 광고 목표: 트래픽 / 동일한 입찰 전략으로 24시간 시행
  - 동일한 청중을 대상으로 독립적으로 시행


average
Black man
Black woman
neutral
white man
white woman
lumber
ai
taxi
lawyer
restaurant
doctor
nurse
secretary
preschool
janitor
supermarket
0
0.25
0.50
0.75
1
Fraction of men in the audience
lumber
ai
secretary
doctor
preschool
nurse
lawyer
supermarket
restaurant
janitor
taxi
0
0.25
0.50
0.75
1
Estimated fraction of white users in the audience

고용 광고

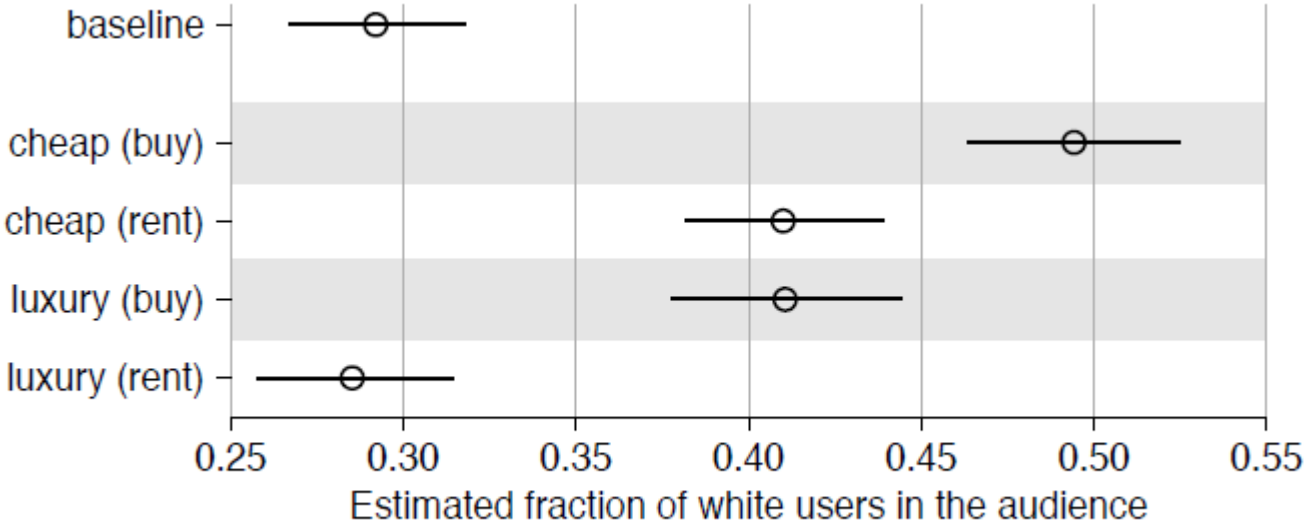
- 시행 결과
  - 광고 전반에 걸쳐 인종 및 성별을 따라 전달 차이
    - 목재업(lumber)에 대한 5개 광고는 총 90% 이상의 남성 및 70% 이상의 백인 이용자에게 전달
    - 환경미화원(janitor)에 대한 5개 광고는 총 65% 이상의 여성과 75% 이상의 흑인 이용자에게 전달
  - 광고 전달 최적화 알고리즘이 광고주 경쟁보다 더욱 큰 영향
    - 5종의 광고들은 상호 경쟁해야 하고 해당 경매 시간에 다른 광고주들과 동일하게 경쟁하였음

주택 광고

- 주택 광고는 차별금지 대상
- 광고 설계
  - 부동산 유형(임대 vs 구매)과 내부 비용(수리 vs 호화) 다양화
    - 가격은 광고 문구와 이미지를 통해 암시
    - 실제 주택 사이트에 있는 노스캐롤라이나 매매/임대 아파트 목록으로 링크
  - 일반(비주택) 텍스트 기준 광고 동시 시행 (구글 링크)
  - 광고 목표: 트래픽 / 동일한 입찰 전략으로 12시간 시행
  - 동일 광고를 A 그룹과 B 그룹에 각 시행 (지리적 요인 고려)

주택 광고

- 시행 결과: 인종별 광고 전달 편향
  - (명품 임대) 백인 27% vs 흑인 72%
  - (저가 구입) 백인 49% vs 흑인 51%



결론

- 전통적인 미디어 광고와 차이점
  - 광고주가 광범위하게 고용/주택 광고를 해도 의도치 않은 광고 전달 편향 (불법적 차별 가능성)
  - 시청자가 본인이 속한 집단을 벗어나 선택하기 어려움 (특히 인종과 자산은 스스로 설정 < 플랫폼이 추론하는 영역)
  - 개인화된 타겟 광고 결과에 대하여 관찰 및 감지, 공공 조사가 어려움
- 자동화된 광고 전달 메커니즘으로 페이스북이 핵심 역할
  - 플랫폼에서 최적화한 광고 전달과 시장 효과의 결합

결론

- 현재 페이스북 정책처럼 개별 광고주가 타겟팅에서 차별하지 않도록 보장하는 것만으로는 규제 기관과 대중이 추구하는 차별 금지 목표 달성 어려워
  - 미국 통신품위법(CDA) 230조는 중개자(온라인 플랫폼 포함)가 제3자 콘텐츠에 대한 책임을 지지 않도록 보호하지만, 온라인 플랫폼의 중립성을 검토해볼 필요성이 있음
  - 정책입안자 및 플랫폼은 디지털 광고 차별 방지를 위해 광고주의 타겟팅 옵션뿐 아니라 광고 전달 최적화 알고리즘의 역할을 검토해야
    - 새로운 알고리즘 및 기계학습 기술 개발 뿐 아니라
    - 광고 플랫폼의 높은 책무성과 투명성이 요구됨

구직 광고 전달 알고리즘의 차별에 대한 감사 (2021)

연구 요약

- (선행 연구) 광고주가 선택하지 않아도 플랫폼 광고 전달 알고리즘의 최적화 기능 → 성별/인종에 따른 편향 발생
  - 블랙박스 알고리즘의 편향 요인을 구분할 필요
- (법적 규제) 성별/인종 등 금지 요인에 의한 편향(법적 차별) vs 실제 자격 편향(법적 차별X) 구분
- (감사 방법론) 알고리즘 자체의 최적화/학습으로 인한 결과 vs 자격 편향으로 인한 결과 구분
- (방법론 실행) 두 개의 주요 구인광고 플랫폼
  : 페이스북 vs 링크드인(LinkedIn)

광고 전달 최적화 알고리즘

- 광고 전달 최적화(ad delivery optimization) 알고리즘
  이 누가 어떤 광고를 볼지 / 광고주 지불액이 얼마인지 결정
  - 광고주: 광고 게시, 목표 청중/캠페인 예산/광고 목표 특정

  → 플랫폼: 알고리즘이 해당 이용자를 타겟팅하는 광고주 간 광고 경매 실행 후 광고 전달
  - 광고 전달 최적화 알고리즘: 광고주가 선택한 매개변수
    + 광고 관련성 점수 고려
    - 관련성 점수: 광고 플랫폼 알고리즘이 개별 이용자의
      예상 참여 수준, 이용자 가치 등을 기반으로 계산 (비공개 영업비밀)

직업 자격

- 美 타이틀 VII: 인종, 피부색, 종교, 성별 및 출신국가를 근거로 한 고용, 주택, 신용 차별 금지. 단, 직업 자격간 차이는 정당
  - 광고 전달 알고리즘 편향 연구 vs 편향의 법적 차별 연구 사이 간극이 해결되어야 할 필요성이 있음
  - 규제기관 법집행 가능성 높여야 할 필요성이 있음
  - 플랫폼의 법적 방어에 사용될 가능성이 있음
- 편향의 요인에 따라 법적 책임에 차이 발생
  - 광고주 선택에 따랐다면 통신품위법 230조에 따라 중개인(플랫폼) 면책
  - 광고주/플랫폼의 편향 기여도 및 그 법적 책임 모호한 상황

광고 전달 편향

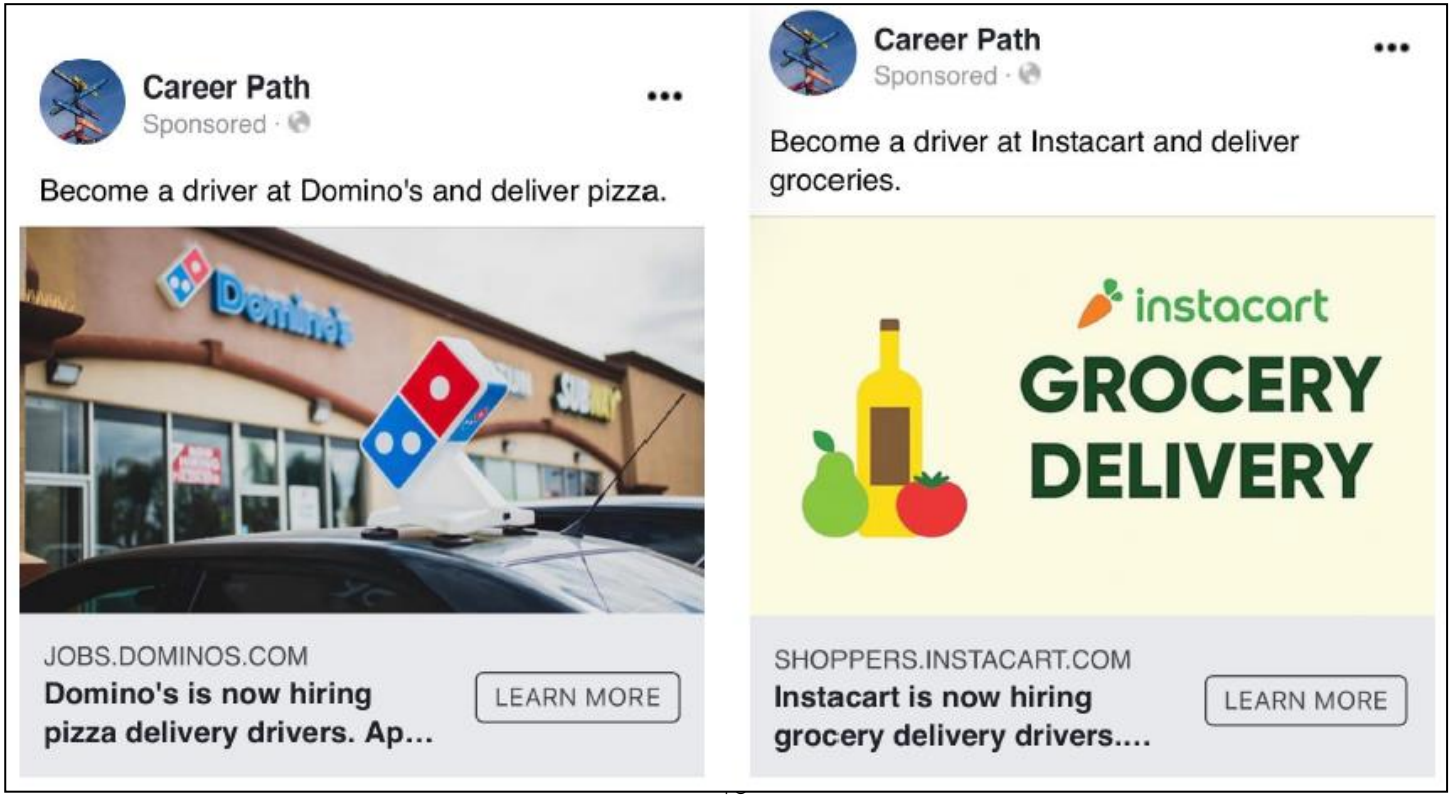
- 광고 전달 편향(skew): 광고 수신자들 사이에서 한 집단의 구성원이 과잉 또는 과소 표출되는 결과를 낳는 광고 플랫폼 알고리즘에 의한 결정
- 가능한 편향 요인 (연구주제)
  - 광고주 타겟팅 매개변수 및 청중: 현재 변경됨 (소송합의)
  - 광고 전달 최적화 알고리즘의 선택: 광고 관련성, 참여도, 광고주 만족도, 수익성 또는 기타 비즈니스 목표 최대화 (자격 요인 문제)
  - 광고주 광고 목표 선택: '도달(reach)' vs '전환(conversion)' 옵션
    - 전환: 페이지 방문, 구매, 회원가입 등 추가적인 동작
  - 기타 교란: 인구집단간 로그인율 차이, 시간대 효과, 광고주간 경쟁률

연구 설계

- (선행 연구) 플랫폼의 선택과 무관한 요인 통제
  < 알고리즘의 역할에 분리 집중
  - 플랫폼 선택과 무관한 요인:
    광고 시행 중 온라인 상태인 인구집단, 타광고주와 경쟁 등
- 자격 요건은 유사 but 성별 분포 현실이 편향적인
  직업 광고쌍을 같은 청중에 대해 동시 시행
  - 편향적 결과가 나타나면 청중의 자격 차이로 인한 것이 아님
- 일반 광고주 이상의 특별한 접근을 가정하지 않음
  - 제3자의 공익적 감사 가정: 알고리즘의 코드나 입력값, 또는 플랫폼 구성원이나 광고주의 데이터나 동작에 대한 접근 없이 조사
  - 규제기관의 역량 한계 극복
  - 플랫폼 제공 기능 의존 극복

같은 자격 / 성별 편향 광고쌍

- 배달기사의 경우:
  도미노(좌, 남편향) vs Instacart(우, 여편향)



연구 대상 플랫폼 - 링크드인과 페이스북

- 광고주가 광고 형식, 입찰 전략, 지불에 대한 옵션 직접 선택
  - 광고 목표: 인지, 고려, 전환
  - 광고 대상: 지역, 나이, 성별에 따른 청중 타겟팅
    - 구인 광고의 경우에는 연령과 성별 비활성화/제한
    - (링크드인) 직업 플랫폼으로 직급, 학력, 경력 별 추가 타겟팅 제공
- 광고주가 연락처 목록 업로드하여 일치하는 청중 맞춤 가능
  - (링크드인) 이름과 성, 이메일 주소
  - (페이스북) + 우편번호, 전화번호 추가 제공
- 보고서 기능: 웹/API를 통한 광고 집행 결과제공
  - (링크드인) 지역, 직급, 업종, 회사 분류
  - (페이스북) 지역, 나이, 성별 분류

감사 방법론

- 1단계: 광고 플랫폼의 기능을 사용하여 광고 수신자의 성별을 추론할 수 있는 맞춤 청중 구성
- 2단계: 청중 집단의 모든 사람이 동등한 자격을 갖지만 구인 회사의 실제 직원의 성별 분포에 차이가 있는 직업 범주를 신중하게 선택하여 직업 자격 요인을 통제
  - 이 방법론은 성별에 따른 편향을 설명하지만 직업 범주 선택에 따라 인종 및 연령과 같은 다른 속성에도 일반화할 수 있음
- 3단계: 각 직업 범주에 대해 쌍을 이루는 광고를 동시에 시행하고 통계를 통해 광고 전달 결과 편향 여부를 평가
  - 구인 광고의 맥락에서 기회의 평등: 어떤 직무에 대해 자격이 있는 인구 집단의 개인이 다른 인구 집단의 동일한 자격을 갖춘 개인과 비교하여 동일한 비율로 긍정적인 결과(광고 노출)를 얻어야 함

맞춤 청중 구성

- 성별 추론
  - (링크드인) 광고 노출의 성별 구분을 제공하지 않음
    → 지역(노스캐롤라이나)을 대리변수로 사용 (카운티별 관찰)
  - 노스캐롤라이나 유권자명부: 이름, 우편번호, 카운티(지역), 성별, 인종, 연령 표시
- 동일한 이용자에 대한 광고간 경쟁을 방지하기 위하여 동일한 직업 범주의 광고는 한 쌍을 동시에 시행
  - 연구내 자기 경쟁을 방지하기 위하여 다른 직업 범주 광고는 다른 시간대 시행
- 테스트-재테스트 편향(테스트 간 알고리즘 학습)을 방지하기 위하여 일반적으로 다른 (그러나 동등한) 청중으로 반복

연구에 사용된 맞춤 청중

(남녀 구성) 특정 카운티의 남성 + 비슷한 인구대 다른 카운티의 여성
Aud #*f는 각각의 역

ID	Size	Males	Females	Match Rate
Aud #0	954,714	477,129	477,585	11.83%
Aud #1	900,000	450,000	450,000	11.6%
Aud #2	950,000	450,000	500,000	11.8%
Aud #0f	850,000	450,000	400,000	11.88%
Aud #1f	800,000	400,000	400,000	12.51%
Aud #2f	790,768	390,768	400,000	12.39%

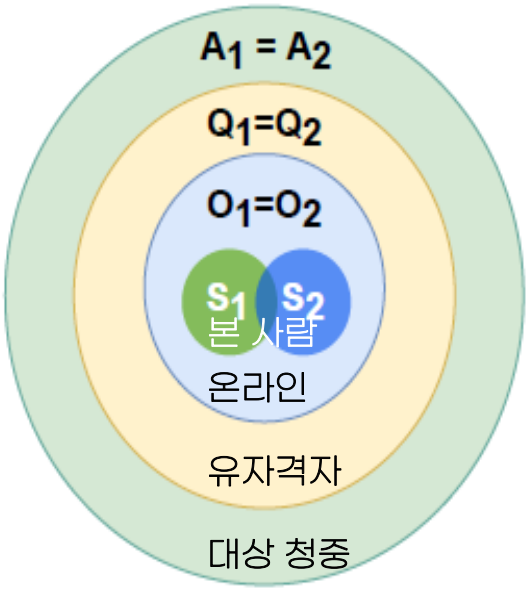
(목록 업로드 후 일치율) 링크드인(12%), 페이스북(미공개)

자격 통제

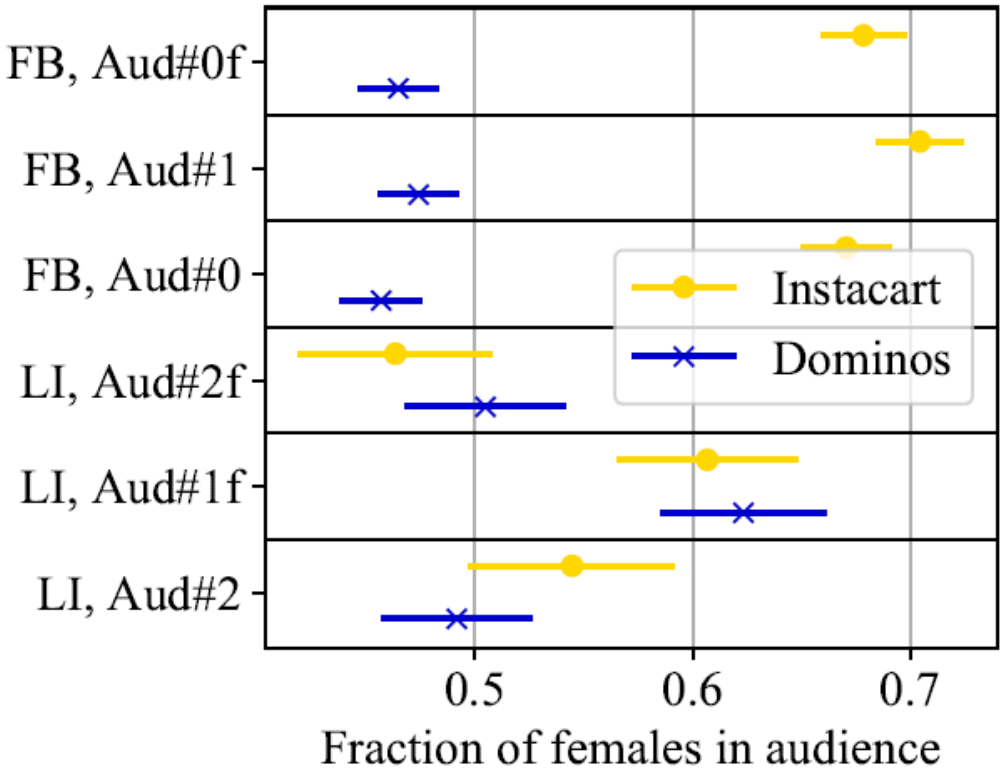
- 한 쌍으로 구성된 두 개의 실제 회사 구인 광고
  - 대상 청중이 유사한 자격 요건을 갖춰야 함
    - 차별적이지 않은 광고 전달의 경우: 수신자 성별 분포가 균등해야
  - 현실 세계에서 둘 사이 편향된 성별 분포를 보여야 함
    - 플랫폼은 참여도 또는 비즈니스 목표에 최적화하기 위해 이러한 과거 편향을 학습하였을 수 있음
- 세 가지 직업 범주에 대한 광고쌍 (플랫폼 이용자별 선호도 통제)
  - 배달기사(저숙련): 도미노(남편향) vs Instacart(여편향)
  - 소프트웨어 엔지니어(고숙련): Nvidia(남편향) vs 넷플릭스(여편향)
  - 영업직(저숙련/인기직): Leith Automotive(자동차 대리점, 남편향) vs Reeds Jewelers(보석 소매상, 여편향)

기타 요인 및 지표 통제

- 중립적 광고 콘텐츠 제작
  - 성중립적 텍스트와 이미지 사용
  - 각 광고는 실제 구직 광고(각 회사 채용 페이지 /링크드인 채용 공고)에 링크
- 링크드인과 페이스북의 맞춤 청중을 대상으로 같은 직업 범주에 대한 광고쌍 동시 시행
  - 광고쌍은 시간대 요인, 타광고주와 경쟁 요인 동일
  - 취업 의지로 인한 편향을 방지하기 위해 한쌍은 동일한 물리적 위치에 있는 직업
- 광고주 선택에 따른 시행 구분
  - 도달(선) vs 전환(후) 옵션 각각 시행
- 캠페인당 $50 예산으로 하루종일/소진시까지
  - 통계적 평가에 합리적인 샘플 크기(모든 광고가 최소 340번 노출)
  - 실험 광고 시행에 총 5,000달러 비용 소요



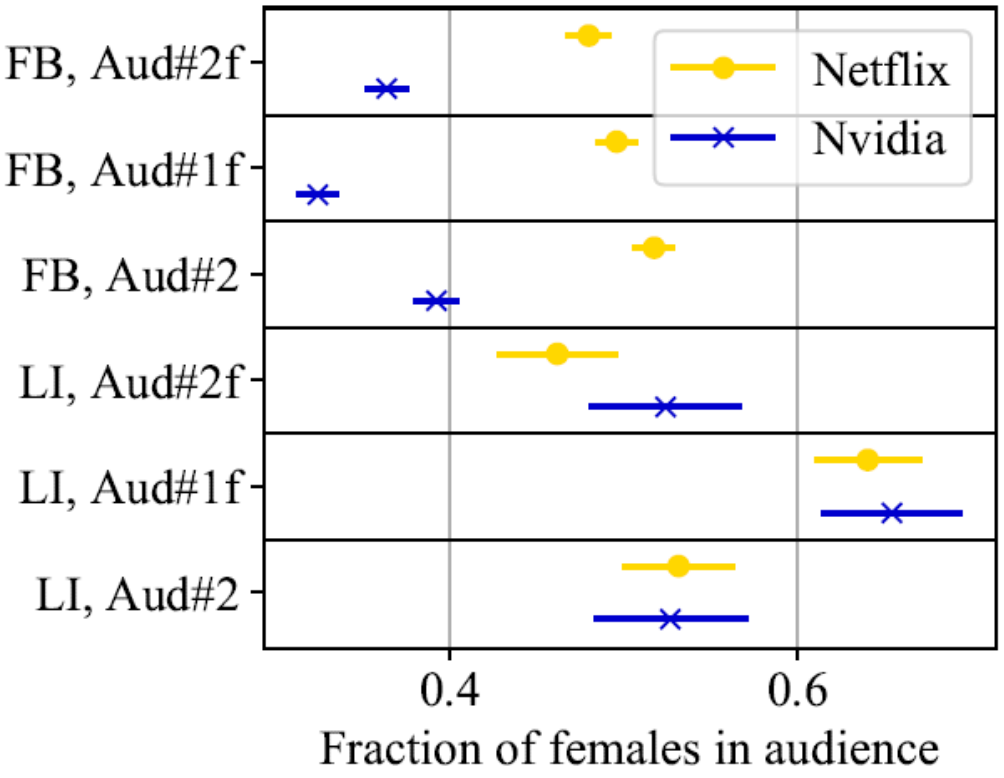
실험 결과 (배달 기사)



(실제) Instacart: 여성 50%
vs 도미노 남성: 98%

(3종의 청중에 광고 노출)
페이스북: 유의미한 편향
vs 링크드인: 성별 편향X

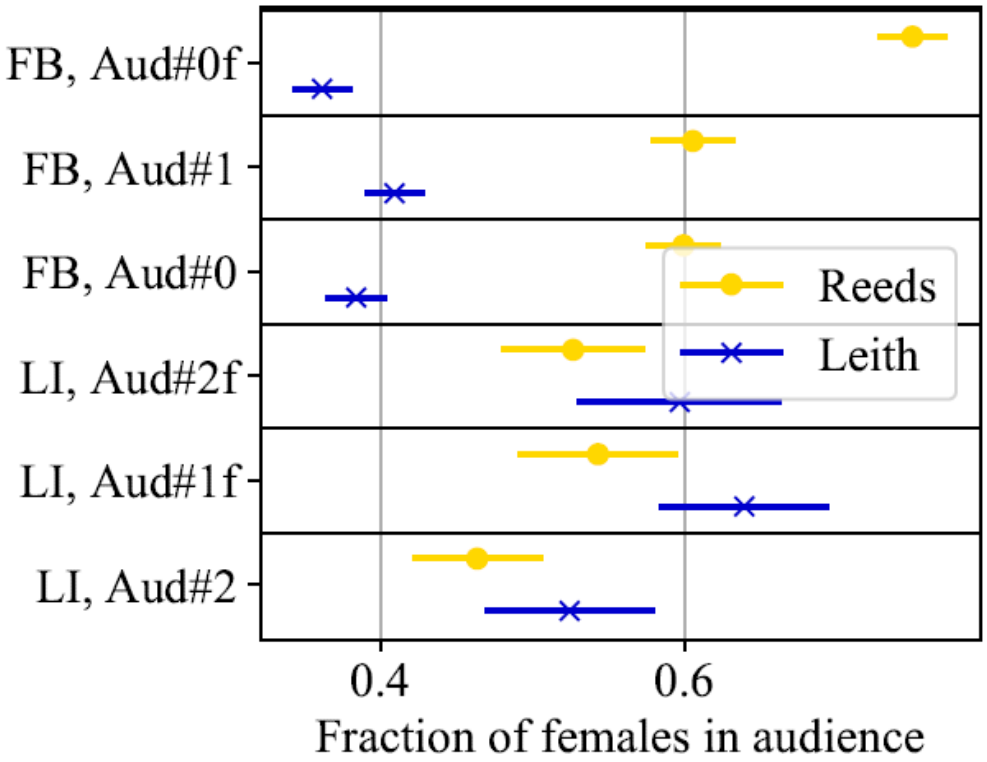
실험 결과 (소프트웨어 엔지니어)



(실제) 넷플릭스: 여성 35%
vs Nvidia: 여성 14~19%

(3종의 청중에 광고 노출)
페이스북: 유의미한 편향
vs 링크드인: 성별 편향X

실험 결과 (영업직)



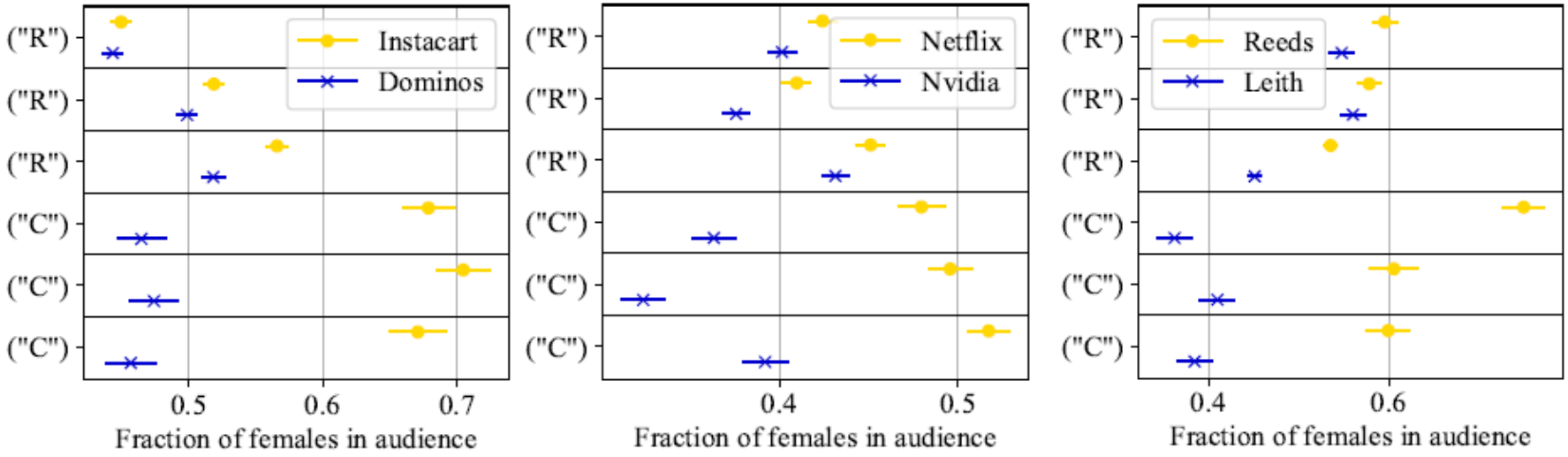
(실제) Reeds: 여성 62%
vs Leith: 여성 17.9%

(3종의 청중에 광고 노출)
페이스북: 유의미한 편향
vs 링크드인: 성별 역편향

실험 결과 (페이스북 도달율)

❖‘전환’(C) 목표 vs 중립적인 ‘도달’(R) 목표로 페이스북 관찰

(좌) 배달기사 (중) 소프트웨어 엔지니어 (우) 영업직



결론

- (배송기사, 영업직) 도달 목표의 경우 3개 실험 중 2개에서 유의미한 편향
- 광고주가 성별 균형적 청중을 대상으로 하는 경우에도 광고 전달 알고리즘이 성별 편향적 전달
- 광고주의 '도달' 목표 선택에서도 편향된 전달이 발생하므로 편향은 광고주 선택이 아닌 플랫폼 알고리즘 선택에 기인함
- 페이스북에 불법적 차별에 대한 법적 책임이 있을 수 있음
  - '전환' 목표 선택에 따른 편향: 광고주의 책임 < 플랫폼 책임

    (최적화 알고리즘 작동 방식과 입력을 결정하는 권한有)

감사합니다