

fast.ai 스터디 3주차 일요일

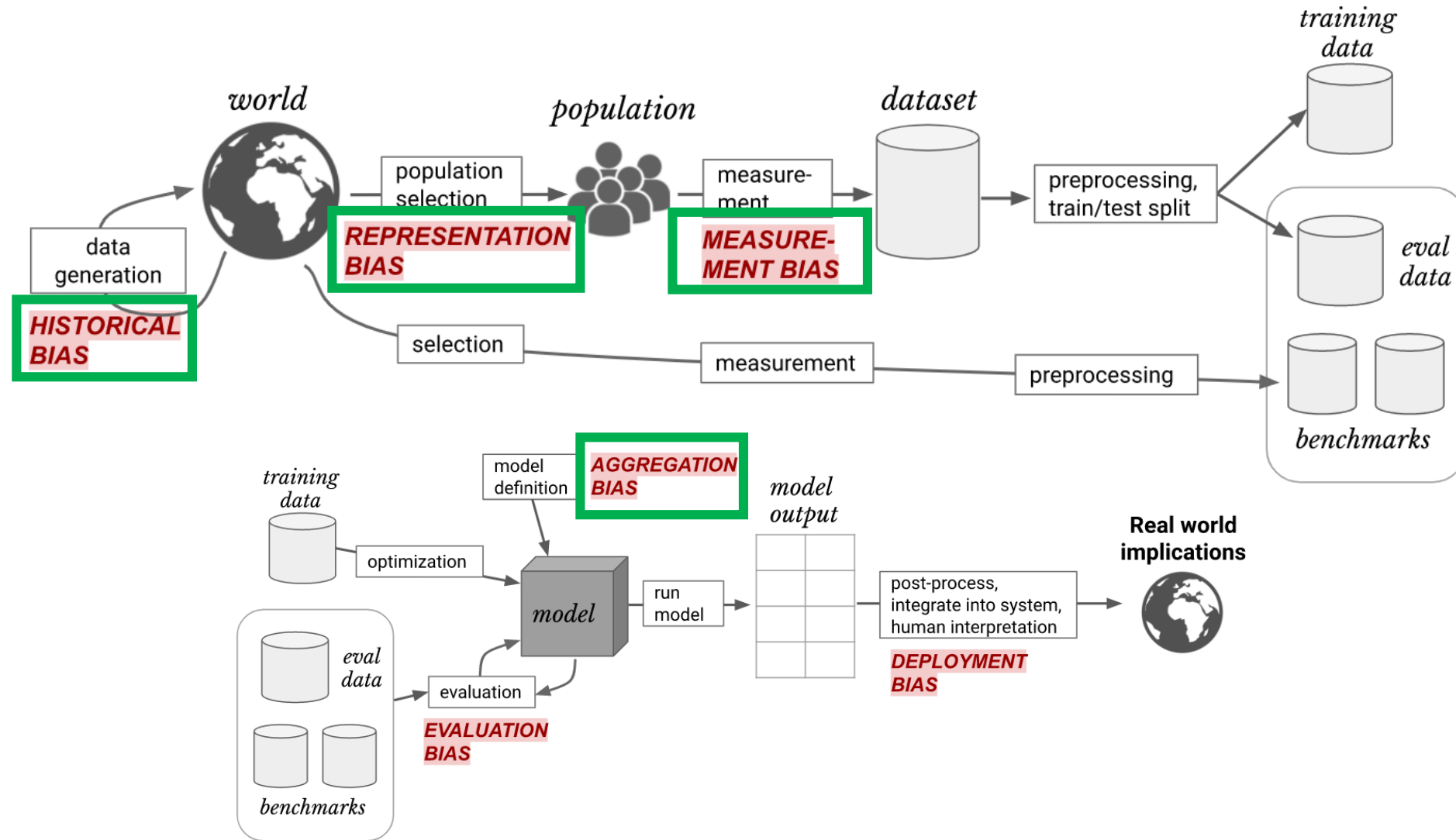
Topics in Data Ethics : Bias

2020. 09. 13. 최민영

Bias

- Bias는 다양한 분야에서 다른 의미로 사용한다.
 - 통계학자
 - 머신러닝 엔지니어
 - 데이터 윤리학자 \leftarrow target

Bias Types



Historical bias - 1

- 역사적 편향은 사람이 편향되고, 과정이 편향되며, 사회가 편향된 데서 비롯된다. 수레쉬와 구트태그는 "역사적 편향은 데이터 생성 프로세스의 첫 단계인 근본적이고 구조적인 문제로서 완벽한 샘플링과 특징 선택에서도 존재 할 수 있다."
- historical race bias in the US, from the New York Times article "Racial Bias, Even When We Have Good Intentions"
 - 의사들이 동일한 파일을 보여주었을 때, 그들은 흑인 환자들에게 심장 카테터 치료(도움이 되는 시술)를 권할 가능성이 훨씬 적었다.
 - 중고차를 협상할 때, 흑인들은 초기 가격을 700달러 더 높게 제시 받았고 훨씬 적은 할인을 받았다.
 - 흑인 이름으로 크레이그리스트의 아파트 임대 광고에 응답한 결과 백인 이름보다 반응이 적었다.
 - 백인 배심원단은 백인 배심원보다 흑인 피고인에게 유죄를 선고할 확률이 16%포인트 높았지만 배심원단이 흑인 피고인을 한 명 두 명 모두 같은 비율로 유죄를 선고했다.

Historical bias - 2

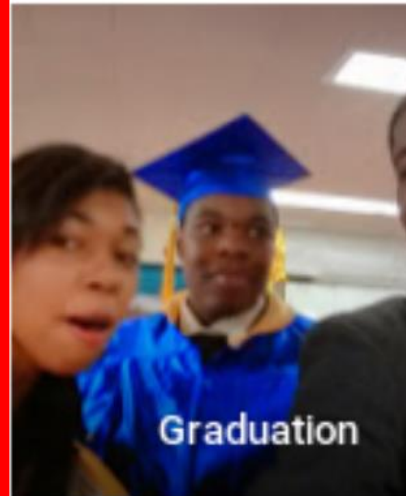
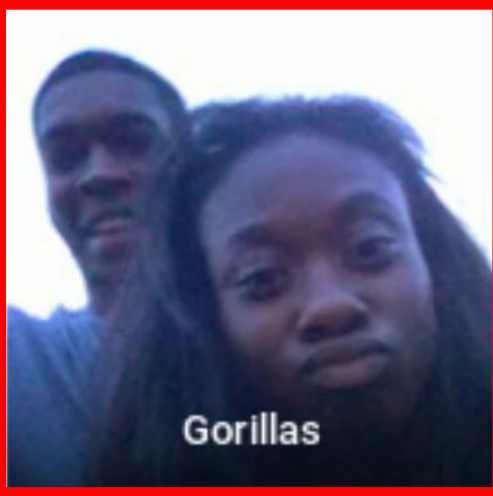
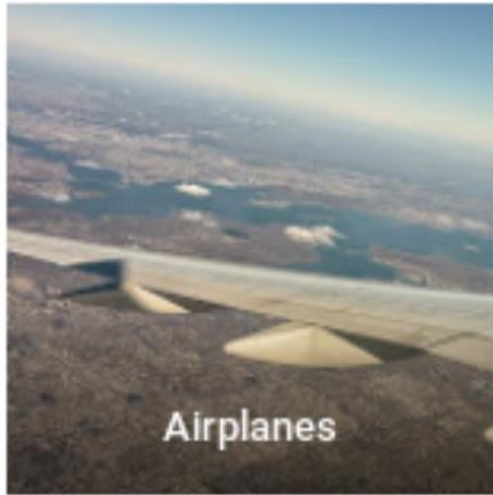
- COMPAS 알고리즘, 양형 및 보석을 결정, racial bias

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%









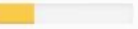





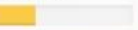
- 전체적으로 노스포인트의 평가 도구는 재범률을 61% 정확하게 예측한다. 그러나 흑인들은 백인들보다 거의 두 배나 더 위험하다는 꼬리표가 붙을 가능성이 높지만 실제로 그렇지 않다. 그것은 백인들 사이에서 정반대의 실수를 저지른다. 그들은 흑인들보다 훨씬 더 낮은 위험으로 분류되지만 다른 범죄를 저지르기 쉽다.

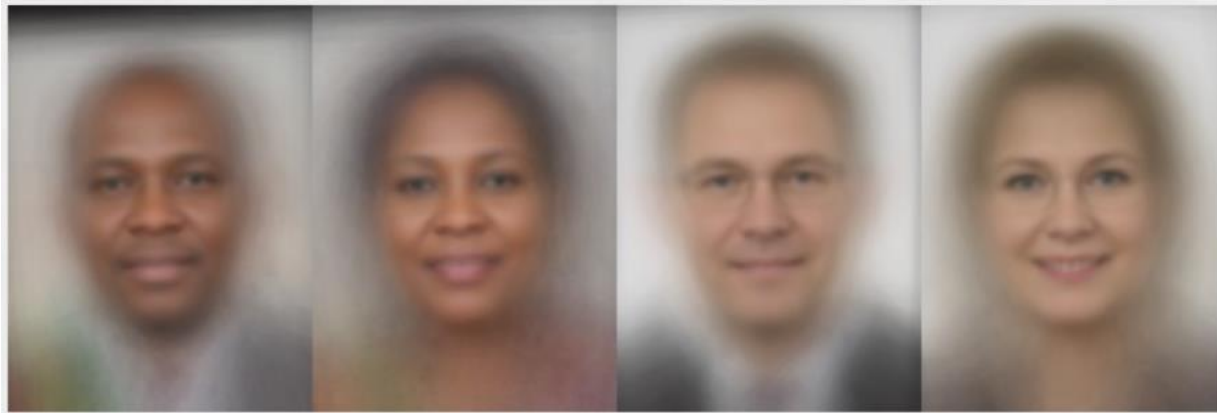
Historical bias - 3



- Google Photos 분류기가 흑인을 고릴라로 분류했다.

Historical bias - 4

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 



- IBM's system, for instance, had a 34.7% error rate for darker females, versus 0.3% for lighter males
- 어두운 피부색을 구별하는게 더 어렵다. 그러나, 1년뒤 백인과 동등하게 성능을 향상 시켰다.
- 이것이 의미하는 것은 데이터셋이 불충분 했거나, 테스트를 실패한 것이다.

Historical bias - 5

- One of the MIT researchers, Joy Buolamwini, warned
 - 우리는 지나치게 자신만만하고 준비가 덜 된 자동화의 시대로 접어들었다. 윤리적이고 포용적인 인공지능을 만들지 못하면 기계 중립을 빙자해 시민의 권리와 양성평등의 이익을 잃을 위험이 있다.
- ["No Classification Without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World"](#) by Shreya Shankar et al. states
 - 우리는 지역 다양성을 평가하기 위해 두 개의 대규모의 공개적으로 이용할 수 있는 이미지 데이터 세트를 분석하며, 이러한 데이터 세트가 관찰 가능한 미국중심과 유럽중심의 표현 편향을 나타내는 것으로 보인다. 또한 이러한 데이터 세트에 대해 훈련된 분류자를 분석하여 이러한 훈련 분포의 영향을 평가하고 서로 다른 로케일의 이미지에 대한 상대적 성능의 강한 차이를 발견한다.

Historical bias - 6



Ground truth: Soap Nepal, 288 \$/month

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment



Ground truth: Soap UK, 1890 \$/month

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave



Ground truth: Spices Phillipines, 262 \$/month

Azure: bottle, beer, counter, drink, open
Clarifai: container, food, bottle, drink, stock
Google: product, yellow, drink, bottle, plastic bottle
Amazon: beverage, beer, alcohol, drink, bottle
Watson: food, larger food supply, pantry, condiment, food seasoning
Tencent: condiment, sauce, flavorer, catsup, hot sauce



Ground truth: Spices USA, 4559 \$/month

Azure: bottle, wall, counter, food
Clarifai: container, food, can, medicine, stock
Google: seasoning, seasoned salt, ingredient, spice, spice rack
Amazon: shelf, tin, pantry, furniture, aluminium
Watson: tin, food, pantry, paint, can
Tencent: spice rack, chili sauce, condiment, canned food, rack

Historical bias - 7

- NLP model

English Turkish Spanish Detect language ▾

↔

English Turkish Spanish ▾

Translate

She is a doctor.
He is a nurse.

31/5000

O bir doktor.
O bir hemşire.

☆ 📄 🔊 🔗

English Turkish Spanish Turkish - detected ▾

↔

English Turkish Spanish ▾

Translate

O bir doktor.
O bir hemşire

28/5000

He is a doctor.
She is a nurse ✓

☆ 📄 🔊 🔗

Measurement bias

- Measurement bias
 - 우리가 잘못된 것을 측정하고 있거나, 잘못된 방법으로 측정하고 있거나, 또는 그 측정을 부적절하게 모델에 통합하고 있기 때문에 우리의 모델이 실수를 할 때 발생한다.
 - e.g.
 - 뇌졸중 예측하기,
 - Prior stroke
 - Cardiovascular disease
 - Accidental injury
 - Benign breast lump
 - Colonoscopy
 - Sinusitis

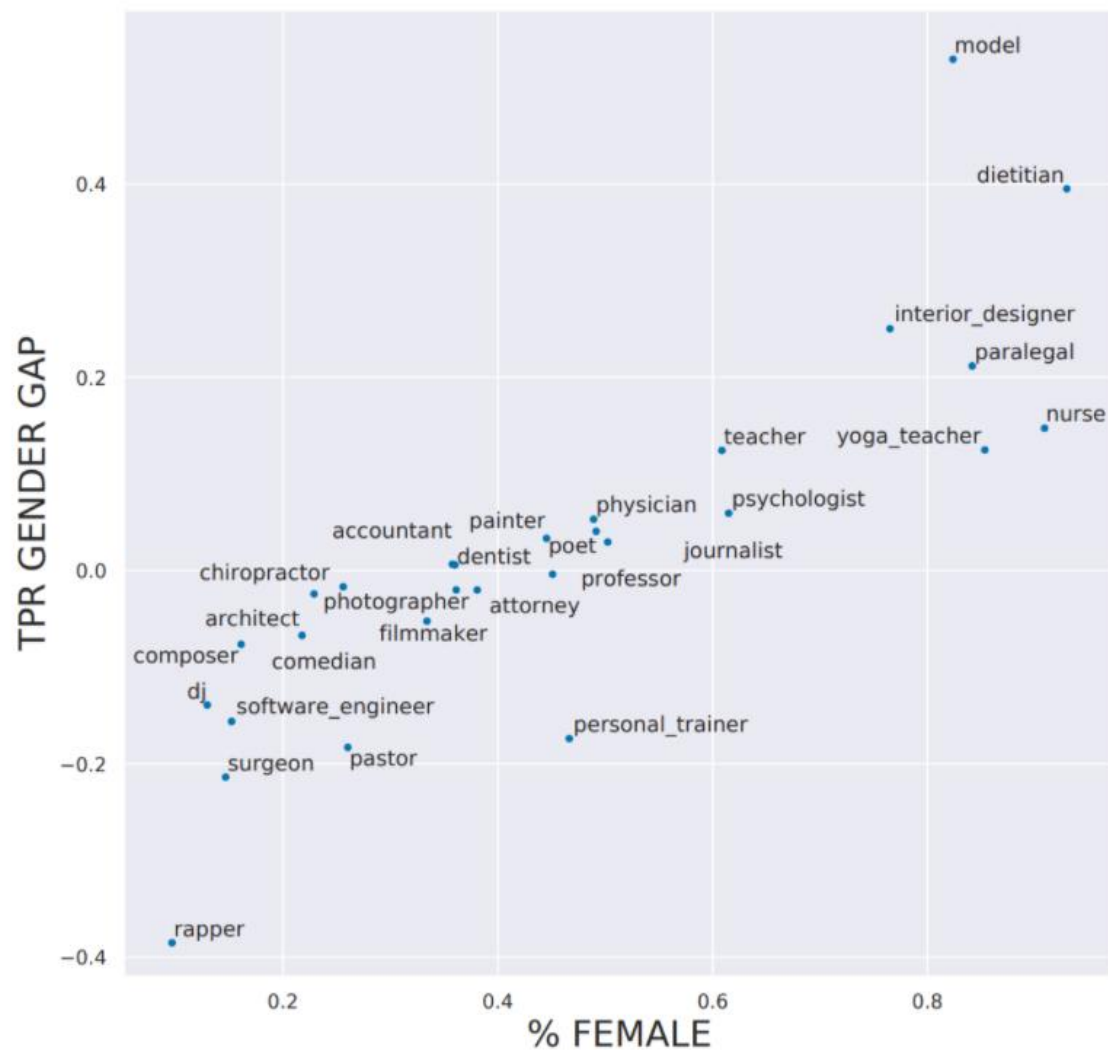
Aggregation bias (집계 편향?)

- Aggregation bias는 모형이 모든 적절한 요인을 통합하는 방식으로 데이터를 집계하지 않거나 모형에 필요 상호작용 항, 비선형성 등을 포함하지 않을 때 발생한다.
- 이것은 특히 의료 환경에서 발생할 수 있다. 예를 들어, 당뇨병을 치료하는 방법은 종종 단순한 일변량 통계와 이질적인 소수 집단을 포함하는 연구에 기초한다. 결과 분석은 종종 다른 민족이나 성별을 고려하지 않는 방식으로 이루어진다. 그러나 당뇨병 환자는 민족마다 합병증이 다르고, HbA1c 수준(당뇨병 진단 및 감시에 광범위하게 사용됨)은 민족과 성별에 따라 복합적으로 차이가 있는 것으로 나타났다. 이것은 의학적인 결정이 이러한 중요한 변수들과 상호작용을 포함하지 않는 모델에 기초하기 때문에 사람들이 잘못 진단되거나 잘못 치료되는 결과를 초래할 수 있다.

Representation bias - 1

- The abstract of the paper ["Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting"](#) by Maria De-Arteaga et al.
 - 직업에 성비 불균형이 존재한다고 언급하면서(예: 여성은 간호사일 가능성이 더 높고 남성은 목사가 될 가능성이 더 높다) "성간 진정한 양성률의 차이는 직업의 기존 성비 불균형과 상관관계가 있어 이러한 불균형을 더욱 심화시킬 수 있다"고 말한다.
 - 다시 말해, 연구원들은 직업을 예측하는 모델들이 기초 인구의 실제 성 불균형을 반영했을 뿐만 아니라 실제로 그것을 증폭시켰다는 것을 알아챘다! 이러한 유형의 표현 편향은 특히 단순한 모델에서 매우 흔하다. 어떤 분명하고 쉽게 볼 수 있는 기초적인 관계가 있을 때, 단순한 모델은 종종 이 관계가 항상 유지된다고 가정할 것이다.

Representation bias - 2



- 예를 들어, 훈련 데이터 집합에서 외과의사의 14.6%가 여성이었지만 모델 예측에서 True Positive의 11.6%만이 여성이었다. 따라서 모델은 훈련 세트에 존재하는 편견을 증폭시키고 있다.

Addressing different types of bias - 1

- 다른 유형의 편향은 완화를 위해 다른 접근법을 요구한다. 좀 더 다양한 데이터 집합을 수집하면 표현 편향을 다룰 수 있지만, 이는 과거 편향이나 측정 편향에는 도움이 되지 않는다. 모든 데이터 집합에는 치우침이 있다. 완전히 타락한 데이터 집합 같은 것은 없다. 이 분야의 많은 연구자들은 특정 데이터세트가 어떻게 그리고 왜 생성되었는지, 어떤 시나리오에서 사용하기에 적합한지, 그리고 어떤 제한사항이 무엇인지에 대한 의사결정, 맥락 및 세부사항을 더 잘 문서화할 수 있도록 일련의 제안에 집중해 왔다. 이런 식으로, 특정 데이터 집합을 사용하는 사람들은 그것의 편견과 한계에 의해 방심하지 않을 것이다.

Addressing different types of bias - 2

- 우리는 자주 그 질문을 듣는다—"인간은 편파적이니까 알고리즘 편향도 중요한가?" 이렇게 자주 나오는 얘기인데, 물어보는 사람들한테는 일리가 있는 추리가 있을 텐데, 우리한테는 별로 논리적으로 안 맞는 것 같아! 이것이 논리적으로 건전한지 여부와는 별개로 알고리즘(특히 기계학습 알고리즘!)과 사람이 다르다는 것을 깨닫는 것이 중요하다. 기계 학습 알고리즘에 대한 다음 사항을 고려하자.

- *Machine learning can create feedback loops*:: Small amounts of bias can rapidly increase exponentially due to feedback loops.
- *Machine learning can amplify bias*:: Human bias can lead to larger amounts of machine learning bias.
- *Algorithms & humans are used differently*:: Human decision makers and algorithmic decision makers are not used in a plug-and-play interchangeable way in practice.
- *Technology is power*:: And with that comes responsibility.

Addressing different types of bias - 3

- 아칸소 건강관리 사례에서 알 수 있듯이 머신러닝은 더 나은 결과로 이어지기 때문이 아니라 더 저렴하고 더 효율적이기 때문에 실제로 구현되는 경우가 많다. 캐시 오닐은 자신의 저서 '수학 파괴 무기'에서 특권층이 사람들에게 의해 처리되는 반면 가난한 사람들은 알고리즘에 의해 처리되는 패턴에 대해 설명했다. 이것은 알고리즘이 인간의 의사결정자들과 다르게 사용되는 여러 방법들 중 하나일 뿐이다. 그 밖의 사항에는 다음이 포함된다.
- People are more likely to assume algorithms are objective or error-free (even if they're given the option of a human override).
- Algorithms are more likely to be implemented with no appeals process in place.
- Algorithms are often used at scale.
- Algorithmic systems are cheap.
- 편견이 없는 경우에도 알고리즘(특히 효과적이고 확장 가능한 알고리즘이기 때문에 딥러닝)은 부정적으로 사회 문제를 일으킬 수 있다.