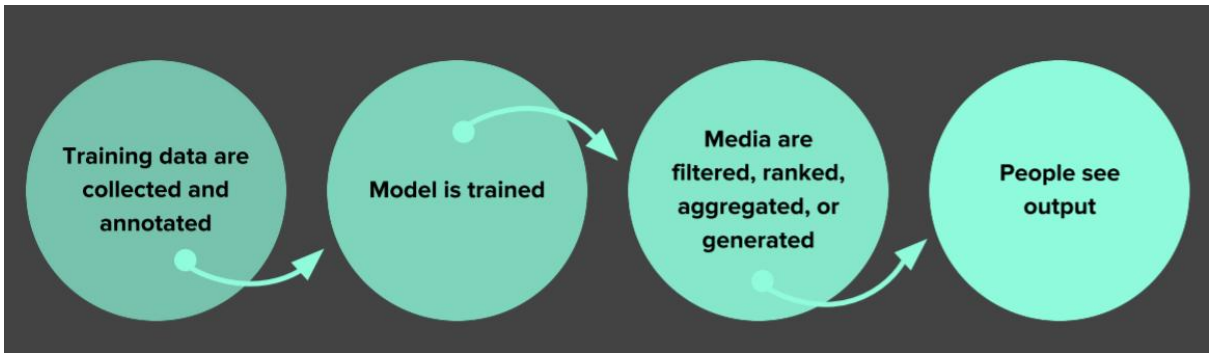
Lecture 19. Bias in AI

Ⅰ. Prototype Theory

Categorization의 목적 중 하나는 행동적으로 그리고 인식적으로 사용가능한 비율에 대한 자극 사이에 무한한 차이를 줄이는 것이다. 하나의 object category에 대해 저장된 일반적인 특징에서 발생한 item들의 중심적이고 원형적인 개념이 있을 수 있다. 예를 들면 test 피실험대상의 대부분은 의사가 여자일 수 있는 확률을 간과한다.

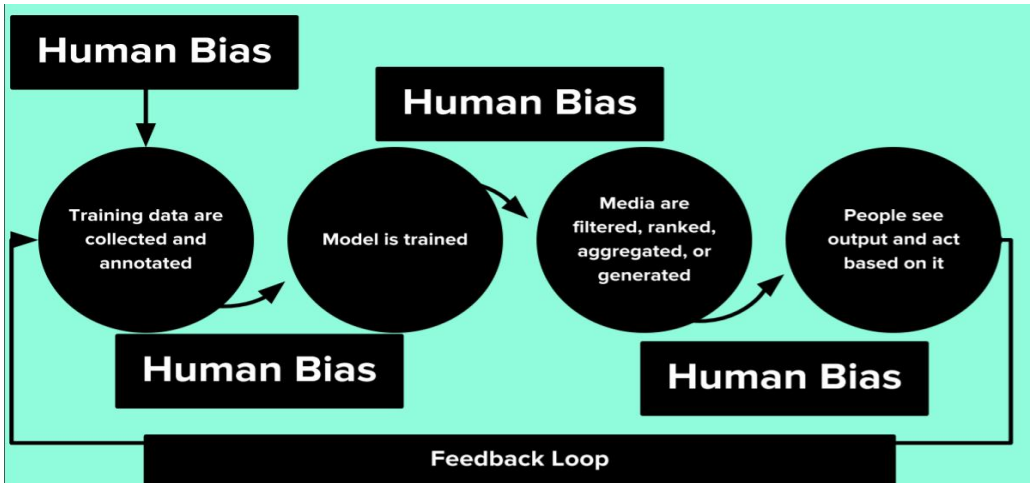
Ⅱ. Human Reporting Bias

행동, 결과 또는 특징에 대해 사람들이 쓰는 것에 대한 것은 현실 세계의 반영이 아니고 또는 하나의 특징이 개인의 집단의 특징인 정도도 아니다.



1. Biases in Data
    1) Reporting bias : 사람들이 공유하는 것은 현실 세계의 반영이 아니다.
    2) Selection bias : selection은 random sample을 반영하지 않는다.
    3) Out-group homogeneity bias : 사람들은 태도, 가치, 개인별 속성, 다른 특징들을 비교할 때 다른 그룹의 멤버를 같은 그룹 멤버들과 더 비슷하게 보는 경향이 있다.
2. Biases in Interpretation
    1) Confirmation bias : 누군가의 이미 존재하는 믿음 또는 가설을 확인하는 방법으로 해석, 선호, 기억해내는 정보를 찾는 경향
    2) Overgeneralization : 너무 일반적이고 충분히 구체적이지 않은 정보에 기반하여 결과를 도출하는 것
    3) Correlation fallacy : 상관관계를 원인으로 혼돈하는 것

4) Automation bias : 사람이 자동화없이 모순된 정보보다 자동화된 의사 결정 시스템의 제안을 선호하는 경향



사람 데이터는 human biases를 인식한다. ML이 human data를 학습할 때, 결과는 bias network effect이다.

3. Bias 장단점

-Bias of an estimator : 우리가 예측하려는 정확한 값과 예측 사이의 차이. y=mx+b에서 b가 bias of an estimator이다.

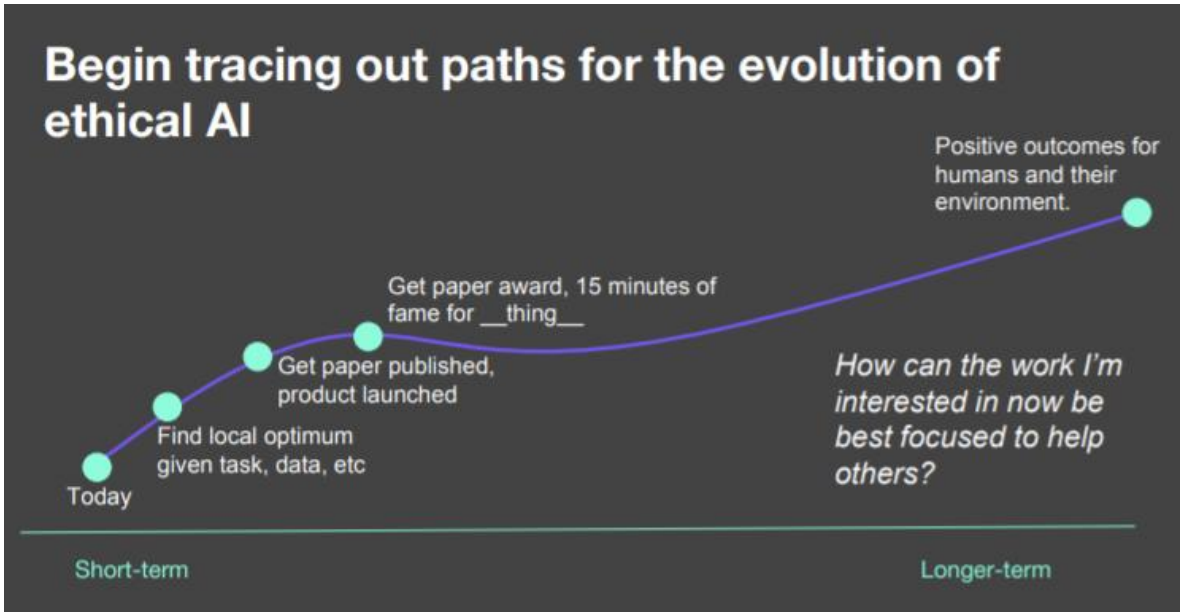
-Cognitive biases : confirmation biases, recency bias, optimism bias

-algorithm bias : 역사적으로 차별과 소외와 연관된 수입, 성적지향, 정교, 성별 등등과 연관된 사람의 불공평하고 불정의하고 편견적인 대우. 그들은 명백하게 algorithm systems 안에 있거나 algorithm이 의사결정을 돕는다.

Ⅲ. Predicting Future Crimminal

Faception : 최초의 기술이며 사람을 프로파일링하고 얼굴 이미지만을 기반으로 하여 그들의 개성을 드러내는 컴퓨터비전과 머신러닝이 합쳐진 소유권

IV. Machine Learning



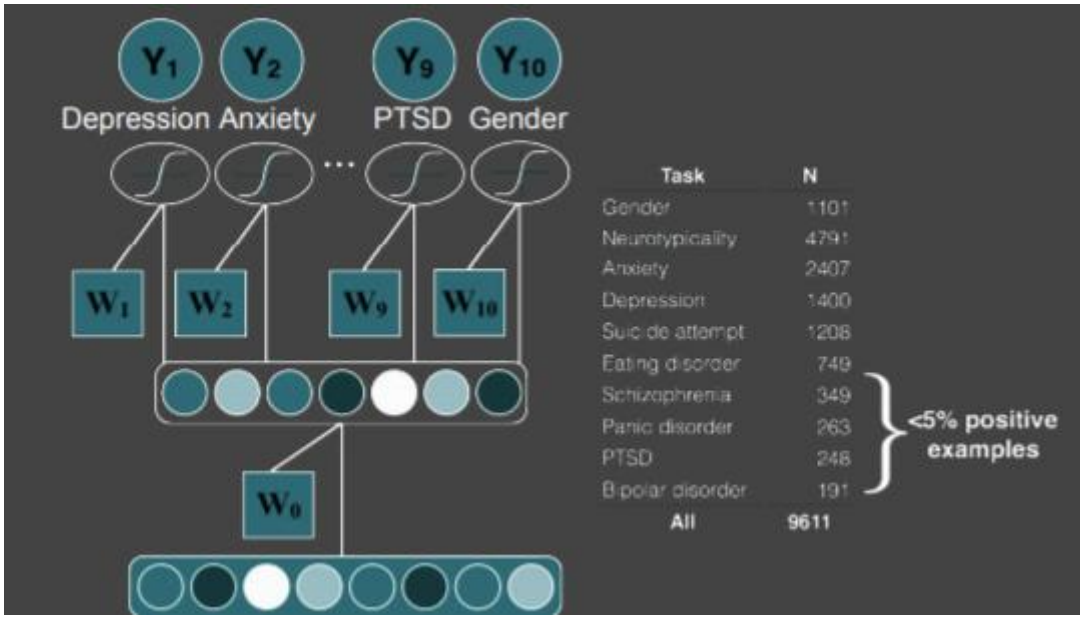
-Bias Mitigation : 문제있는 output에 대한 신호 제거(Stereotyping, sexism, racism, *-ism, debiasing)

-inclusion : 바람직한 변수에 대한 신호 추가(모델수행증가, 서브 그룹에 집중하거나 나쁜 수행으로 데이터 나누기)

1. Multiple Tasks + Deep Learning for Inclusion

- internal data : 전자 건강 기록

-proxy data : twitter media data



2. Multitask Adversarial Learning

결합하여 예측한다. Output decision D, decision Z로부터 제거하고 싶은 특성, 바람직하지 않은 특성의 효과 협상