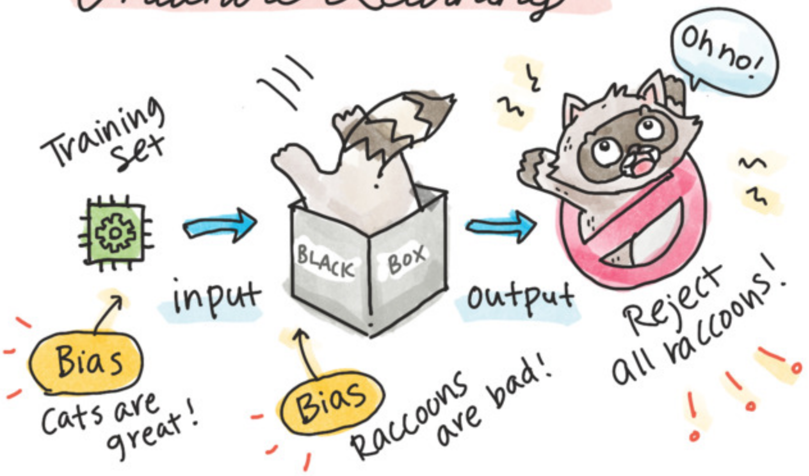
Fairness in Machine Learning



Fairness-related harms

Unfairness = negative impacts for group of people such as those defined in terms of

- race
- gender
- age
- disability status

Harms:

- ☆ Allocation
- ☆ Quality of service
- ☆ Stereotyping
- ☆ Denigration
- ☆ over- / under-representation

Complex sociotechnical challenges

✗ purely social

✗ purely technical

Assessment & mitigation

Fairlearn

fairlearn.github.io

False negatives

False positives

- ♡ Identify the harm (+benefits)
- ♡ Identify the affected groups
- ♡ Define fairness metrics

	False-	False+	Counts
men	0.35	0.27	6239
women	0.29	0.35	3124

@AzureAdvocates
@girlie_mac