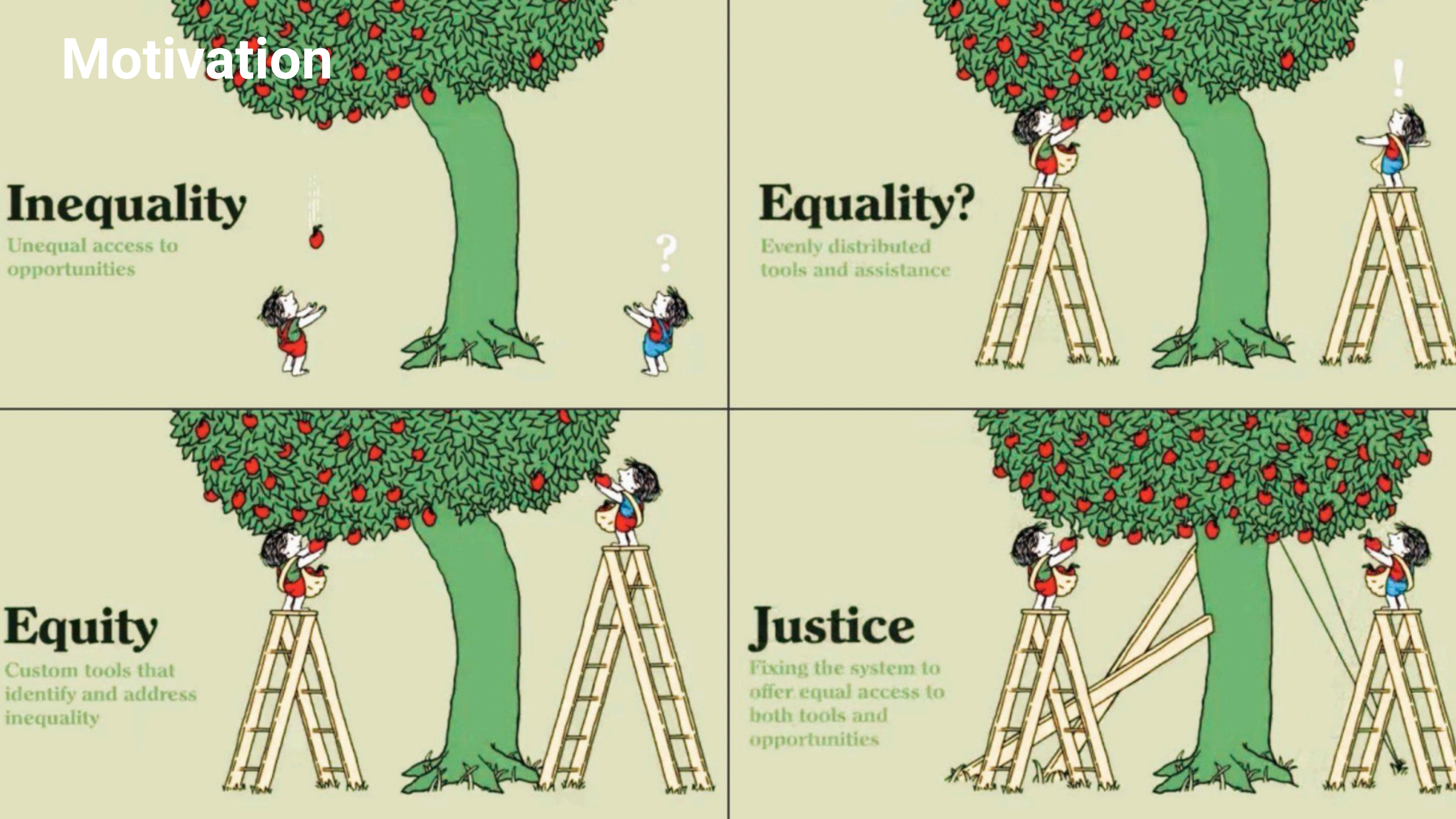
EVERYTHING IS TERRIBLE!

WHAT TO DO IN PRACTICE

Statistics are good - as indicators

- Use (multiple) fairness measures to evaluate the models
  - If you find large discrepancies **debug** the model
  - Useful in spotting issues (and preventing misconceptions)
- Do **not** just optimize fairness criteria when training and hope that the problem gets better (you might increase discrimination)
  - Dropping attributes makes classifier less accurate
  - “Affirmative action” might reduce diversity via stereotyping (see e.g. Lipton, 2019 study for student admissions)

Motivation


Inequality
Unequal access to opportunities
?
Equality?
Evenly distributed tools and assistance
!
Equity
Custom tools that identify and address inequality
Justice
Fixing the system to offer equal access to both tools and opportunities

Data & Biases

Check for data collection bias, e.g.

- Bias is problem specific (e.g. using gender for medical data vs. gender discrimination for credit applications).
- Population (e.g. many white actors in Celebface)
- Different demographics behave differently
- Cultural stereotypes inherent e.g. in textual data for large language models (female nurses vs. male doctors)
- Temporal bias (e.g. initial user base of a social network)

TL;DR - Talk to other humans / stakeholders

Things that might help

- Diverse team (helps catch more issues)
- Stakeholder feedback
- Ask where the data came from
- Look for potential issues (rather than being reactive)
- If things look strange, they probably are

- Continue testing model even after deployment

USE YOUR

COMMON SENSE:

ADDRESSED TO THE

INHABITANTS

OF

AMERICA,

Example - Poison Needle in the Haystack

Debug and fix

External system testers will find its weaknesses

- Break security (e.g. voice / face ID)
- Generate awful text
  - Tay (Microsoft chat client) started spouting racist tweets
  - AI Dungeon (GPT-2 text adventure) started generating child sexual abuse dialog
- Find images where system fails
  - Humans vs monkeys for Google image classification
  - Parliament Pilot Benchmark study

Example - Risk vs. MLE decoding

- We get a box of mushrooms that are 99% safe to eat
  - MLE estimator will eat the mushrooms
  - Common sense suggests we throw them out

$$\hat{y}(x) = \underset{y'}{\operatorname{argmin}} \sum_{y} \hat{p}(y \mid x) R[y' \mid y]$$

- Risk score $R[y' \mid y]$ denotes cost for making an error
  - $R[\text{edible} \mid \text{poison}] = 10^6$ but $R[\text{poison} \mid \text{edible}] = 1$
  - $R[\text{monkey} \mid \text{human}] = 10^6$ (encode this for decisions)

Summary

- **Examples**
- **Law**
- **Algorithmic Fairness**
  - Evaluating estimators
  - Fairness criteria
  - Impossibility results
- **In Practice**
  - Human evaluation
  - Poison needle in the haystack
  - Decisions, risk and estimates

Use your common sense and try to understand the problem!