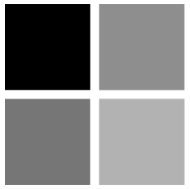
Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs

Solon Barocas[1], Anhong Guo[2], Ece Kamar[1], Jacquelyn Krones[1], Meredith Ringel Morris[1], Jennifer Wortman Vaughan[1], W. Duncan Wadsworth[1], Hanna Wallach[1]

[1] Microsoft, New York City, NY & Redmond, WA, USA

[2] University of Michigan, Ann Arbor, MI, USA



Microsoft

Disaggregated Evaluations of AI Systems

- AI systems can perform differently for different groups of people, often exhibiting especially poor performance for already disadvantaged groups
- A ***disaggregated evaluation*** assesses and reports an AI system's performance separately for different groups of people
- Well-known examples:
  - Recidivism prediction (ProPublica 2016)
  - Gender classification (Gender Shades 2018)
  - Speech recognition (Koenecke et al. 2020)
  - Health risk prediction (Obermeyer et al. 2019)

Conceptually simple, however…

Disaggregated Evaluations of AI Systems

- Conceptually simple, however… they involve a variety of choices
- These choices influence:
  - The results that will be obtained, and thus the conclusions that can be drawn
  - The impacts of these disaggregated evaluations, both beneficial and harmful, on people

Note: we intentionally avoid using the word "audit," since an audit involves formal roles, responsibilities, and expectations, as well as considering procedures and documentation in addition to system outputs

- *A disaggregated evaluation would likely be only one component of an audit*

Who? *When?*

Why? **Choices, Considerations, & Tradeoffs** *Where?*

How? *What?*

Choices, Considerations, and Tradeoffs

- **What** is the goal of the evaluation?
- **Who** will design and conduct the evaluation?
- **When** will the evaluation be conducted?
- **What** system or component(s) will be evaluated?
- **Where** will the evaluation occur?
- **What** are the factors and groups of interest?
- **Which** additional factors will be accounted for and **how** will they be accounted for?
- **How** will the evaluation dataset be created?
- **Which** performance metric(s) will be used?
- **How** will performance be analyzed?
- **How** transparent will the evaluation be?

Example: What is the goal of the evaluation?

- **Consideration 1:**
  - Intended to show the existence or absence of performance disparities?
  - Or intended to uncover potential causes of performance disparities?

Example: What is the goal of the evaluation?

- **Consideration 1:**
  - Intended to show the existence or absence of performance disparities?
  - Or intended to uncover potential causes of performance disparities?
- **Consideration 2:**
  - Intended to focus on specific set of people for their past encounters?
  - Or focus on people in past or future encounters in general?

Example: What is the goal of the evaluation?

- **Consideration 1:**
  - Intended to show the existence or absence of performance disparities?
  - Or intended to uncover potential causes of performance disparities?
- **Consideration 2:**
  - Intended to focus on specific set of people for their past encounters?
  - Or focus on people in past or future encounters in general?
- **Consideration 3:**
  - Intended to be confirmatory or exploratory?

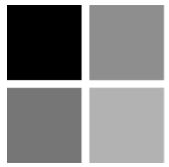
Takeaway Message:

When designing a disaggregated evaluation, it is important to carefully consider
why, who, when, where, what, and **how.**

Solon Barocas[1], Anhong Guo[2], Ece Kamar[1], Jacquelyn Krones[1], Meredith Ringel Morris[1], Jennifer Wortman Vaughan[1], W. Duncan Wadsworth[1], Hanna Wallach[1]

[1] Microsoft, New York City, NY & Redmond, WA, USA

[2] University of Michigan, Ann Arbor, MI, USA



Microsoft