AI systems perform differently for **different groups of people**. **Many choices** must be carefully considered to fully understand **performance disparities**.

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]

What is the goal of the evaluation?

To demonstrate the existence or absence of performance disparities? Or to uncover potential causes of performance disparities? Will it focus on actual disparities experienced by specific people? Or on potential disparities that may (have) generally affect(ed) people? Will it be confirmatory or exploratory?

Who will design and conduct it?

The development team(s) responsible for the system? Or outside parties, including consultants, researchers, etc.?

When will it be conducted?

Before system deployment? Or after?

What will be evaluated?

The system as a whole? Or one or more of its constituent components?

Where will the evaluation occur?

"In the laboratory?" Or "in situ?"

What are the factors and groups?

Should the evaluation focus on social constructs (e.g., race)? Or observable properties? Should the groups be based on single factors? Or multiple factors?

Which additional factors will be accounted for and how?

Demographic factors? Sociocultural factors? Behavioral factors? Morphological factors? Environmental factors? Will their values be held constant? Or will a range of values be considered?

How will the dataset be created?

Reuse an existing dataset? Create a new one using scraped data? Create a new one using data from the system's context of use? Create a new one by collecting data from data subjects?

Which metric(s) will be used?

One metric? Multiple metrics? For example, are false positives and false negatives equally harmful to people? Or not?

How will performance be analyzed?

Are statistical methods or ML methods more appropriate? Is uncertainty accounted for via p-values and confidence intervals? What about overfitting?

How transparent will the evaluation (e.g., choices, dataset, results) be?

Fully? Partially? Or not at all?

Definitions

Disaggregated evaluations: AI systems can perform differently for different groups of people, often exhibiting especially poor performance for already disadvantaged groups. Disaggregated evaluations assess and report system performance separately for different groups of people, providing way to understand performance disparities.

Confirmatory evaluations: Confirmatory evaluations are intended to provide conclusive evidence about performance disparities. Like scientific experiments, they must posit clear hypotheses to be tested and they must be designed carefully so as to minimize the risk of drawing incorrect conclusions. Confirmatory evaluations are most feasible when assessing and reporting system performance for a small number of groups in scenarios where there are only a few additional factors that can affect system performance.

Exploratory evaluations: Exploratory evaluations are not intended to provide conclusive evidence about performance disparities so there is much more flexibility in their design. Because it is so difficult to design confirmatory evaluations, most well-known disaggregated evaluations are best understood as exploratory evaluations. Exploratory evaluations can be used to inform the design of subsequent confirmatory evaluations.

Factors and groups: There are many different groups of people for which AI systems exhibit poor performance, including groups based on demographic factors, sociocultural factors, behavioral factors, morphological factors, and environmental factors. For example, race, gender, age, facial hair, hairstyle, glasses, facial expression, pose, and skin tone have all been shown to affect the performance of face-based AI systems.

arXiv link

← Download the paper

Affiliations

[1]Microsoft

[2]University of Michigan