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

Solon Barocas
solon@microsoft.com
Microsoft
New York City, NY, USA

Anhong Guo
anhong@umich.edu
University of Michigan
Ann Arbor, MI, USA

Ece Kamar
eckamar@microsoft.com
Microsoft
Redmond, WA, USA

Jacquelyn Krones
jakrones@microsoft.com
Microsoft
Redmond, WA, USA

Meredith Ringel Morris
merrie@microsoft.com
Microsoft
Redmond, WA, USA

Jennifer Wortman Vaughan
jenn@microsoft.com
Microsoft
New York City, NY, USA

W. Duncan Wadsworth
duwads@microsoft.com
Microsoft
Redmond, WA, USA

Hanna Wallach
wallach@microsoft.com
Microsoft
New York City, NY, USA

ABSTRACT

Disaggregated evaluations of AI systems, in which system performance is assessed and reported separately for different groups of people, are conceptually simple. However, their design involves a variety of choices. Some of these choices influence the results that will be obtained, and thus the conclusions that can be drawn; others influence the impacts—both beneficial and harmful—that a disaggregated evaluation will have on people, including the people whose data is used to conduct the evaluation. We argue that a deeper understanding of these choices will enable researchers and practitioners to design careful and conclusive disaggregated evaluations. We also argue that better documentation of these choices, along with the underlying considerations and tradeoffs that have been made, will help others when interpreting an evaluation's results and conclusions.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Social and professional topics** → **User characteristics**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

KEYWORDS

evaluations, disaggregated evaluations, fairness, artificial intelligence, machine learning

ACM Reference Format:
Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462610

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462610

1 INTRODUCTION

AI systems can perform differently for different groups of people, often exhibiting especially poor performance for already disadvantaged groups [e.g., 3, 13, 15, 21, 49, 60, 62, 67, 70, 75]. Several pieces of work have uncovered such performance disparities by conducting disaggregated evaluations of AI systems, in which system performance is assessed and reported separately for different groups of people, such as those based on race and gender [58]. Such evaluations can provide a way to hold development teams and system owners accountable for system performance, to decide whether to use or keep using a system, or to identify potential system modifications that would make it acceptable to use or keep using a system.

In this paper, we draw attention to the choices that must be made when designing a disaggregated evaluation. At a high level, these choices can be thought of as roughly spanning "why," "who" (both who will design and conduct the evaluation and for which groups of people will system performance will be assessed and reported), "when," "what," "where," and "how." Some of these choices influence the results that will be obtained, and thus the conclusions that can be drawn; others influence the impacts—both beneficial and harmful—that a disaggregated evaluation will have on people, including the people whose data is used to conduct the evaluation.

Using face-based AI systems[1] as a running example, we argue that a deeper understanding of these choices will enable researchers and practitioners to design careful and conclusive disaggregated evaluations, better enabling themselves and others to understand the ways in which AI systems perform differently for different groups of people. To that end, we highlight some of the key considerations that underlie the choices that must be made when designing a disaggregated evaluation. We emphasize that these considerations

[1] We use the phrase "face-based AI systems" to refer to face-detection systems, face-characterization systems (e.g., gender or age classifiers), face-verification systems, and face-identification systems; the latter two are types of face-recognition systems.

are not independent; designing a disaggregated evaluation means making tradeoffs between considerations. These tradeoffs must be clearly articulated by the evaluation's designers, so that they and others know how to interpret its results and conclusions.

Although we use face-based AI systems as a running example, the choices, considerations, and tradeoffs that we discuss are common to disaggregated evaluations of many other types of AI systems as well. However, face-based AI systems have, in many ways, become a bellwether for AI systems in general. Face-based AI systems are used throughout society in comparatively low-stakes domains like advertising [28] and digital media management [30], as well as in high-stakes domains like education [65], employment [41], healthcare [76], security [4, 6, 57], and criminal justice [23]. Despite their growing prevalence, their use remains controversial [74]. Much of this debate has been spurred by a number of high-profile disaggregated evaluations, most notably the Gender Shades study [15, 67]. In turn, this debate has helped to highlight some key limitations of disaggregated evaluations and the dangers of focusing narrowly on performance disparities [38, 68]. For these reasons, we believe that face-based AI systems make a particularly compelling running example. That said, our focus on face-based AI systems should not be viewed as an endorsement of their use.

Throughout this paper, we intentionally avoid using the word "audit" to refer to the process of assessing an AI system for performance disparities. In other industries, the word "audit" refers to an official examination with mutually agreed-upon actors, responsibilities, and expectations. Audits typically consider procedures and documentation, as well as considering system outputs. Although a disaggregated evaluation could be a component of an audit of an AI system, it would likely not be the only component. Developing a clear definition of what it means to audit an AI system is an extremely important and much-needed research direction. We are particularly encouraged by a recent paper by Raji et al. [69], which proposes the "Scoping, Mapping, Artifact Collection, Testing, and Reflection" (SMACTR) framework for conducting comprehensive internal audits of AI systems via a rigorous, multi-stage process. Disaggregated evaluations would most likely fit within the "Testing" stage of the SMACTR framework.

Lastly, we note that performance disparities are just one type of fairness-related harm. Researchers have highlighted many other types of fairness-related harms, such as system outputs that stereotype, demean, or lead to erasure [7]. Evaluations that focus on these types of harms are outside the scope of this paper.

In the next section, we provide an overview of disaggregated evaluations and their role to date in the context of AI systems. We then describe the choices that must be made when designing a disaggregated evaluation. We highlight some of the key considerations that underlie these choices, as well as the tradeoffs between these considerations. Finally, we conclude with a short discussion.

2 DISAGGREGATED EVALUATIONS

AI systems are typically evaluated by assessing and reporting one or more aggregate performance metrics—such as accuracy, precision, recall, word error rate, perplexity, and root-mean-square error—calculated using an evaluation dataset. For example, the National Institute of Standards and Technology (NIST) has conducted vendor tests of commercially available face-recognition systems for decades, assessing and reporting aggregate performance metrics like false positive and false negative rates using evaluation datasets that are comprised of front-facing mugshots, side-view images, webcam images, and images taken by photo journalists and amateur photographers [33, 34]. However, aggregate performance metrics can obscure poor performance for groups of people that are not well represented in an evaluation dataset. For example, consider an evaluation dataset that contains 100 data points, where 90 data points are associated with group A and 10 data points are associated with group B. If a system makes correct predictions for the data points associated with group A and incorrect predictions for the data points associated with group B, then the aggregate accuracy of the system will be 90% when evaluated using the dataset. But this aggregate accuracy obscures the fact that there is an absolute performance disparity of 100% because the accuracy for the data points associated with group B is zero.

For this reason, researchers and practitioners seeking to uncover performance disparities exhibited by AI systems often conduct disaggregated evaluations [58]. This practice draws on and parallels similar practices in other industries. For example, the U.S. Food and Drug Administration mandates that clinical trial results be assessed and reported separately for groups based on race, gender, and age. Disaggregated evaluations have proven to be remarkably effective at uncovering the ways in which AI systems perform differently for different groups of people. For instance, Obermeyer et al. [62] demonstrated that a system used to enroll patients in a high-risk care management program assigned different risk scores to Black and white patients with comparable health statuses, leading to a large disparity in the proportions of Black and white patients identified for enrollment; DeVries et al. [21] compared the accuracies of six object-classification systems using images of household objects from fifty countries, finding that all six systems had substantially lower accuracies for images from lower-income countries and households; and Koenecke et al. [49] showed that five commercially available speech-recognition systems had much higher word error rates for Black people than for white people.

In the context of face-based AI systems, disaggregated evaluations have been used by researchers and practitioners to assess and report the performance of face-recognition systems since the early 2000s, focusing primarily on groups based on environmental factors like pose and lighting conditions [e.g., 10, 47, 54] and on groups based on race, gender, and age [e.g., 9, 18, 50]. In 2019, NIST finally conducted its own disaggregated evaluation of commercially available face-recognition systems, focusing on false positive and false negative rates for groups based on race, sex, age, and country of birth [35]. The resulting report provides an overview of the literature on performance disparities exhibited by face-based AI systems.

The most notable such piece of work is the Gender Shades study, which used a disaggregated evaluation to show that three commercially available gender classifiers, a type of face-characterization system, had higher error rates for women with darker skin tones than for women or for people with darker skin tones overall [15]. By focusing on intersectional groups based on skin tone and gender, the study demonstrated the need to specifically assess and report system performance for groups based on multiple factors, drawing

on and highlighting the importance of Crenshaw's work on intersectionality [20], which showed that the experiences of Black women differ from the experiences of women or of Black people overall.

As well as being widely cited in the research community, Gender Shades contributed to ongoing changes to the industry around face-based AI systems. Follow-up work revealed that the study was effective at getting the companies responsible for the three gender classifiers to address the performance disparities [67]; media coverage of the study led to greater public awareness of the societal impacts of face-based AI systems [14, 55], spurring calls to action [e.g., 1] and legislation [24]; and a company responsible for one of the three gender classifiers announced that it would no longer develop any face-based AI systems [42], while another announced that it would not sell face-recognition systems to police departments in the U.S. [32].

3 CHOICES, CONSIDERATIONS, AND TRADEOFFS

Although disaggregated evaluations are conceptually simple, their results, conclusions, and impacts depend on a variety of choices. In this section, we describe these choices—which can be thought of as roughly spanning "why," "who," "when," "what," "where," and "how"—and highlight some of the key considerations that underlie them. We emphasize that these considerations are not independent, and discuss some of the tradeoffs that must therefore be made when designing a disaggregated evaluation.

3.1 What is the goal of the evaluation?

When designing a disaggregated evaluation, the first choice that must be made is the goal of the evaluation. There are three considerations that underlie this choice. First, is the evaluation intended to demonstrate the existence or absence of performance disparities? Or is it intended to uncover potential causes of performance disparities? For example, the ACLU focused on demonstrating the existence of performance disparities exhibited by Amazon's Rekognition [73], while Koenecke et al. [49] attempted to understand *why* speech-recognition systems had much higher word error rates for Black people than for white people.

Second, will the evaluation focus on actual performance disparities experienced by a specific set of people who encountered the system in the past? Or will it focus on potential performance disparities that may have generally affected people who encountered the system in the past or that may generally affect people who will encounter the system in the future? For example, Angwin et al. [3] investigated the actual risk scores assigned to specific defendants by Northpointe's COMPAS recidivism-prediction system, while NIST's disaggregated evaluation of commercially available face-recognition systems focused on potential performance disparities [35]. Designing evaluations that are focused on a specific set of people is usually easier than designing evaluations that are intended to be general.

Finally, will the evaluation be confirmatory or exploratory? Confirmatory evaluations are intended to provide conclusive evidence about performance disparities, while exploratory evaluations are not. By analogy, confirmatory evaluations are like scientific experiments—that is, they must posit clear hypotheses to be tested and they must be designed very carefully so as to minimize the risk of drawing incorrect conclusions. For example, confirmatory evaluations must account for all factors that can affect system performance, including demographic factors, environmental factors, and behavioral factors. As we describe in Section 3.7, this can be challenging. However, a failure to do so can yield results that seem to indicate the existence of meaningful performance disparities, when in fact these disparities are simply due to spurious correlations. Confirmatory evaluations are therefore most feasible when assessing and reporting system performance for a small number of particularly salient groups in scenarios where there are only a few additional factors that can affect system performance. In contrast, exploratory evaluations are not intended to provide conclusive evidence, so there is much more flexibility in their design. That said, they can be used to inform the design of subsequent confirmatory evaluations. Because it is so difficult to design confirmatory evaluations, most well-known disaggregated evaluations are best understood as exploratory evaluations.

3.2 Who will design and conduct the evaluation?

A disaggregated evaluation can be designed and conducted by the development team(s) responsible for the system or by outside parties, including consultants, researchers, and journalists. When an evaluation will be designed and conducted by outside parties, this can be done in collaboration with the development team(s) or it can be done without their help or knowledge. In some cases, outside parties can even conduct a disaggregated evaluation when the development team(s) would prefer that they not.

Many well-known disaggregated evaluations have been designed and conducted by outside parties. For example, ProPublica, an independent newsroom, evaluated Northpointe's COMPAS recidivism-prediction system [3]; researchers from Stanford evaluated speech-recognition systems from Amazon, Apple, Google, IBM, and Microsoft [49]; and the Gender Shades study was conducted by a researcher from MIT and a researcher from a company responsible for one of the three gender classifiers [15]. Comparatively few well-known disaggregated evaluations have been performed by development teams, perhaps because evaluations of this sort are not usually publicly disclosed, though this situation may change if documentation approaches like model cards continue to gain traction [58].

Given that a disaggregated evaluation's results and conclusions can be quite troubling—and, in some cases, damning for the the development team(s) responsible for the system—the independence (either perceived or real) of an evaluation might affect whether its results and conclusions are seen as trustworthy. As a result, disaggregated evaluations that are designed and conducted by outside parties may carry more credibility. That said, outside parties, especially those that are not working in collaboration with development teams, might possess fewer resources and may therefore struggle to design and conduct evaluations that are as comprehensive.

There are also practical limits to what can be achieved by outside parties. In most cases, outside parties will only be able to engage with the system as a black box—that is, observe the system's outputs when presented with different inputs. Moreover, some systems do not produce publicly accessible outputs, making it impossible for outside parties to evaluate these systems. In contrast, development

teams will benefit from the fact that they have a deep understanding of their own systems, including the components that make up their systems and how these components fit together, the performance metrics that are already used to evaluate their systems, and their system's intended use cases and deployment contexts. Development teams are also better positioned to design and conduct evaluations that are focused on identifying immediate opportunities for improvement. That said, outside parties can possess expertise or perspectives that development teams lack, allowing them to uncover performance disparities that might otherwise be overlooked.

3.3 When will the evaluation be conducted?

Many well-known disaggregated evaluations have focused on AI systems that are already commercially available and, in some cases, widely used—in large part because these evaluations were designed and conducted by outside parties, without collaboration from the development teams responsible for the systems. However, disaggregated evaluations can also take place before system deployment. Indeed, development teams—potentially in collaboration with outside parties who have been provided with pre-deployment access—can use disaggregated evaluations to determine whether their systems are ready for deployment, reducing the likelihood of harms.

3.4 What system or component(s) will be evaluated?

AI systems often consist of multiple components, such as machine-learning models, whose inputs and outputs are interrelated. For example, a speech-recognition system consists of an acoustic model that models the acoustics of speech and a language model that models relationships between words. A disaggregated evaluation can focus on the performance of a system as a whole or on the performance of one or more of the system's constituent components. Focusing on the performance of the system as a whole can make it easier to uncover performance disparities that will lead directly to harms when the system is deployed. In contrast, focusing on the performance of one or more components can make it easier to uncover potential causes of any system-level performance disparities.

If a disaggregated evaluation will be designed and conducted by outside parties without collaboration from the development team(s) responsible for the system, then it may not be possible to assess the performance of its constituent components unless they are also accessible by outside parties. That said, some design choices may allow component-level performance disparities to be inferred from system-level performance disparities. For example, by asking Black and white people to speak the same sequence of words, the researchers who evaluated five commercially available speech-recognition systems could be fairly sure that the higher word error rates for Black people were due to the acoustic models [49].

Some AI systems depend on the outputs of other systems. A disaggregated evaluation can yield distorted results if it does not account for such dependencies. For example, a face cannot be identified by a face-identification system unless it is first detected by a face-detection system. Evaluating the performance of a face-identification system using only images in which faces have previously been detected will therefore fail to uncover performance disparities due to face-detection errors. For this reason, NIST recommends assessing and reporting performance disparities for every component of a face-recognition system, as well as any other systems on which it depends [34].

3.5 Where will the evaluation occur?

A disaggregated evaluation can take place "in the laboratory" or "in situ"—that is, in the system's context of use. For example, it is possible to evaluate a face-verification system that is used to grant workplace access to employees by presenting the system with a set of images or by asking employees and non-employees to attempt to gain access to the workplace using the system. In the former scenario, the images may not accurately reflect environmental factors like lighting conditions or behavioral factors like pose. Moreover, if the system requires an operator to make decisions (e.g., grant or deny access) based on the system's outputs, then the behavior of the operator will not be reflected in the evaluation [64]. In contrast, conducting a disaggregated evaluation in situ allows for the performance of the entire sociotechnical system, of which the AI system may be just one component, to be evaluated [69]. Indeed, the performance of an AI system in isolation may not reveal much about the ultimate performance of a sociotechnical system that also involves human discretion and judgment [72]. However, conducting a disaggregated evaluation in situ can be expensive and may not be possible if the evaluation will be designed and conducted by outside parties. In addition, some system components may be difficult to evaluate in situ because they cannot be isolated from the system as a whole in a meaningful way. However, as we described in Section 3.4, some design choices may allow component-level performance disparities to be inferred from system-level performance disparities.

3.6 What are the factors and groups of interest?

There are many different groups of people for which AI systems exhibit poor performance, including groups based on demographic factors, sociocultural factors, behavioral factors, and morphological 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 [33, 34]. Many well-known disaggregated evaluations have assessed and reported system performance for a small number of particularly salient groups based on one or two factors—often groups that are already disadvantaged. This is because performance disparities involving such groups may compound existing injustices [40]. For example, ProPublica focused on one factor (race) and two groups of people based on that factor (Black and white defendants) when evaluating Northpointe's COMPAS recidivism-prediction system [3], while the researchers who conducted the Gender Shades study focused on two factors (skin tone and gender) and multiple intersectional groups based on those factors [15].

The latter example raises an important consideration—namely whether disaggregated evaluations should focus on social constructs, such as race and gender, or on observable properties, such as skin tone and hairstyle. Unlike skin tone and hairstyle, race and gender are not objective, inherent properties of people; they are categories constructed by humans that, by social convention, serve as the basis for social differentiation. These social constructs so deeply

structure people's understanding of others and of themselves that they are frequently taken for granted. Yet they are historically and culturally specific, unstable and contested, and often bound up with unjust social hierarchies [e.g., 8, 36]. Even when focusing on observable properties, some properties may be of particular interest because they are thought to serve as proxies for social constructs (e.g., skin tone as a proxy for race). Focusing on these properties therefore raises some of the same challenges as focusing on social constructs. In contrast, other observable properties (e.g., glasses) may be of interest in their own right.

Focusing on social constructs (or their proxies) can be advantageous because social constructs affect people's lives in ways that are both profound and mundane. For example, some of the groups that are most disadvantaged within society are groups based on race and gender. However, this approach raises challenging questions about the status of social constructs and the implications of using them to conduct evaluations of AI systems. Determining which social construct applies to a person can be both practically difficult and ethically fraught. In many cases, social constructs like gender are simply ascribed to people by institutions of authority, as is the case with much official documentation (e.g., government-issued identification). Relying on such ascriptions assumes that they are valid and appropriate. In some cases, it may be possible to ask people about their group membership—an approach that is especially important when self-reported information may conflict with official documentation (e.g., when a person does not identify with the gender ascribed to them at birth). Yet, in other cases, self-reported information may not be accessible or available at all. In these scenarios, it may be tempting to infer group membership from observable properties. However, making decisions about which observable properties can serve as reliable proxies for social constructs is an activity that can be viewed as essentializing, stigmatizing, or alienating, especially if this is done in a way that suggests that these properties are common to all members—or only members—of particular groups based on those social constructs. It also runs the risk of inaccuracies, especially if group membership is difficult to infer from observable properties. Moreover, inferring group membership also raises questions about the ethics of imposing labels on people [2, 12]. Lastly, we emphasize that people may not want a disaggregated evaluation to be designed and conducted, no matter the potential benefits to uncovering performance disparities involving the groups to which they belong. In some cases, this is because the existence of evaluation datasets that contain information about group membership can be actively harmful [68, 71]. This raises questions about who gets to decide whether an evaluation will or will not take place.

In contrast, focusing on observable properties sidesteps some of the difficulties presented by social constructs, provided those properties are not assumed to serve as reliable proxies for social constructs. However, it can still be challenging to obtain accurate information about group membership. For example, how short does a person's hair need to be to be described as "short" and what normative reasons might there be to be interested this factor if not its relationship to gender? This approach also makes it difficult to conclude anything about performance disparities involving social constructs like gender—yet, as we mentioned above, some of the groups that are most disadvantaged within society are groups based on social constructs.

Just as focusing on aggregate performance metrics can obscure poor performance for groups of people that are not well represented in an evaluation dataset, focusing on groups based on single factors can obscure poor performance for people belonging to intersectional groups. For example, the Gender Shades study demonstrated that three commercially available gender classifiers had higher error rates for women with darker skin tones than for women or for people with darker skin tones overall [15]. Disaggregated evaluations should therefore assess and report system performance for people belonging to both intersectional and non-intersectional groups.

Regardless of whether a disaggregated evaluation focuses on social constructs or on observable properties, it must be possible to create an evaluation dataset that can support the goal of the evaluation. For example, if an evaluation will be confirmatory and focused on potential performance disparities that may generally affect people who will encounter the system in the future, then the evaluation dataset must be roughly balanced across the different groups of interest, with sufficient data about each. In practice, this can be difficult to achieve, especially if there are many groups of interest (as is often the case when focusing on intersectional groups). As a result, it is much easier to design disaggregated evaluations that focus on a small number of particularly salient groups based on a small number of factors.

3.7 Which additional factors will be accounted for and how will they be accounted for?

In practice, there are many factors that can affect system performance beyond the factors of interest. Because some of these additional factors may be correlated—even spuriously—with the factors of interest, a failure to appropriately account for them can make it difficult to interpret a disaggregated evaluation's results. For example, although NIST's disaggregated evaluation of face-recognition systems revealed a higher false negative rate for images of Asian people than for images of Black or white people, the resulting report notes that this performance disparity may actually be due to between-group differences in the time elapsed between each pair of images [35]. As another example, a face-recognition system might perform worse for some genders than for others because it exhibits poor performance for people with particular hairstyles that are thought to be meaningfully correlated with gender. From a normative perspective, attributing these performance disparities to gender is reasonable—after all, the correlation is not spurious—though a failure to account for hairstyle will make it more difficult to uncover the causes of these performance disparities.

There are three ways to account for additional factors. The first is to make sure that the evaluation dataset is reflective of the population of interest—that is, people who encountered the system in the past or people who will encounter the system in the future—and the environmental factors, behavioral factors, and other factors found in situ. If this is the case, then provided there are no spurious correlations with the factors of interest—a big assumption—any performance disparities can be interpreted as being reflective of those that either affected or will affect the population of interest. In practice, though, there may well be spurious correlations, making it difficult to be certain that any performance disparities are meaningful (or to uncover their potential causes). As a result, this approach is suitable

for exploratory, but not confirmatory, evaluations. Despite these limitations, many well-known disaggregated evaluations have used this approach as it is comparatively easy to implement [e.g., 3, 62].

The second way to account for additional factors is to hold their values constant. For example, when evaluating a face-recognition system, one way to account for glasses is to only assess and report performance for people without glasses. Although this approach means that any performance disparities are more likely to be genuinely due to the factors of interest, it obscures performance disparities that occur in the context of factor values other than the ones considered. Conclusively determining the absence of performance disparities is therefore especially difficult when using this approach, making it unsuitable for some confirmatory evaluations. For example, glasses might have a greater negative effect on system performance when they are worn by women than when they are worn by men. A failure to assess and report performance for people wearing glasses will obscure this performance disparity. As another example, poor lighting conditions have a greater negative effect on the performance of face-detection systems for people with darker skin tones than for people with lighter skin tones. Conducting a disaggregated evaluation in only good lighting conditions will obscure this performance disparity. In some cases, it may not even be possible to hold the values of additional factors constant, as is the case with evaluations that are focused on actual performance disparities experienced by a specific set of people who encountered the system in the past.

The third way to account for additional factors is to consider a range of values for each such factor. By assessing and reporting performance separately for each group of interest and combination of additional factor values, this approach yields a granular and high-dimensional view of system performance. Moreover, if the distributions over additional factor values are the same for each group of interest, then any performance disparities are more likely to be genuinely due to the factors of interest. In practice, though, it may be difficult to ensure that the distributions over additional factor values are the same for each group of interest. For example, some hairstyles are more common among some genders than among others. However, even if the distributions over additional factor values are not the same, it is still possible to account for them using statistical techniques like generalized linear mixed-effect regression models [e.g., 11, 29], as we describe in Section 3.10. We note that it can be difficult to identify all relevant additional factors or to choose an appropriate range of values for each such factor. That said, this approach will most conclusively demonstrate the existence or absence of performance disparities, making it particularly appropriate for confirmatory evaluations.

Lastly, we emphasize that no group is a monolith, so there will always be additional factors that exhibit within-group variation. For example, people belonging to any group based on race will exhibit a wide range of skin tones, not to mention genders and ages. If there is any chance that these factors can affect system performance, then they should be accounted for by taking the third approach—that is, by considering a range of values for each such factor.

3.8 How will the evaluation dataset be created?

Because a disaggregated evaluation depends so heavily on the evaluation dataset used to conduct it, this is one of the most important choices, with many underlying considerations and tradeoffs. There are four main approaches to creating an evaluation dataset. The first is to reuse an existing dataset that was previously created for some other purpose; the second is to create a new dataset using data scraped from the web or from other data sources; the third is to create a new dataset using data from the system's context of use; and the fourth is to create a new dataset by collecting data directly from data subjects.

For each approach, we discuss the following considerations and the tradeoffs between them: time, financial cost, suitability for the evaluation's goal, representation of the groups of interest, labeling of group membership, representation of additional factors, labeling of additional factor values, licensing, consent, and compensation. Some of these considerations (e.g., time, financial cost, and licensing) are standard when reusing or creating a dataset, while others (e.g., representation of additional factors) influence a disaggregated evaluation's results and conclusions. Others still (e.g., labeling of group membership, consent, and compensation) influence the impacts that a disaggregated evaluation will have on people, including the people whose data is used to conduct the evaluation.

3.8.1 Approach 1: Reuse an existing dataset.

Reusing an existing dataset, such as the IARPA Janus Benchmark A (IJB-A) dataset for evaluating face-detection and face-recognition AI systems [77] or the Adience dataset for evaluating age and gender classifiers [22], is less time consuming and less expensive than creating a new dataset. However, existing datasets may not contain sufficient data about the groups of interest, especially if these groups are intersectional or rare. For example, the researchers who conducted the Gender Shades study [15] analyzed both the IJB-A dataset and the Adience dataset, finding that both are predominantly composed of images of people with lighter skin tones and both contain very few images of women with darker skin tones (7.4% of the IJB-A dataset and 4.4% of the Adience dataset). As a result, the researchers were unable to use either dataset to conduct their disaggregated evaluation, spurring them to create their own dataset—the Pilot Parliaments Benchmark (PPB) dataset. The Gender Shades study motivated the creation of other general-purpose datasets that have better representation of some groups based on race, skin tone, gender, and age [e.g., 45]. However, other groups, especially groups that are intersectional or rare, are still underrepresented in these datasets.

Even if existing datasets do have sufficient data about the groups of interest, the data points may not be labeled so as to indicate group membership. Combined with the fact that existing datasets rarely provide mechanisms for contacting data subjects, this means that reusing an existing data set may require group membership to be inferred, perhaps by asking crowdworkers or by using the outputs of other AI systems. However, as we mentioned in Section 3.6, inferring group membership can be problematic due to the possibility of inaccuracies and due to questions about the ethics of imposing labels on people [2, 12].

Existing datasets may also lack appropriate representation of additional factors. There are several ways in which this might be the case, depending on the approach used to account for additional factors. First, the dataset may not be reflective of the population of interest and the environmental factors, behavioral factors, and other factors found in situ. If this is the case, then it is not possible to

implicitly account for additional factors via the argument that any performance disparities are reflective of those that either affected or will affect the population of interest. For example, when used by a retailer to count in-store customers, a face-detection system might rely on images taken by a wall-mounted camera. Evaluating such a system using a dataset that is composed of well-lit, front-facing images will likely yield results that are not reflective of the system's performance in practice. Second, the data points may have different values for the additional factors, making it impossible to account for them by holding their values constant. Even if the data points do have the same value for each additional factor, this value may not reflect the values typically found in situ. Third, if additional factors are to be accounted for by considering a range of values for each such factor, the data points may not span an appropriate range for each factor or there may not be sufficient data for some combinations of factor values. In all three cases, the data points may not be labeled so as to indicate the values of the additional factors. This is especially likely for additional factors that are not typically considered "interesting" in and of themselves. Taken together, these limitations mean that most existing datasets are not appropriate for conducting confirmatory evaluations.

In some cases, licensing restrictions may prevent the reuse of existing datasets, especially for development teams in companies who wish to understand the performance of their own AI systems. And, given that existing datasets rarely provide mechanisms for contacting data subjects, reusing an existing dataset makes it difficult to request data subjects' consent or to compensate them fairly for contributing their data [5].

Aside from these considerations, reusing an existing dataset can also lead to problems if the dataset is not well documented, for example, with a datasheet [26]. This is because the dataset may have been collected or preprocessed in ways that might affect an evaluation's results but are not immediately apparent.

3.8.2 Approach 2: Create a new dataset using scraped data.

When an appropriate existing dataset does not exist, an alternative is to create a new dataset using data scraped from the web or from other sources. This approach is more time consuming than reusing an existing dataset, and can be more or less expensive. Provided the data sources are selected carefully, scraping data can make it easier to ensure sufficient data about the groups of interest. For example, when creating the PPB dataset, the researchers who conducted the Gender Shades study opted to scrape images from the parliamentary websites of six different countries, which were chosen because their parliaments were roughly balanced in terms of gender. To ensure sufficient data for both darker and lighter skin tones, three of the countries (Rwanda, Senegal, and South Africa) were in Africa, while the other three (Iceland, Finland, and Sweden) were in Europe. This process resulted in a dataset that is roughly balanced across intersectional groups based on skin tone and gender. We note that creating a new dataset using data scraped from the web or from other data sources often requires labeling the data points so as to indicate group membership, again raising the possibility of inaccuracies, as well as questions about the ethics of imposing labels on people.

Scraping data can sometimes make it easier to ensure appropriate representation of additional factors, though most datasets created using scraped data will still not be capable of supporting confirmatory evaluations. For example, by scraping images from parliamentary websites, the researchers who conducted the Gender Shades study also ensured that the values of some additional factors (i.e., lighting conditions and pose) were held reasonably constant; the images were not, however, labeled so as to indicate the values of these factors. Other additional factors, such as age, facial expression, facial hair, hairstyle, glasses, and image resolution, were not accounted for in any way.

Using data scraped from the web or from other data sources also raises a number of considerations involving licensing, consent, and compensation [37]. First, it can be difficult to determine whether a data source is licensed in a such a way that data can be legally scraped from it and used to conduct a disaggregated evaluation. Second, even if a data source is appropriately licensed, this does not mean that the data subjects have consented to their data being used in this way [68]. For example, in the context of face-based AI systems, image licenses are typically chosen by the photographer, not the people depicted in them. A failure to obtain consent from data subjects may even result in legal action in some jurisdictions. For instance, the Illinois Biometrics Information Privacy Act has generated significant litigation over whether and how companies may collect or use Illinois residents' biometric information [53]. Lastly, even if a data source is appropriately licensed and all data subjects have consented to their data being used to conduct a disaggregated evaluation, most data sources do not provide mechanisms for contacting data subjects, making it very difficult to compensate them fairly for contributing their data [5].

3.8.3 Approach 3: Create a new dataset using data from the system's context of use.

The third approach is to create a new dataset using data from the system's context of use. If the system has already been deployed, this can mean reusing deployment data from the system itself. For example, when evaluating a high-risk care management enrollment system, Obermeyer et al. [62] used data about patients that had already interacted with the system. If, however, the evaluation will take place before system deployment, then the data must be obtained via other methods. For example, when evaluating child welfare risk models prior to their potential deployment in Allegheny County, Chouldechova et al. [17] used records that were previously collected by Allegheny County to assess and report different models' area under the curve for groups based on race.

Creating a new dataset using data from the system's context of use can be more time consuming than reusing an existing dataset, and more or less time consuming than creating a new dataset using scraped data. The financial cost can vary. We note that if an evaluation will be focused on actual performance disparities experienced by a specific set of people who encountered the system in the past, then using data from the system's context of use is the only option.

Data from the system's context of use will be reflective of the population of interest and the environmental factors, behavioral factors, and other factors found in situ, provided there are no substantial changes over time. However, there may not be sufficient data about some groups of interest and the data points may not be labeled so as to indicate group membership. In some cases, it may be possible to contact data subjects.

Because data from the system's context of use will be reflective of the population of interest and the environmental factors, behavioral factors, and other factors found in situ, it is possible to implicitly account for additional factors via the argument that any performance disparities are reflective of those that either affected or will affect the population of interest. However, accounting for additional factors by holding their values constant or by considering a range of values for each such factor is likely not possible, making it difficult to conduct confirmatory evaluations using such datasets. As with existing datasets and new datasets created using scraped data, the data points may not be labeled so as to indicate the values of the additional factors, although they may be in some cases.

Data from the system's context of use is unlikely to be subject to licensing restrictions that prevent it from being used by the system's development team(s). However, there may be licensing restrictions that prevent it from being used by outside parties. Using data from the system's context of use also raises questions about consent and compensation. Although it may be possible (though not necessarily easy) to contact data subjects to request their consent or to compensate them fairly for contributing their data, the perceived need to do this may be weaker than when creating a new dataset by collecting data directly from data subjects.

3.8.4 Approach 4: Create a new dataset by collecting data.

By far the most flexible approach is to create a new dataset by collecting data directly from data subjects. However, this approach is much more time consuming and much more expensive than the approaches described above. Indeed, the financial costs can be so great that this approach may only be feasible for development teams with extensive resources.

New datasets created by collecting data directly from data subjects can support evaluations that will take place in the laboratory or in situ. One way to achieve the latter is to conduct an in-situ pilot, in which data subjects interact with a fully or partially functioning system in tightly controlled conditions. This ensures that the resulting dataset reflects the system's intended use cases and deployment contexts, as well as capturing some of the sociotechnical dynamics and their effects on system performance [43, 68]. That said, in-situ pilots can be time consuming and expensive, limiting the number of use cases and deployment contexts that can be considered in practice.

The flexibility afforded by collecting new data directly from data subjects means that it is possible to ensure sufficient data about the groups of interest—even if those groups are intersectional or rare—and to ask data subjects about their group membership. However, we note that it is important to make sure that the data collection mechanisms are sufficiently inclusive of the groups of interest. For example, Park et al. [63] found that many people with disabilities were unable to capture and upload video or speech samples that could improve system performance for people with similar disabilities. It is also possible to account for additional factors by holding their values constant or by considering a range of values for each such factor. And, in either case, the data points can be labeled so as to indicate the values of the additional factors. New datasets created by collecting data directly from data subjects are therefore most capable of supporting confirmatory evaluations. Finally, this approach also provides complete control over licensing, and makes it easy to request data subjects' consent and to compensate them fairly for contributing their data. For example, Facebook AI released a dataset of videos featuring paid actors who had agreed to participate and had explicitly provided labels indicating their age and gender [39]. However, as explained by Raji et al. [68], this approach may place a greater burden on members of already disadvantaged groups by requiring their participation as data subjects—itself a form of injustice.

3.9 Which performance metric(s) will be used?

The performance of many AI systems cannot be usefully summarized via a single metric. For example, the performance of a face-recognition system might refer to its false positive rate or its false negative rate. A disaggregated evaluation that assesses and reports only false positive rates will fail to uncover disparities involving false negative rates, and vice versa. Moreover, false positives and false negatives can cause different harms to different stakeholders, depending on the use case. If a face-identification system is used to grant workplace access to employees, then false positives may cause harms to the employer (e.g., loss of property) by granting access to non-employees, while false negatives will cause harms to individual employees (e.g., being unable to do their jobs or feeling singled out or discriminated against) by denying them access to their workplace. NIST's disaggregated evaluation of face-recognition systems therefore assessed and reported false positive and false negative rates.

The performance metric(s) that will be used to conduct a disaggregated evaluation are primarily determined by what will be evaluated (i.e., the system as a whole or one or more of its constituent components) and where the evaluation will take place. For example, when evaluating commercially available speech-recognition systems, Koenecke et al. [49] used word error rate because they were only able to access the systems as a whole. Had they been able to access the systems' language models, they might instead have used perplexity. As another example, if a face-verification system is evaluated in situ, then it is possible to use performance metrics that depend on the operator's decisions.

We also note that considerations relating to the way in which system performance will be reported can influence the performance metric(s) that will be used. For example, a system's false positive rate is one minus its true negative rate. Therefore assessing a system's false positive rate is equivalent to assessing its true negative rate. However, as noted by NIST in a report on its vendor tests of commercially available face-recognition systems [33], "readers don't perceive differences in numbers near 100% well, becoming inured to the 'high-nineties' effect where numbers close to 100 are perceived indifferently." Moreover, the harms experienced by people whose faces are or are not recognized are directly related to false positive and false negative rates. In contrast, they are inversely related to true negative and true positive rates. The report cites an example that is similar to the workplace access use case described above and explains how doubling the false negative rate doubles the number of employees that will be harmed. Reporting the system's false negative rate, rather than its true positive rate, therefore emphasizes the direct connection between poor performance and harms.

Performance disparities can be reported in absolute or relative terms [46]. Reporting each group's performance relative to that of the best-performing group can make it easier to immediately grasp the extent of any performance disparities. For example, if a system

has a false negative rate of 0.01 for women and a false negative rate of 0.1 for people who are non-binary, then the absolute performance disparity is only 0.09, but false negatives are ten times more likely for people who are non-binary than for women.

Lastly, we emphasize that any set of quantitative performance metrics may fail to capture aspects of system performance that are more subjective, context-dependent, or otherwise harder to quantify. For this reason, different people may experience an AI system differently, even when a disaggregated evaluation appears to demonstrate the absence of any performance disparities involving the groups to which they belong.

3.10 How will performance be analyzed?

The way in which performance will be analyzed depends on many of the choices that we described above, including the goal of the evaluation, the way in which additional factors are accounted for, and the performance metric(s) that will be used.

The simplest way to analyze performance is to assess each performance metric for each group of interest and then compare the resulting values (i.e., point estimates) for each metric. Although this approach is easy to implement, it does not account for any randomness in the evaluation dataset and therefore raises the possibility of misleading results. In contrast, using statistical techniques that incorporate uncertainty (e.g., p-values, confidence intervals, posterior probabilities, permutation tests) can better mitigate this possibility. Moreover, some of these techniques make it easier to detect subtle differences, reducing the likelihood that performance disparities will go undetected. For confirmatory evaluations, it is especially important to use techniques that incorporate uncertainty so as to minimize the risk of drawing incorrect conclusions.

In the context of face-based AI systems, generalized linear mixed-effect regression models have been used to conduct both confirmatory and exploratory evaluations [e.g., 11, 29]. As well as incorporating uncertainty, these models provide a way to account for additional factors when the distributions over factor values are not the same for each group of interest. They are also particularly appropriate when the data points in the evaluation dataset are grouped (e.g., multiple utterances from the same speaker) and they are well suited to analyzing performance for intersectional groups [27]. However, they often rely on strong distributional assumptions about the evaluation dataset that may not hold.

When conducting an exploratory evaluation, decision trees or other partitioning methods can be used to construct a granular and high-dimensional view of system performance. This approach is particularly effective at uncovering potential causes of performance disparities that can then be further investigated via subsequent confirmatory evaluations. For example, researchers have used high-dimensional performance analyses to understand the effects of demographic factors, sociocultural factors, and environmental factors on face-based AI systems [51, 61]. High-dimensional performance analyses can also find intersectional groups for which a system exhibits poor performance using variable importance ranking or other variable selection methods. That said, without efforts to mitigate overfitting, they can yield misleading results.

It is important to note that all of these approaches will struggle to distinguish between the effects of factors that are highly correlated. For example, when assessing and reporting the performance of a face-recognition system, if all images of people wearing glasses are poorly lit and all poorly lit images are of people wearing glasses, but the system only actually exhibits poor performance for one of these groups, then it will be difficult to conclusively determine which one. This highlights the benefit of accounting for additional factors by considering a range of values for each such factor.

There are also non-model-based ways to uncover potential causes of performance disparities. For example, Muthukumar et al. [59] used computer vision post-processing techniques and post-hoc explanation methods to uncover potential causes of the performance disparities that were found in the Gender Shades study.

3.11 How transparent will the evaluation be?

Disaggregated evaluations can vary considerably in their level of transparency. ProPublica made all aspects of their evaluation of Northpointe's COMPAS recidivism-prediction system publicly available, including their results and conclusions [3], a full description of their design choices [52], their evaluation dataset, and the source code that they used to analyze system performance [66]. In contrast, development teams who design and conduct disaggregated evaluations to understand the performance of their own AI systems often choose not to disclose any details, though some have disclosed results and other information as a way to inform others about their systems' characteristics and limitations [e.g., 31, 56].

If all aspects of a disaggregated evaluation are made publicly available, then others can easily repeat the evaluation to verify its results and conclusions. They can also use the evaluation dataset to conduct disaggregated evaluations of other AI systems. For example, many researchers have drawn on ProPublica's design choices, evaluation dataset, and source code to further interrogate its results and conclusions [e.g., 16, 19, 48]. People may be more likely to trust a disaggregated evaluation if all aspects are made publicly available, especially if the evaluation's results are favorable. However, development teams may be reluctant to do this if making aspects of the evaluation available could provide others with a competitive advantage.

Making evaluation datasets publicly available raises the possibility of dataset misuse, such as using evaluation datasets to develop AI systems that cause new harms to already disadvantaged groups. The researchers who conducted the Gender Shades study attempted to mitigate this possibility by restricting access to the PPB dataset to researchers who wish to use it for non-commercial purposes. However, this means that development teams in companies who wish to understand the performance of their own gender classifiers must recreate the PPB dataset themselves, and it is not possible do this in a way that yields an identical dataset.

4 DISCUSSION AND CONCLUSION

We have drawn attention to the variety of choices that must be made when designing a disaggregated evaluation, as well as some of the key considerations that underlie these choices and the trade-offs between these considerations. Making these choices is rarely an easy task and their ramifications can be hard to predict. Some of these choices influence the results that will be obtained, and thus the conclusions that will be drawn; others influence a disaggregated evaluation's impacts—both beneficial and harmful—on people, including the people whose data is used to conduct the evaluation.

Our paper highlights the importance of taking a careful approach to designing disaggregated evaluations and can serve as a road map for evaluation designers. The time and effort spent on dataset construction, including labeling the values of additional factors that can affect system performance, will lead to more conclusive—and more actionable—results. Similarly, ensuring that an evaluation dataset reflects a system's intended use cases and deployment contexts will yield results that better reflect the system's performance in practice. Of course, there are tradeoffs too. For example, tailoring a disaggregated evaluation too closely to one use case or deployment context may lead to results that do not generalize.

Our paper can also serve as a road map for interpreting a disaggregated evaluation's results and conclusions. Have all factors that can affect system performance been accounted for? Does the evaluation dataset reflect the population of interest and the environmental factors, behavioral factors, and other factors found in situ? Was the evaluation confirmatory—that is, did it posit clear hypotheses to be tested and was it designed very carefully so as to minimize the risk of drawing incorrect conclusions? Or is it best understood as an exploratory evaluation? However, for our paper to serve as such a road map, evaluation designers must document their choices, along with the underlying considerations and the tradeoffs that they have made. And, as part of that process, they should clearly communicate any limitations [68]. Existing documentation approaches, like datasheets for datasets [26] and model cards [58], may be of value here.

There are several questions that we have left unaddressed. For example, we have assumed that an evaluation's designers have access to "ground truth" labels for the task that the system is intended to perform. However, in many scenarios, such labels may be inaccurate, either due to measurement issues [44] or due to discrimination [25]. When this is the case and when there is no way to identify the extent or nature of the mislabeling, a disaggregated evaluation's results will be meaningless, no matter how carefully it was designed. We also emphasize that although uncovering performance disparities is essential to the responsible deployment of AI systems, *reducing* performance disparities can be a fraught task. For example, well-known impossibility results imply that performance disparities according to one metric may originate from a decision to reduce performance disparities according to another metric [16, 48]. In addition, improving performance may not always be a desirable outcome. This is especially true for face-recognition systems, which may be perceived as a threat whether they perform well or poorly [38]. Although disaggregated evaluations can add urgency to normative debates about AI systems, there are many other considerations beyond performance that can determine a system's desirability, trustworthiness, or acceptance.

5 ACKNOWLEDGMENTS

We thank Natasha Crampton, Todd Glass, Eric Horvitz, Sasa Junuzovic, Kristen Laird, and Besmira Nushi for their input and feedback.

REFERENCES

[1] Algorithmic Justice League and the Center on Privacy and Technology at Georgetown Law. 2018. Safe Face Pledge. https://www.safefacepledge.org/.

[2] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proc. of the Conf. on Fairness, Accountability, and Transparency (FAccT)*.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. In *ProPublica*.

[4] Apple Inc. 2019. About Face ID advanced technology. https://support.apple.com/en-us/HT208108.

[5] Imanol Arrieta-Ibarra, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E. Glen Weyl. 2018. Should We Treat Data as Labor? Moving Beyond "Free". *American Economic Association Papers and Proceedings* 108 (2018), 38–42.

[6] Australian Border Force. 2019. SmartGates. https://www.abf.gov.au/entering-and-leaving-australia/smartgates.

[7] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing. *Information and Society (SIGCIS)* 2 (2017).

[8] Sebastian Benthhall and Bruce D. Haynes. 2019. Racial categories in machine learning. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT*)*. 289–298.

[9] Lacey Best-Rowden and Anil K. Jain. 2018. Longitudinal study of automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 1 (2018), 148–162.

[10] J. Ross Beveridge, Geof H. Givens, P. Jonathon Phillips, and Bruce A. Draper. 2009. Factors that influence algorithm performance in the Face Recognition Grand Challenge. *Computer Vision and Image Understanding* 113, 6 (2009), 750–762.

[11] J. Ross Beveridge, Geof H. Givens, P. Jonathon Phillips, Bruce A. Draper, David S. Bolme, and Yui Man Lui. 2010. FRVT 2006: Quo Vadis face quality. *Image and Vision Computing* 28, 5 (2010), 732–743.

[12] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination. *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT*)* (2020).

[13] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[14] Joy Buolamwini. 2018. When the Robot Doesn't See Dark Skin. *The New York Times* (2018).

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT*)*. 77–91.

[16] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[17] Alexandra Chouldechova, Emily Putnam-Hornstein, Diana Benavides Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT*)*.

[18] Cynthia Cook, John Howard, Yevgeniy Sirotin, Jerry Tipton, and Arun Vemury. 2019. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2019).

[19] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post* (October 2016).

[20] Kimberlé Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* (1989).

[21] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone? https://research.fb.com/wp-content/uploads/2019/06/Does-Object-Recognition-Work-for-Everyone.pdf.

[22] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2170–2179. https://doi.org/10.1109/TIFS.2014.2359646

[23] Federal Bureau of Investigation. 2019. Next Generation Identification (NGI). https://www.fbi.gov/services/cjis/fingerprints-and-other-biometrics/ngi.

[24] Electronic Frontier Foundation. 2019. Bans, Bills and Moratoria. https://www.eff.org/aboutface/bans-bills-and-moratoria.

[25] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. CoRR arXiv:1609.07236.

[26] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. CoRR arXiv:1803.09010.

[27] Andrew Gelman. 2006. Multilevel (hierarchical) modeling: What It can and cannot do. *Technometrics* 48, 3 (2006), 432–435. https://doi.org/10.1198/004017005000000661

[28] Eden Gillespie. 2019. Are you being scanned? How facial recognition technology follows you, even as you shop. https://www.theguardian.com/technology/2019/feb/24/are-you-being-scanned-how-facial-recognition-technology-follows-you-even-as-you-shop.

[29] G. H. Givens, J. R. Beveridge, P. J. Phillips, B. Draper, Y. M. Lui, and D. Bolme. 2013. Introduction to face recognition and evaluation of algorithm performance.

Computational Statistics and Data Analysis 67 (2013), 236–247.

[30] Google. 2019. Google Photos Help – Search by people, things, & places in your photos. https://support.google.com/photos/answer/6128838.

[31] Google. 2020. Google Cloud Model Cards: Face Detection Model Card v0. https://modelcards.withgoogle.com/face-detection.

[32] Jay Greene. 2020. Microsoft won't sell police its facial-recognition technology, following similar moves by Amazon and IBM. *The Washington Post* (2020). https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/

[33] Patrick Grother, Mei Ngan, and Kayee Hanaoka. 2019. *Face Recognition Vendor Test (FRVT) Part 1: Verification.* Interagency Report DRAFT. National Institute of Standards and Technology (NIST). https://nist.gov/programs-projects/frvt-11-verification

[34] Patrick Grother, Mei Ngan, and Kayee Hanaoka. 2019. *Face Recognition Vendor Test (FRVT) Part 2: Identification.* Interagency Report 8271. National Institute of Standards and Technology (NIST).

[35] Patrick Grother, Mei Ngan, and Kayee Hanaoka. 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Retrieved on at https://doi.org/10.6028/NIST.IR.8280.

[36] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proc. of the Conference on Fairness, Accountability, and Transparency (FAT*).* 501–512.

[37] Jules. Harvey, Adam. LaPlace. 2021. *Exposing.ai.* https://exposing.ai

[38] Nabil Hassein. 2017. Against Black Inclusion in Facial Recognition. https://digitaltalkingdrum.com/2017/08/15/against-black-inclusion-in-facial-recognition/.

[39] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Towards measuring fairness in AI: the Casual Conversations dataset. https://ai.facebook.com/research/publications/towards-measuring-fairness-in-ai-the-casual-conversations-dataset.

[40] Deborah Hellman. 2018. Indirect Discrimination and the Duty to Avoid Compounding Injustice. In *Foundations of Indirect Discrimination Law*, Hugh Collins and Tarunabh Khaitan (Eds.). Hart.

[41] HireVue. 2019. HireVue - Hiring Intelligence | Assessment & Video Interview Software. https://www.hirevue.com.

[42] IBM. 2020. IBM CEO's Letter to Congress on Racial Justice Reform. https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/.

[43] Lucas D Introna and Helen Nissenbaum. 2009. Facial Recognition Technology: A Survey of Policy and Implementation Issues. *Center for Catastrophe Preparedness and Response, New York University* (2009).

[44] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proc. of the Conf. on Fairness, Accountability, and Transparency (FAccT).*

[45] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv preprint arXiv:1908.04913* (2019).

[46] Kenneth Keppel, Elsie Pamuk, John Lynch, Olivia Carter-Pokras, Insun Kim, Vickie Mays, Jeffrey Pearcy, Victor Schoenbach, and Joel S. Weissman. 2005. Methodological issues in measuring health disparities. *Vital and Health Statistics, Series 2: Data Evaluation and Methods Research* 2, 141 (2005), 1–16.

[47] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

[48] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conf. (ITCS 2017).*

[49] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proc. of the National Academy Sciences* 117, 14 (2020), 7684–7689.

[50] K. S. Krishnapriya, Kushal Vangara, Michael C. King, Vitor Albiero, and Kevin Bowyer. 2019. Characterizing the variability in face recognition accuracy relative to race. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops.*

[51] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. *31st AAAI Conf. on Artificial Intelligence, AAAI 2017* (2017), 2124–2132. arXiv:1610.09064

[52] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (May 2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=SV45W9VHgigYbUE-m7o9xnvExqobnjcg

[53] Law 360. 2021. Illinois Cases To Watch In 2021. https://www.law360.com/articles/1336065/illinois-cases-to-watch-in-2021.

[54] Kuang Chih Lee, Jeffrey Ho, and David J. Kriegman. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 5 (2005), 684–698. https://doi.org/10.1109/TPAMI.2005.92

[55] Steve Lohr. 2018. Facial Recognition Is Accurate, if You're a White Guy. *The New York Times* (2018).

[56] Microsoft. 2019. Transparency Note: Azure Cognitive Services Face API. https://azure.microsoft.com/en-us/resources/transparency-note-azure-cognitive-services-face-api/.

[57] Microsoft. 2019. Windows Hello: Discover facial recognition on Windows 10. https://www.microsoft.com/en-us/windows/windows-hello.

[58] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proc. of the Conf. on Fairness, Accountability, and Transparency (FAT*).* ACM, 220–229.

[59] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R. Varshney. 2019. Understanding Unequal Gender Classification Accuracy from Face Images. CoRR arXiv:1812.00099.

[60] Antony Nicol, Chris Casey, and Stuart MacFarlane. 2002. Children are ready for speech technology-but is the technology ready for them. *Interaction Design and Children, Eindhoven, The Netherlands* (2002).

[61] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proc. of the AAAI Conf. on Human Computation and Crowdsourcing (HCOMP).*

[62] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (2019), 447–453.

[63] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proc. of the Conf. on Fairness, Accountability, and Transparency (FAccT).*

[64] Forough Poursabzi-Sangdeh, Samira Samadi, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. A Human in the Loop is Not Enough: The Need for Human-Subject Experiments in Facial Recognition. In *CHI Workshop on Human-Centered Approaches to Fair and Responsible AI.*

[65] Proctorio. 2021. A Comprehensive Learning Integrity Platform - Proctorio. https://proctorio.com/.

[66] ProPublica. 2016. Data and analysis for 'Machine Bias'. https://github.com/propublica/compas-analysis.

[67] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proc. of the AAAI/ACM Conf. on Artificial Intelligence, Ethics, and Society (AIES).*

[68] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society.* 145–151.

[69] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proc. of the Conf. on Fairness, Accountability, and Transparency (FAT*).* 33–44.

[70] James A Rodger and Parag C Pendharkar. 2004. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies* 60, 5-6 (2004), 529–544.

[71] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. *Proc. of the ACM on Human–Computer Interaction* (2019).

[72] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proc. of the Conf. on Fairness, Accountability, and Transparency (FAT*).* 59–68.

[73] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

[74] Luke Stark. 2019. Facial Recognition is the Plutonium of AI. *ACM XRDS* 25, 3 (2019), 50–55.

[75] Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proc. of the First ACL Workshop on Ethics in Natural Language Processing.* 53–59.

[76] Sepsis Watch. 2021. Sepsis Watch: the implementation of a Duke-specific early warning system for sepsis - Duke Institute for Health Innovation. https://dihi.org/project/sepsiswatch/.

[77] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K. Jain, James A. Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother. 2017. IARPA Janus Benchmark-B Face Dataset. *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops* 2017-July (2017), 592–600. https://doi.org/10.1109/CVPRW.2017.87