Measuring stereotype harm from machine learning errors requires understanding who is being harmed by which errors in what ways

ANGELINA WANG, Princeton University, USA
XUECHUNZI BAI, Princeton University, USA
SOLON BAROCAS, Microsoft Research, USA
SU LIN BLODGETT, Microsoft Research, USA

Many forms of machine learning fairness measurement treat errors on all target labels as equally liable to cause fairness-related harms. In this work, we consider that errors associated with a stereotype are more likely to cause harm than other errors when incorrectly classified. This observation has far-reaching implications, as any fairness mitigation technique which does not take this into consideration may inadvertently increase the number of harmful errors at the expense of reducing the overall number of errors. We propose that stereotypes are an important distinction to make between different kinds of errors, and map out a concrete subspace of the relevant harms that these stereotype-associated errors may lead to. Through the use of human studies on our case study of gender stereotypes in object recognition, we empirically find that stereotype-associated errors are indeed more harmful than neutral ones, but for both stereotype-reinforcing and stereotype-violating reasons. We conclude that the presence of harm alone cannot be the sole guide in dictating which errors should be prioritized in fairness mitigation, and offer a more nuanced perspective that depends on who it is that is experiencing the harm.[1]

ACM Reference Format:
Angelina Wang, Xuechunzi Bai, Solon Barocas, and Su Lin Blodgett. 2023. Measuring stereotype harm from machine learning errors requires understanding who is being harmed by which errors in what ways. In *2023 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO).* ACM, New York, NY, USA, 27 pages.

1 INTRODUCTION

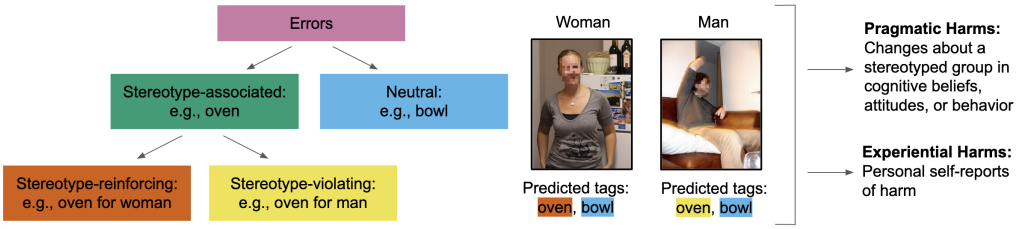

Fig. 1. We distinguish between types of machine learning errors which are likely to have different harms. The first split is between errors made on labels which are associated with stereotypes, e.g., ovens, and those which are more neutral, e.g., bowl. These examples of oven and bowl come from the results of a human study described in Sec. 4.1.1. The second split is within stereotype-associated errors between those which either reinforce or violate the stereotype. This is a simplification as there may be stereotype-associated errors which neither reinforce nor violate the stereotype, and while we rely on the social categories of men and women in this work due to the prevalence of stereotypes about both groups, we do not endorse the binarization of gender. For stereotype-associated errors, we propose two forms of relevant harms: pragmatic and experiential.

[1]This current version is an earlier draft of our work. An updated draft is available at https://angelina-wang.github.io/files/ml_stereotype_harm.pdf

Recent work on the measurement of fairness-related harms in machine learning seeks to be more precise about what we are measuring when we say "bias," and how it might cause harm [12, 18, 19, 29, 73, 137]. In classification tasks, erroneous predictions are straightforwardly treated as more harmful than accurate predictions, but different types of errors are often implicitly treated as either equally liable to cause harm or harmful for the same reasons [15, 36, 122, 141]. This can be at least partially attributed to a focus on equality instead of equity, and therefore failing to recognize that the same error may have different harm if made towards one demographic group compared to another. However, there are cases where certain classification errors have been interpreted as more normatively concerning than others because they are reflecting stereotypes [119, 132, 152]. For example, prior work has raised concerns about object recognition models that amplify the degree to which labels for kitchen items like "knife, fork, and spoon" are incorrectly assigned to photos featuring women and labels for technology-related items like "keyboard and mouse" are incorrectly assigned to photos featuring men [152]. Like others, these authors identified these errors by hand [94], pulling out misclassifications that they believe will reinforce existing stereotypes and thus cause harm to those so stereotyped. Based on these select examples, researchers have tended to extrapolate further, assuming that any other errors that might amplify apparent associations between gender groups and objects must also be stereotype-reinforcing. But it is not a given that the actual people exposed to any such errors will share researchers' views on whether they implicate stereotypes and whether they are harmful; and in fact, we find in our work that they do not on any of these named objects.

In this work, we take a more theoretically-grounded approach, supported by empirical findings, to study the relationship between errors, stereotypes, and harms. We make three primary conceptual contributions (Fig. 1).[2] First, we distinguish errors that are associated with stereotypes from those that are not, and further disentangle within the former category those that are stereotype-reinforcing from those that are stereotype-violating. Second, in considering how these errors might give rise to harm, we distinguish two broad types of harm: *pragmatic* harms involve measurable changes in someone's cognitive beliefs, attitudes, or behaviors towards the group being stereotyped, while *experiential* ones involve self-reports of negative affect. By explicitly disentangling the types of stereotypical associations and types of harm, we enable a fuller understanding of how different errors give rise to different harms, and allow for more grounded mitigation strategies. And finally, we expand the space of considered stereotypes in the literature beyond just associations with occupations and traits.

As a concrete case study, we apply our conceptual contributions to the popular machine learning task of object recognition as used in photo search engines, and consider gender stereotypes. We hypothesize that stereotype-reinforcing errors will give rise to more harms than neutral errors, challenging the often implicit assumption that the harm caused by an error depends more on the relevant group experiencing the error (e.g., as in performance disparities), rather than the type of error (e.g., stereotype-reinforcing or neutral). We design surveys and experiments with online participants from the United States to empirically study which of our defined harms arises from which kinds of errors. We find little immediate evidence of pragmatic harms, and discuss the implications for work aiming to measure such harm, but do find empirical support that stereotype-reinforcing errors (e.g., incorrectly reporting that an oven appears in a photo of a woman) are more experientially harmful than neutral errors (e.g., incorrectly reporting that a bowl appears in a photo of a woman). We also conduct a novel exploration into stereotype-violating errors, which receive scarce attention in the literature, and find that while the stereotyped group mostly finds it more harmful for the error to reinforce rather than violate an existing stereotype, this is not true when it comes to clothing-related items typically associated with women (e.g., cosmetics, necklace, etc.). Here, we see a backlash towards violations of the norms around gender

[2]By scoping this problem to be within errors that machine learning models make, we leave out correct predictions that may or may not be made for stereotypical reasons, i.e., "being right for the wrong reasons."

presentation [112], calling into question the idea that it is always normatively desirable to reduce errors perceived as more harmful due to their relationship to stereotypes. By bringing greater clarity to the types of machine learning errors which are more harmful than others, we urge researchers and practitioners to more carefully consider different kinds of classification errors, potential harms, and the relevant relationships between them. Without doing so, mitigation approaches may actually inadvertently *increase the number of harmful errors* in a well-intentioned but ultimately misguided attempt to reduce the overall number of errors.

Our contributions in this work are both conceptual and empirical. Conceptually, we: (1) propose a distinction between different errors based on their association to a stereotype, (2) map out a concrete subspace of harms (pragmatic and experiential) relevant to this distinction, and (3) expand the narrow scope of what has previously been considered a stereotype beyond just occupations and traits to objects. Applying these conceptual contributions, we then empirically find that: (1) our lab setting does not capture pragmatic harms for stereotype-reinforcing errors; (2) stereotype-associated errors are more experientially harmful than neutral ones, and (3) gender stereotype-reinforcing errors are more experientially harmful to women than gender stereotype-violating errors, except in the case of errors on traditionally feminine clothing items—a finding we discuss which adds nuance to the framework of experiential harms.

2 BACKGROUND

2.1 Stereotypes in Social Psychology

Stereotypes are abstract knowledge structures linking a social group to a set of traits or behavioral characteristics [4, 65, 70, 81, 90]. As abstract knowledge, stereotypes serve as heuristics that help people efficiently make good-enough decisions. Nonetheless, stereotypes can be problematic because they often tend to be over-generalized beliefs which may be inaccurate and even harmful [50, 93]. When considering stereotypes, people do not simply differentiate between which group is good or bad, but rather hold more complex representations. Two primary dimensions of these representations are thought to be warmth (i.e., friendliness and morality) and competence (i.e., ability and assertiveness) [2, 49].

While stereotypes themselves are *cognitive beliefs* that exist in people's minds, they often correlate with people's *attitudes*, i.e., prejudice, and *behavioral tendencies*, i.e., discrimination [49]. These three components of cognitive beliefs, attitudes, and behavioral tendencies comprise a broader framework of bias [47]. Under this framework, people may have *cognitive beliefs* that women are more warm but less competent, and thus *emotionally* express protective attitudes and pity for women [58]. People then *behave* in ways that maintain women's warmth and discount their competence, such as being less likely to promote women to leadership positions [41, 44, 49, 67]. People acquire stereotypes from social learning, such as through family members, friends, news, and mass media [51, 54, 76, 87, 97, 108, 145]. Under this view, we hypothesize that machine learning errors serve as another source for stereotype acquisition.

2.2 Stereotypes in Machine Learning

Stereotyping is a common (often implicit) operationalization of "bias" that is studied in machine learning fairness [10, 16]. Abbasi et al. [1] mathematically formalize stereotyping and show how models trained on stereotyped data can lead to unfair outcomes. Bias amplification is a different statistical notion that rests on the idea that any amplification of an existing bias is undesirable [64, 138, 140, 152]. We show in our work that these approximations of the type of error that is more harmful is often poorly correlated with human judgment, and provide a detailed comparison in Appendix G. Work in natural language processing consider stereotypes reflected in word embeddings [20, 24] and language models [25, 88, 127], with critiques pointed out in how fairness benchmarks fail to operationalize a meaningful

notion of stereotyping [19]. There is generally a large reliance on occupation data from the American Bureau of Labor Statistics, e.g., WinoBias [153], as a data source for stereotypes. One large limitation of this usage (in addition to it only representing American occupation data) is that it only captures descriptive stereotypes, e.g., overrepresentations of groups in an occupation, and misses prescriptive stereotypes, e.g., beliefs about what occupations people of different groups should be in; these two types can often differ in practice [22, 92].

Image Search. We concretize our task of object recognition by constructing human studies about the application of image search. There is a large body of work on biases in image search engines [82, 98, 105, 106, 136], as well as search algorithms and media systems more broadly [37, 61, 72, 103].

Most findings on image search engines have focused on auditing the search engine itself for biased results [82, 98, 105, 106, 136], leaving more speculatory the implications of these results for feeding back into the stereotypes people hold. However, components also attempt to understand the influence that biased search results can have on users. Kay et al. [82] and Metaxa et al. [98] both find that exposure to gender biased image search results can influence an individual's opinion on the gender representation of that occupation (accounting for 7% of a person's opinion in the former and roughly 5% increases, with great variance, when representation shifts from 10% to 50% to 90% in the latter). Metaxa et al. [98] also find an effect of .146 (95% CI [-.16, .45]) on a 7-point Likert scale for the impact of gender representation on occupation inclusivity, with no effect for interest in occupation. Notably, these feelings of inclusivity depend on the gender of the participant. Vlasceanu and Amodio [136] take a different approach by studying occupations (e.g., peruker, lapidary) for which there are very little preconceived notions of stereotypes. In our work, we focus on the reinforcing of existing stereotypes, rather than trying to induce new ones. Contrary to much of this existing work, we do not audit an existing image search engine, instead focusing on downstream concerns that may arise from biases in the search results.

Crowdsourcing Opinions on Harm. We solicit our labels of stereotypes and harm from crowdworkers, similar to how prior work has elicited stereotypes [20, 127] and fairness notions in different contexts [63, 129, 149]. Woodruff et al. [147] specifically focused on the opinions of individuals from marginalized communities who are likely to be adversely affected by algorithms. In a critique [75] of crowd-solicited opinions on the trolley problem [6], the point is made that aggregations of individual opinions on morality follow a flawed framework of methodological individualism, and neglect the structural considerations at play—an important observation we also discover in our findings.

3 CONCEPTUAL CONTRIBUTIONS

Here, we present the conceptual contributions we make that our human studies are then built upon. First, in Sec. 3.1 we introduce a categorization of errors, based on stereotypes, that is useful for thinking about harm. Then, in Sec. 3.2 we present two distinct forms of harm that are relevant to these stereotype errors. Finally, in Sec. 3.3 we explain how this way of thinking is relevant for a broad variety of tasks, including our case study of object recognition in image search.

3.1 Different Kinds of Errors: Based on Association to Stereotype

In quantitative fairness evaluation, errors of any kind are often treated as equally harmful. However, in reality different errors are liable to cause different levels of harm. The distinctions we make among errors are based on their association with a stereotype. As seen in Fig. 1, the first distinction we make is between errors made on labels which are associated with a gender stereotype, e.g., ovens because women are often seen to be in domestic roles [9], and on labels which are more neutral, e.g., bowls because they do not tend to have a stereotypical connotation. This distinction is motivated by the intuition that errors on labels which are more stereotypically-loaded are liable to cause more harm. Then, we further make the distinction between whether the error of the stereotype-associated label is made on an individual such that

the stereotype is reinforced (e.g., oven is falsely predicted on a woman) or violated (e.g., oven is falsely predicted on a man). While both errors are associated with a stereotype, we hypothesize that their implications will be quite different.

3.2 Relevant Harms: Pragmatic and Experiential

Next, we propose two distinct types of harm that are relevant to this categorization of stereotype-associated errors. Stereotypes are not inherently harmful in their own right; for example, a widely-held stereotype is that children are bad drivers — this is not necessarily a harmful belief to hold.[3] Inspired by prior thinking in the space [13, 28, 68], we propose two different conceptualizations of relevant harms: pragmatic harms and experiential harms. Loosely speaking, pragmatic harms are more defined by the view that others hold towards a stereotyped group, whereas experiential harms are those directly experienced by a subject.

Pragmatic harms take the form of a negative change towards the stereotyped group in any of the three components of bias: cognitive beliefs, attitudes, and behavioral tendencies [47]. This framework and the way we measure each component, as we will describe in Sec. 4.1, is carefully grounded in the social psychology literature, allowing us to be systematic in the way we go about our measurements.

For experiential harms, we draw from the idea of microaggressions. Rini [110] defines microaggressions to be "a small act of insult or indignity, relating to a person's membership in a socially oppressed group, which seems minor on its own but plays a part in significant systemic harm." The reason behind the action does not matter and can often be unintentional; similar to errors made by machine learning systems, microaggressions are "easily interpretable as inadvertent errors rather than as malevolent actions" [7]. Because of this, we need to move away from the motivation of the error to focus on the experiential judgment of the subject. We know from standpoint epistemology [46, 104, 148] that we should look to the reported experiences of the individuals being stereotyped, and the difficulty in establishing the legitimacy of this as a measure of harm thus far can be at least partially attributed to testimonial injustice [52]. Emotion has long been discounted as a legitimate source of knowledge, especially when it comes from social groups such as women which are associated with it [74], so we find this measure an important way of conceptualizing harm. We reflexively even found this bias in ourselves as researchers, as we initially conceived only of pragmatic harms in seeking to have an "objective" measure, and neglected to consider self-reports of emotions.

3.3 Relevance of Stereotypes Beyond Occupations and Traits: Objects

In order to concretely apply our conceptual contributions and perform human studies, we select the machine learning task of object recognition as our case study. This is one of the most popular tasks in machine learning [45, 114], and can be used in practical applications such as image search engines and search in personal photo albums on smart phones. We use the latter as our concrete application. An implicit assumption we have been making thus far is that an object can be associated with a stereotype. While not necessarily a surprising statement, it is a relatively novel one as prior work on stereotypes, across disciplines, has largely only focused on occupations and traits — a focus that is certainly well-grounded, given the powerful role that social roles and occupations play in shaping stereotypes [41]. However, in our work, we expand the scope of stereotypes to consider objects, and show how stereotypes are at play beyond just the traditional domains of occupations and traits. While we do not test for it, we hypothesize that stereotypes are also

[3] Beeghly [13] has argued that stereotypes are not intrinsically morally nor epistemically wrong, and it is the context extrinsic to the stereotype which determines whether it may be wrong. Here, we draw on our initial scoping to be within the space of machine learning *errors*. Thus, given that all errors are by definition epistemically wrong, we study what it is that makes certain errors also *harmful* in addition to wrong.

at play in other machine learning tasks as well, e.g., action recognition, pose estimation, and encourage researchers to consider the role of stereotypes in domains more broadly.

3.4 Research Questions

With these conceptual contributions in hand, we formulate two primary research questions: *(1) Are errors which reinforce stereotypes more harmful than neutral ones?* and *(2) What is the role of stereotype-violating errors?*

4 STUDY OVERVIEW

To answer our research questions, we conduct five studies (Tbl. 1). The concrete errors we use are based on the object labels from two popular object recognition datasets: Common Objects in Context (COCO) [89] and OpenImages [85]. COCO has 80 objects annotated across the images, and OpenImages has 600; both datasets have annotations for perceived binary gender annotations of individuals [118, 151].

4.1 Methods

We have established in Sec. 3.1 that we should distinguish errors between those which are stereotype-reinforcing, stereotype-violating, or neutral, and in Sec. 3.2 that the relevant harms we are concerned with between these errors are pragmatic and experiential ones. Now, we need a way to concretely make these distinctions amongst object recognition errors as well as actually measure these harms.

4.1.1 Distinguishing Errors by Stereotype (Studies 1, 4). We recruit human participants to annotate which object labels in the datasets are stereotypes. Drawing from Devine and Elliot [34]'s critique of the Princeton trilogy studies ([57, 79, 81]) which studied racial stereotypes, we asked participants about their perception of stereotypes held by Americans, rather than for their personal beliefs. These stereotype labels are then used in the rest of our studies. Because our work is novel in how we consider objects to be stereotypes, as opposed to traits like *nurturing* or occupations like *mechanic*, on the smaller dataset of COCO we also explore the reasons that participants have for marking object to be stereotypes, and whether or not they find it harmful. Thus, if a participant marks an object as a stereotype, they are further asked the follow-up questions of a) why they marked this object as a stereotype, b) whether they find this stereotype harmful, and c) why they find this stereotype harmful or not.

4.1.2 Measuring Pragmatic Harms (Study 2). We conduct a between-subjects survey experiment on two groups of participants who are exposed to an image search result page that contains either stereotype-reinforcing or neutral errors, and then measure their *cognitive beliefs*, *attitudes*, and *behaviors* (details in Appendix B). In this section, we will use as examples oven and women for the stereotype-reinforcing error and bowl and women for the neutral one. Each question we ask is carefully grounded in the social psychology literature.

For *cognitive beliefs*, we ask three sets of questions which span the spectrum of stereotype-specific to more generically about gendered beliefs. Concretely, we ask about: estimations of who uses ovens and bowls more between men and women; estimations of who tends to be the homemaker more between men and women; and perceived levels of warmth and competence [49] of women. To assess *attitude*, we ask two sets of questions. The first is about how participants feel about women in terms of four emotional components that are believed to mediate interactions between cognitive beliefs and behaviors: a) respect or admiration, b) pity or sympathy, c) disgust or sickening, and d) jealousy or envy [28, 48, 121].

The second asks about sexist attitudes via a shortened scale focused on benevolent sexism [58, 59, 111].[4] Finally, for *behavioral* measures, we ask participants to undertake a realistic task they are liable to encounter which can cause harm: data labeling [134]. We chose this behavior measure because online participants are often the source of training labels in large-scale machine learning datasets. We ask participants to perform two common types of labeling on image data: tagging and captioning (Fig. 2). In the tagging task, we ask participants to label the top three most relevant tags in an image which contains both the stereotype object (e.g., oven) and neutral object (e.g., bowl). We alter the perceived gender of the person to assess whether this changes what is tagged in the image. For the captioning task we show two people, one who looks masculine and another feminine, and swap whether there is a bowl or oven present in the image. This is to understand if the annotators will differently describe who is interacting with the object depending on whether it is stereotypically associated with women or not. All images are generated and/or manipulated by DALL-E 2.

Fig. 2. To measure behavioral tendencies, we ask participants to complete a realistic data annotation task on images which are created and manipulated by DALL-E2. The left pair is for the annotation of image tags, and the right pair is for image captions. Each participant is shown one image from each pair, and then we perform a between-subjects analysis to understand whether perceived gender expression affects the tags, and whether object shown influences how people of different perceived genders are described.

4.1.3 Measuring Experiential Harms (Studies 3, 5). In this operationalization of harm, rather than showing participants a search result page where it is up to the viewer to interpret what kinds of errors are being made, we explicitly ask participants to rate how personally harmful they find different kinds of errors, on a scale from 0 to 9. This is analogous to situations where one reads in the news about the types of errors that artificial intelligence systems make [124], notices such a pattern of errors themselves, or is informed by a friend. For our question phrasing, we draw from the terms used by the Positive and Negative Affect Schedule (PANAS) [26, 143] and select the more relevant of the 10 negative affects, asking if participants experience harms such as feeling upset, irritated, ashamed, or distressed.[5]

4.2 Participants

Human perceptions and judgments are crucial to our study, and we collect judgments from Amazon Mechanical Turk participants through Cloud Research [91]. Different annotators bring different subjective experiences in their labeling of data [30, 31, 102, 142], and there can be strong associations between annotator identity and annotations [115]. While men and women generally tend to hold the same gender stereotypes [42, 69, 92, 144], we still collect equal numbers

[4]We ask questions from the Ambivalent Sexism Inventory [58] about benevolent sexism, as opposed to hostile sexism, because the latter is believed to suffer heavily from social desirability bias.

[5]For the first version of this study done on COCO, we did not use the PANAS terms and simply asked about personal harm through discomfort. In this first version, in addition to personal discomfort, we also ask about societal harm, so that even if the participant does not personally feel harmed, they may feel it on behalf of the stereotyped group. However, we find that participants' responses to both personal and societal harm are extremely correlated, and leave the results for the latter in Appendix C.

Table 1. Summary of all the studies we run and their key findings.

Study	Goal	Dataset	Key Findings
1	Stereotype labeling	COCO	Annotated object labels by stereotype (e.g., oven and women) to use for Studies 2 and 3.
2	Pragmatic harms for stereotype-reinforcing errors	COCO	Stereotype-reinforcing errors do not lead to the pragmatic harms we measure in this lab setting.
3	Experiential harms for stereotype-reinforcing errors	COCO	Stereotype-reinforcing errors are more experientially harmful for women.
4	Stereotype labeling	OpenImages	Annotated object labels by stereotype to use for Study 5.
5	Experiential harms on stereotype-reinforcing and stereotype-violating errors	OpenImages	Stereotype-associated errors are more experientially harmful for all; stereotype-reinforcing errors are more experientially harmful than stereotype-violating ones for women (see Sec. 5.2 for more).

of participants who identify as men and women, and use this variable as a covariate throughout. When we do not see meaningful differences across participant gender, we report aggregated results. Due to limitations in the survey platform which only allow us to specify gender as "male" or "female," this formulation excludes people who identify as non-binary, which is a harmful limitation. Because we do not control for race in the recruitment of participants, our sample diverges from a nationally representative sample. For the scope of our current work which focuses on gender stereotypes, we find this to be an acceptable limitation, especially given that one defining feature of stereotypes is they are largely shared through a cultural consensus [81]. However, we discuss this further as well as other considerations in Sec. 6.3. The timing, pay ($15/hour), and racial demographics for our studies are reported in Tbl. 6 in the Appendix, along with our sample size selection logic. We received IRB approval for all studies.

4.3 Analysis Plan

We use a mixture of qualitative and regression analyses to report our findings. For our within-subjects surveys, we regress with a mixed-effect model whose parameter estimations are adjusted by the group random effects for each individual. We report the coefficients from our regression analyses, which represent the effect size of that independent variable. For example, if we want to understand the effect of the stereotype-reinforcing stimulus over the neutral stimulus for the amount of competence women are believed to have, we would fit: $competence = b \cdot stereotype\_reinforcing + a$, where $stereotype\_reinforcing$ is a binary variable indicating which condition the participant is in, and report the value of b with 95% confidence intervals. We preregistered our study plans on Open Science Framework.

5 EMPIRICAL RESULTS

In this section, we present the results of our human studies, with a summary of each in Tbl. 1.

5.1 RQ1: Are errors which reinforce stereotypes more harmful than neutral ones?

We first establish which errors reinforce stereotypes in Study 1, and then use these to measure for differential pragmatic (Study 2) and experiential (Study 3) harms between the two types of errors.

Study 1: Labeling COCO objects as stereotypes. More than half the participants marked 13 of the 80 objects in COCO (e.g., handbag with women, wine glass with women, tie with men, truck with men) to be a stereotype associated with a gender group. Zero participants marked 18 of the 80 objects (e.g., keyboard, carrot, traffic light) to be a stereotype.

Men and women largely agreed in their annotations, and we present aggregated results in Fig. 3, showing the percentage of participants marking an object to be a gender stereotype on the x-axis.

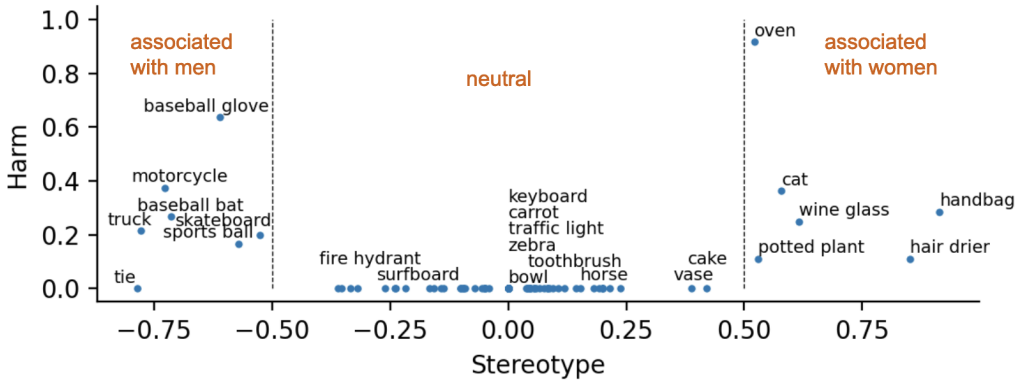

Fig. 3. Participant responses for 80 objects in COCO dataset. The x-axis indicates the percentage of participants who indicated an object is a stereotype, where negative numbers indicate it is a stereotype about men, and positive numbers about women. For objects where more than half of the respondents indicate it is a stereotype, the y-axis indicates the percentage who marked it to be harmful.

One of the authors coded the free responses of why an object was marked to be a stereotype, and found diverse responses grouped into roughly six categories. The most prevalent reasons were: descriptive (45%), e.g., for handbag and women: "women are often seen wearing handbags and buying them"; occupation/role (22%), e.g., for oven and women: "women are stereotyped to always be in the kitchen cooking while the men go out and work"; trait (11%), e.g., for chair and men: "sometimes men would be seen as coming home and just being lazy and lounging in their chair." The full analysis is in Appendix A. It is interesting to note that an object's association to a stereotype is frequently mediated by its connection to a role or trait. We also found that associations between a group and an object can exist through a number of paths. For example, explanations for stereotypical associations between cats and women include: "cat lady," "women are called *kitten*," "women like cats more than dogs," "cats are a feminine animal," and "women are called *cougars*."

In terms of harm, for the 13 objects marked by more than half the participants to be stereotypes, the y-axis of Fig. 3 shows the percentage of participants marking them as harmful. We see large variations within stereotypical objects for whether the association is perceived to be harmful. When asked why or why not, many respondents simply reiterated that the object was a stereotype. Dropping those responses, one of the authors coded the free responses of why a stereotype was marked to be harmful into seven categories, with the top three being: proscriptive (40%), e.g., for dining table and women: "it makes it looked down upon if a man cooks dinner"; prescriptive (26%), e.g., for dining table and women: "I think it puts women in a box that says they must prepare dinner"; negative trait (13%), e.g., for handbag and women: "it is harmful because it implies that women cares more about looks and their appearance." The remaining response categories are in Appendix A. Here, there is a difference in response depending on the participant's gender for who they perceive the harm to be towards. When women specify which of the men group or women group are harmed, they say it is the women group 79% (95% CI [.67, 88]) of the time, while men say it is the women group only 67% (95% CI [.51, .80]) of the time.

Table 2. Our measure of experiential harms show that women find errors associated to stereotypes about both men and women to be more harmful than neutral errors. However, men do not have this kind of differential experiential harm. Bolded results are statistically significant at $p < .05$, and 95% confidence intervals are provided.

Participant Gender	Stereotypical Gender of Object	Effect Size for Stereotype-Reinforcing Error
Women	Women	**1.060 [0.493, 1.627]**
	Men	**1.050 [0.467, 1.633]**
Men	Women	.330 [-.225, .885]
	Men	.030 [-.480, .540]

From these stereotypes, we select a few to use in Studies 2 and 3. For Study 2 we select women and oven because it was marked to be the most harmful, and women and hair dryer because it was marked to be less harmful.[6] For each, we select neutral objects which co-occur in the same setting to serve as controls (bowl for oven and toothbrush for hair dryer). In Study 3, we expand the stereotypes we consider from the two about women to include two more about men (baseball glove and necktie).

Study 2: Pragmatic harms for stereotype-reinforcing errors on COCO. Contrary to what we had initially expected, after performing multiple significance testing with the Benjamini-Hochberg Procedure [14], we do not find any statistically significant changes along the lines of behavior, attitude, or cognitive beliefs for people who experienced stereotype-reinforcing errors as compared to neutral errors (full results in Appendix C and E).

Unrelated to the stimulus, we find a gender effect in how participants from different gender groups respond to certain questions. For the cognitive question asking who tends to be the homemaker, which is on a scale from -10 (mostly men) to 10 (mostly women), participants who identify as women are more likely to associate women with homemakers by an average difference of 1.98 (95% CI [1.28, 2.68]). For the attitude question assessing ambivalent sexism, which is on a scale from 1 (low) to 5 (high), participants who identify as women have a lower sexism score by an average of .28 (95% CI [-.49, -.07]). Similar to our results from Study 1 on which gender group a harm is towards, we again see that participants' own gender identity influences their annotations.

Study 3: Experiential harms for stereotype-reinforcing errors on COCO. In Tbl. 2 we see that participants who identify as women find stereotype-reinforcing errors about both men (b = 1.050, 95% CI [.467, 1.633]) and women (b = 1.060, 95% CI [.493, 1.627]) to be more harmful than neutral errors. However, participants who identify as men do not.

Discussion of Studies 1-3. We can contextualize the current results with respect to prior work in biased image search results. In terms of gendered occupations and cognitive belief changes about the representation of those occupations, Kay et al. [82] and Metaxa et al. [98] both found very small effects, which we do not see on our stimulus condition of oven. This may be due to our smaller level of precision, i.e., discrete values on a slider, compared to their solicitation of a precise percentage number. Then, related to Metaxa et al. [98]'s investigation into feelings of inclusivity in gender biased occupations, when we investigate a related notion of experiential harm depending on the type of error being made, we find that members of the oppressed group (i.e., women) report a significant effect.

Overall we find that the pragmatic harms we mapped out as stemming from repeated stereotypical errors were not able to be measured in our lab setting, likely due to the fact that the effects of these harms are too diffuse and

[6]In our work we focus on harms towards women rather than men because, as expressed by Frye [53], "We hear that oppressing is oppressive to those who oppress as well as to those they oppress... Barriers have different meanings to those on opposite sides of them, even though they are barriers to both... But that barrier is erected and maintained by men, for the benefit of men." We thus focus on stereotypes about women in this study even though stereotypes about men can have harms towards women as well, e.g., through excluding women. This is because given how we constructed our experimental conditions, those errors can be less salient.

long-term, impacted by all of the facets of society we encounter in our lives. Importantly, however, we find that for experiential harms, which can cause stress for oppressed groups [3, 66, 126, 146], women find errors which reinforce gender stereotypes to be more harmful, even though men do not. That there were different reactions speaks not only to the well-known importance of disaggregating survey results, but also to weighting them differently. As discussed in Sec. 3.2, we should likely prioritize the responses of women when addressing gender stereotypes, and take seriously these reports of experiential harms. Our own limitation here is the use of the gender binary. Because of the results we found in these studies, in the next section we focus only on experiential harms.

5.2 RQ2: What is the role of stereotype-violating errors?

We now turn to asking about stereotype-violating errors (e.g., misclassifying oven for men rather than women), and whether there are harms are associated with them.[7] Prior fairness evaluation work tends to only consider the implications of errors that reinforce stereotypes, which is relatively more intuitive to think of as harmful. However, both practically and normatively, it is important to understand the implications of stereotype-violating errors. Practically, mitigations deployed to counteract stereotype-reinforcing errors which act upon the target label will necessarily impact the number of stereotype-violating errors as well. It is important to know, when acting on the information that a particular label is more prone to harm, how concerned we should be with the demographic group that the error is made with respect to: in other words, when correcting for fewer misclassifications of oven than bowl, whether the focus should be only on images of women, or also images of men. This would involve more targeted mitigation, potentially necessitating the collection of demographic information. And normatively, there can be questions of whether stereotype-violating errors may even play a role in reducing stereotypical associations by counteracting them.

Study 4: Labeling OpenImages objects as stereotypes. We follow our method from Study 1 and find 249 of the 600 objects in OpenImages are marked as a stereotype.[8] From this, we compile a list of 40 stereotypical objects (20 about men and 20 about women) and 20 neutral objects, and use this larger set of objects for Study 5.

Study 5: Exploring experiential harms with stereotype-violating errors. We study both stereotype-reinforcing and stereotype-violating errors on this larger set of stereotypes from OpenImages.

Takeaway 1: Not only are stereotype-reinforcing errors more harmful than neutral ones (as we found in Study 3), but so are stereotype-violating errors. In the replication of Study 3 on this larger set of objects we find that, like before, women report a higher level of harm with a coefficient for stereotype-reinforcing errors of .67 (95% CI [.55, .79]) compared to men reporting .33 (95% CI [.21, .45]). For stereotype-violating errors we find the same increased level of harm relative to neutral errors, though with men reporting a higher effect than women at .67 (95% CI [.56, .78]) compared to .42 (95% CI [.31, .53]). In terms of stereotype-associated errors at large, the gendered differences balance out and participants report these to be more harmful than neutral errors with an effect size of .55 (95% CI [.44, .65]).

Next, we dig deeper into the difference in harm between stereotype-associated objects depending on if they are stereotype-reinforcing or stereotype-violating. We draw from theories that differentiate between the costume (i.e., body and appearance) and script (i.e., behavior, traits, and preferences) aspects of gender performance, which has found that reactions to androgynous or conventionally contradictory components can differ depending on which of the two it manifests in [23, 60, 100, 131]. Although this analysis is not preregistered, we introduce an additional independent variable into our regression for whether the stereotyped object can be worn or not. For example, stereotype-associated

[7] We also performed Studies 2-3 on stereotype-violating errors as well as stereotype-reinforcing ones, but because those findings are a subset of what we find here, we leave those results for the Appendix C.

[8] The discrepancy in percentage of objects marked to be a stereotype from COCO is likely due to the fact that we did not ask follow-up questions if an object was marked as a stereotype, so respondents were more willing to mark an object without incurring extra work.

Table 3. Study 5 coefficients on understanding the differential harms of stereotype-associated errors which reinforce or violate stereotypes, and the difference it makes if the object can be worn as "clothing." Statistically significant ($p < .05$) results are bolded, and 95% confidence intervals are provided.

Participant Gender	Stereotypical Gender of Object	Reinforcing vs violating	Clothing	Clothing · (Reinforcing vs violating)
Women	Women	**.316 [.077, .555]**	**.486 [.232, .739]**	**-.734 [-1.077, -.391]**
	Men	**.605 [.419, .792]**	.023 [-.277, .323]	-0.281 [-.688, .126]
Men	Women	-.018 [-.255, .219]	**.726 [.472, .979]**	**-1.018 [-1.361, -.675]**
	Men	-0.033 [-.217, .151]	.037 [-.275, .348]	**-.742 [-1.164, -.319]**

objects that can be worn include ties, suits, lipstick, brassieres, and handbags, and those that cannot include trucks and wine glasses.

Takeaway 2: While women find stereotype-reinforcing errors more harmful than stereotype-violating ones, this reverses if the error is on a more traditionally feminine clothing item, where all participants find these errors more harmful when they violate stereotypes instead of reinforce them. Tbl. 3 shows all of the effect sizes, and we can see that women find stereotype-reinforcing errors more harmful than stereotype-violating ones for both feminine objects (b=.316, 95% CI [.077, .555]) and masculine objects (b=.605, 95% CI [.419, .792]). More generally, we also see that errors on stereotypically feminine clothing objects are deemed more harmful, regardless of who the error is made on, by both women (b=.486, 95% CI [.232, .739]) and men (b=.726, 95% CI [.472, .979]). However, when it comes to our interaction variable of clothing items which reinforce or violate stereotypes, we see a reversal of our initial finding that stereotype-reinforcing errors are more harmful than stereotype-violating ones. Here, both women (b=-.734, 95% CI [-1.077, -.391]) and men (b=-1.018, 95% CI [-1.361, -.675]) find it more harmful for a feminine clothing object to be misclassified on a man than a woman, with the men exhibiting a higher effect size. Men also feel this way about misclassified masculine clothing objects (b=-.742, 95% [-1.164, -.319]), though women do not exhibit this pattern.

More than just a result of stereotype-backlash effects [113], a likely interpretation of these results is as a manifestation of both precarious manhood (i.e., the notion that manhood is precarious and needs continuous social validation such that threats to masculinity can inspire anxiety from men) [135] as well as transphobia, (i.e., a negative reaction to apparent incongruity between a person's perceived gender and a wearable gender presentation item) [23, 100]. The effect size of the results for participants of different genders is also supported by findings that transphobia is higher amongst cisgender men when judging transgender women due to the perceived threat to masculinity [95, 101]. Thus, we find the expansion of scope from stereotype-reinforcing to stereotype-violating errors in this study unveils the limits of our framework for thinking about harm through the lens of experiential harms, because it may encompass intolerances we do not wish to support.

6 DISCUSSION

Taking stock of our studies, we have three primary findings. First, stereotype-reinforcing errors do not result in any of the pragmatic harms we measured. We believe this is because the cumulative effect of being exposed to errors along these lines over a long period of time are extremely hard to measure in the lab setting, likely due to the diffuse and long-term effects that reinforcing stereotypes can have. Long-term observational studies are likely more well-suited to measure these kinds of impacts [51, 54, 76, 87, 97, 108, 145]. Second, in terms of experiential harms, stereotype-reinforcing errors — and in fact stereotype-associated errors more broadly — are more harmful to all participants than neutral errors. This

supports our initial hypothesis that certain labels are more liable for harm than others, and deserve a special focus when trying to understand the representational harms that arise in a machine learning fairness setting.[9] Third, only women find stereotype-associated errors to be more harmful when they reinforce stereotypes rather than violate them, except in the case when the stereotype is about a conventionally feminine clothing item, in which case both men and women find the error to be more harmful when it is made on an image of a man than woman. This final point warrants an especially nuanced discussion, as we find ourselves qualifying a prior claim that we should take people's words at face value when they indicate something is personally harmful. To resolve this conflict, we return to the notions of epistemic injustice [52] and standpoint epistemology [46, 104, 148]. If we understand the negative reaction to misclassifications of stereotypically feminine clothing items on men as a manifestation of precarious manhood [135] or transphobia [23], then we want to downweight these concerns. Respecting people's experiential harms may not be as simple as accepting them at face value for use as a direct guide for measurement, but rather involves understanding which groups of people are likely to be harmed by each kind of error, and prioritizing the experiential harms of certain marginalized groups.

In the next few sections, we describe implications and additional observations that fall out of our findings.

6.1 Implications for Machine Learning

The implications of our findings are significant, and call for us to reconsider fairness measurement in supervised machine learning tasks, such as in leveraging human-driven insights in determining how we train our models [21]. Labels that are not occupations or traits can give rise to stereotype-associated harms, and the finding that not only are certain labels more liable to cause harm than others, but that it matters for *which* demographic group that label is misclassified, suggests that generic approaches like having a higher threshold for the classification of certain labels are insufficient. Instead, more nuanced fairness-through-awareness approaches [40] will need to be taken. This also means that mitigation approaches focused on equality rather than equity which attempt to reduce the maximum discrepancy of errors between two groups may be more appropriate if they instead aim to reduce specific kinds of errors. In other words, performance metrics which do not differentiate between the harm level of different errors may inadvertently prioritize an overall decrease of errors at the potential expense of even increasing the number of harmful errors. While adopting simply a cost-sensitive framework [43, 84] (e.g., where different costs are associated with false positives and false negatives) is a reductive way of interpreting our findings, it is certainly a good starting point to begin with, as one grapples with the questions of whose levels of harms we would prioritize reducing in a bias mitigation framework.

Additionally, our finding from Study 1 that stereotypical associations between a single group and object can emerge from many paths (e.g., the many reasonings behind the association between cat and women), each with different normative valences, illustrates what an oversimplification it is to only label an association as "good" or "bad," and the limitations of mitigations simply aiming to sever the associations deemed "bad." This underscores the importance of work about diversity in annotators' perspectives [30, 31, 39, 77, 102, 142], and how much complexity is reduced by the use of discrete labels. By asking qualitative follow-up questions about why a particular stereotype is held, in addition to the discrete choice of whether something is a stereotype, we gained rich information about the associations we studied, which can in turn be utilized to better inform our understanding of the harms of different errors for different groups. Lack of consensus here is not a weakness or artifact to be averaged out, but rather a point for deeper inquiry on how to prioritize differential experiences of harm.

[9]Related but different is work considering normative arguments for labels on which classifications should not performed at all: lewd labels [27, 107, 150], non-imageable properties (e.g., vegetarian) [150], emotions [8, 11, 86, 120], gender [83, 116, 117], and physical attractiveness [11].

6.2 Can We Automatically Discover Which Labels are Stereotypes?

Evaluation is sometimes considered secondary to algorithm development, and thus rapid and fully-automated evaluations are often prioritized over those requiring human input. Thus, one might imagine trying to automate the determination of which labels are stereotypes, rather than soliciting judgments from human annotators. To test the limits of this approach, we train a variety of models (Support Vector Machine, Random Forest, and Multi-Layer Perceptron) with hyperparameter search over the number of features and find the highest ROC AUC for predicting whether an OpenImages object is a stereotype given an input of BERT word embeddings [35] to be 74%. Given that an object is a stereotype, the highest ROC AUC at predicting which gender is being stereotyped is 85%. These inadequate performance rates indicate that stereotypes are highly contextual, and even with the use of powerful word embeddings which capture bias and social context [55], they are insufficient without human input. Even if the growing power of large language models enables us to predict with higher accuracy which objects are stereotypes, we likely still may want to ensure these annotations come from people themselves [5, 71], thus allowing more room for explanation and critical reflection.

6.3 Limitations

The primary limitations of our study fall along two themes. First and foremost is regarding both our focus on gender stereotypes, and explicit recruitment of participants who identify as men and women. While we believe our general finding about the relative harm of stereotype-associated errors compared to neutral errors generalize beyond gender stereotypes (e.g., to racial stereotypes) this is unlikely to remain true for the gender-specific findings about stereotype-associated errors which reinforce or violate stereotypes, and further work will need to be done to understand the implications in domains other than gender. In terms of participant gender, our choice to use a recruiting platform to have equal numbers of participants who identify as men as women excludes those who do not fall into this gender binary. Especially given our findings which indicate the likeliness of transphobia, it would have been especially important to collect responses from those who do not identify within the gender binary. In future work, it would also be useful to collect information from respondents regarding their held beliefs and worldviews. Another related facet to this set of limitations is that gender stereotypes typically represent stereotypes of the majority subset within that group, e.g., stereotypes about "men" are often those of "cis straight white men" [56]. Further, by setting a threshold of 50% for respondents indicating an object is a stereotype, we are in some senses privileging the opinion of the majority, which may further reify marked stereotypes to be those for the majority subset [99].

Another theme of limitations in our study is that we have relied on surveys in our work, and have not covered the full scope of harms that stereotypes in machine learning errors can have. We have merely focused on two possibilities (i.e., pragmatic and experiential harm), but other spaces for harm include stereotype threat [128, 130] or self-stereotyping [125]. Additionally, most of the changes to cognitive beliefs and attitudes that we measure are explicit, and not through implicit scales such as the Implicit Association Test [62]. Due to this choice, we also risk social desirability bias where respondents answer in a way to represent themselves more favorably. And of course, by formulating the problem under the lens of a rather narrow intervention point, i.e., evaluation, this study likely excludes other manifestations of stereotyping that occur throughout the entire process of the machine learning pipeline [133].

6.4 Broadening the Scope: Machine Learning without Ground-Truth

We scoped our work to machine learning tasks which have a clear notion of error, i.e., ground-truth labels. Here, we consider the implications of our findings for other machine learning tasks which do not have such a clear notion of a error, for example in text generation.

Bolukbasi et al. [20] and Caliskan et al. [24] first brought to light that word embeddings mirror stereotypes in our society, such as about occupations and attributes from the Implicit Association Test [62]. Garg et al. [55] continue this thread of work on how word embeddings reflect societal stereotypes by using the shifting of word embedding biases over time to track societal stereotype changes. However, these works do not provide much insight as to what that means for our use of these word embeddings, and potential feedback loop effects for their implications for society.

Instead what we see in the follow-up work in this space is a logical fallacy where the harm of one type of error, e.g., a correlation of some set of stereotypical occupations to gender, is extended to *all* unnamed errors. The nuance is lost when biases of all kinds in word embeddings are equated to stereotypes, and most notions of gender are targeted to be removed from word embeddings. To put this into perspective, in the large body of literature that has followed the discovery of gender biases in the embedding space [32, 33, 78, 80, 96, 109, 123, 154], all eight of these works would, as far as we can tell, attempt just as much to debias words like "table" and "apple" as they would "homemaker" and "doll." While it is not clear exactly is the desired state of debiasing, e.g., describing the world as it is, prescribing the world as it ought to be, aligning with people's existing stereotypic expectations [139], it surely seems that more thinking should be done on the different implications of debiasing stereotypes as opposed to debiasing more neutral concepts.

7 CONCLUSION

In this work, we use the lens of stereotypes from social psychology to distinguish between errors which cause different levels of harm. To measure this harm, we concretely map out a subspace of harms relevant to this delineation of errors: pragmatic and experiential. We widen the scope of what has typically been considered a stereotype beyond just occupations and traits to also include objects, and perform human studies to validate the utilty of our distinction amongst errors. We empirically find that stereotype-associated errors are more harmful than neutral ones, for both stereotype-reinforcing and stereotype-violating reasons that warrant nuanced consideration regarding how to address. We hope that practitioners will take away from our findings that the differential harms among possible errors should warrant a rethinking of fairness evaluations which treat all labels as equally liable to cause harm.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship to Angelina Wang. We are grateful to funding from the Data-Driven Social Science Initiative at Princeton University. We thank Molly Crockett for suggesting the framing of microaggressions, and Orly Bareket, Sunnie S. Y. Kim, Anne Kohlbrenner, Danaë Metaxa, Vikram V. Ramaswamy, Olga Russakovsky, Hanna Wallach, and members of the Visual AI Lab at Princeton, Fiske Lab at Princeton, and Perception and Judgment Lab at the University of Chicago for feedback.

REFERENCES

[1] Mohsen Abbasi, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in representation: quantifying stereotyping as a representational harm. *Siam International Conference on Data Mining* (2019).

[2] Andrea E. Abele, Naomi Ellemers, Susan T. Fiske, Alex Koch, and Vincent Yzerbyt. 2021. Navigating the Social World: Toward an Integrated Framework for Evaluating Self, Individuals, and Groups. *Psychological Review* 128 (2021). Issue 2.

[3] Kevin W. Allison. 1998. Stress and Oppressed Social Category Membership. *Prejudice: The Target's Perspective* (1998).

[4] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice. (1954).

[5] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* (2023).

[6] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* (2018).

[7] Carla Bagnoli. 2022. Feeling Wronged: The Value and Deontic Power of Moral Distress. *Ethical Theory and Moral Practice* 25 (2022).

[8] Jennifer Bard. 2020. Developing a Legal Framework for Regulating Emotion AI. *University of Florida Levin College of Law Research Paper* (2020).

[9] Orly Bareket, Nurit Shnabel, Anna Kende, Nadine Knab, and Yoav Bar-Anan. 2021. Need some help, honey? Dependency-oriented helping relations between women and men in the domestic sphere. *Journal of Personality and Social Psychology* (2021).

[10] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (2021).

[11] Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. Social B(eye)as: Human and Machine Descriptions of People Images. *Proceedings of the International AAAI Conference on Web and Social Media* (2019).

[12] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *Proceedings of SIGCIS*. Philadelphia, PA.

[13] Erin Marie Beeghly. 2014. Seeing Difference: The Ethics and Epistemology of Stereotyping. *UCBerkeley PhD Thesis* (2014).

[14] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* 57 (1995).

[15] Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578* (2019).

[16] Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (2019).

[17] Yochanan E. Bigman, Desman Wilson, Mads N. Arnestad, Adam Waytz, and Kurt Gray. 2022. Algorithmic Discrimination Causes Less Moral Outrage Than Human Discrimination. *Journal of Experimental Psychology: General* (2022).

[18] Su Lin Blodgett, Solon Barocas, Hal Daume III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Association for Computational Linguistics (ACL)* (2020).

[19] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (2021).

[20] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Conference on Neural Information Processing Systems (NeurIPS)* (2016).

[21] C. Malik Boykin, Sophia T. Dasch, Vincent Rice Jr., Venkat R. Lakshminarayanan, Taiwo A. Togun, and Sarah M. Brown. 2021. Opportunities for a More Interdisciplinary Approach to Measuring Perceptions of Fairness in Machine Learning. *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)* (2021).

[22] Diana Burgess and Eugene Borgida. 1999. Who women are, who women should be: descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, Public Policy, and Law* 5 (1999). Issue 3.

[23] Judith Butler. 1990. Gender Trouble: Feminism and the Subversion of Identity. *Routledge* (1990).

[24] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* (2017).

[25] Yang Cao, Anna Sotnikova, Rachel Rudinger Hal Daumé III, and Linda Zou. 2022. Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2022).

[26] John R. Crawford and Julie D. Henry. 2004. The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology* (2004).

[27] Kate Crawford and Trevor Paglen. 2019. Excavating AI: the politics of images in machine learning training sets. *Excavating AI* (2019).

[28] Amy J. C. Cuddy, Susan T. Fiske, and Peter Glick. 2007. The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology* 92 (2007). Issue 4.

[29] David Danks and Alex John London. 2017. Algorithmic bias in autonomous systems. *International Joint Conference on Artificial Intelligence (IJCAI)* (2017).

[30] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics* (2022).

[31] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation. *NeurIPS 2021 Workshop on Data-Centric AI* (2021).

[32] Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. *AAAI Technical Track: Natural Language Processing* (2020).

[33] Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word Vectors. *International Conference on Artificial Intelligence and Statistics* (2019).

[34] Patricia G. Devine and Andrew J. Elliot. 1995. Are Racial Stereotypes Really Fading? The Princeton Trilogy Revisited. *Personality and Social Psychology Bulletin* 21 (1995). Issue 11.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* (2019).

[36] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone?

[37] Alejandro M. Diaz. 2008. Through the Google Goggles: Sociopolitical Bias in Search Engine Design. *Information Science and Knowledge Management* (2008).

[38] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)* (2021).

[39] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing Ambiguity in Crowdsourcing Frame Disambiguation. *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)* (2018).

[40] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2012. Fairness Through Awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (2012).

[41] Alice H. Eagly. 1987. Sex differences in social behavior: A social-role interpretation. *Lawrence Erlbaum Associates, Inc.* (1987).

[42] Alice H. Eagly, Christa Nater, David I. Miller, Michèle Kaufmann, and Sabine Sczesny. 2020. Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist* (2020).

[43] Charles Elkan. 2001. The Foundations of Cost-Sensitive Learning. *Proceedings of the international joint conference on artificial intelligence* (2001).

[44] Naomi Ellemers et al. 2018. Gender stereotypes. *Annual review of psychology* 69 (2018), 275–298.

[45] Mark R. Everingham, Luc J van Gool, Christopher Kenneth Ingle Williams, John M. Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)* (2010).

[46] Saba Fatima. 2020. I Know What Happened to Me: The Epistemic Harms of Microaggression. *Microaggressions and Philosophy* (2020).

[47] Susan T. Fiske. 1998. Stereotyping, prejudice, and discrimination. *McGraw-Hill* (1998).

[48] Susan T. Fiske, Amy J. C. Cuddy, and Peter Glick. 2002. Emotions Up and Down: Intergroup Emotions Result from Status and Competition. *Prejudice to Intergroup Emotions: Differentiated Reactions to Social Groups* (2002).

[49] Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* 82 (2002). Issue 6.

[50] Susan T Fiske and Steven L Neuberg. 1990. A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology*. Vol. 23. Elsevier, 1–74.

[51] Thomas E. Ford. 1997. Effects of Stereotypical Television Portrayals of African-Americans on Person Perception. *Social Psychology Quarterly* (1997).

[52] Miranda Fricker. 2009. Epistemic Injustice: Power and the Ethics of Knowing. *Oxford University Press* (2009).

[53] Marilyn Frye. 1983. Oppression. *The Politics of Reality* (1983).

[54] Yuki Fujioka. 1999. Television Portrayals and African-American Stereotypes: Examination of Television Effects when Direct Contact is Lacking. *Journalism and Mass Communication Quarterly* (1999).

[55] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zhou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* (2018).

[56] Negin Ghavami and Letitia Anne Peplau. 2012. An Intersectional Analysis of Gender and Ethnic Stereotypes: Testing Three Hypotheses. *Psychology of Women Quarterly* 37 (2012). Issue 1.

[57] Gustave Mark Gilbert. 1951. Stereotype persistence and change among college students. *The Journal of Abnormal and Social Psychology* 46 (1951). Issue 2.

[58] Peter Glick and Susan T. Fiske. 1996. The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology* 70 (1996). Issue 3.

[59] Peter Glick and Jessica Whitehead. 2010. Hostility Toward Men and the Perceived Stability of Male Dominance. *Social Psychology* 41 (2010). Issue 3.

[60] Erving Goffman. 1959. The Presentation of Self in Everyday Life. *Doubleday* (1959).

[61] Sherryl Browne Graves. 1999. Television and prejudice reduction: When does television as a vicarious experience make a difference?m. *Journal of Social Issues* 55 (1999). Issue 4.

[62] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* (1998).

[63] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. *Proceedings of the Web Conference (WWW)* (2018).

[64] Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. 2022. A Systematic Study of Bias Amplification. *arXiv:2201.11706* (2022).

[65] David L Hamilton and Jeffrey W Sherman. 2014. Stereotypes. In *Handbook of social cognition*. Psychology Press, 17–84.

[66] Shelly P. Harrell. 2010. A Multidimensional Conceptualization of Racism-Related Stress: Implications for the Well-Being of People of Color. *American Journal of Orthopsychiatry* (2010).

[67] Madeline E Heilman. 2001. Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of social issues* 57, 4 (2001), 657–674.

[68] Deborah Hellman. 2011. When Is Discrimination Wrong? *Harvard University Press* (2011).

[69] Tanja Hentschel, Madeline E. Heilman, and Claudia V. Peus. 2019. The Multiple Dimensions of Gender Stereotypes: A Current Look at Men's and Women's Characterizations of Others and Themselves. *Frontiers in Psychology* (2019).

[70] James L Hilton and William Von Hippel. 1996. Stereotypes. *Annual review of psychology* 47, 1 (1996), 237–271.

[71] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. *Conference on Human Factors in Computing Systems (CHI)* (2023).

[72] Lucas D. Introna and Helen Nissenbaum. 2000. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society* (2000).

[73] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. *Conference on Fairness, Accountability and Transparency (FAccT)* (2021).

[74] Alison M. Jaggar. 1989. Love and knowledge: Emotion in feminist epistemology. *Inquiry* (1989).

[75] Abby Everett Jaques. 2019. Why the moral machine is a monster. *We Robot* (2019).

[76] Joyce Jennings-Walstedt, Florence L. Geis, and Virginia Brown. 1980. Influence of television commercials on women's self-confidence and independent judgment. *Journal of Personality and Social Psychology* (1980).

[77] Sanjay Kairam and Jeffrey Heer. 2016. Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)* (2016).

[78] Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. *Annual Conference of the Association for Computational Linguistics (ACL)* (2019).

[79] Marvin Karlins, Thomas L. Coffman, and Gary Walters. 1969. On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology* 13 (1969). Issue 1.

[80] Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor Debiasing of Word Representations Evaluated on WEAT. *arXiv:1906.05993* (2019).

[81] Daniel Katz and Kenneth Braly. 1933. Racial Stereotypes of One Hundred College Students. *The Journal of Abnormal and Social Psychology* 28 (1933). Issue 3.

[82] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. *Conference on Human Factors in Computing Systems (CHI)* (2015).

[83] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)* (2018).

[84] Matjaž Kukar and Igor Kononenko. 1998. Cost-Sensitive Learning with Neural Networks. *European Conference on Artificial Intelligence* (1998).

[85] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Shahab Kamali Jordi Pont-Tuset, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)* (2020).

[86] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *Proceedings of the International AAAI Conference on Web and Social Media* (2019).

[87] Gloria Ladson-Billings. 2009. 'Who you callin' nappy-headed?' A critical race theory look at the construction of Black women. *Race Ethnicity and Education* (2009).

[88] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. *arXiv:2211.09110* (2022).

[89] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)* (2014).

[90] Walter Lippmann. 1922. Public opinion. (1922).

[91] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 42 (2017).

[92] Mercedes López-Sáez and Ana Lisbona. 2014. Descriptive and prescriptive features of gender stereotyping. Relationships among its components. *International Journal of Social Psychology* 24 (2014). Issue 3.

[93] C Neil Macrae, Alan B Milne, and Galen V Bodenhausen. 1994. Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of personality and Social Psychology* 66, 1 (1994), 37.

[94] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)* (2022).

[95] Arti P. Makwana, Kristof Dhont, Jonas De keersmaecker, Parisa Akhlaghi-Ghaffarokh, Marine Masure, and Arne Roets. 2018. The Motivated Cognitive Basis of Transphobia: The Roles of Right-Wing Ideologies and Gender Role Beliefs. *Sex Roles* 79 (2018).

[96] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (2019).

[97] Dana Mastro, Elizabeth Behm-Morawitz, and Michelle Ortiz. 2007. The Cultivation of Social Perceptions of Latinos: A Mental Models Approach. *Media Psychology* (2007).

[98] Danaë Metaxa, Michelle A. Gan, Su Goh, Jeff Hancock, and James A. Landay. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *ACM Conference on Human-Computer Interaction (CSCW)* (2021).

[99] John Stuart Mill. 1859. On Liberty. *Longman, Roberts, Green Co.* (1859).

[100] Thekla Morgenroth and Michelle K. Ryan. 2020. The Effects of Gender Trouble: An Integrative Theoretical Framework of the Perpetuation and Disruption of the Gender/Sex Binary. *Perspectives on Psychological Science* 16 (2020).

[101] Craig T. Nagoshi, J. Raven Cloud, Louis M. Lindley, Julie L. Nagoshi, and Lucas J. Lothamer . 2019. A Test of the Three-Component Model of Gender-Based Prejudices: Homophobia and Transphobia Are Affected by Raters' and Targets' Assigned Sex at Birth. *Sex Roles* 80 (2019).

[102] Jennifer A. Noble. 2012. Minority voices of crowdsourcing: why we should pay attention to every member of the crowd. *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)* (2012).

[103] Safiya Umoja Noble. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. *NYU Press* (2018).

[104] Omaith O'Dowd. 2018. Microaggressions: A Kantian Account. *Ethical Theory and Moral Practice* 21 (2018).

[105] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. *Conference on Human Factors in Computing Systems (CHI)* (2017).

[106] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)* (2018).

[107] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv:2006.16923* (2020).

[108] Narissra M. Punyanunt-Carter. 2008. The Perceived Realism of African American Portrayals on Television. *Howard Journal of Communications* (2008).

[109] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *Annual Conference of the Association for Computational Linguistics (ACL)* (2020).

[110] Regina Rini. 2020. The Ethics of Microaggression. *Routledge Taylr & Francis Group* (2020).

[111] Chiara Rollero, Peter Glick, and Stefano Tartaglia. 2014. Psychometric properties of short versions of the Ambivalent Sexism Inventory and Ambivalence Toward Men Inventory. *TPM-Testing, Psychometrics, Methodology in Applied Psychology* 21 (2014). Issue 2.

[112] Laurie A. Rudman, Corinne A. Moss-Racusin, Peter Glick, and Julie E. Phelan. 2012. Reactions to Vanguards: Advances in Backlash Theory. *Advances in experimental social psychology* 45 (2012).

[113] Laurie A. Rudman, Corinne A. Moss-Racusin, Julie E. Phelan, and Sanne Nauts. 2012. Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology* 48 (2012).

[114] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Zhiheng Huang Sean Ma, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015).

[115] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2022).

[116] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)* (2019).

[117] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)* (2020).

[118] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. 2021. A Step Toward More Inclusive People Annotations for Fairness. *ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (2021).

[119] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. 2020. Diagnosing gender bias in image recognition systems. *Socius* (2020).

[120] Elaine Sedenberg and John Chuang. 2017. Smile for the Camera: Privacy and Policy Implications of Emotion AI. *arXiv:1709.00396* (2017).

[121] Charles R. Seger, Ishani Banerji, Sang Hee Park, Eliot R. Smith, and Diane M. Mackie. 2017. Specific emotions as mediators of the effect of intergroup contact on prejudice: findings across multiple participant and target groups. *Cognition and Emotion* 31 (2017). Issue 5.

[122] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.

[123] Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing Gender Bias in Word Embedding with Latent Disentanglement and Counterfactual Generation. *Findings of EMNLP* (2020).

[124] Tom Simonite. 2018. When It Comes to Gorillas, Google Photos Remains Blind. *Wired, January* (2018).

[125] Stacey Sinclair, Curtis D. Hardin, and Brian S. Lowery. 2006. Self-stereotyping in the context of multiple social identities. *Journal of Personality and Social Psychology* 90 (2006). Issue 4.

[126] William A. Smith, Man Hung, and Jeremy D. Franklin. 2011. Racial Battle Fatigue and the MisEducation of Black Men: Racial Microaggressions, Societal Problems, and Environmental Stress. *The Journal of Negro Education* 80 (2011).

[127] Anna Sotnikova, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. 2021. Analyzing Stereotypes in Generative Text Inference Tasks. *Findings of the Association for Computational Linguistics: ACL-IJCNLP* (2021).

[128] Steven J. Spencer, Christine Logel, and Paul G. Davies. 2015. Stereotype Threat. *Annual Review of Psychology* 67 (2015).

[129] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2019).

[130] Claude M. Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology* 69 (1995). Issue 5.

[131] Chadly Stern and Nicholas O. Rule. 2017. Physical Androgyny and Categorization Difficulty Shape Political Conservatives' Attitudes Toward Transgender People. *Social Psychological and Personality Science* (2017).

[132] Pierre Stock and Moustapha Cisse. 2018. ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. *arXiv:1711.11443* (2018).

[133] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2022).

[134] Emiel van Miltenburg. 2016. Stereotyping and Bias in the Flickr30K Dataset. *Proceedings of the Workshop on Multimodal Corpora* (2016).

[135] Joseph A. Vandello, Jennifer K. Bosson, Dov Cohen, Burnaford Rochelle M, and Jonathan R. Weaver. 2008. Precarious manhood. *Journal of Personality and Social Psychology* (2008).

[136] Madalinea Vlasceanu and David M. Amodio. 2022. Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 119 (2022).

[137] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring Representational Harms in Image Captioning. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2022).

[138] Angelina Wang and Olga Russakovsky. 2021. Directional Bias Amplification. *International Conference on Machine Learning (ICML)* (2021).

[139] Clarice Wang, Kathryn Wang, Andrew Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2021. User Acceptance of Gender Stereotypes in Automated Career Recommendations. *arXiv:2106.07112* (2021).

[140] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *International Conference on Computer Vision (ICCV)* (2019).

[141] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation.

[142] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science* (2016).

[143] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology* 54 (1988).

[144] John E. Williams and Deborah L. Best. 1977. Sex Stereotypes and Trait Favorability on the Adjective Check List. *Educational and Psychological Measurement* 37 (1977). Issue 1.

[145] Clint C. Wilson, Féliz Gutiérrez, and Lena M. Chao. 2003. Racism, Sexism, and the Media: Multicultural Issues Into the New Communications Age. *SAGE* (2003).

[146] Gloria Wong-Padoongpatt, Nolan Zane, Sumie Okazaki, and Anne Saw. 2017. Decreases in implicit self-esteem explain the racial impact of microaggressions among Asian Americans. *Journal of Counseling Psychology* 64 (2017).

[147] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. *Conference on Human Factors in Computing Systems (CHI)* (2018).

[148] Alison Wylie. 2003. Why Standpoint Matters. *Science and Other Cultures: Issues in Philosophies of Science and Technology* (2003).

[149] Mohammad Yaghini, Andreas Krause, and Hoda Heidari. 2021. A Human-in-the-loop Framework to Construct Context-aware Mathematical Notions of Outcome Fairness. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (2021).

[150] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2020).

[151] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. *International Conference on Computer Vision (ICCV)* (2021).

[152] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2017).

[153] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *North American Chapter of the Association for Computational Linguistics* (2018).

[154] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. *Empirical Methods in Natural Language Processing (EMNLP)* (2018).

A ADDITIONAL RESULTS FROM STUDY 1

Here we present the full analyses we perform on the open-ended responses we received in Study 1 regarding why participants believe an object is a stereotype, and if so, why they find it harmful or not.

Our categorization for why an object is a stereotype or not are as follows (some responses did not fall into any of the categories):

- Descriptive (45%), e.g., for handbag and women: "women are often seen wearing handbags and buying them"
- Occupation/role (22%), e.g., for oven and women: "Women are stereotyped to always be in the kitchen cooking while the men go out and work"
- Trait (11%), e.g., for chair and men: "sometimes men would be seen as coming home and just being lazy and lounging in their chair"
- Pop culture (8%), e.g., for cow and women: "Most people who describe a women as a cow are being harmful and hurtful"
- Connection to another object (5%), e.g., for vase and women: "I think women are seen as liking flowers, which are often put into a vase"
- Prescriptive (3%), e.g., for handbag and women: "society generally believes that only women should carry handbags"

Our categorization for why a stereotypical object is harmful is as follows (some responses did not fall into any of the categories):

- Proscriptive, i.e., excluding (40%), e.g., for dining table and women: "it makes it looked down upon if a man cooks dinner."
- Prescriptive, i.e., restricting (26%), e.g., for dining table and women: "I think it puts women in a box that says they must prepare dinner"
- Negative Trait (13%), e.g., for handbag and women: "It is harmful because it implies that women cares more about looks and their appearance."
- Demeans (10%), e.g., for cow and women: "cow is a typical insult for a women a man doesn't like ('stupid cow')"
- Objectifies (5%), e.g., for cup and women: "It is harmful because a cup is an object and it's comparing it to a woman"
- Sexism (3%), e.g., for sandwich and women: "make me a sandwich meme sexism"
- Incorrect (3%), e.g., for sandwich and women: "It's an old, tired stereotype that holds no merit."

The following are the two reasons respondents listed a stereotype to not be harmful: not negative (96%), e.g., for tie and men: "I don't think it's harmful because it's just a fashion choice"; positive stereotype (4%), e.g., for cake and women: "cake can be used to describe a woman as sweet and nice looking. For that reason I don't find it harmful."

B ADDITIONAL SETUP FOR STUDY 2

In Study 2 we first ask the participant to engage with this search result by posing a hypothetical situation where they are trying to find a particular image of someone with an oven or bowl. We ask them to describe the result in 3-4 sentences.

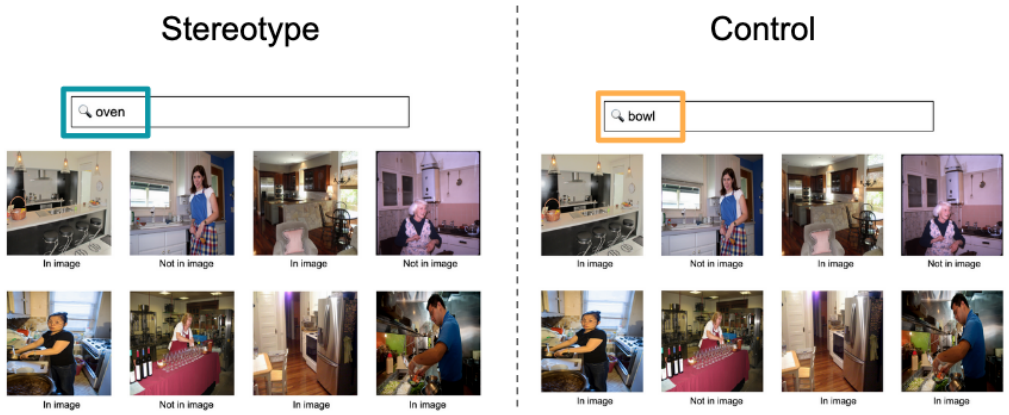

Fig. 4. Our two stimuli from each condition are shown. They are both an image search result containing the same set of images, each of which indicates whether the search query is pictured in the image, i.e., if the image search retrieval was correct or not. The teal and orange squares indicate that the only difference between the stimuli is the search term, as all images which contain an oven also contain a bowl, and all which do not contain an oven also do not contain a bowl. This was a deliberate choice to control for all potential confounding factors from the images in the study.

We then ask them the behavior questions, then re-expose them to the stimulus before asking them the cognitive belief and attitude questions.

The stimuli take the form of an image search result and are pictured in Fig. 4 with teal and orange colored boxes around the component of the image that changes between conditions. The search bar contains the search query, and then eight images that may or may not be correctly retrieved are shown. Each of the eight images is annotated with either "In image" or "Not in image" to make it clear to the participant which images are correct or not. The stereotype condition on the left contains the search query of "oven" with five correctly identified ovens, and three false positive images that all contain women. In other words, this classifier erroneously (and stereotypically) assumes there are ovens in images of women. The control condition contains all of the exact same images, with the only change being that the search query is now "bowl" instead of "oven." This is because the five correct images were deliberately chosen to contain both bowls and ovens, which allows us to control for the variance between the different search conditions. All false positive images were selected from the actual errors of a Vision Transformer (ViT) model [38] trained on COCO so that they are as realistic as possible to a computer vision model's errors, and not completely egregious, e.g., a picture of a woman in a sports field as a false positive for "oven" or "bowl."

C ADDITIONAL RESULTS FROM STUDIES 2-3

Here we report additional results from Studies 2 and 3 that did not make it into the main text. Fig. 5 shows the lack of statistically significant responses in our study on COCO between stereotype-reinforcing and neutral errors. Below, notably, we present comparison between stereotype-reinforcing and stereotype-violating conditions. In Fig. 6 we present the regression coefficients between the stereotype-reinforcing and stereotype-violating conditions, and like the analysis presented in the main text, find no statistically significant changes.

In Tbl. 4 we present the full results from Study 3, including responses from participants about whether an error was personally harmful or societally harmful. There is a high correlation between the responses to these two versions of the question for each error.

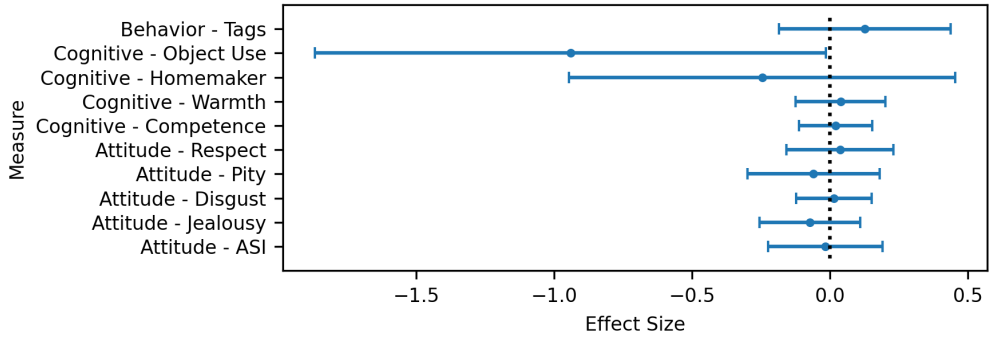

Fig. 5. The effect sizes and 95% confidence intervals are reported for 10 of our 11 measures of pragmatic harm (for the behavior measure of captioning, we provide a descriptive analysis in Appendix E). Deviations from zero indicate that exposure to the stereotype-reinforcing stimulus resulted in a behavior, cognitive, or attitude change compared to exposure to the neutral stimulus. After correcting for the p-values using the Benjamini-Hochberg Procedure [14], we see no statistically significant effects.

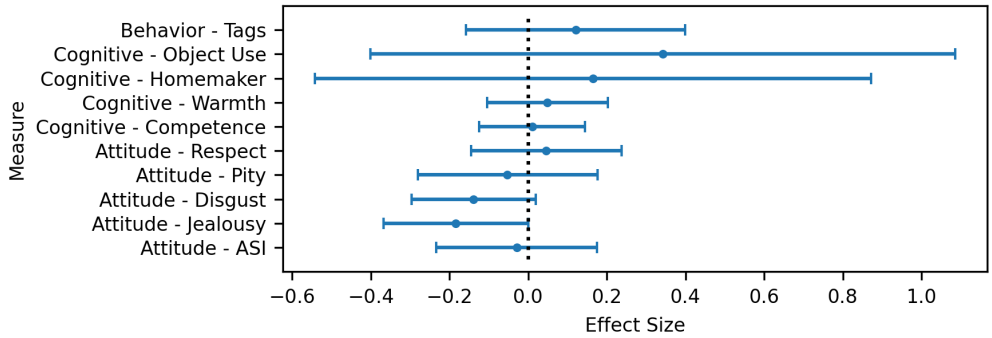

Fig. 6. The results of our regression for 10 of our 11 measures of pragmatic harm (for the behavior measure of captioning, we provide a descriptive analysis in Appendix E) between the conditions of stereotype-reinforcing and stereotype-violating, and find no statistically significant differences.

We were somewhat surprised to find that participants who identified as men did not even think stereotype-reinforcing errors would be more harmful towards society than those which are not, but this also follows from prior work which has found that algorithmic discrimination (as opposed to human discrimination) tends not to lead to moral outrage, which as an unfortunate consequence of not associating intent with machine learning errors, and often leads to a lack of collective action against such harms [17].

D OPENIMAGES OBJECTS

In OpenImages, of the 600 objects, we select the 20 marked with the most agreement to be stereotypically associated with men, and the 20 marked with the most agreement to be stereotypically associated with women. We then randomly select amongst 20 objects that are marked to have no gender stereotypes associated with them. Left out of this are all

Table 4. Full results from Study 3 that asks participants whether they find certain errors personally harmful or societally harmful. Statistically significant ($p < .05$) effects are in bold.

Participant Gender	Stereotypical Gender of Object	Question Asked	Associated vs neutral	Reinforcing vs Violating
Women	Women	Personal	**1.06 [.49, 1.63]**	**.99 [0.39, 1.59]**
		Societal	**1.24 [.67, 1.81]**	**1.05 [0.45, 1.65]**
	Men	Personal	**1.05 [0.47, 1.63]**	.46 [-0.11, 1.03]
		Societal	**1.06 [0.48, 1.64]**	.27 [-.30 .84]
Men	Women	Personal	.33 [-0.23, 0.89]	-0.04 [-.62, .54]
		Societal	.40 [-.16 , .96]	.33 [-.25, .91]
	Men	Personal	.03 [-.48, .54]	-.28 [-.84, .28]
		Societal	**.66 [.15, 1.17]**	**.19 [-.37, .75]**

human-related categories: *boy*, *girl*, *human eye*, *human face*, *human body*, *human ear*, *human arm*, *human board*, *human hand*; as well as *invertebrate* because there was confusion amongst pilot testers about what this word meant.

The 20 objects stereotyped about men are: *football helmet*, *football*, *cowboy hat*, *hammer*, *sports equipment*, *jet ski*, *truck*, *tie*, *golf ball*, *beer*, *skateboard*, *briefcase*, *plumbing fixture*, *tire*, *wrench*, *suit*, *missile*, *tool*, *rifle*, *shotgun*. The 4 that we consider "clothing," i.e., able to be worn, are *football helmet*, *cowboy hat*, *tie*, *suit*.

The 20 objects stereotyped about women are: *ladybug*, *doll*, *hair spray*, *lily*, *hair dryer*, *perfume*, *kitchenware*, *cat*, *wine glass*, *fashion accessory*, *necklace*, *flower*, *handbag*, *lipstick*, *bathtub*, *face powder*, *cosmetics*, *rose*, *oven*, *brassiere*. The 9 that we consider "clothing," i.e., able to be worn, are *necklace*, *face powder*, *fashion accessory*, *lipstick*, *brassiere*, *cosmetics*, *hair spray*, *handbag*, *perfume*.

The 20 neutral objects are: *pillow*, *owl*, *giraffe*, *balloon*, *jellyfish*, *stop sign*, *french fries*, *eraser*, *shower*, *orange*, *chopsticks*, *window*, *personal flotation device*, *bed*, *goldfish*, *zebra*, *raccoon*, *sea lion*, *microphone*, *popcorn*.

E INDICES FOR CHANGES IN COGNITIVE BELIEFS, ATTITUDES, AND BEHAVIORS

For most of our measurements, we simply use the measure directly (e.g., the value for competence of women) as the index to regress on. For the measurements that we do something more complicated, we describe below.

E.1 Behavior - Tags

Each participant has a set of three ordered tags associated with an image of a feminine-presenting person and a set associated with a counterfactual image of a masculine-presenting person. We convert this set of tags by scoring the presence of the object in question, e.g., "hair dryer" (along with common misspellings such as "hair drier") based on its position in the ordered list of tags. When the word is present in the first spot it is given 3 points, second spot 2 points, third spot 1 point, otherwise no points.

The index is the score of both the stereotyped-associated and neutral object on the feminine-presenting person. This is intended to capture whether the stereotype condition is able to increase the presence of the stereotype tag more than just the priming effect captured by the neutral object.

Table 5. Descriptive statistics about the captions annotated as a part of Study 2's behavior measure for the stereotype of women and hair dryer, where toothbrush serves as the control neutral object.

Condition	Mention of Hair Dryer / Mention of Toothbrush		Warmth		Competence	
Gender of Person Being Described	Women	Men	Women	Men	Women	Men
Stereotype	1.095 (0.679-1.800)	1.125 (0.679-1.800)	6.750 ± 0.386	1.107 ± 0.327	1.222 ± 0.263	0.933 ± 0.225
Control	0.913 (0.562-1.450)	0.571 (0.100-1.750)	1.300 ± 0.425	0.769 ± 0.300	0.783 ± 0.293	1.182 ± 0.229

E.2 Behavior - Captions

Here, we offer some descriptive statistics about the captions in Tbl. 5. This analysis was mostly exploratory, and we do not find any statistically significant differences. We first ran Study 2 looking at pragmatic harms on the stereotype of women and oven (with bowl as the control). In this iteration, we asked that respondents please describe each person in the image in separate sentences. However, there was too much noise in how respondents interpreted this set of instructions, such that the data became hard to interpret. Thus, in our second iteration of this study using the stereotype of women and hair dryer (with toothbrush as the control), we have two separate text entry boxes to caption each person in the image. We only present the results of this iteration in the table, as we were unable to parse anything differentiating in the first iteration.

E.3 Cognitive - Object Use

In this measurement, we have a value from -10 (mostly men) to 10 (mostly women) for both the stereotyped-associated neutral object.

The index is the summation of both values. Again, this is intended to capture whether the stereotype condition is able to change the value of its associated object more than the control condition is able to.

F PARTICIPANT INFORMATION

Tbl. 6 contains for each of our studies the time, pay, and racial distribution of each gender. Our sample size selection method is recorded on Open Science Framework and is done as follows: Study 1 we selected the number such that each COCO object was labeled by 10 participants from each gender; once we saw there was sufficient consensus from Study 1, for Study 4 we selected the number such that each OpenImages object was labeled by 5 participants from each gender; Study 2 we had three stimulus conditions across two objects, so for this between-subjects study selected the number to have 50 participants from each gender for each object-condition setting; Study 3 we had three stimulus conditions across four objects, but this is a within-subjects study so each participant sees all possible scenarios, and thus we again selected the number to have 50 participants from each gender; Study 5 we had 40 objects and as our last study ended up having the budget to have around 37.5 participants per object.

We did not use quality check questions in any of our surveys, because our pilot studies showed high quality responses. Instead, we used filters on Cloud Research to only recruit participants who have had at least 50 HITs approved, and have a HIT approval rate of 98%.

Table 6. The time, pay, and reported races of the participants for each of our five studies.

Study	Time (min)	Pay ($)	Participant Gender	American Indian or Alaska Native	Asian	Black or African American	Hispanic or Latinx	Native Hawaiian or Other Pacific Islander	White	Multi-Racial / Other	Prefer not to say	**Total**
1: COCO Objects	7	1.75	Women	0	3	5	0	0	25	6	1	40
			Men	1	4	2	2	0	30	1	0	40
2: Pragmatic Harms	10	2.50	Women	1	11	32	8	0	229	19	0	300
			Men	0	19	35	10	1	211	22	2	300
3: Experiential Harms	5	1.25	Women	0	4	7	3	1	35	5	0	50
			Men	0	4	2	3	1	35	5	0	50
4: OpenImages Objects	4	1	Women	0	5	8	0	0	42	4	1	60
			Men	0	2	6	5	1	44	2	0	60
5: Experiential Harms Extension	5	1.25	Women	0	5	15	1	0	120	7	2	150
			Men	1	9	17	6	1	107	9	0	150

G BIAS AMPLIFICATION

Bias amplification is a popular statistical metric of evaluating fairness to implicitly capture stereotypes. In this line of work, a "bias" is measured in the dataset, e.g., that women are correlated with object A, and so any amplification of this in the model's test-time predictions is considered undesirable, and likely the application of something like a stereotype. This "bias" is determined statistically, and two possible formulations come from Zhao et al. [152] (Bias Amp) and Wang and Russakovsky [138] (Directional Bias Amp). As an example, Zhao et al. [152] measures oven, wine glass, and potted plant, to all be biased towards men. From our human annotations, we find all these of these objects to be biased towards women. Thus, mitigation algorithms directed at reducing either of these formulations of bias amplification would actually likely *increase* certain types of harmful errors in an attempt to reduce overall bias amplification. This formulation also assumes that every label is biased in a way such that one direction of error is worse than another, missing that many labels can be neutral in certain respects, e.g., bowl and table.

We quantify two aspects of each bias amplification metric, which are its abilities to identify either objects as stereotypes (measured by calculating the percentage overlap between the top-n "biased" objects and n stereotypes) or the gender direction of the stereotype's alignment (measured by calculating the gender direction on the n stereotyped objects). In Tbl. 7 we can see that while both bias amplification metrics are able to approximate the gender that a stereotyped object is correlated with in the COCO dataset reasonably well, this is not true for identifying which objects are stereotypes, nor the gender alignment in the larger OpenImages dataset. Thus, attempts to reduce either metric of bias amplification are likely to inadvertently increase the number of stereotypical errors in an attempt to reduce a "bias amplification" error that may not actually be stereotypically harmful.

Table 7. Comparison of how correlated to human perceptions prior measures of bias amplification, which approximate the directions of bias that are more harmful, are.

Dataset	Metric	Pearson R	Identification of Stereotypes	Alignment of Stereotypes
COCO	Bias Amp [152]	.5722	7/13 (54%)	10/13 (77%)
	Directional Bias Amp [138]	.6507	6/13 (46%)	13/13 (100%)
OpenImages	Bias Amp	.3912	124/249 (50%)	141/249 (57%)
	Directional Bias Amp	.1502	120/249 (48%)	153/249 (61%)