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

Angelina Wang[*], Xuechunzi Bai[†], Solon Barocas[‡§], Su Lin Blodgett[§]

Abstract

As machine learning applications proliferate, we need a clear understanding of their potential for harm. In this work, we look to psychological experiences to understand what makes certain classification errors more or less harmful. Specifically, we posit that errors that reinforce stereotypes are more likely to cause harm than errors that violate stereotypes or errors that are more stereotype-neutral. 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 when aiming to reduce the *overall* number of errors. Rather than gesturing at possible harm, we actually map out concrete harms that these stereotype-reinforcing errors may lead to. Through the use of human studies on our case study of gender stereotypes in an image search engine, we not only annotate which misclassifications are stereotypes, but also which misclassifications lead to harm. We empirically find that stereotype-reinforcing errors are indeed more harmful than neutral ones, but the results become more complex when we consider stereotype-violating errors due to potential reasons like the stereotype-backlash effect. We conclude that the presence of harm alone cannot be the sole guide in dictating which errors should be prioritized in fairness mitigation, and propose a more nuanced perspective that depends on who it is that is experiencing the harm and why.

[*]Department of Computer Science, Princeton University
[†]Department of Psychology, Princeton University
[‡]Department of Information Science, Cornell University
[§]Microsoft Research

Introduction

Machine learning systems are increasingly playing a central role in everyday life. They are revolutionizing the ways in which humans communicate information, generate ideas in arts and science, and make decisions in hiring, education, medical diagnosis, and beyond. Accompanying this rapid proliferation is an increasing attention on the potential harm these systems may cause, and movements toward developing them in a fair, ethical, and inclusive way [7, 8]. The first step in reducing harm is to be specific about who exactly experiences what kind of harm and why [7, 10, 11, 22, 51, 98]. Despite the importance of human psychological experiences in thinking about harm, we know relatively little about how humans react, evaluate, and reason about machine learning outputs and their potential harms outside the context of an allocation of resources or opportunities [19]. Drawing on psychological theories of bias and harm from stereotypes, this paper empirically demonstrates the complexity of experiences of harm that can occur when machine learning models inevitably make mistakes.

Stereotypes are frequently invoked to explain why certain machine learning classifications are more harmful than others [1, 6, 9, 99]. However these explanations often stem from a researcher's own assumptions and worldviews. For example, an object recognition model that was found to amplify bias was described as harmful because the model amplifies 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 [107]. Not only is this justification then also extended to even more neutral objects like table, but these remarked-upon errors themselves may not even be genuinely harmful — while computer scientists working in the male-dominated technology space are prone to finding technology-related items like keyboards highly male-stereotyped, broader audiences do not actually share this idea, as we find in our study. Other studies on stereotypes from machine learning models have a large reliance on occupation

data from the American Bureau of Labor Statistics, e.g., WinoBias [108]. While more grounded than relying on a set of researchers' worldviews, one large limitation of this usage (in addition to 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) — and these two types can often differ in practice [14, 66]. In our work, not only do we use human participants for the annotation of stereotypes, which some prior work has also done [12, 17], but we also use human participants for the annotation of harm, which is generally just assumed to equally stem from any classification that is wrong and/or stereotypical. Our study does the necessary work of explicitly connecting stereotypes to harm, because without greater clarity grounded in psychological experiences, bias mitigation may inadvertently *increase* the number of harmful errors in a well-intentioned but ultimately misguided attempt to reduce other kinds of errors.

Our first major conceptual contribution is in differentiating between machine learning errors which are stereotype-reinforcing, stereotype-violating, or neutral. While stereotypes are cognitive beliefs in people's minds, they can have an influence on attitudes (i.e., prejudice) and behaviors (i.e., discrimination) [2, 47, 49, 57, 64]. For example, people may have *cognitive beliefs* that women are more warm but less competent, and thus *emotionally* express protective attitudes and pity for women [42]. 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 [33, 36]. Therefore, stereotypes of certain social groups can lead to changes in attitudes and behaviors which cause harm to the stereotyped group. Although the mediating role of stereotypes in machine learning harm is implicitly used by researchers, here we draw on psychological studies of stereotypes to provide a systematic assessment.

Our second major conceptual contribution comes from our definition of harm. Prior work in the machine learning fairness space has rarely been concrete about what harm actually means [10]. We distinguish between two types of harm as the most likely to result from stereotype-reinforcing errors: *pragmatic harms* involve measurable changes in someone's cognitive beliefs, attitudes, or behaviors toward the group being stereotyped, while *experiential harms* involve self-reports of negative affect. Pragmatic harms are motivated by prior research showing that, for example, people express envy and passively harm groups that are stereotyped as competent but untrustworthy (e.g., lawyers), or express contempt and actively attack groups that are stereotyped as incompetent and unreliable (e.g., homeless [21]). In the domain of machine learning, prior work has considered components of this framework and found that exposure to gender-biased image search results can influence an individual's opinion (i.e., cognitive belief) on the gender representation of that occupation as well as feelings of inclusivity [58, 69]. To examine if people experience pragmatic harms, we measure cognitive, emotional, and behavioral changes between people who experience machine learning outputs that contain stereotype-reinforcing errors compared to those who experience stereotype-neutral or stereotype-violating errors, hypothesizing that the former will result in pragmatic harm.

In contrast to pragmatic harms which focus on external impositions towards a stereotyped group, experiential harms consider the subjective feeling of harm directly experienced by the stereotyped group member [103]. Subjective experiences of emotion have long been discounted as a legitimate source of knowledge, especially when it comes from social groups such as women who are associated with it [52]. Additionally, these feelings can also influence one's own behaviors. When women are given a math exam and told that the exam is diagnostic of their own intellectual abilities, stereotypes of women as less capable of math negatively impact their performance on the exam [91]. In conceptualizing the experiential harm of machine learning errors which

may individually seem minor, we also draw a parallel to microaggressions, which are "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" [80]. Just like how a machine learning model's classification error (e.g., of an `oven` on an image of a woman) may seem small on its own, and are "easily interpretable as inadvertent errors rather than as malevolent actions," their negative effects on the target are real and should not be neglected [80]. Important in this measure of harm is who the respondent is: standpoint epistemology emphasizes the importance of the 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 [34, 38, 75, 105]. Hence, we hypothesize greater reports of experiential harm on stereotype-reinforcing errors for the stereotyped group.

Our third and more nuanced conceptual contribution is a call for an increased appreciation of the diversity of reasons that can lead to the same measured harm. While prior work often uses human judgments, they do not always incorporate the potential divergent reasons that individuals have which may lead to the same annotation. In our work, we find complexity in what people find to be stereotypical and harmful. This complements prior work studying how human annotators bring different subjective experiences in their labeling of data [23, 24, 74, 101], introducing strong associations between annotator identity and annotations [84]. In more subjective tasks such as labeling text as toxic or not, annotations are often divergent. While taking the majority vote is a common way of reconciling differences in annotations, there is a growing consensus to use a more representative system [30, 45, 54, 76]. Understanding harm faces similar nuances of incorporating differing perspectives on the same issue. However, simply incorporating representative annotations is not enough; it misses the differing personalized reasonings behind each response. For example, in a heterosexual gender normative society, some people think that men wearing skirts is harmful and

should be regulated [15, 82]. Careless incorporation of this perspective could lead to a system which treats misclassifications of skirt on men as a harmful error on par with those which reinforce sexist stereotypes. If not carefully examined, naive additions of more voices may even lead to increased future bias.

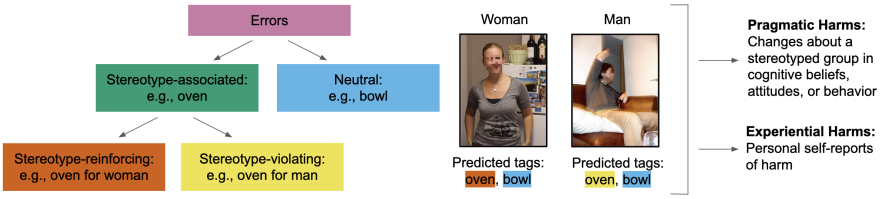

Fig. 1 Types of error and types of harm. We distinguish between types of machine learning errors which are likely to have different harms. The first split is between errors made on labels which are associated with stereotypes, e.g., ovens, and those which are more neutral, e.g., bowl. These examples of oven and bowl come from the results from Study 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.

Our primary conceptual contributions are in drawing on the psychological science of stereotypes to trouble assumptions and draw attention to three under-examined aspects of literature on bias and harms in machine learning systems: first, in differentiating between errors that are stereotype-reinforcing, stereotype-violating, or neutral in order to recognize that machine learning errors are not all equal in harm; second, in clarifying what we mean by "harm" by distinguishing and measuring two distinct types of pragmatic and experiential (Fig. 1); third, in bringing nuance to an interpretation of the reasoning behind annotations. We conduct human studies to concretely measure the presence of harms when people experience machine learning errors. As a concrete application to ground our human studies in, we consider gender stereotypes in the popular machine learning task of object recognition as used in photo search engines. We use the COCO [63] and OpenImages [62] datasets (Fig. 2), and design survey experiments with online American participants from Amazon Mechanical Turk through Cloud Research [65] to empirically study how stereotypes relating

to different classification errors result in different kinds of harms (Methods). We first ask participants whether different objects are stereotypically associated with different gender groups in order to distinguish which kinds of errors are stereotype-reinforcing, stereotype-violating, or neutral (Results - Study 1). Using those unveiled distinctions, we then expose participants to synthesized search result pages we create which contain different kinds of errors. We find little immediate evidence of pragmatic harms, but sizable evidence that stereotype-reinforcing errors are experientially harmful – a finding that is more pronounced among participants who identify as women compared to those who identify as men (Results - Study 2). In addition to stereotype-reinforcing errors (e.g., `oven` on women), we also explore stereotype-violating errors (e.g., `oven` on men), which have received scarce attention in the machine learning fairness literature. We find that while the stereotyped group (e.g., women) generally finds it more harmful for the error to reinforce rather than violate stereotypes, this is not true when it comes to clothing-related items typically associated with women (e.g., `cosmetics`, `necklaces`) being misclassified on men. Here, we see a backlash towards violations of the norms around gender presentation where men tend to find these misclassifications of, e.g., `cosmetics`, more harmful on men rather than women, calling into question the idea that it is always normatively desirable to reduce errors perceived as more harmful due to their relationship to stereotypes (Results - Study 3). Finally, our qualitative analysis reveals the plurality of why participants think certain objects are stereotypes, and why those stereotypes may be harmful or not (Results - Study 4).

All studies are approved by Princeton IRB, protocol number 14738. Studies 1 (https://osf.io/cpyn4), 2 (https://osf.io/m9akd, https://osf.io/v2w4m), and part of Study 3 (https://osf.io/xpv5j) are pre-registered on OSF, while Study 4 is more exploratory. By bringing greater clarity to different types of machine learning errors based on their relationship to a stereotype and embracing the rich psychological experiences behind them, we urge researchers and practitioners to more carefully consider

Fig. 2 COCO and OpenImages object recognition datasets. We use two popular image recognition datasets in our work to represent the application of a photo search engine. Both datasets contain annotations for perceived binary gender expression of the people in the images as well as the objects present in each image.

different kinds of classification errors, potential harms, and the relevant relationships between them. We believe that identifying psychological experiences with machine learning outputs is critical to understanding the potential harm of a system, and in turn, mitigating it. Without doing so, we may inadvertently prioritize an overall decrease of errors at the expense of increasing the number of harmful errors.

Results

The setting we explore in this work is a popular task in machine learning: object recognition. To make it concrete for our human studies, we conceive of it as used in a smart phone's photo search engine, and consider gender stereotypes. The specific machine learning classification we consider is an error of the form of a false positive: when an object is predicted to be present in an image when it is in fact not there. This causes the image with a false positive to be wrongly surfaced on an image search results page.[1] In our work, we are only concerned with the effect of the misclassification, and

[1] We note that false negatives are subsumed in this setting because enough false positives will crowd out the results page and ultimately have a similar effect as false negatives on images of the gender that does not have false positives.

do not look into why the model may have made the mistake, or what the participant thinks is the reason the model made the mistake.

Study 1: Distinguishing which machine learning errors reflect social stereotypes

To understand the social stereotypes held by American society relevant to our machine learning task, we first elicit human judgments ($N = 80$) on a popular object recognition dataset: Common Objects in Context (COCO) [63]. COCO has 80 objects and perceived binary gender expression of pictured people annotated across the images. In the survey, we ask the participants whether each object (e.g., `keyboard`, `zebra`) is stereotypically associated with men, women, or neither. As expected, not all objects reflect gender stereotypes. We note that this is already an impactful explicit finding to make as prior work in ML fairness has sometimes assumed that any difference at all between groups amplifies a stereotype [11]. Among 80 objects, 13 objects are marked as stereotypes by more than half of the participants (Figs. 3, 4). Some examples of stereotypically gendered objects are `handbag` with women, `wine glass` with women, `tie` with men, and `truck` with men. Among the remaining objects, 18 objects (e.g., `keyboard`, `carrot`, `traffic light`) are marked by zero participants as stereotypes with any gender group. If an object was marked to be a stereotype, we also asked participants whether they believed it was harmful. We show the results on the y-axis of Fig. 3 for the 13 objects marked as stereotypes. We see large variations within stereotypical objects for whether the association is perceived to be harmful. From these stereotypes, we select a few to use as errors in subsequent studies. In Study 2a the stereotype-reinforcing condition includes women and `oven` (marked to be most harmful), women and `hair dryer` (marked to be least harmful), and the associated control conditions include women and `bowl`, women and `toothbrush`. In Study 2b we also include in the stereotype-reinforcing conditions of men and `baseball glove` (marked

to be more harmful) and men and `necktie` (marked to be less harmful) with the control conditions of men and `bench` and men and `cup`.

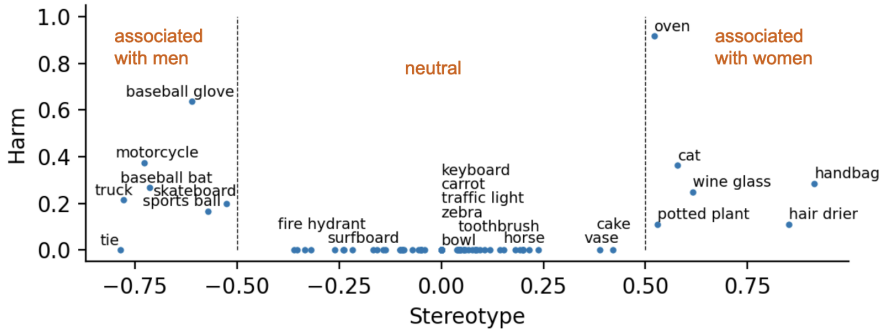

Fig. 3 Study 1 Results. Participant responses for 80 objects in COCO dataset. The x-axis indicates the percentage of participants who indicated an object is a stereotype, where negative numbers indicate it is a stereotype about men, and positive numbers about women. For objects where more than half of the respondents indicate it is a stereotype, the y-axis indicates the percentage who marked it to be harmful.

Stereotyped with Women

handbag 21/23	hair drier 17/20	wine glass 8/13	cat 11/19	potted plant 9/17	oven 12/23	cake 8/19	vase 7/18
dining table 5/21	tennis racket 3/14	cow 3/15	sink 4/20	teddy bear 4/20	horse 5/26	bird 4/22	person 4/26
sandwich 3/23	umbrella 3/25	parking meter 2/19	toaster 2/21	refrigerator 2/23	sheep 2/24	suitcase 2/25	backpack 2/26
bench 1/14	apple 1/17	snowboard 1/18	frisbee 1/18	scissors 1/18	baseball glove 1/18	bicycle 1/20	bottle 1/20
dog 1/21	book 1/21	knife 1/22	remote 1/22	cup 1/22	mouse 1/22	toothbrush 1/24	chair 1/24
orange 1/25	fork 1/26	cell phone 1/30	boat 0/14	bowl 0/18	bus 0/16	skis 0/25	toilet 0/15
stop sign 0/21	tie 0/14	keyboard 0/15	sports ball 0/21	baseball bat 0/21	spoon 0/17	carrot 0/25	donut 0/19
couch 0/22	train 0/20	kite 0/17	clock 0/24	giraffe 0/19	pizza 0/18	zebra 0/19	truck 0/18
traffic light 0/17	motorcycle 0/11	skateboard 0/19	microwave 0/12	car 0/23	bed 0/15	laptop 0/24	elephant 0/20
broccoli 0/14	bear 0/17	banana 0/26	hot dog 0/14	surfboard 0/21	fire hydrant 0/25	airplane 0/18	tv 0/21

Stereotyped with Men

tie 11/14	truck 14/18	motorcycle 8/11	baseball bat 15/21	baseball glove 11/18	sports ball 12/21	skateboard 10/19	fire hydrant 9/25
bear 6/17	snowboard 6/18	remote 7/22	car 6/23	suitcase 6/25	surfboard 5/21	sandwich 5/23	microwave 2/12
donut 3/19	person 4/26	boat 2/14	tennis racket 2/14	bicycle 2/20	bottle 2/20	tv 2/21	dog 2/21
couch 2/22	knife 2/22	sheep 2/24	backpack 2/26	bench 1/14	hot dog 1/14	vase 1/18	frisbee 1/18
pizza 1/18	cake 1/19	hair drier 1/20	teddy bear 1/20	train 1/20	elephant 1/20	toaster 1/21	stop sign 1/21
chair 1/24	toothbrush 1/24	umbrella 1/25	skis 1/25	horse 1/26	cell phone 1/30	potted plant 0/17	cup 0/22
scissors 0/18	traffic light 0/17	dining table 0/21	fork 0/26	book 0/21	orange 0/25	toilet 0/15	mouse 0/22
cat 0/19	refrigerator 0/23	spoon 0/17	bus 0/16	oven 0/23	zebra 0/19	carrot 0/25	giraffe 0/19
bed 0/15	laptop 0/24	sink 0/20	broccoli 0/14	banana 0/26	clock 0/24	bowl 0/18	wine glass 0/13
parking meter 0/19	airplane 0/18	kite 0/17	handbag 0/23	cow 0/15	keyboard 0/15	bird 0/22	apple 0/17

Fig. 4 Study 1 Object Results. Detailed participant responses for each of the 80 objects in COCO dataset. Fraction indicates number of participants asked about each object who marked it as stereotypically related to the gender group of women or men.

Study 2a: Stereotype-reinforcing errors show no pragmatic harm compared to both the stereotype-violating and neutral conditions

To test pragmatic harm in stereotype-reinforcing errors, we conduct a between-subject survey experiment, using the stereotype-violating and neutral errors as control conditions. The cover story instructs participants to look at our synthesized search result page, imagining it is their personal phone photo album, and find a picture they had taken of someone they saw with a particular object. The search result page looks different for each randomized condition. We randomly assign participants to one of the three conditions ($N = 600$): the stereotype-reinforcing condition exposes an image search result page with stereotype-reinforcing errors, e.g., false positive of `oven` on images of women; the stereotype-violating condition contains the same for stereotype-violating errors, e.g., false positive of `oven` on images of men; the stereotype-neutral condition contains neutral errors, e.g., false positive of bowl on images of women. We then measure participants' cognitive beliefs, attitudes, and behaviors to see if there are any changes because of such exposure (Methods). The behavioral measure is of particular interest, as we ask participants to undertake a realistic task they are liable to encounter by virtue of their jobs as online annotators: data labeling. We choose this 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. If stereotype-reinforcing errors have an influence on participants' cognitive representations, attitudes, and tagging or captioning behaviors, we should expect to see a statistically significant difference between participants who are exposed to search results with `oven`-women and those who are exposed to search results with `oven`-men or `bowl`-women. Contrary to what we had expected, after adjusting for multiple comparisons we do not find hypothesized statistically significant differences. We run an Ordinary-Least-Square (OLS) regression

with the control condition coded as 0 and the experimental condition coded as 1, composite scores for beliefs, attitudes, and behaviors respectively as the dependent variables. Results are shown in Fig. 5 with further details of the descriptive analysis of the captioning task in the Supplementary Material.

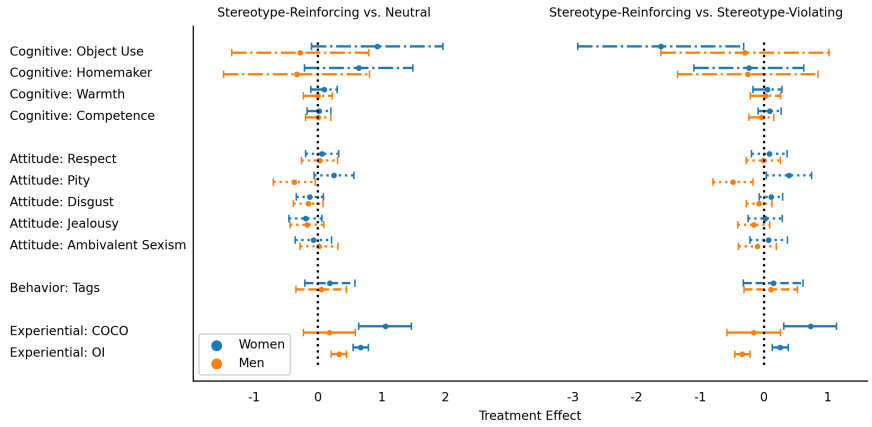

Fig. 5 Study 2, 3 Results The effect sizes and 95% confidence intervals are reported for 10 of our 11 measures of pragmatic harm (for the behavior measure of captioning, we provide a descriptive analysis), experiential harm on COCO, and experiential harm on our larger dataset of OpenImages. Deviations from zero indicate that exposure to the stereotype-reinforcing stimulus resulted in our measured harm compared to exposure to the control condition.

Study 2b: Stereotype-reinforcing errors show statistically significant experiential harm compared to both the stereotype-violating and neutral conditions

In terms of experiential harm, we design a within-subjects experiment ($N = 100$). We operationalize experiential harm by explicitly asking participants to rate how personally harmful they find different kinds of errors (which are stereotype-reinforcing, stereotype-violating, or neutral), on a scale from 0 (not at all) to 9 (extremely). This experience of error is analogous to situations where one reads in the news about the types of errors that artificial intelligence systems make [89], notices such a pattern of errors themselves, or is informed by a friend.

Comparing stereotype-reinforcing against neutral errors, an OLS regression shows participants rate stereotype-reinforcing errors to be more harmful than neutral ones ($b = .62$, 95% CI [.32, .91], $p < .001$). However, when disaggregating by gender this effect is only present among women participants (women: $b = 1.06$, 95% CI [.64, 1.47], $p < .001$; men: $b = .18$, 95% CI [-.23, .59], $p = .393$). When we use the stereotype-violating error as the control condition rather than the neutral error, we again find participants rate stereotype-reinforcing errors to be more harmful, though to a smaller degree, ($b = .28$, 95% CI [-.01, .58], $p = .062$), with once again an effect only for women participants (women: $b = .73$, 95% CI [.31, 1.14], $p = .001$; men: $b = -.16$, 95% CI [-.58, .26], $p = .453$). Results are in Fig. 5.

In short, while we find little immediate evidence of pragmatic harms, we do find the existence of experiential harms resulting from stereotype-reinforcing errors, compared to both stereotype-violating and neutral errors. However, this pattern is present only among woman participants, and not men participants.

Prior work looking at a subset of what we call pragmatic harm has found very small effects in terms of cognitive belief changes about the representation of gendered occupations [58, 69], but we do not see the effects here, potentially because we have a coarser scale of measurement. Another line of work that finds a cognitive effect takes a different approach by studying occupations (e.g., peruker, lapidary) for which there are very few preconceived notions of stereotypes [97]. In our work, we focus on the activation of existing stereotypes, rather than the induction of novel stereotypes. Overall we find that the pragmatic harms are not measurable after exposure from repeated stereotypical errors in the current survey experiment, likely due to the fact that the effects of these harms are too diffuse and long-term, impacted by all of the facets of society we encounter in our lives. However, we find consistent evidence that members of the oppressed group report a significant experiential harm in the form of negative affect on stereotypical errors made on them, consistent with the feelings of

inclusivity in gender-biased occupations [69]. This result warrants significant attention, especially because of the fact that women find errors that reinforce gender stereotypes to be more harmful even though men do not.

Study 3: Stereotype-violating errors can be perceived as harmful too, but for system-justifying reasons

In this study, we first test the generalizability of the previous findings by using a popular dataset in object recognition tasks which is much larger: OpenImages [62]. We then explore a new hypothesis about gender presentation-aligned objects, e.g., clothing, to dive deeper into our findings. OpenImages has 600 objects, annotated with perceived binary genders of people present in the image if applicable. Following the same procedure as in the COCO dataset with new online participants ($N = 120$), we find 249 of the 600 objects are marked as stereotypes by more than half participants, replicating the finding that not all objects are perceived as stereotypes (see more in Supplementary Materials). We then compile a list of 40 stereotypical objects (20 about men: e.g., `football`, `tool`; 20 about women: e.g., `doll`, `lipstick`), and 20 neutral objects (e.g., `balloon`, `goldfish`) for this study.

To test whether participants experience more experiential harm when they are exposed to stereotype-reinforcing (e.g., `skirt` on women), stereotype-violating (e.g., `skirt` on men), and neutral (e.g., `toothbrush` on women) errors, we use a similar procedure as in Study 2b. Rather than asking simply about "personal harm" as we did in Study 2b, here we draw from the Positive and Negative Affect Schedule (PANAS; [18, 102]) and provide more details by asking about if they experience harm such as feeling upset, irritated, ashamed, or distressed. We conduct a within-subjects study and ask participants ($N = 300$) to report their subjective experiences on a Likert scale from 0 to 9 for a variety of errors (see more in Methods). The analysis uses a mixed-effects regression with experimental conditions as the independent variable, a

composite score of experiential harm as the dependent variable, participants' gender as the covariate variable, and error terms clustered at the individual level.

Replicating Study 2b, we find that participants experience stereotype-reinforcing errors to be more harmful than neutral ones (b = .50, 95% CI[.42, .59], $p < .001$). Again, this pattern is more pronounced among women participants (b = .67, 95% CI [.55, .79], $p < .001$), with now a small effect among men participants (b = .33, 95% CI [.21, .45], $p < .001$). Different from Study 2b, we do not see differences in experiential harm between stereotype-reinforcing and stereotype-violating conditions (b = -.04, 95% CI [-.13, .05], p = .338). The effect is canceled out by the opposite effects for women (b = .25, 95% CI [.13, .38], $p < .001$) and men (b = -.34, 95% CI [-.46, -.22], $p < .001$) participants. In other words, while women participants feel upset, irritated, ashamed, and distressed when they see stereotype-reinforcing errors (e.g., skirt on women), men participants feel that way when they see stereotype-violating errors (e.g., skirt on men). Results are included in Fig. 5.

To better understand this finding, we conduct an exploratory analysis that digs deeper into the 40 stereotypical objects to understand why stereotype-violating errors are sometimes perceived to be more experientially harmful than stereotype-reinforcing ones. According to the gender trouble framework, costume (i.e., body and appearance) and script (i.e., behavior, traits, and preferences) are two aspects of gender performance, and reactions to androgynous or conventionally contradictory components can differ depending on which of the two it manifests in [15, 44, 72, 94]. We thus hypothesize that in our study, conventionally contradictory costume objects may be evoking a more negative reaction compared to conventionally contradictory script objects [82]. So, we add an additional independent variable we call "wearable." We determined the value of this variable by manually marking 13 of the 40 stereotypical objects to be conventionally wearable by a person. These include objects like `football helmet` and `necktie` for men, and `necklace` and `lipstick` for women. These do not include

objects like `truck` or `wine glass`. After introducing this independent variable, we find that overall participants do rate stereotype-reinforcing errors to be more harmful than stereotype-violating ones (b = .23 95% CI [.12, .34], $p < .001$), though again this is true of women participants (b = .49, 95% CI [.34, .64], $p < .001$) rather than men participants (b = -.03, 95% CI [-.18, .12], $p = .726$). We also find a higher experiential harm for the object being "wearable," regardless of if the error is stereotype-reinforcing or -violating at b = .36 95% CI [.22, .49], $p < .001$ (slightly higher for men at b = .42 95% CI [.23, .61], $p < .001$ compared to women at b = .31, 95% CI [12, .50], $p = .001$). Very interestingly, for the interaction effect of a "wearable" object with the condition type, we find that wearable stereotype-violating errors have more higher experiential harm than wearable stereotype-reinforcing errors (b=.80, 95% CI [.62, .99], $p < .001$), which is higher for men participants (b=.94, 95% CI [.67, 1.12], $p < .001$) than women participants (b=.69, 95% CI [.43, .94], $p < .001$).

More than just a result of stereotype-backlash effects [83], 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) [96] as well as transphobia, (i.e., a negative reaction to apparent incongruity between a person's perceived gender and a wearable gender presentation item) [15, 72]. 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 [67, 73]. This analysis pushes us to consider how we should think about reducing experiential harm, because it may encompass intolerances we do not wish to support.

Study 4: Plurality of stereotypes and harms with image recognition objects

Finally, we report qualitative analyses on open-ended responses from online participants' annotations on why they think certain objects are stereotypes and harmful or not. Because the objects in common image recognition tasks contain physical objects, such as `oven`, `hair dryer`, which are distinct from prior stereotype work on occupations or personality traits, we also use this exploratory analysis to better understand how objects can be associated with stereotypes. When a participant from Study 1 responds that an object is a stereotype, we follow up and ask: "Please describe in 1-2 sentences a) why you marked the above as a stereotype, and b) why you found it to be harmful or not."

One of the authors coded the responses for why an object is a stereotype 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 the Supplementary Material. It is interesting to note that an object's association to a stereotype is frequently mediated by its connection to a role or trait, which are the more common sites of inquiry when it comes to stereotypes. 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*."

When asked why a stereotype was harmful or 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 the Supplementary Material. 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.

Building on Study 1's finding that participants do not even all agree on whether an object is a stereotype or not (and if it is, whether it is harmful), this analysis further shows that even when participants are in agreement that an object is a stereotype, they are not necessarily in agreement about why. The same holds true for whether a stereotype is harmful. One potential implication of this is considering whether different reasonings should lead to different bias mitigation. For example, if the reason an object is a stereotype is descriptive, then mitigation should aim to change the cognitive representations of people. To change these descriptive statistics themselves, while we can also work to alter the model outputs, we should also work to change society, the burden of which falls on a much larger group than just machine learning practitioners, e.g., policymakers. On the other hand, if particular stereotypes are deemed harmful because they are proscriptive and seem to restrict people from various avenues, we can consider ways to break free of gender norms.

Discussion

Taking stock of our studies, we have three primary findings regarding our three conceptual contributions: a meaningful distinction between machine learning errors is whether they are stereotype-reinforcing, stereotype-violating, or neutral; harm can be pragmatic or experiential; and harm annotations can be measured for a diversity of reasons that need to be critically engaged with. First, we find that harm is different depending on a machine learning error's relation to a stereotype. Second, while stereotype-reinforcing errors do lead to more experiential harm for women, they do not lead to 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 [37, 39, 53] In contrast, that stereotype-reinforcing errors are consistently found to be more experientially harmful for women supports that certain errors 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.[2] Third, stereotype-violating errors are also found to be experientially harmful, especially in cases where the error is with respect to a wearable item, which is likely associated with gender presentation. This effect is stronger for participants who identify as men than those who identify as women. 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 [38] and standpoint epistemology [34, 75, 105]. If we understand the negative reaction to misclassifications of stereotypically feminine clothing items on

[2]Related but different is work considering normative arguments for labels on which classifications should not performed at all: lewd labels [20, 77, 106], non-imageable properties (e.g., vegetarian) [106], emotions [4] gender [59, 85, 86], and physical attractiveness [5].

men as a manifestation of precarious manhood [96] or transphobia [15], 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 why, and prioritizing the experiential harms of certain marginalized groups.

In the next few sections, we describe some implications and limitations of our findings.

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 how to leverage human-driven insights in determining how we train and evaluate our models [13]. Fairness evaluations which incorporate stereotypes have tended to be restricted to occupations or traits, but we have expanded this idea by showing that other labels such as objects can also give rise to such harms. Additionally, most prior work has only considered 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. And normatively, there can be questions of whether stereotype-violating errors may even play a role in reducing stereotypical associations by counteracting them. This 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 [31] 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 [61] (e.g., 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.

Understanding whose levels of harms we should prioritize, and why, will come from stronger understandings of the psychological basis and reasoning of different harms. Our finding from Study 4 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 [23, 24, 30, 54, 74, 101], 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 were able to gain rich information about the associations, 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.

In thinking about different experiences of harm, stereotypes are often defined to be shared cultural knowledge [57], yet perceptions and justifications of their harm can

be different depending on an individual's place within a society [60, 78]. For example, some people may find misclassifications which associate women with oven to be harmful because it reinforces gender stereotypes about women being homemakers. Meanwhile, others may find misclassifications which associate men with skirts to be harmful because it violates gender stereotypes about how they believe men should present themselves [83]. While the motivation of the former is discomfort over antiquated stereotypes, that of the latter is about upholding the norms of a patriarchal society. It is an open challenge to machine learning researchers on whether and how to incorporate participant responses that stem from divergent motivations.

We can also 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. Prior work brought to light that word embeddings mirror stereotypes in our society [12, 16], such as about occupations and attributes from the Implicit Association Test [46]. Since then, most follow-up work in this space seeks to remove nearly all gendered associations in text, conflating each such association with harmful "bias." Again we see a similar pattern to the logical fallacy of the harm of one type of error, e.g., a correlation of some set of stereotypical occupations to gender, extending to *all* errors. The nuance is lost when gendered associations of all kinds in word embeddings are equated to stereotypes, and most notions of gender are targeted to be removed from the embeddings. To put this into perspective, in the large body of literature that has followed the discovery of gender biases in the embedding space [25, 26, 55, 56, 68, 79, 88, 109], 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 what 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 [100]) it surely seems that more thinking should be done on the different implications of debiasing stereotypes as opposed to debiasing more neutral concepts.

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 [28] 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 [40], 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 [3, 50], thus allowing room for positionality, explanation, and critical reflection.

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-reinforcing errors compared to stereotype-violating and neutral errors generalize beyond gender stereotypes (e.g., to racial stereotypes) this is unlikely to remain true for the gender- and culture-specific findings about wearable items in stereotype-violating errors, 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 and 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. 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" [41]. 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 [70].

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 [92, 93] or self-stereotyping [90]. 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 [46]. 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 [95].

Methods

Analysis

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

Table 1 The time, pay, and reported races of the participants for each of our five studies. The full column names of races from left to right are: American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latinx, Native Hawaiian or Other Pacific Islander, White, Multi-Racial / Other, and Prefer not to say.

Study	Time (min)	Pay ($)	Gender	AI/AN	Asian	Black	H/L	NHOPI	White	MR/O	PNTS	**Total**
1 and 4	7	1.75	Women	0	3	5	0	0	25	6	1	40
			Men	1	4	2	2	0	30	1	0	40
2a	10	2.50	Women	1	11	32	8	0	229	19	0	300
			Men	0	19	35	10	1	211	22	2	300
2b	5	1.25	Women	0	4	7	3	1	35	5	0	50
			Men	0	4	2	3	1	35	5	0	50
3	4	1	Women	0	5	8	0	0	42	4	1	60
(Labeling)			Men	0	2	6	5	1	44	2	0	60
3	5	1.25	Women	0	5	15	1	0	120	7	2	150
(Harms)			Men	1	9	17	6	1	107	9	0	150

the coefficients from our regression analyses, which represent the effect size of that independent variable.

Participants

While men and women generally tend to hold the same gender stereotypes [32, 48, 66, 104], we still collect equal numbers of participants who identify as men and women, and use this variable as a covariate throughout. 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 gender stereotype scope of our current work, 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 [57].

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%.

Studies 1, 3: Distinguishing Errors by Stereotype

When asking about which machine learning errors are stereotypes, we make sure to ask participants about their perception of stereotypes held by Americans, rather than for their personal beliefs [27].

Study 2a: Measuring Pragmatic Harm

We conduct a between-subjects survey experiment on participants who are exposed to an image search result page that contain one of three types of errors: stereotype-reinforcing, stereotype-violating, or neutral (Fig. 6).[3] To have the participants engage with these results we ask them to describe it in 3-4 sentences. Next, we ask them the behavior questions, then re-expose them to the stimulus before asking them the cognitive belief and attitude questions. We analyze changes in cognitive beliefs, attitudes, and behaviors as pragamatic harms resulting from stereotype-reinforcing errors compared to the two other conditions as controls. In this section when describing our method, we will use as examples `oven` and women for the stereotype-reinforcing error, `oven` and men for the stereotype-violating error, and `bowl` and women for the neutral one. Each question we ask is carefully grounded in the social psychology literature.

The stimuli take the form of an image search result and are pictured in Fig. 6 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-reinforcing 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 stereotype-violating

[3] The people pictured in our search results pages are predominantly White, which is the majority group in the dataset we employ.

condition contains the same search query, but the mistakes are replaced with false positive images that all contain men. The neutral condition contains all of the exact same images as the stereotype-reinforcing condition, 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 [29] 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."

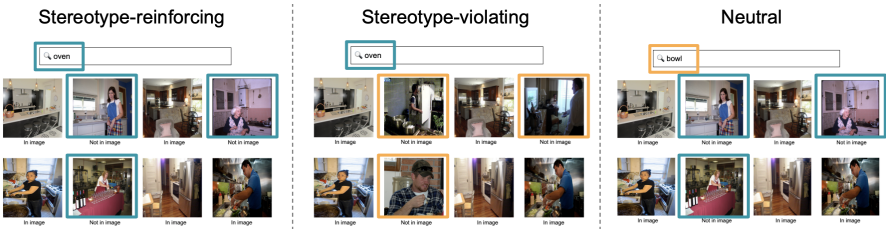

Fig. 6 Study 2 Stimuli. Our three different stimuli are shown for the conditions: stereotype-reinforcing, stereotype-violating, and neutral. They are all image search results containing minimal changes from each other, each of which indicates whether the search query is pictured in the image, i.e., if the image search retrieval was correct or not. The teal and orange squares indicate that the only difference between the stimuli, 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.

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 [36] 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 [21, 35, 87]. The second asks about sexist attitudes via a shortened scale focused on benevolent sexism [42, 43, 81].[4] Finally, for *behavioral* measures, we ask participants to undertake a realistic task they are liable to encounter which can cause harm: data labeling [71]. 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. 7). 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. 7 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] We ask questions from the Ambivalent Sexism Inventory [42] about benevolent sexism, as opposed to hostile sexism, because the latter is believed to suffer heavily from social desirability bias.

Dependent Variables

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

Behavior - Tags. Each participant produces 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 dependent variable is the score of both the stereotypical and neutral object on the feminine-presenting person. This is intended to capture whether the stereotype-reinforcing condition is able to increase the presence of the stereotype tag more than just the priming effect captured by the neutral object.

Behavior - Captions. We offer some descriptive statistics about the captions in the Supplementary Material. This analysis was mostly exploratory, and we do not find any statistically significant differences. We first ran Study 2a 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.

Cognitive - Object Use. In this measurement, we have a value from -10 (mostly men) to 10 (mostly women) for both the stereotypical and neutral object. The

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

Study 2b, 3: Measuring Experimental Harm

In Study 2b, in addition to personal discomfort, we also ask about societal harm. This way, 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 the Supplementary Material.

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] Abbasi M, Friedler SA, Scheidegger C, et al (2019) Fairness in representation: quantifying stereotyping as a representational harm. Siam International Conference on Data Mining

[2] Allport GW, Clark K, Pettigrew T (1954) The nature of prejudice

[3] Argyle LP, Busby EC, Fulda N, et al (2023) Out of one, many: Using language models to simulate human samples. Political Analysis

[4] Bard J (2020) Developing a legal framework for regulating emotion ai. University of Florida Levin College of Law Research Paper

[5] Barlas P, Kyriakou K, Kleanthous S, et al (2019) Social b(eye)as: Human and machine descriptions of people images. Proceedings of the International AAAI Conference on Web and Social Media

[6] Barlas P, Kyriakou K, Guest O, et al (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)

[7] Barocas S, Crawford K, Shapiro A, et al (2017) The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In: Proceedings of SIGCIS, Philadelphia, PA

[8] Barocas S, Hardt M, Narayanan A (2019) Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org, http://www.fairmlbook.org

[9] Bhaskaran J, Bhallamudi I (2019) Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. Proceedings of the First Workshop on Gender Bias in Natural Language Processing

[10] Blodgett SL, Barocas S, III HD, et al (2020) Language (technology) is power: A critical survey of "bias" in nlp. Association for Computational Linguistics (ACL)

[11] Blodgett SL, Lopez G, Olteanu A, et al (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

[12] Bolukbasi T, Chang KW, Zou J, et al (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Conference on Neural Information Processing Systems (NeurIPS)

[13] Boykin CM, Dasch ST, Jr. VR, et al (2021) Opportunities for a more interdisciplinary approach to measuring perceptions of fairness in machine learning. Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)

[14] Burgess D, Borgida E (1999) Who women are, who women should be: descriptive and prescriptive gender stereotyping in sex discrimination. Psychology, Public Policy, and Law 5

[15] Butler J (1990) Gender trouble: Feminism and the subversion of identity. Routledge

[16] Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science

[17] Cao Y, Sotnikova A, Hal Daumé III RR, et al (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

[18] Crawford JR, Henry JD (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

[19] Crawford K (2017) The trouble with bias. Conference on Neural Information Processing Systems Keynote

[20] Crawford K, Paglen T (2019) Excavating AI: the politics of images in machine learning training sets. Excavating AI

[21] Cuddy AJC, Fiske ST, Glick P (2007) The BIAS map: behaviors from intergroup affect and stereotypes. Journal of Personality and Social Psychology 92

[22] Danks D, London AJ (2017) Algorithmic bias in autonomous systems. International Joint Conference on Artificial Intelligence (IJCAI)

[23] Davani AM, Díaz M, Prabhakaran V (2022) Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics

[24] Denton E, Díaz M, Kivlichan I, et al (2021) Whose ground truth? accounting for individual and collective identities underlying dataset annotation. NeurIPS 2021 Workshop on Data-Centric AI

[25] Dev S, Phillips J (2019) Attenuating bias in word vectors. International Conference on Artificial Intelligence and Statistics

[26] Dev S, Li T, Phillips J, et al (2020) On measuring and mitigating biased inferences of word embeddings. AAAI Technical Track: Natural Language Processing

[27] Devine PG, Elliot AJ (1995) Are racial stereotypes really fading? the princeton trilogy revisited. Personality and Social Psychology Bulletin 21

[28] Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT

[29] Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on

Learning Representations (ICLR)

[30] Dumitrache A, Aroyo L, Welty C (2018) Capturing ambiguity in crowdsourcing frame disambiguation. AAAI Conference on Human Computation and Crowdsourcing (HCOMP)

[31] Dwork C, Hardt M, Pitassi T, et al (2012) Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference

[32] Eagly AH, Nater C, Miller DI, et al (2020) Gender stereotypes have changed: A cross-temporal meta-analysis of u.s. public opinion polls from 1946 to 2018. American Psychologist

[33] Ellemers N, et al (2018) Gender stereotypes. Annual review of psychology 69:275–298

[34] Fatima S (2020) I know what happened to me: The epistemic harms of microaggression. Microaggressions and Philosophy

[35] Fiske ST, Cuddy AJC, Glick P (2002) Emotions up and down: Intergroup emotions result from status and competition. Prejudice to Intergroup Emotions: Differentiated Reactions to Social Groups

[36] Fiske ST, Cuddy AJC, Glick P, et al (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

[37] Ford TE (1997) Effects of stereotypical television portrayals of african-americans on person perception. Social Psychology Quarterly

[38] Fricker M (2009) Epistemic injustice: Power and the ethics of knowing. Oxford University Press

[39] Fujioka Y (1999) Television portrayals and african-american stereotypes: Examination of television effects when direct contact is lacking. Journalism and Mass Communication Quarterly

[40] Garg N, Schiebinger L, Jurafsky D, et al (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)

[41] Ghavami N, Peplau LA (2012) An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. Psychology of Women Quarterly 37

[42] Glick P, Fiske ST (1996) The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. Journal of Personality and Social Psychology 70

[43] Glick P, Whitehead J (2010) Hostility toward men and the perceived stability of male dominance. Social Psychology 41

[44] Goffman E (1959) The presentation of self in everyday life. Doubleday

[45] Gordon ML, Lam MS, Park JS, et al (2022) Jury learning: Integrating dissenting voices into machine learning models. Conference on Human Factors in Computing Systems (CHI)

[46] Greenwald AG, McGhee DE, Schwartz JLK (1998) Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology

[47] Hamilton DL, Sherman JW (2014) Stereotypes. In: Handbook of social cognition. Psychology Press, p 17–84

[48] Hentschel T, Heilman ME, Peus CV (2019) The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others

and themselves. Frontiers in Psychology

[49] Hilton JL, Von Hippel W (1996) Stereotypes. Annual review of psychology 47(1):237–271

[50] Hämäläinen P, Tavast M, Kunnari A (2023) Evaluating large language models in generating synthetic hci research data: a case study. Conference on Human Factors in Computing Systems (CHI)

[51] Jacobs AZ, Wallach H (2021) Measurement and fairness. Conference on Fairness, Accountability and Transparency (FAccT)

[52] Jaggar AM (1989) Love and knowledge: Emotion in feminist epistemology. Inquiry

[53] Jennings-Walstedt J, Geis FL, Brown V (1980) Influence of television commercials on women's self-confidence and independent judgment. Journal of Personality and Social Psychology

[54] Kairam S, Heer J (2016) Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)

[55] Kaneko M, Bollegala D (2019) Gender-preserving debiasing for pre-trained word embeddings. Annual Conference of the Association for Computational Linguistics (ACL)

[56] Karve S, Ungar L, Sedoc J (2019) Conceptor debiasing of word representations evaluated on weat. arXiv:190605993

[57] Katz D, Braly K (1933) Racial stereotypes of one hundred college students. The Journal of Abnormal and Social Psychology 28

[58] Kay M, Matuszek C, Munson SA (2015) Unequal representation and gender stereotypes in image search results for occupations. Conference on Human Factors in Computing Systems (CHI)

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

[60] Kraus MW, Piff PK, Mendoza-Denton R, et al (2012) Social class, solipsism, and contextualism: how the rich are different from the poor. Psychological review 119(3):546

[61] Kukar M, Kononenko I (1998) Cost-sensitive learning with neural networks. European Conference on Artificial Intelligence

[62] Kuznetsova A, Rom H, Alldrin N, et al (2020) The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision (IJCV)

[63] Lin TY, Maire M, Belongie S, et al (2014) Microsoft COCO: Common objects in context. European Conference on Computer Vision (ECCV)

[64] Lippmann W (1922) Public opinion.

[65] Litman L, Robinson J, Abberbock T (2017) Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behavior Research Methods 42

[66] López-Sáez M, Lisbona A (2014) Descriptive and prescriptive features of gender stereotyping. relationships among its components. International Journal of Social Psychology 24

[67] Makwana AP, Dhont K, keersmaecker JD, et al (2018) The motivated cognitive basis of transphobia: The roles of right-wing ideologies and gender role beliefs. Sex Roles 79

[68] Manzini T, Lim YC, Tsvetkov Y, et al (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)

[69] Metaxa D, Gan MA, Goh S, et al (2021) An image of society: Gender and racial representation and impact in image search results for occupations. ACM Conference on Human-Computer Interaction (CSCW)

[70] Mill JS (1859) On liberty. Longman, Roberts, Green Co

[71] van Miltenburg E (2016) Stereotyping and bias in the flickr30k dataset. Proceedings of the Workshop on Multimodal Corpora

[72] Morgenroth T, Ryan MK (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

[73] Nagoshi CT, Cloud JR, Lindley LM, et al (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

[74] Noble JA (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)

[75] O'Dowd O (2018) Microaggressions: A kantian account. Ethical Theory and Moral Practice 21

[76] Peterson JC, Battleday RM, Griffiths TL, et al (2019) Human uncertainty makes classification more robust. International Conference on Computer Vision (ICCV)

[77] Prabhu VU, Birhane A (2020) Large image datasets: A pyrrhic win for computer vision? arXiv:200616923

[78] Pratto F, Sidanius J, Stallworth LM, et al (1994) Social dominance orientation: A personality variable predicting social and political attitudes. Journal of personality and social psychology 67(4):741

[79] Ravfogel S, Elazar Y, Gonen H, et al (2020) Null it out: Guarding protected attributes by iterative nullspace projection. Annual Conference of the Association for Computational Linguistics (ACL)

[80] Rini R (2020) The ethics of microaggression. Routledge Taylr & Francis Group

[81] Rollero C, Glick P, Tartaglia S (2014) Psychometric properties of short versions of the ambivalent sexism inventory and ambivalence toward men inventory. TPM-Testing, Psychometrics, Methodology in Applied Psychology 21

[82] Rudman LA, Moss-Racusin CA, Glick P, et al (2012) Reactions to vanguards: Advances in backlash theory. Advances in experimental social psychology 45

[83] Rudman LA, Moss-Racusin CA, Phelan JE, et al (2012) Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. Journal of Experimental Social Psychology 48

[84] Sap M, Swayamdipta S, Vianna L, et al (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

[85] Scheuerman MK, Paul JM, Brubaker JR (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)

[86] Scheuerman MK, Wade K, Lustig C, et al (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)

[87] Seger CR, Banerji I, Park SH, et al (2017) Specific emotions as mediators of the effect of intergroup contact on prejudice: findings across multiple participant and target groups. Cognition and Emotion 31

[88] Shin S, Song K, Jang J, et al (2020) Neutralizing gender bias in word embedding with latent disentanglement and counterfactual generation. Findings of EMNLP

[89] Simonite T (2018) When It Comes to Gorillas, Google Photos Remains Blind. Wired, January

[90] Sinclair S, Hardin CD, Lowery BS (2006) Self-stereotyping in the context of multiple social identities. Journal of Personality and Social Psychology 90

[91] Spencer SJ, Steele CM, Quinn DM (1999) Stereotype threat and women's math performance. Journal of Experimental Social Psychology 35

[92] Spencer SJ, Logel C, Davies PG (2015) Stereotype threat. Annual Review of Psychology 67

[93] Steele CM, Aronson J (1995) Stereotype threat and the intellectual test performance of african americans. Journal of Personality and Social Psychology 69

[94] Stern C, Rule NO (2017) Physical androgyny and categorization difficulty shape political conservatives' attitudes toward transgender people. Social Psychological and Personality Science

[95] Suresh H, Movva R, Dogan AL, et al (2022) Towards intersectional feminist and participatory ml: A case study in supporting feminicide counterdata collection. ACM Conference on Fairness, Accountability, and Transparency (FAccT)

[96] Vandello JA, Bosson JK, Cohen D, et al (2008) Precarious manhood. Journal of Personality and Social Psychology

[97] Vlasceanu M, Amodio DM (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

[98] Wang A, Barocas S, Laird K, et al (2022) Measuring representational harms in image captioning. ACM Conference on Fairness, Accountability, and Transparency (FAccT)

[99] Wang A, Liu A, Zhang R, et al (2022) REVISE: A tool for measuring and mitigating bias in visual datasets. International Journal of Computer Vision (IJCV)

[100] Wang C, Wang K, Bian A, et al (2021) User acceptance of gender stereotypes in automated career recommendations. arXiv:210607112

[101] Waseem Z (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

[102] Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the panas scales. Journal of Personality and Social Psychology 54

[103] Wenzel K, Devireddy N, Davidson C, et al (2023) Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition. Conference on Human Factors in Computing Systems (CHI)

[104] Williams JE, Best DL (1977) Sex stereotypes and trait favorability on the adjective check list. Educational and Psychological Measurement 37

[105] Wylie A (2003) Why standpoint matters. Science and Other Cultures: Issues in Philosophies of Science and Technology

[106] Yang K, Qinami K, Fei-Fei L, et al (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)

[107] Zhao J, Wang T, Yatskar M, et al (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)

[108] Zhao J, Wang T, Yatskar M, et al (2018) Gender bias in coreference resolution: Evaluation and debiasing methods. North American Chapter of the Association for Computational Linguistics

[109] Zhao J, Zhou Y, Li Z, et al (2018) Learning gender-neutral word embeddings. Empirical Methods in Natural Language Processing (EMNLP)