# Measuring machine learning harms from stereotypes: 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 an understanding of their potential for harm. However, current fairness metrics are rarely grounded in human psychological experiences of harm. Drawing on the social psychology of stereotypes, we use a case study of gender stereotypes in image search to examine how people react to machine learning errors. First, we use survey studies to show that not all machine learning errors reflect stereotypes nor are equally harmful. Then, in experimental studies we randomly expose participants to stereotype-reinforcing, -violating, and -neutral machine learning errors. We find stereotype-reinforcing errors induce more experientially (i.e., subjectively) harmful experiences, while having minimal changes to cognitive beliefs, attitudes, or behaviors. This experiential harm impacts women more than men. However, certain stereotype-violating errors are more experientially harmful for men, potentially due to perceived threats to masculinity. We conclude that harm cannot be the sole guide in fairness mitigation, and propose a nuanced perspective depending on who is experiencing what harm and why.

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

## Introduction

Over the past decade, researchers have demonstrated that machine learning models run the risk of learning stereotypical associations. For example, natural language processing (NLP) models have been shown to associate women with homemakers and men with programmers [9], while computer vision models have been shown to associate women with shopping and men with driving [102]. These associations can cause machine learning models to make systematic errors, mistakenly reporting, for instance, that female doctors are nurses and that male nurses are doctors [81]. A rich literature has developed offering many more such examples, spurring calls to address the risk that machine learning models might make mistakes that perpetuate harmful stereotypes.

Unfortunately, portions of this literature have suffered from three main limitations that impede effective intervention. First, while some important prior work has asked crowd workers to annotate when errors invoke stereotypes or drawn on pre-existing inventories of stereotype from the psychology literature [7, 9, 12, 13, 84, 91], machine learning researchers often rely on their own moral intuitions to determine which categories of associations to investigate (e.g., associations between gender and occupation) and to judge which specific associations (e.g., the association between women and nursing) are stereotypes. As a result, certain categories of associations that broader populations might perceive as stereotypical have not been investigated and many discovered associations within these categories are assumed to be stereotypical, even if broader populations would not perceive them to be so. For example, researchers identified an object recognition model as reproducing stereotypes because it 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 [102]. However, these intuitions might not always be shared by the broader population; indeed, as we'll show later in this paper, some of these associations are not commonly seen as invoking stereotypes among a sample of people in the United States.

Second, prior work has tended to ignore whether errors reinforce or violate stereotypes, treating each type of error as equally harmful. Researchers have even occasionally treated all spurious correlations between objects and specific social groups as harmful, even if such correlations may not be perceived as reinforcing or violating any stereotype. Blodgett et al [8] has documented many instances of this, for example how remarks following a statement about "Norwegian salmon" are nonsensically used to assess stereotypes. As a result, it remains unclear the degree to which the harmfulness of errors depends on whether and how these errors invoke stereotypes.

Third, while stereotypes are frequently invoked to explain why some machine learning errors are more harmful than others [1, 4, 6, 93], little has been done to establish the actual impact that these

stereotypes have on people in practice. Rather than measuring the harms that stereotypes bring about and identifying the particular mechanisms by which they do so, much of the literature simply presumes that stereotypes have negative impacts on the people so stereotyped or on society more generally. While some prior work has found that exposure to gender-biased image search results can lead both to more biased estimations of the representation of different gender groups in certain occupations and to a decreased sense of belonging [51, 59], there remains a paucity of empirical work that seeks to measure the psychological and practical effects of exposure to such errors and to characterize the nature of these harms.

In this paper, we draw on the psychological literature on social stereotypes to try to overcome each of these limitations, performing a series of empirical studies in which human subjects are exposed to machine learning errors. As a concrete application in which to ground these studies, we focus on gender stereotypes in the machine learning task of object recognition, which is now a common feature in photo management software like Apple Photos and Google Photos.

First, we investigate which associations people view as stereotypical. Recognizing from the psychological literature that the space of stereotypical associations is potentially much broader than what has been examined in the machine learning literature [27, 30, 43, 65], we make no a priori decisions about the categories of association worth investigating. We expand our scope of analysis beyond commonly focused-on categories—for example, occupations or activities—to include all objects that might appear in an image. We do so via two studies. In the first (Study 1), we present participants with images from popular computer vision datasets that include a range of objects (COCO [54] and OpenImages [53] datasets (Fig. 2)), and ask whether these objects are stereotypically associated with different gender groups. We further expand on this in a second study (Study 2), in which we conduct qualitative analysis on participants' open-ended responses to the questions in Study 1 to better understand why certain objects are seen as stereotypes. We find that participants do not all agree on whether particular objects reflect stereotypical associations, and that even when objects are seen as reflecting stereotypes, the reasons that participants provide are varied. For example, the association between women and cats is variously explained by "cat lady," "women are called *kitten*," and "women are called *cougars*." We additionally ask participants whether and why they find these stereotypical associations to be harmful. We find that while many participants describe stereotypes as self-evidently harmful, others differentiate reasons such as whether the error invokes a stereotypical association perceived as prescriptive (i.e., how members of certain gender groups should behave) or proscriptive (i.e., how they should not behave), providing a richer set of psychological reasons [72, 76].

Then, to overcome the second and third limitations of imprecision about which kind of stereotypical errors may cause what kind of harm, we attempt to causally assess the effect that exposure to these stereotypes have on people in practice. To do so, we draw on psychological theories of

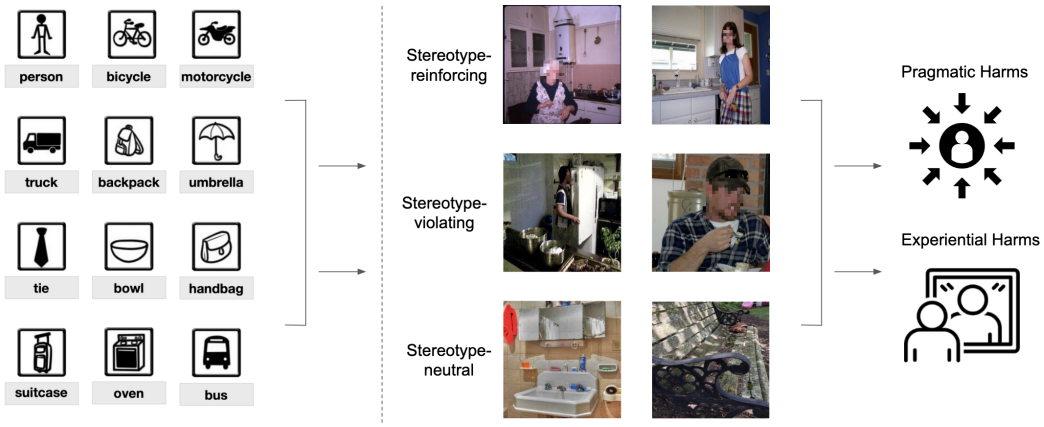


**Fig. 1 Summary of our studies.** The left side represents studies 1 and 2, where we ask human participants to mark which of the relevant objects in our application are stereotypically associated with which gender groups, as well as to qualitatively explain why that is and why it is harmful or not. The right side represents studies 3 and 4 where we randomly expose participants to machine learning errors which are stereotype-reinforcing, stereotype-violating, or stereotype-neutral—as determined by the annotations from study 1. Then, we measure two forms of harm we introduce: pragmatic (measurable changes in someone's cognitive beliefs, attitudes, or behaviors toward the group being stereotyped) and experiential (self-reports of negative affect). The images shown are examples of misclassifications of `oven` where stereotype-reinforcing errors are when it is falsely predicted on a woman, stereotype-violating when on a man, and stereotype-neutral when no stereotypes are invoked.

stereotypes to conceptualize harm as well as the relationship between stereotype and harm. We then investigate concretely the degree to which stereotypes cause such harm (Fig. 1).

The psychology literature describes stereotypes as cognitive beliefs in people's minds, which can have an influence on attitudes (i.e., prejudice) and behaviors (i.e., discrimination) [2, 41, 43, 49, 55]. For example, people may have *cognitive beliefs* that women are more warm but less competent than men, and thus *express* protective attitudes and pity for women [15, 36]. 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 [27, 30]. Members of stereotyped social groups may also experience changes in their own beliefs and attitudes, causing them to behave differently; for example, 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 [85]. Therefore, stereotypes of certain social groups can prompt shifts in attitudes and behaviors that can ultimately harm the stereotyped group.

Beyond these changes in beliefs, attitudes, and behaviors, members of the stereotyped social group may also feel disrespected, demeaned, or otherwise discounted. Such experiences can be thought of as dignitary harms that bring about negative affect in members of the group so stereotyped [50]. While dignitary harms are frequently treated as less important than the concrete effects of changes in beliefs, attitudes, and behaviors, they may impose a substantial emotional toll, akin to that of microaggressions [74]. We therefore introduce a distinction between *pragmatic harms*, which involve measurable changes in someone's cognitive beliefs, attitudes, or behaviors toward the group being stereotyped, and *experiential harms*, which involve self-reports of negative affect (Fig. 1). We additionally differentiate between errors that are stereotype-reinforcing, stereotype-violating, or stereotype-neutral.

Errors that invoke stereotypes may do so in different ways and may therefore have different effects. With these in place, we set out to measure the degree to which different kinds of errors bring about pragmatic and experiential harm. Concretely, we randomly assign participants to synthesized search result pages which contain different kinds of errors and we investigate three related hypotheses. First, we hypothesize that errors that reinforce social stereotypes will be perceived as more harmful than those that do not. Second, we hypothesize that stereotype-reinforcing errors will result in pragmatic harm, while stereotype-neutral or stereotype-violating errors will not. Third, we hypothesize greater reports of experiential harm on stereotype-reinforcing errors for the stereotyped group.

To examine if people experience pragmatic harms, we measure cognitive, attitudinal, and behavioral changes between people who experience machine learning outputs containing stereotypes, varying whether the errors reinforce or violate stereotypes. To measure experiential harms, we ask people to self-report negative affect when exposed to machine learning outputs containing stereotypes, again varying whether these errors reinforce or violate stereotypes.

We find little immediate causal effect 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 (Study 3). We find that while the stereotyped group (e.g., women) generally finds it more experientially 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. This last observation calls into question the idea that it is always normatively desirable to reduce errors perceived as more harmful due to their relationship to stereotypes (Study 4).

All studies are approved by our institution IRB, protocol number 14738. Studies 1 (https://osf.io/cpyn4), 3 (https://osf.io/m9akd, https://osf.io/v2w4m), and part of Study 4 (https://osf.io/xpv5j) are pre-registered on OSF, while Study 2 is 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 different kinds of machine learning errors, potential harms, and the relevant relationships between them. Investigating the psychological experiences that people have when encountering machine learning errors is critical to understanding the potential harm of a system, and in turn, mitigating it.

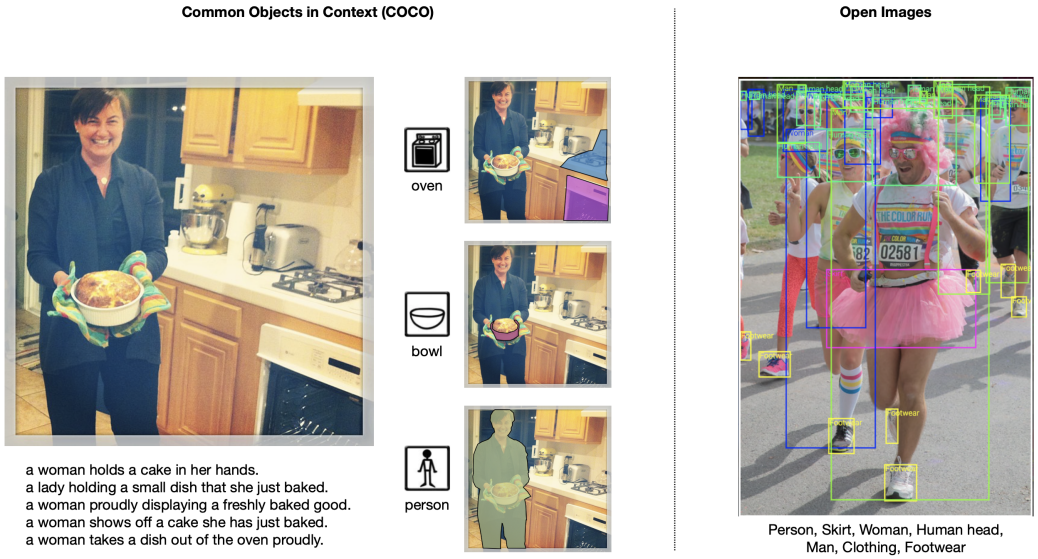


**Fig. 2 COCO and Open Images object recognition datasets.** We use two commonly used image recognition datasets to represent the application of a photo search engine. Both datasets contain annotations for perceived binary gender expression of the people in the images and the objects present in each image. The left panel shows one example figure from COCO annotated with objects such as `oven` and `bowl`. The right panel shows one example figure from Open Images annotated with objects such as `person` and `skirt`.

# Results

We explore a popular task in machine learning known as object recognition (i.e., classifying the objects present in an image). To make it concrete for our human studies, we use it in the context of a smart phone's photo search engine, and examine gender stereotypes. Specifically, we consider one type of machine learning error called 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 not why the model may have made the mistake, or what the participant thinks is the reason the model made the mistake. Unlike prior work auditing search engines [51, 59, 69, 70, 90], our sole focus is on tracing the concrete effects that search results can have.

## 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 Common Objects in Context (COCO) [54]. COCO has 80 objects and perceived binary gender expression of pictured people annotated across the images [101]. In the study, 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. This

[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.

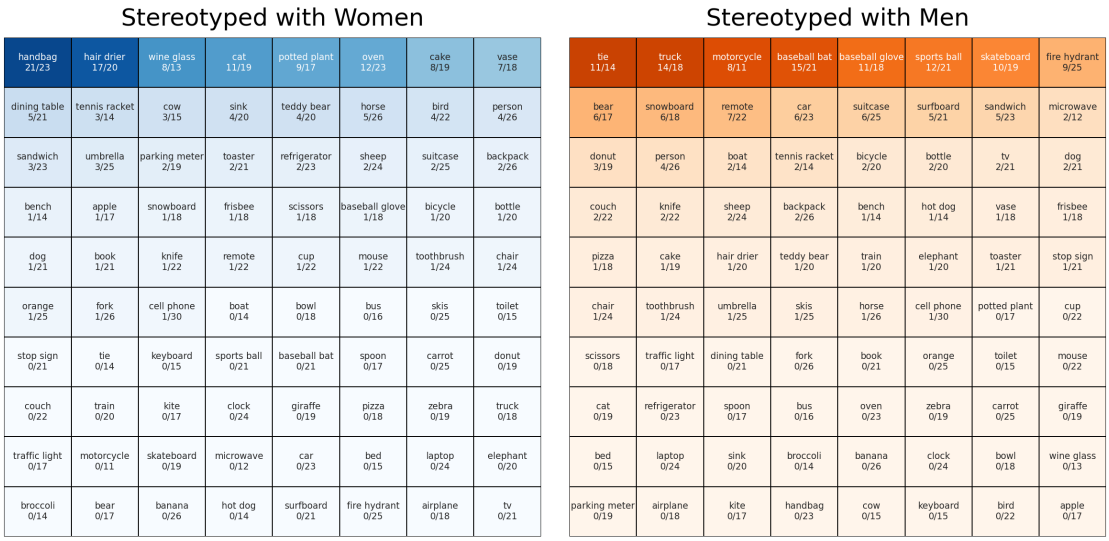


**Fig. 3 Study 1 Object Results.** Detailed participant responses for each of the 80 objects in COCO dataset. Fraction indicates number of participants asked about each object who marked it as stereotypically related to the gender group of women or men.

is already in contrast to a somewhat common assumption in ML fairness research that *any* difference between groups is an amplification of a stereotype [8].

Among 80 objects, 13 objects are marked as stereotypes by more than half of the participants (Figs. 3). 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, challenging prior assumptions on what is seen as a stereotype [102]. If an object was marked to be a stereotype, we also asked participants whether they believed it was harmful in the abstract. Complete results are in the Supplementary Material, but we use these initial findings to select experimental stimuli in subsequent studies. In Study 3a 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 3b 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`.

## Study 2: Plurality of stereotypes and harms with image recognition objects

Next, we report qualitative analyses on open-ended responses from participants' annotations, where they explain why certain objects are seen as stereotypes and harmful or not. While prior work in gender stereotypes has often focused on social roles and traits [25, 36], our data provides insights as to how objects (e.g., `oven`, `hair dryer`) can also be associated with stereotypes. This is an important departure because it expands the scope of machine learning tasks for which stereotypes are relevant

beyond its current more narrow framing. Specifically, 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 [25, 27, 30]. 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. There seems to be a disparity in responses based on the participant's gender regarding whom they believe is harmed. 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, while we can 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.

## Study 3a: 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 [30] 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. 4 with further details of the descriptive analysis of the captioning task in the Supplementary Material.

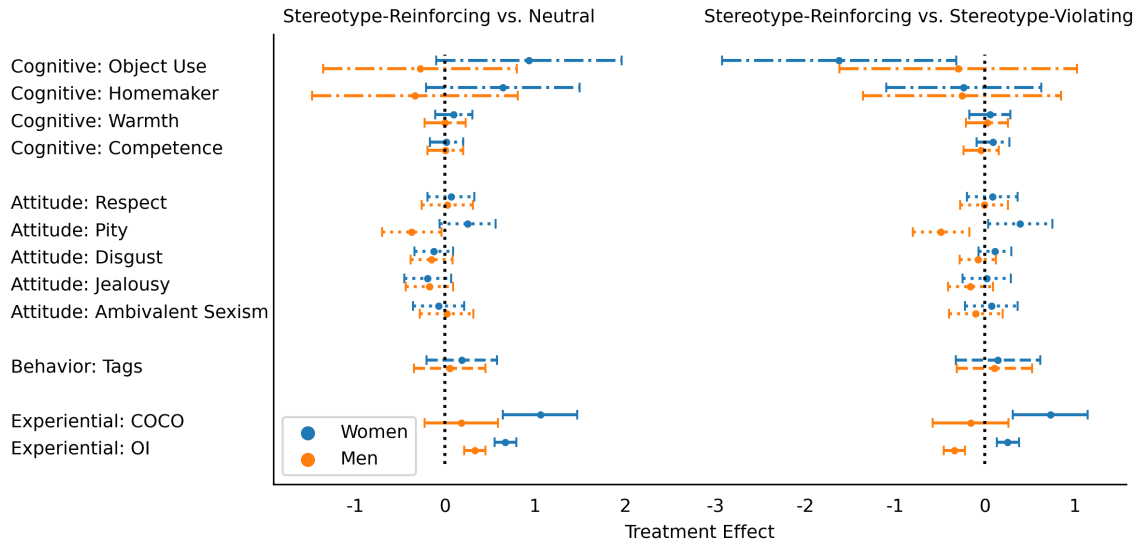


**Fig. 4 Study 3, 4 Results** The effect sizes and 95% confidence intervals are reported for 10 of our 11 measures of pragmatic harm (for the behavior measure of captioning, we provide a descriptive analysis), experiential harm on COCO, and experiential harm on our larger dataset of OpenImages. Deviations from zero indicate that exposure to the stereotype-reinforcing stimulus resulted in our measured harm compared to exposure to the control condition.

## Study 3b: 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 survey 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 [83], 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. 4.

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 [51, 59]. 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 [90]. 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 [67]. Long-term observational studies are likely more well-suited to measure these kinds of impacts [31, 33, 45]. However, we do 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 [59].

## Study 4: Stereotype-violating errors can be perceived as harmful too

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 [53]. 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 [79]. 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 of the 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 3b. Rather than asking simply about "personal harm" as we did in Study 3b, here we draw from the Positive and Negative Affect Schedule (PANAS; [14, 97]) 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 3b, 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], $p < .001$). Unlike Study 3b, 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 in Fig. 4.

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 [11, 38, 62, 63, 88]. We thus hypothesize that conventionally contradictory costume objects may evoke more negative reactions compared to conventionally contradictory script objects [77]. To test this hypothesis, we explore 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 `lipstick`, and exclude those like `truck` or `wine glass`. With this "wearable" distinction, we find that 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 effect exists in women participants (b = .49, 95% CI [.34, .64], $p < .001$) rather than men participants (b = -.03, 95% CI [-.18, .12], $p = .726$). Notably, for the interaction effect of a "wearable" object with the condition type, we find that wearable stereotype-violating errors have 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], $p < .001$). In other words, men participants tend to find it more harmful than women participants do when lipstick is misclassified on a man than on a woman.

Stereotype-violating errors seem to cause harm too, possibly through different mechanisms. In addition to this result being a consequence of backlash effects [78], we raise two more possible mechanisms. First, it could be seen as an expression of precarious manhood; a concept that suggests

manhood is precarious and needs continuous social validation such that threats to traditional masculinity can provoke anxiety in men, thus resulting in higher reports of harm [89]. Second, these results may reflect elements of transphobia, which involves a negative reaction to the apparent incongruity between a person's perceived gender and a wearable gender presentation item [11, 63]. The divergent effect between men and women participants aligns with research indicating that transphobia is higher amongst cisgender men when judging transgender women due to the perceived threat to masculinity [57, 64]. This analysis pushes us to reevaluate how we should think about reducing experiential harm, as it may encompass intolerances we do not wish to support.

## Discussion

In summary, our studies have three key contributions: we investigate the kinds of associations people believe to be stereotypical; we distinguish between machine learning errors that are stereotype-reinforcing, stereotype-violating, or stereotype-neutral; we formulate harm as pragmatic or experiential to empirically study the effect of stereotypes. Overall, while stereotype-reinforcing errors do not lead to more pragmatic harm in the lab setting we use, we do find that stereotype-reinforcing errors are consistently found to be more experientially harmful. Such experiential harm is unequally distributed, impacting more participants who are women than who are men. Formulating concrete notions of harm as we have done has implications beyond just machine learning: legal documents like the European AI Act is beginning to incorporate notions of psychological harm but lacking definitions to ground regulation in [5, 71]. We also find stereotype-violating errors to be experientially harmful, especially when these errors pertain to wearable items associated with gender presentation. This effect is stronger for participants who identify as men compared to 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 navigate this complexity, we turn to the notions of epistemic injustice [32] and standpoint epistemology [28, 68, 99]. If we interpret the negative reactions to misclassifications of stereotypically feminine clothing items on men as a manifestation of precarious manhood [89] or transphobia [11], then we should down weight these concerns in designing mitigation algorithms. Respecting people's experiential harms may not be as simple as accepting them at face value for direct measurement, but rather involves understanding which groups are likely to be harmed by what kinds of errors and why.

Our findings call for reconsidering fairness measurement in supervised machine learning tasks. This involves considering how we can leverage human-driven insights to inform model training and evaluation [10]. Traditionally, fairness evaluations tend to focus on stereotypes only in relation to occupations or traits. However our work expands this idea by showing that 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, strategies aimed at mitigating stereotype-reinforcing errors which act upon the target label will inevitably impact the occurence of stereotype-violating errors as well. And normatively, there are also questions about 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 [24] will need to be taken. While adopting simply a cost-sensitive framework [52] (e.g., different costs are associated with false positives and false negatives) is a simplified interpretation of our findings, it could be a starting point 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 2 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 [16, 17, 23, 46, 66, 96], and how much complexity is reduced by the use of discrete labels. Qualitative follow-up questions supplemented our annotations, where a lack of consensus 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. This also indicates that 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, 44, 100], thus allowing room for positionality, explanation, and critical reflection.

Our findings are limited by the methodological choices we made: First, we focused on gender stereotypes as a case study. We do not know to what extent this finding generalizes to other groups such as race and age. Second, we recruited online participants who identify as men and women and who speak English without an extensive inclusion of non-binary participants or who come from a different cultural background. Given that stereotypes are culture-sensitive, and our work also shows that the harm perception is identity-sensitive, future work needs to study the interaction between participants' identity, culture, and harm perceptions. Third, by setting a threshold of 50% for respondents indicating an object is a stereotype, we are in some senses privileging the majority

opinion which may further reify marked stereotypes to be those for the majority subset [35, 60]. Fourth, the survey experiment does not capture harms beyond the two we measure (e.g., stereotype-threat [86, 87]), nor the longitudinal effects of machine learning effects. Future work needs to capture not only the plurality in harm of machine learning errors but also how its' effect emerges and endures over time.

Overall, our work offers a rigorous empirical study connecting machine learning outputs to concrete harms by understanding the impact of stereotypical misclassifications. Rather than gesturing at harm as a justification for fairness measurement, we are very concrete in our analysis of the effects on people. Our finding that stereotype-reinforcing errors are experientially harmful for women underscores the importance for machine learning fairness interventions to be more rooted in social contexts, moving beyond objectives like just achieving equal prediction performance across groups. The diversity of responses we've presented, each influenced by participants' unique rationales, suggests the need for greater exploration of human psychological experiences in understanding how machine learning can cause harm.

# 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 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 [26, 42, 56, 98], 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 [49].

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 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 2 | 7 | 1.75 | Women | 0 | 3 | 5 | 0 | 0 | 25 | 6 | 1 | 40 |
| | | | Men | 1 | 4 | 2 | 2 | 0 | 30 | 1 | 0 | 40 |
| 3a | 10 | 2.50 | Women | 1 | 11 | 32 | 8 | 0 | 229 | 19 | 0 | 300 |
| | | | Men | 0 | 19 | 35 | 10 | 1 | 211 | 22 | 2 | 300 |
| 3b | 5 | 1.25 | Women | 0 | 4 | 7 | 3 | 1 | 35 | 5 | 0 | 50 |
| | | | Men | 0 | 4 | 2 | 3 | 1 | 35 | 5 | 0 | 50 |
| 4 | 4 | 1 | Women | 0 | 5 | 8 | 0 | 0 | 42 | 4 | 1 | 60 |
| (Labeling) | | | Men | 0 | 2 | 6 | 5 | 1 | 44 | 2 | 0 | 60 |
| 4 | 5 | 1.25 | Women | 0 | 5 | 15 | 1 | 0 | 120 | 7 | 2 | 150 |
| (Harms) | | | Men | 1 | 9 | 17 | 6 | 1 | 107 | 9 | 0 | 150 |

## Studies 1, 4: 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 [20].

## Study 3a: 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. 5).[2] 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. 5 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 condition contains the same search query,

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

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 [22] 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. 5 Study 3 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 [30] 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 [15, 29, 80]. The second asks about sexist attitudes via a shortened scale focused on benevolent sexism [36, 37, 75].[3] Finally, for *behavioral* measures, we ask participants to undertake a realistic task they are liable to encounter which can cause harm: data labeling [61]. 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. 6). 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

[3]We ask questions from the Ambivalent Sexism Inventory [36] about benevolent sexism, as opposed to hostile sexism, because the latter is believed to suffer heavily from social desirability bias.

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. 6** 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.

## 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 3a 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 3b, 4: Measuring Experimental Harm

In Study 3b, 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 Orly Bareket, Molly Crockett, 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] 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)

[5] Bayefsky R (2016) Psychological harm and constitutional standing. Brooklyn Law Review

[6] 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

[7] Bianchi F, Kalluri P, Durmus E, et al (2023) Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. ACM Conference on Fairness, Accountability, and Transparency (FAccT)

[8] 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

[9] 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)

[10] 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)

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

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

[13] Cao Y, Sotnikova A, III HD, 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

[14] 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

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

[16] 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

[17] 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

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

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

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

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

[22] 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)

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

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

[25] Eagly AH (1987) Sex differences in social behavior: A social-role interpretation. Lawrence Erlbaum Associates, Inc

[26] 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

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

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

[29] 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

[30] 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

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

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

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

[34] 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)

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

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

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

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

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

[40] Hall M, van der Maaten L, Gustafson L, et al (2022) A systematic study of bias amplification. arXiv:220111706

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

[42] 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

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

[44] 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)

[45] 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

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

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

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

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

[50] Katzman J, Wang A, Scheuerman M, et al (2023) Taxonomizing and measuring representational harms: A look at image tagging. AAAI Conference on Artificial Intelligence

[51] 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)

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

[53] 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)

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

[55] Lippmann W (1922) Public opinion.

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

[57] 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

[58] 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)

[59] 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)

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

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

[62] Morgenroth T, Ryan MK (2018) Gender trouble in social psychology: How can butler's work inform experimental social psychologists' conceptualization of gender? Frontiers in Psychology

[63] 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

[64] 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

[65] Nicolas G, Bai X, Fiske ST (2022) A spontaneous stereotype content model: Taxonomy, properties, and prediction. Journal of personality and social psychology

[66] 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)

[67] Noble SU (2018) Algorithms of oppression: How search engines reinforce racism. NYU Press

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

[69] Otterbacher J, Bates J, Clough P (2017) Competent men and warm women: Gender stereotypes and backlash in image search results. Conference on Human Factors in Computing Systems (CHI)

[70] Otterbacher J, Checco A, Demartini G, et al (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)

[71] Pałka P (2023) Ai, consumers & psychological harm. Cambridge University Press

[72] Prentice DA, Carranza E (2002) What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. Psychology of women quarterly 26(4):269–281

[73] 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)

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

[75] 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

[76] Rudman LA, Glick P (2001) Prescriptive gender stereotypes and backlash toward agentic women. Journal of social issues 57(4):743–762

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

[78] 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

[79] Schumann C, Ricco S, Prabhu U, et al (2021) A step toward more inclusive people annotations for fairness. ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)

[80] 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

[81] Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626

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

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

[84] Sotnikova A, Cao YT, III HD, et al (2021) Analyzing stereotypes in generative text inference tasks. Findings of the Association for Computational Linguistics: ACL-IJCNLP

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

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

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

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

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

[90] 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

[91] Wan Y, Pu G, Sun J, et al (2023) "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. Conference on Empirical Methods in Natural Language Processing (EMNLP Findings)

[92] Wang A, Russakovsky O (2021) Directional bias amplification. International Conference on Machine Learning (ICML)

[93] 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)

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

[95] Wang T, Zhao J, Yatskar M, et al (2019) Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. International Conference on Computer Vision (ICCV)

[96] 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

[97] 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

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

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

[100] Yaghini M, Krause A, Heidari H (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

[101] Zhao D, Wang A, Russakovsky O (2021) Understanding and evaluating racial biases in image captioning. International Conference on Computer Vision (ICCV)

[102] 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)

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

# Appendix A Sample Size Justification

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 the first part of Study 4 where we labeled OpenImages objects, we selected the number such that each OpenImages object was labeled by 5 participants from each gender; Study 3a 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 3b 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 4 we had 40 objects and as our last study ended up having the budget to have around 37.5 participants per object.

# Appendix B Additional Results from Study 1

We show the results of harm annotation in the abstract on the y-axis of Fig. B1 for the 13 objects marked as stereotypes. We see large variations within stereotypical objects for whether the association is perceived to be harmful when it is disconnected from a particular impact.

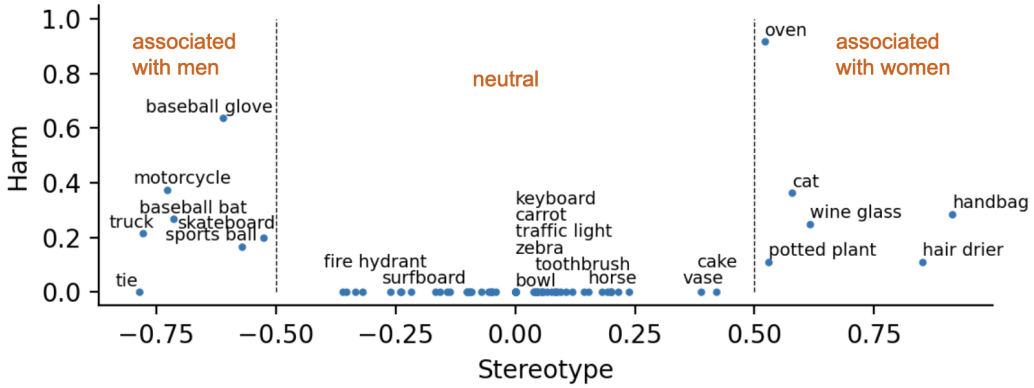


**Fig. B1 Study 1 Results.** Participant responses for 80 objects in COCO dataset. The x-axis indicates the percentage of participants who indicated an object is a stereotype, where negative numbers indicate it is a stereotype about men, and positive numbers about women. For objects where more than half of the respondents indicate it is a stereotype, the y-axis indicates the percentage who marked it to be harmful.

# Appendix C Additional Results from Study 2

Here we present the full analyses we perform on the open-ended responses we received in Study 2 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"

**Table D1** Descriptive statistics about the captions annotated as a part of Study 3a'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-reinforcing | 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 |
| Stereotype-violating | .810 (0.458-1.286) | .857 (0.222-2.500) | 1.182 ± 0.417 | 0.905 ± 0.384 | 0.632 ± 0.265 | 0.750 ± 0.248 |
| Neutral | 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 |

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."

# Appendix D Behavior Caption Analysis from Study 3a

Here we present an exploratory analysis of the captions produced by participants for the behavior task in Study 3a. In Tbl. D1 we compare the captions generated by participants across conditions, and find no statistically significant results.

# Appendix E Additional Results from Study 3b

For Study 3b not only did we ask for personal experiential harm from an error, but also societal harm, so that even if the participant does not personally feel harmed, they may feel it on behalf of the stereotyped group. There is a high correlation between the responses to these two versions of the question for each error, which is why we only reported results from personal harm in the main text.

We regress on harm value with the independent variables of stereotype condition, personal harm compared to societal harm, and their interaction effect. We find that for stereotype-reinforcing vs neutral errors, the coefficient for personal or societal harm is $b = .22$ (95% $CI$ [-.07, .51], $p = .143$), and that for the interaction effect to be $b = -.18$ (95% $CI$ [-.59, .23], $p = .398$). For stereotype-reinforcing vs stereotype-violating errors, the coefficient for personal or societal harm is $b = .26$ (95% $CI$ [-.03, .55], $p = .076$), and that for the interaction effect to be $b = -.22$ (95% $CI$ [-.63, .19], $p = .288$). In other words, there is no statistically significant difference in the results for societal harm compared to personal harm.

# Appendix F Object Selection from OpenImages

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 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*.

**Table G2** 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 [102] | .5722 | 7/13 (54%) | 10/13 (77%) |
| | Directional Bias Amp [92] | .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%) |

# Appendix G Bias Amplification

Bias amplification is a statistical notion that rests on the idea that any amplification of an existing bias is undesirable, and often used to implicitly capture stereotypes [40, 92, 95, 102]. 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 [102] (Bias Amp) and Wang and Russakovsky [92] (Directional Bias Amp). As an example, Zhao et al [102] 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. G2 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.

# Appendix H Automatic Discovery of 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 [21] to be 74%. Erroneous predictions include `carnivore` (is stereotype), `tennis ball` (is not stereotype), `infant bed` (is stereotype), `soap dispenser` (is not stereotype), `handbag` (is stereotype). Given that an object is a stereotype, the highest ROC AUC at predicting which gender is being stereotyped is 85%. Erroneous predictions include `mixer` (stereotyped with women), `doughnut` (stereotyped with men), `houseplant` (stereotyped with women), `wheel` (stereotyped with men). 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 [34], they are insufficient without human input. As we note in the main text, 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, 44], thus allowing room for positionality, explanation, and critical reflection.

# Appendix I Connection to Open-Ended ML Tasks without Correctness

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. 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 [9, 12], such as about occupations and attributes from the Implicit Association Test [39]. 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 [18, 19, 47, 48, 58, 73, 82, 103], 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 [94]) it surely seems that more thinking should be done on the different implications of debiasing stereotypes as opposed to debiasing more neutral concepts.