Measuring Representational Harms in Image Captioning: Supplementary Material

In this supplementary material we first present measurement approaches for harms that we could not fit in the main text: denying the opportunity to self-identify in Section A, reifying social groups in Section B, erasure in Section C, and additional stereotyping approaches we could not fit in the main text in Section D. Then, in Section E we provide a table of summary statistics for each measurement approach on the COCO and CC datasets. References to numerical sections refer to the main text.

A HARM: DENYING OPPORTUNITY TO SELF-IDENTIFY

Katzman et al. [6] argue that image tagging systems that deny people the opportunity to self-identify as members of particular social groups is its own kind of representational harm. Our measurement for this harm involves detecting any mention of group identity, assuming that the people in the images have not communicated the particular groups with which they self-identify to the labeler. We know this holds true in the COCO dataset's human-generated labels and captions as well as all system-generated predictions we consider. However, it is possible that in the CC dataset some labeled identity groups may have been self-defined. Without knowing who created the alt-text tags the captions were generated from, we cannot know for sure. While this was the most straightforward way we conceived of to operationalize the harm of denying people the opportunity to self-identify, we want to emphasize that there are no doubt many other approaches to this. For example, perhaps another option would be asking the people pictured in each of the images whether they felt the caption applied to their image denied them the opportunity to self-identify.

A.1 Social Group Identity Words

Social groups come up throughout many of our measurements, and in the context of our work we consider them along four axes: gender, age, ethnicity, and NORP (nationalities or religious or political groups).[1] The axes to consider and ways of detecting each involve difficult decisions without a clear solution, though having a specific application context in mind will certainly help guide the choices, and make it clearer than trying to determine these in the abstract [1, 2]. For example, age is not always considered a sensitive attribute, and race may be more applicable than ethnicity. To detect mentions of these social groups, we do the following: for gender we use a list that consists of the words covered by the hyponyms under the *female* and *male* WordNet synsets, which is limited in its scope by the focus on binary genders; for ethnicity and age we use adjective word lists curated from prior work [16]; for NORP we leverage SpaCy's entity extractor [5]. As we discuss in Section 3.5.1, these detection mechanisms are imperfect and likely to under- and over-count instances of identity. For example, words like *white* and *Black* can refer to both a social group as well as descriptive color, and would be liable to being over-counted if included as part of a word list. We thus exclude both from the ethnicity word list. However, as racial descriptors are more likely to be used when describing non-white people because of a perception as being non-normative [10, 14], this choice likely results in more under-counting for non-white people than white people.

In Table 1 we compare measurements at and across stage 3 (human-generated caption) and stage 4 (system-generated caption). For system-generated captions, in COCO the most common words in each identity category are *man*, *woman*,

[1] This is one category because SpaCy's entity extractor does not differentiate within NORP, and for our purposes the detection of this singular category is sufficient.

Table 1. Counts of social group mentions across human-generated captions, system-generated captions, and captions where the human has not assigned identity but the system has. Although the captioning system generally labels social group less than humans, that they do not assign them to the same images as humans indicates an arbitrariness in application.

Stage	Data	Gender	Age	Ethn.	NORP	Data	Gender	Age	Ethn.	NORP
3: Human-generated caption	COCO	6959	3385	312	189	CC	1883	1106	116	72
4: System-generated caption		4905	1009	0	45		2179	1190	92	57
(3) → (4)		818	284	0	15		1349	876	63	47

men for gender; *young*, *little*, *baby* for age; none for ethnicity; and *french*, *poles* for NORP. We note that all instances of *french* are in fact referring to *french fries*, and *poles* referring to *ski poles*, indicating broken assumptions in our reliance on external detection mechanisms. For system-generated captions in CC the most common words in each identity category are *woman*, *man*, *girl* for gender; *young*, *old*, *little* for age; *american*, *european*, *polish* for ethnicity; and *western*, *roman catholic*, *australian* for NORP. With a few exceptions, system-generated captions generally apply identity labels at a lower rate than human captions. This is likely due to the overall lower level of descriptiveness of model captions as compared to human captions [7, 15]. However, we can also observe from the measurement of stage 4 (system-generated captions) using stage 3 (human-generated captions) as the "ground-truth," i.e., counting the individual captions in which humans have not assigned identity but the system has, that even though the model assigns identity less, it is not necessarily on the same images that humans are assigning identity to. This can be concerning because it implies an arbitrariness in the system's application of social group mentions, even to images where humans felt they could not.

B HARM: REIFYING SOCIAL GROUPS

Building on the previous harm, reifying social groups is defined to be an additional representational harm that comes from treating social groups as fixed categories rather than social constructs and from assigning people to these groups on the basis of visual appearances [6]. In this operationalization, we identify correlations between facial attributes and externally ascribed gender.

One manifestation of this harm that arises in image tagging is during the problem formulation stage where the space of output labels can be restrictive, e.g., only including two gender identities. In image captioning, which does not suffer from such a technical restriction of limited outputs, the same limitation of available identities is encoded into the system through the exclusion of words in the training data; e.g., even if a system is technically able to produce the word *non-binary*, the system will not know how to apply it unless it was used in the training data.

B.1 Correlations between identity and facial attributes

Prior work on harms in automated image captioning systems have used interpretability techniques to understand where an image captioning system is looking when assigning gender [4, 13], and considering it problematic if it is the non-person parts of the image. While the measurement techniques they employed are very useful for understanding which visual parts of the image are being reified as indicative of gender, their mitigation strategy of teaching the captioning systems to look at the person when assigning gender are flawed. This merely shifts the visual cues that determine gender from the scene context to the physical appearance of the person, continuing to reify a relationship between gender and visual features. To understand which such physical appearance traits and social groups are being reified in our COCO demographic annotations [19], we call Microsoft Azure's Face API [9] to detect physical attributes in each face, noting that these APIs frequently come with their own set of biased outputs that over- and under-predict certain characteristics for people of different skin tones and perceived gender expressions. With these limitations in mind, we map the detected attributes to the annotator-inferred gender labels. On the 1,977 faces that both have gender annotated as either *female* or *male* and were detected by the API, we find statistically significant differences on all ten of the facial attributes we consider. We present demonstrative empirical results for the first example of each gender, and find that people who are labeled as *male* are more likely to have a moustache (0.183 ± 0.198 compared to $0.002 + -0.018$), beard, sideburns, baldness, reading glasses, and headwear. People who are labeled as *female* are more likely to have a smile (0.499 ± 0.459 compared to $0.392 + -0.449$), eye makeup, and lip makeup. For hair color, people labeled *male* have more black and gray hair, while people labeled *female* have more blonde and brown hair. By ascribing gender solely based off of physical appearance, various correlations are reified.

We note that with this measurement we do not mean to imply that the goal should be to sever all associations between gender expression and physical appearance, as the performance of gender can at times be a valuable way for individuals to signal their identity. Rather we want to make existing reifications transparent such that people, and systems if assigning identity, can adjust their beliefs such that individuals do not feel the need to conform to a narrow set of prescribed mappings, e.g., people with moustaches are *male*.

C HARM: ERASURE

The harm of erasure, as defined in Katzman et al. [6], refers to when people, attributes, or artifacts belonging to certain social groups are not recognized by a system. The examples they provide are when a *menorah* is misnamed as *candlesticks*, women suffragists marching are tagged with *people*, *walking*, and *street*, and a Holocaust memorial is tagged with *field* and *sculpture*. By neglecting to mention important parts of the image context, captions have the potential to cause harmful erasures. On the other hand, this can be in tension with other harms such as denying the opportunity to self-identify when social group words are important to the image context and should be mentioned.

All of these harms have to do with some kind of a false negative, and we break it down into five different operationalizations. The first measures systematic errors such as for the menorah, the next three measure individual types of errors such as for the examples of suffragists and a Holocaust memorial, and the last captures the remainder of the false negatives that do not fall into any of the defined types. These three defined types are for individual instances of erasure, and specified based on pre-conceived notions we have about what kinds of omitted words are likely to be most harmful: 1) when a social group mention (e.g., *woman*) is mentioned in the human-generated caption but not the system-generated caption, 2) when a named entity (e.g., *9/11 Memorial*) is mentioned in the human-generated caption but not the system-generated caption, and 3) when a word with more vagueness than needed is used in the system-generated caption as compared to the human-generated caption (e.g., in the context of women suffragists *walking* instead of *marching*). We note that there is some overlap in these measures, e.g., *people* is a more vague word for *women*, but *women* is also considered an identity word.

C.1 Consistent misnaming

For the first measurement of systematic erasures where something is consistently misnamed, we collect all instances in which a synset that is present in an image is not mentioned (a false negative), and examine what other synset that is not present is mentioned instead (a false positive). We assume a high-enough quality in our annotation such that erasure has not already occurred at stage 1 (human-generated labels), otherwise we would not have a prexisting way of knowing of an object's existence in the first place.

Although most other names for *menorah* that are not *menorah* would be a erasure harm, here we focus on the systematic misapplication of the same word, e.g., *candlestick*. This is motivated by the types of measurements possible in an image captioning system, because we do not know which object instance in an image is being specifically referred to by a particular word. For example, in the COCO dataset we find that for captions with images of women that fail to mention *suit* (i.e., a false negative), the most common false positives are *surfboard*, *wave*, and *ocean*. It is not the case that the model is mistakenly remarking upon a *suit* with the words *surfboard* or *wave*, but rather the type of images where *suit* is missed is also the type of images where *surfboard* and *wave* are hallucinated. In fact, upon closer inspection we find that these images are of people at the beach, and *suit* is referring to a swimsuit that was not mentioned by the caption.

Thus, by focusing on systematic misapplications of the same word, we are able to more confidently discern when the false positive is replacing the false negative. We surface cases where the 95% confidence interval of a specific replacement word being used is fully above 50%, and in COCO while using stage 1 (human-generated labels) as the "ground-truth," find two instances in stage 3 (human-generated captions) and no instances in stage 4 (system-generated captions). The two instances in human captions are that 70.8% of the time *skateboard* is not captioned, *trick* is, and 100% of the time *toaster oven* is not captioned, *kitchen* is. While neither of these erased objects are associated with a particular social group and are not harmful, this method would allow us to surface cases such as the erasure of *menorah*.

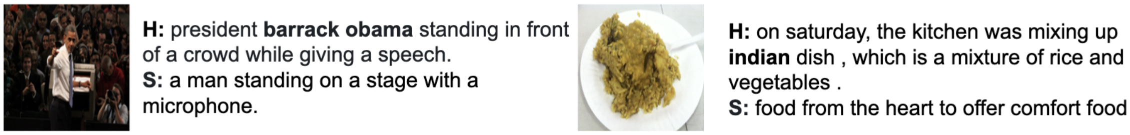
C.2 Social group not mentioned

This measurement aims to capture instances in which a social group is critical to appropriately understanding the context of the image, but the system does not provide it, and in doing so, erases the relevance of that particular social group. We detect instances in which the social group is mentioned in the human-generated caption, but not the system-generated caption. When measured on COCO, we count 5,680 out of 17,360 captions to fit this criterion. We find that, from most to least common, the social group categories named in the human-generated caption but not system-generated caption are: gender at 4,462 (e.g., *man*, *woman*, *girl*), age at 3,872 (e.g., *young*, *old*, *little*), nationality/religion at 159 (e.g., *mans*, *poles*, *french*), and ethnicity at 106 (e.g., *asian*, *foreign*, *oriental*). The numbers are quite high because the COCO human-generated captions frequently infer gender even when it may not be necessary,[2] so our assumption of the optimality of ground-truth captions is violated. If our human-generated captions were the gold standard, this measure would likely be more accurate. Additionally, we can see that *mans* and *poles* are detected to be identity words in the NORP category, but we note that *poles* may be referring to something like *ski poles* and this may be misreported. On CC we measure 1,687 of 14,560 captions, seeing a similar distribution over the social group categories of gender at 1005, age at 817, ethnicity at 88, and NORP at 62.

C.3 Named entity not named

For the next measurement, a very similar method is employed with named entities detected by spaCy [5] instead of social group mentions. A similar motivation follows as well, where there is the notion that if a named entity (e.g., the 9/11 Memorial) is explicitly named as such rather than a more generic term such as *building*, then the specific name provides added necessary context. For system-generated captions we measure 855 on COCO, and 255 on CC. The top

[2] We acknowledge that determining when social group is necessary to mention is a difficult task, without always a clear answer. However, given the content of an image, there may be certain cases where it is clear the significance that social group has to its meaning.



H: president **barrack obama** standing in front of a crowd while giving a speech.
S: a man standing on a stage with a microphone.

H: on saturday, the kitchen was mixing up **indian** dish , which is a mixture of rice and vegetables .
S: food from the heart to offer comfort food

Fig. 1. Examples of erasure harms, where the bolded word is discovered by our measurement approach to be a named entity that is important to include. The human-generated caption is denoted by the prefix **H** and the system-generated caption is denoted by the prefix **S**. The left example is from COCO, and the right from CC.

four named entities ignored for the former are *london*, *bush*, *teddy*, *marina*, and those for the latter are *footballer*, *uk*, *businessperson*, and *los angeles.* We can probably decide that some of these words, like *marina* and *footballer*, were incorrectly tagged by the named entity extractor. Some qualitative examples from both datasets where the named entity that is ignored may be important for contextual understanding of the image are shown in Figure 1. It is also likely that the captioning system may be better at recognizing named entities from certain countries and regions over those from others, similar to the findings from DeVries et al. [3] on household objects. Because erasure is especially salient and harmful when applied to historically under-represented groups, this kind of a deficiency can further amplify these harmful erasures by failing to detect them.

C.4 Too vague

The next aspect we measure is using too vague of a term. The motivation behind this measurement is that when a more specific term could be used to describe something but is not, there is room for an erasure harm that comes from missing the necessary specificity. Not only can we measure the additional vagueness introduced by the system-generated caption on top of that by the human-generated caption, but if we have reason to believe our human-generated labels are at a more desired level of specificity than our human-generated captions, then we can measure the difference there as well. Taking this measurement at stage 4 (system-generated caption) using stage 3 (human-generated caption) as "ground truth," we find that for COCO 918 captions become more vague, and for CC 862 captions become more vague. Common examples of terms that become more vague include man→person, batter→player, stand→be, and pony→mammal. From these examples, it is clear that one of the biggest violated assumptions is that which claims that a more vague use of a term when a specific one is applicable is erasure. The process of automating the differentiation in harms of mistaking *marching* to be *walking*, as opposed to simply calling a *bear* an *animal*, remains a difficult task.

C.5 Ignoring

Our final operationalization of erasure catches the remainder of the false negatives. These are measured as what is missed in stage 4 (system-generated caption) but mentioned in stage 3 (human-generated caption). Our heuristic involves filtering out all false negatives that involve something present in over 1% of all images, because anything that prevalent is unlikely to be susceptible to an erasure harm; this included synsets such as *inside* and *picture.* Then, we ranked tuples by false negative rate, hypothesizing again similarly to as in stereotyping that the more systematic erasures are likely to be more harmful than the rarer ones. We do not find especially harmful erasures here by the model, and it is likely that we missed erasure harms because the stage 1 (human-generated labels) committed the same erasure of not naming something in the first place, so we were unable to detect it.

D HARM: STEREOTYPING

In addition to the measurement approaches listed in the main text, here we include a few more potential operationalizations.

D.1 Different distributions for different group identities and bias amplification

To make measuring correlations of social groups and different synsets tractable, we employ taxonomies to categorize objects, attributes, and relationships. For objects, we use the list of *supercategories* from COCO [8] which assigns objects like *table* into a higher-level category of *furniture*. Because this taxonomy only categorizes the 80 labeled objects in the original COCO dataset (and does not include VisualGenome objects), we employ a word embedding similarity method from prior work to categorize all of the object synsets [17]. For attributes and relationships we similarly employ existing taxonomies [11, 16]. We then use the chi-squared test to check if the distributions of each tuple's categories differs between people of different social groups. Overall, we find the greatest distribution differences between images of people who are labeled male or female to be in stage 1 (human-generated labels), and the least distribution difference to be in stage 4 (system-generated captions). This indicates that perhaps the image distribution portrays people of different genders in stereotypical scenes, and the system-generated captions actually somewhat flattens these stereotypes in what it chooses to describe, perhaps again largely due to the general simplicity and lack of descriptiveness of model captions.

To mechanize the comparison between stages beyond manual inspection, we incorporate the notion of bias amplification [18, 20]. Bias amplification measures the amplification of correlations in one stage relative to those that already exist in another. It captures that if the correlation of a social group-tuple category pairing (e.g., male and vehicle) increases from one stage to another, this is only meaningful if this correlation existed in the original data; if it did not, then the increase is not considered problematic. Thus, similar to the motivation for the heuristic we used for measuring stereotypes in Section 4.1.1, the assumption made by this metric is that problematic correlations are those which are captured in the existing dataset. Some categories we find to be amplified when measuring stage 4 (system-generated captions) treating stage 3 (human-generated captions) as the "ground truth" in COCO is *sports* for people with lighter skin tones and *vehicle* for people with darker skin tones.

E SUMMARY STATISTICS

With our caveats in mind about the limitations of aggregate numbers that summarize a harm, we acknowledge there are benefits of digestibility that come with such a number. Thus, we ultimately provide a set of higher-level numbers to summarize each of the measurements, emphasizing that these numbers are not to be taken at face value, but rather as a high-level entry point to begin the guided understanding of the identified harms. This is in Table 2.

Generally, the measurements tend to over-count rather than under-count because we see the latter to be more problematic by missing harms. We consider our measurements more as heuristics for focusing attention, rather than a complete reduction of an entire harm to a small set of numbers. Thus, missing a harm completely is more problematic than giving users more cases to inspect (although of course too much of an overcount can make the entire endeavor unwieldy). The problem of differentiating between a more innocuous error which manifests technically in the same way as a harmful error is difficult, and not always possible to be automated; this limits our ability to lower certain over-counts we have. For example, captioning the presence of an apple when there is not one is fairly innocuous, but captioning the presence of a stove in a picture of a woman when there is not one is a much more loaded error due to societal stereotypes about women in kitchens.

Table 2. Summary statistics that serve as guiding heuristics for understanding what is measured by each of our approaches. COCO has 17,360 captions considered, and CC has 14,560, which are the denominators used for converting the absolute numbers to fractions.

Harm	Measurement	COCO [8]	CC [12]
Denying Opportunity to Self-Identify (Sec. A)	Group identity words	5,160 (29.7%)	2,821 (19.4%)
Reifying Social Groups (Sec. B)	Correlations between facial attributes and identity	N/A	N/A
Stereotyping (Sec. 4.1 and Sec. D)	Incorrectly uses a word	11,328 (65.3%)	11,539 (79.3%)
	What is correctly mentioned	gender: 23 of 156 objects; skin tone: 20 of 148 objects	N/A
	Different distributions for different group identities	gender: obj, att, rel; skin tone: obj	N/A
	Bias amplification	gender: att; skin tone: att	N/A
Demeaning (Sec. 4.2)	Demeaning words	0 (.0%)	29 (.2%)
	Different mentions of people	.0018 ± .0074 for gender, .0113 ± .0157 for skin tone	1,924 (13.2%)
	Context-specific	27 (.2%)	45 (.3%)
	Identity adjective as noun	0 (.0%)	2 (.0%)
Erasing (Sec. C)	Consistent misnaming	0 objects	0 objects
	Group identity not used	6,371 (36.7%)	1,908 (13.1%)
	Named entity not named	855 (4.9%)	255 (1.8%)
	Too vague	913 (5.3%)	862 (5.9%)
	Ignoring	16,226 (93.5%)	12,234 (84.0%)

REFERENCES

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

[2] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (2021).

[3] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019).

[4] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. *European Conference on Computer Vision (ECCV)* (2018).

[5] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python.* https://doi.org/10.5281/zenodo.1212303

[6] Jared Katzman, Solon Barocas, Su Lin Blodgett, Kristen Laird, Morgan Klaus Scheuerman, and Hanna Wallach. 2021. Representational Harms in Image Tagging. *Beyond Fair Computer Vision Workshop at CVPR 2021* (2021). https://drive.google.com/file/d/1oJp8CqNpYEsOlO8cwv4cTnHGbOjWxEZ-/view

[7] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (2016).

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

[9] Microsoft. 2021. Microsoft Azure Face API. (2021). https://azure.microsoft.com/en-us/services/cognitive-services/face/

[10] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. 2019. How Do We Talk about Other People? Group (Un)Fairness in Natural Language Image Descriptions. *The Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP-19)* (2019).

[11] Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing Common Human Visual Actions in Images. *Proceedings of the British Machine Vision Conference (BMVC)* (2015).

[12] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (2018).

[13] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, and Xia Hu. 2020. Mitigating Gender Bias in Captioning Systems. *arXiv:2006.08315* (2020).

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

[15] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the Diversity of Automatic Image Descriptions. *International Conference on Computational Linguistics* (2018).

[16] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Talking about other people: an endless range of possibilities. *International Natural Language Generation Conference* (2018).

[17] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. 2020. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *European Conference on Computer Vision (ECCV)* (2020).

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

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

[20] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2017).