Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark

Sina Mohseni
sina.mohseni@tamu.edu
Texas A&M University
College Station, Texas

Jeremy E. Block
Eric D. Ragan
j.block@ufl.edu
eragan@ufl.edu
University of Florida
Gainesville, Florida

ABSTRACT

Research in interpretable machine learning proposes different computational and human subject approaches to evaluate model saliency explanations. These approaches measure different qualities of explanations to achieve diverse goals in designing interpretable machine learning systems. In this paper, we propose a benchmark for image and text domains using multi-layer human attention masks aggregated from multiple human annotators. We then present an evaluation study to compare model saliency explanations obtained using Grad-cam and LIME techniques to human understanding and acceptance. We demonstrate our benchmark's utility for quantitative evaluation of model explanations by comparing it with human subjective ratings and ground-truth single-layer segmentation masks evaluations. Our study results show that our threshold agnostic evaluation method with the human attention baseline is more effective than single-layer object segmentation masks to ground truth. Our experiments also reveal user biases in the subjective rating of model saliency explanations.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**.

KEYWORDS

machine learning explanations, explanation evaluation, explanation benchmark, data annotation

ACM Reference Format:
Sina Mohseni, Jeremy E. Block, and Eric D. Ragan. 2021. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397481.3450689

1 INTRODUCTION

With the recent and continuing advancements in robust deep neural networks (DNN), the prominence of machine learning techniques for automated decision-making is growing. In such cases, expert users, operators, and decision-makers can also take advantage of advanced machine learning techniques to account for latent features and assist in taking real-world actions. However, because of the disparity between the sense-making process in humans and the computational feature learning of machine learning models, people require model transparency to be able to understand and trust machine learning models. Thus, for more effective human-AI collaboration, advancements in model explainability are needed to support user understanding. This is the primary goal of recent interdisciplinary research thrusts in *Explainable Artificial Intelligence* (XAI). While a multi-faceted topic, the ultimate goal is for people to understand machine models, and it is therefore essential to involve user feedback and reasoning as a requisite component for design and evaluation of XAI systems [25].

Research on interpretable algorithms has recently proposed various techniques to design inherently interpretable models [37] and generate explanations for black-box models [33]. Interpretability techniques enable user review of model reasoning and learning representations for their correctness in accordance to design goals, law and regulations, and safety requirements. Such evaluations could potentially prevent adverse outcomes of AI-based systems—such as unfair and discriminatory decision-making when performing real-world tasks. However, with the complexity of interpretability techniques and human cognitive biases, the question remains: how should we assess the correctness and completeness of the evaluation methods for machine learning explanations?

Different approaches have been proposed for evaluating interpretable models and XAI systems at different stages of system design [7]. In machine learning research, various computational methods are used to measure the fidelity of interpretability techniques with respect to the underlying black-box model [1, 13]. On the other hand, in the field of human-computer interaction, human-grounded evaluation approaches measure human factors such as user satisfaction, mental model, and trust in XAI systems designed for different tasks. However, there are fundamental differences between these evaluation approaches. Computational methods set a precedent to objectively evaluate the model against a baseline ground truth, yet they lack the ability to quantify human interpretations. On the other hand, while more descriptive in nature, human subject studies tend to be more costly, imprecise, and subjective to the task. Another major difference between these evaluation methods is that once the human user is exposed to the evaluation study setup, she can not unlearn the experience for another round of evaluation. These differences raise the need to study the

IUI '21, April 14–17, 2021, College Station, TX, USA

ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450689

trade-off between objective ground-truth evaluation and subjective human-judgment of explanations.

In this paper, we propose a human-attention baseline to quantitatively evaluate model saliency explanations. Our publicly available[1] human-grounded benchmark enables fast, replicable, and objective evaluation of model saliency explanations. Using this benchmark, we study the relationship between subjective and objective evaluation of saliency explanations by comparing our benchmark with both binary feature mask ground truth (i.e., objective measure) and user rating (i.e., subjective measure) evaluations. Specifically, we are looking into the following two research questions:

- **RQ1:** What is the relationship between saliency explanations' evaluation results with objective ground truth and subjective users rating?
- **RQ2:** How do user biases affect subjective rating of saliency explanations?

We measure and report the correctness and completeness of explanations based on the feature-wise mean absolute error between model saliency map and our ground truth baseline. Our study results reveal the trade-off between objective ground-truth evaluations and subjective human-judgments of explanations. Our experiments also reveal user biases toward different model error types and explanation visualizations in their subjective rating of explanations.

2 BACKGROUND

The evaluation of model explanations and interpretability techniques can be categorized in different ways [7, 25]. For instance, previous works have examined the fidelity of interpretability techniques to the black-box model [1, 13], evaluated correctness of model explanations with ground-truth [8], as well as the usefulness of explanations in different tasks and domains [16].

Recently, inspired by the call to understand how XAI impacts user trust [10], Hoffman et al. [12] compiled a set of methods uniquely situated for understanding the fuzzy conceptualization of trust in these systems. Clearly, the trustworthiness of an explanation goes beyond computer science literature and draws heavily from the social sciences [15, 24] to study the similarities between the user trust in machines and interpersonal trust [11]. As Shneiderman discusses [34], XAI research can help dissolve the ambiguity among researchers as it refers to human trust, reliance and safety.

However, in contrast to the context of machine learning trustworthiness, this paper focuses on evaluating the correctness and completeness of explanations with the assumption of having high-fidelity explanations. Also, we only focus on model saliency explanations (e.g., gradient-based [33] and backpropagation-based [36] techniques) which present a map of feature importance for individual inputs as presented in Figure 3). The rest of this section presents a review of related work on the two *human judgment* and *ground-truth* based evaluation approaches.

2.1 Evaluation with Ground Truth

An objective way to quantify the correctness of model saliency explanation is to examine it against a ground truth baseline. The ground truth is often defined by human annotation of representative features (i.e., feature mask) and provides a baseline for quantitative evaluation of explanations quality. Examples include annotations of the object's "segmentation mask" in natural datasets, e.g., [9], and synthesized datasets, e.g., [27], that represent specific features associated with the target class. Different similarity metrics, such as Intersection over Union (IoU) (also called Jaccard index) and mean Average Precision (mAP), are used to quantify the quality of model saliency explanations or bounding boxes compared to the ground truth. For instance, Li et al. [22] used IoU, between the model saliency map from a Convolutional Neural Network (CNN) and the ground truth segmentation mask from the validation set, to measure their quality as a weakly-supervised semantic segmentation task. In another work, Du et al. [8] calculate the mAP between the bounding boxes of an objects' saliency mask and the ground truth bounding boxes to evaluate their interpretability technique as an object localization task. Similarly, in the text domain, direct comparison of model attention explanations with human-annotated sentences, e.g., evidence supporting the target label [38], provides an explanation quality score [20]. However, the relationship between the evaluation of machine learning explanations and these auxiliary tasks (i.e., binary object localization and semantic segmentation) is not clear yet. Our studies reveal some aspects of *user feedback* that are missing in the ground truth baselines but would complement the evaluation of machine learning explanations.

In a review of limitations in threshold-based evaluations for model saliency map, Choe et al. [4] present an evaluation protocol to include a hyperparameter search for the τ threshold for generating objects' "binary mask" from the saliency score map. However, unlike our proposed evaluation protocol, they do not consider the pixel-wise evaluation of saliency score maps in the first place. Aside from object segmentation mask baselines that annotate entire features associated with the target class. Perhaps the closest works to our human attention benchmark are Huang et al.'s [14] and Das et al.'s [5] baselines for evaluating saliency maps for different machine learning algorithms. Huang et al.'s [14] gamified approach asked participants to suggest regions of interest before then guessing which areas previous users had suggested on the same images as a method to capture more accurate user attention maps. Das et al. [5] proposed the VQA-HAT baseline for evaluating saliency maps for visual question and answering models. They test multiple game-inspired methods for attention annotation by asking participants to sharpen regions of a blurred image to answer a given question. The resulting baseline is a human attention map that enabled object identification by individual participants.

Similar to the previous benchmarks, our benchmark assembles annotations from multiple unique participants to increase diversity in responses. We go further by also calculating and assigning importance scores to features for creating generalizable human-attention maps. In the next sections, we present a series of evaluation experiments and argue that our proposed multi-layer human-attention baseline is able to evaluate the completeness (i.e., the existence of false-negative explanation errors) and correctness (i.e., the existence of false-positive explanation errors) of model saliency maps.

[1] https://github.com/SinaMohseni/ML-Interpretability-Evaluation-Benchmark

2.2 Human Judgment Evaluation

A common approach for evaluating machine learning explanations is the direct review of model explanations with end-users for their subjective feedback. Multiple papers have reported measurements of users' understanding of explanations as a proxy for usefulness and interpretability of explanations [17, 29]. Others have measured user-reported trust as a proxy for explanation goodness. For example, Nourani et al. [26] and Papenmeier et al. [28] studied the effects of explanation meaningfulness and ad-hoc explainer fidelity on user reliance. Both studies show that model accuracy and explanation fidelity impact users' trust in the model and conclude that providing nonsensical explanations (i.e., those that do not align with users' expectations) may harm users' reported trust and observed reliance on the system. With a crowdsourced evaluation approach, Schmidt and Biessmann [31] present quantitative measures for system interpretability and human trust. They propose that analyzing user interaction time can serve as a proxy for users' understanding of the explanation. They recommend that model explanations need to enhance the information transfer rate to users, help users establish an intuitive understanding of system performance and perform well independent from the user task. Along similar lines, Hoffman et al. [12], describe how temporality, or the amount of time one spends with a particular tool, is considered one of the most significant factors in leading to accurate representations of trust, system capabilities and expectations. Taking a different perspective, Schneider et al. [32] inspected the effects of deceptive model explanations in a user study. Their findings indicate that explanations that are unfaithful to the black-box model can fool users in accepting wrong predictions. Following a similar goal, Lakkaraju et al. [18] present an approach to generate misleading explanations in a case study with law and criminal justice domain experts. Their study results found that misleading explanations were able to significantly increase users' trust. Conclusively, a more robust scale is needed to evaluate explanation correctness and completeness as various research show unjustified user trust can be developed with user biases and deceptive systems.

Different papers have run user studies to evaluate the human understanding of saliency map explanations from DNNs. For example, Alqaraawi et al. [2] showed that instance explanations carry new information to users, but model behavior remained largely unpredictable for participants. In other work, Zhang et al. [39] compared saliency explanations from multiple networks with human explanations of objects in images. They performed a large crowdsourced study to directly compare machine learning and human explanations and human feedback on model explanations. Their results indicate that the features learned by some DNN models are more similar to human intuition. To address the limitations in human judgment evaluation studies, Lertvittayakumjorn and Toni [21] defined a set of objective evaluation tasks for quantitative evaluation of model explanations with respect to different explanatory purposes. They used three human-grounded tasks to evaluate local explanation methods for their ability to reveal model behavior, justify model predictions, and help users investigate uncertain predictions. Overall, the review of previous research indicates that the dissonance between machine learning models' goal to *learn discriminant features* and human expectation for *common sense explanations* (as presented in [39]) undermines the choice of using human judgment for evaluation of machine learning models.

Table 1: Details of the evaluation benchmark for human attention masks in different datasets.

Domain	Image		Text	
Dataset	PASCAL VOC	ImageNet	20 Newsgroup	IMDB 50K
Number of classes	20	20	2	2
Samples per class	50	5	100	100
Total annotation sample size	1000	100	200	200

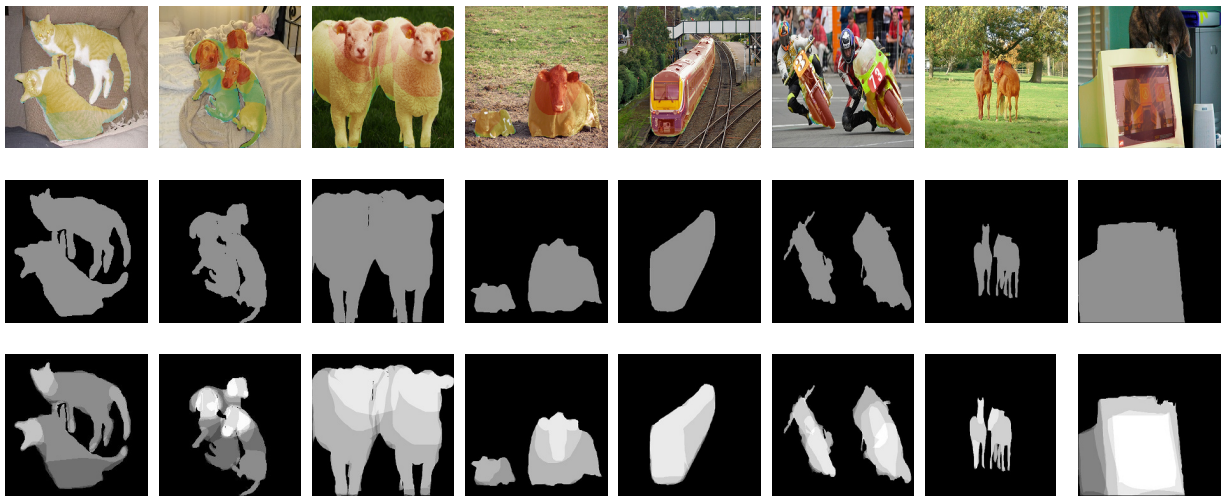
3 HUMAN-ATTENTION BENCHMARK

We captured human annotations of salient features to create this human-grounded benchmark. Participants were prompted to select relevant regions in images and phrases in text documents that they felt most represented the target subject or topic, respectively. Figure 1 show examples of the resulting multi-layer masks derived from aggregating annotations from multiple unique annotators for each image. In comparison to the single-layer object's segmentation map, the human-attention benchmark allows for a higher level of granularity in the evaluation of saliency maps and reflects the features most salient to human attention. Also, compared to human judgment rating evaluations, the human attention benchmark enables reproducible and cost-efficient evaluation. The following reviews the details of the benchmark specification, annotation procedure, and data processing.

3.1 Benchmark Specifications

The benchmark presents multi-layer masks representing what features humans expect to be the most important representations of a particular class. For each sample, we collect annotations from 10 unique annotators from the Amazon Mechanical Turk platform that were instructed to select areas (in images) or words (in documents) that they deem most relevant to the target class. The multi-layer mask generated by aggregating annotations for each individual sample provides a more granular representation of attributed features compared to the single-layer mask. Note that our method—collecting multiple user annotations for human-attention masks—balances the trade-off between objective annotation of precise feature-masks (i.e., segmentation mask) and subjective human judgment of the representative features. Also, it is important to mention that this human-attention baseline evaluates the explanations' correctness or trustworthiness of saliency explanations and does not intend to measure the fidelity of ad-hoc interpretability techniques to the black box models.

The development of this benchmark consists of a validation subset from *ImageNet* [6] and *PASCAL VOC* [9] image datasets and *20 Newsgroup* [19] and *IMDB* [23] text datasets. Table 1 presents details for the number of classes and annotated samples from the four datasets in our explanation evaluation benchmark. For the PASCAL VOC dataset, 50 randomly selected samples from all 20 classes are annotated including Vehicles (airplane, bicycle, boat, bus, car, motorbike, train), Households (bottle, chair, dining table,



(a) Cat (b) Dog (c) Sheep (d) Cow (e) Train (f) Motorbike (g) Horse (h) TV/Monitor

Figure 1: Examples of human annotations of salient features on images with the target class in the caption. (Top) Input images with human-attention mask heatmap overlay. (Middle) Single-layer object's segmentation mask for the target class. (Bottom) Resulting multi-layer human attention mask. Each image is annotated by 10 unique participants. The salient areas are selected by more participants and visualized with lighter shades.

potted plant, sofa, TV/monitor), and Animals (bird, cat, cow, dog, horse, sheep) and other (person). To create a validation set from the ImageNet dataset, we randomly selected five images from 20 classes including living things (man, woman, cat, dog, bird, ant, elephant, shark, zebra, flower, tree), indoor objects (chair, computer, ball, book, phone), and outdoor objects (car, ship, airplane, house). The images included in the benchmark cover a broad consideration of complex scenes, where, for example, the target object may co-occur in the same image at varying scales and/or lighting conditions.

For the text domain datasets, 100 randomly selected movie reviews from each positive and negative classes of *IMDB* dataset are selected. Similarly, 100 randomly selected text documents (with the headers removed from samples) from the *20 Newsgroup* dataset are selected from two categories of medical (*sci.med*) and electronic (*sci.elect*).

3.2 Annotation Interface and Procedure

To generate multi-layer human-attention explanations, we ask annotators to provide their interpretations of the salient features that are most meaningful for the specific class from the data set. Each sample is annotated by 10 unique annotators recruited from Amazon Mechanical Turk (AMT). Recruitment advertisement for Human Intelligence Task (HIT) required participants to have at least 1000 previously approved HITs in AMT platform with the HIT approval rate of above 95%. Recruited participants were walked through a training slideshow of the task instructions and interface controls at the beginning of their HIT. As a control, each training slide was displayed on screen for two seconds before participants were able to continue to the next slide. Afterward, they were asked to agree to the IRB approved information sheet for data collection, and continued to a set of 12 images or documents for annotation. Participants were paid $0.40 for the image and text annotation HITs to reach an average hourly pay rate of $10 an hour.

We designed two fundamentally similar human annotation interfaces to capture human feedback for all image and text datasets. Annotators used an interface with basic annotation tools in which each document or image was presented individually. Task instructions prompted participants to select relevant regions or words that they found "most representative" of the target object or topic. Each annotation HIT started with the same two samples to serve as the attention check and help the annotator adjust to the interface and task. These were then followed by 10 samples from the main validation set. For image annotations, the annotators were specifically asked to use their mouse to lasso "salient area(s) that explain the target *"object"* in the image". Similarly, for text annotations, participants were prompted to select relevant words in text documents that they felt most representative of the target topic or class. For example, for the movie review IMDB dataset, the annotators were explicitly asked to "select words and phrases which explain the positive (or negative) sentiment of the movie review".

3.3 Data Processing and Storage

To generate multi-layer feature masks from multiple user annotations, we run a union operation on all individual annotation that displays what areas are most frequently selected by the annotators. Figure 1 presents examples of resulting human-attention masks for different images. Although specified in annotation task instructions, we also applied the exact segmentation mask of the target object's true pixels (only for image datasets) to remove the impact of participants' imprecision or hand jitter that might have included the background pixels. The exact segmentation masks for images are created by two authors and included in the benchmark. Human attention masks for image datasets are stored as grayscale masks

with the same size as the original images. The human attention masks for text datasets are JSON objects with lists of index-word pairs with human-attention scores in the range of 0 to 1.0. We did not perform any feature filtering for text annotation samples. The benchmark is stored in a public domain and free for research use.

4 EVALUATION EXPERIMENTS

In this section, we present multiple evaluation experiments on image data to validate the proposed benchmark with empirical results. These experiments compare three baselines: 1) human-attention mask (our approach) as the ground truth, 2) segmentation mask as the ground truth, and 3) human-judgment rating for evaluating model saliency explanations. Our goal is to understand the relationship between the three evaluation methods and communicate the benefits of the proposed benchmark over other common evaluation methods in the literature. We limit our experiments to the image benchmark and the series of experiments are based on a subset of 100 validation samples from the two classes of cat and dog in PASCAL VOC dataset. Grad-CAM [33] technique is used to generate saliency maps from a VGG-19 [35] image classifier. The VGG network is pre-trained on ImageNet-1k and tuned on PASCAL VOC for the purpose of this evaluation.

Evaluation Criteria: We used pixel-wise Mean Absolute Error (MAE) between the model saliency score map and the ground truth mask as the quantitative measure for error in model explanations. We also looked into False Positive (FP) and False Negative (FN) saliency explanation errors individually. We calculate FP saliency error as pixel-wise MAE for the model saliency map scores outside the object's segmentation mask (i.e., selecting background pixels) as a representation of **explanations correctness**. We calculated FN error as the pixel-wise MAE for model saliency map scores inside the ground truth mask (i.e., non-selected target pixels) to represent **explanations completeness**. In the following subsections, we review details and share evaluation results from two evaluation methods.

4.1 Comparison to Segmentation Mask

In the first evaluation experiment, we compare our proposed human-attention benchmark (multi-layer feature mask) with the segmentation mask (single-layer feature mask) as the evaluation ground truth for the set of saliency maps from Grad-CAM technique. Given the lack of granularity for distinguishing important features in the segmentation mask, we hypothesize that the two baselines would result in different evaluation scores for the same set of inputs.

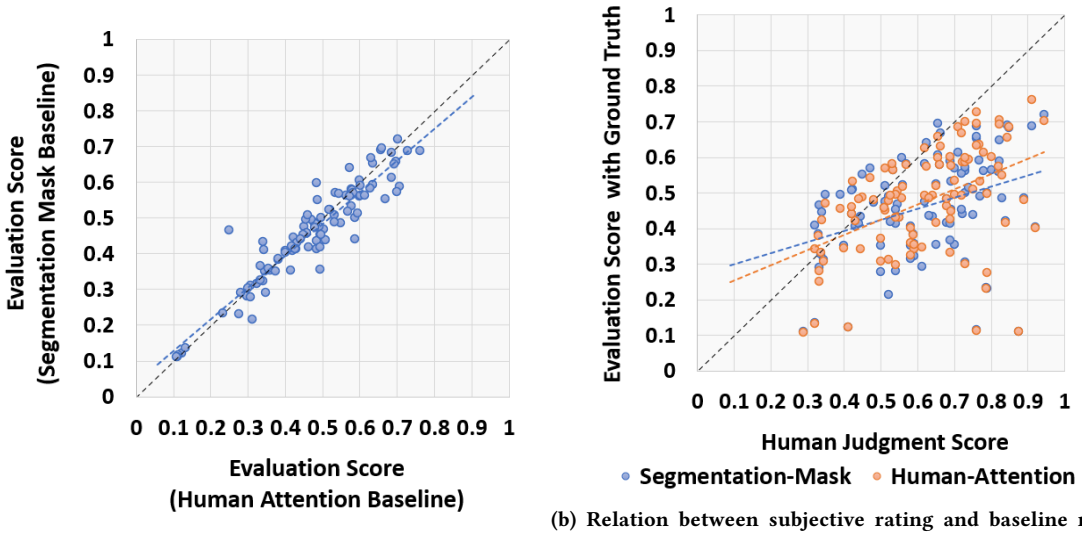
Intuitively, the difference between the two baselines is that unlike the segmentation mask, which scores all target features equally, the human-attention mask gradates the "salient" features more than others. To identify the difference between the two evaluation baselines, we calculate evaluation scores using both baselines for direct comparison. Specifically, we first normalize both ground truth masks and model saliency maps and then calculate the pixel-wise MAE error between the model saliency map and the ground truth baseline. For example, a saliency map identical to its human attention mask results in zero MAE error. In the opposite situation, with cases having no overlap between the ground truth mask and the model saliency map, the MAE error would be 1.0. Note that MAE is a threshold agnostic metric that—unlike Intersection over Union (IoU)—does not require choosing the τ hyperparameter for generating objects' binary masks or bounding boxes, see [4] for more discussion. Also, evaluating the saliency score map (without converting to a binary mask) retains the granular information in the model explanation.

We chose correlation analysis followed by the test for homogeneity of regression slopes to compare the evaluation results for the two baselines. We hypothesize that if the two measures are equal, then the test should find the regression slopes to be homogeneous. Figure 2-(a) shows the scatter plot of evaluation scores (1.0 − MAE) between human-attention and segmentation mask baselines. A Pearson correlation test shows that the two evaluation scores are statistically significantly ($r = 0.896$, $p < 0.001$) correlated, as expected. Using a linear regression test, we find a regression slope of $w = 0.896$ and intercept of $b = 0.48$. As seen in Figure 2-(a), this weight and bias result in different evaluation scores between the two ground truth, especially in the higher and lower range of scores. To examine the statistical significance of the difference between two ground truth evaluations, we use an ANCOVA test with a custom model to the test for homogeneity of regression slopes between the calculated regression model and the ideal of slope 1.0 with a zero intercept. The test for homogeneity of regression slopes fails with a significant difference ($p < 0.001$) between the two lines indicating that the two evaluation baselines are not equal. This indicates that the proposed human-attention baseline contains additional information as compared to the objective binary ground truth mask.

4.2 Comparison to Human Judgment

In the second evaluation experiment, we compare explanation evaluation scores using the two ground truth baselines with the human ratings of explanation goodness. Subjective human ratings of the model explanations are commonly used as a direct approach for evaluating machine learning explanations by providing a numerical rating of explanations goodness using a simple quantitative measure such as Likert scales. However, subjective measures typically lack precision and may include user bias. We hypothesize that results from human-judgment scores will be significantly different for both (human-attention mask and object segmentation mask) ground-truth evaluations. We use the same subset of images and saliency map explanations from Grad-CAM technique similar to the previous section for the purposes of this human-subjects study. Figure 3-(Top) shows examples of heatmap overlays from the Grad-CAM technique used in the user study.

4.2.1 Human Judgment Interface and Data Collection. We designed a simple interface to collect user feedback about the quality of heatmap overlays from the Grad-CAM saliency explanation technique. The participants started by reading task instructions followed by a series of images to rate. Given an image from the test set, the target classification, and a heatmap overlay, participants were instructed to "review and rate the heatmaps which explain what parts the AI used to make its classification decision" and for each image they were specifically asked "Please rate 'how good' the AI is explaining the "object" in this image" in which the word "object" was substituted with the target in each image. A total of 200 unique



(a) Relation between two ground truth measures

(b) Relation between subjective rating and baseline measures

Figure 2: Comparison of averaged evaluation scores (1.0− MAE) between two ground truth baselines and human judgment ratings for each image sample. Evaluation scores are not normalized and the black dashed lines shows the ideal regression line with the slope equal to 1.0 and intercept of zero. (a) Scatterplot of evaluation scores based on segmentation mask (vertical axis) and human-attention mask (horizontal axis). (b) Scatterplot of evaluation score based on two ground truth baselines and human judgment rating.

Figure 3: Examples of heat-map overlay of saliency maps using (Top) Grad-cam [33] and (Bottom) LIME [30] techniques used in the user study for human judgment.

participants' were recruited from Amazon Mechanical Turk and paid $0.20 per HIT to review and rate 14 images on the scale of 1–10. The first four image ratings were identical for all participants and used as training and attention check examples; these early images were discarded for data collection.

4.2.2 Results: We now compare evaluation results from the user study with both ground truth measurements. Figure 2-b shows a scatterplot of the evaluation results between human judgment ratings and two ground truth scores (1.0 − MAE) from objects' segmentation masks and human attention masks. The two regression lines for human-attention ground truth (in orange) and segmentation mask (in blue) show both baselines produce different evaluation scores from the user rating scores. To test for the statistical significance of observed differences, we first normalize user ratings across participants by subtracting each participant's mean rating. Then, we use a Pearson's correlation test and linear regression test to compare the human judgment rating scores and the two ground truth scores. The user ratings show a moderate-strength correlation with human-attention baseline ($r = 0.428$, $p = 0.002$) and small correlation with object segmentation baseline ($r = 0.268$, $p < 0.001$). We also observe signs of user bias, noting that none of the participants rated any of the saliency map instances in the test set below 3-stars even though there are multiple examples with scores below 0.3 for both ground truth evaluation types. These cases were specifically from the examples with multiple occurrences of the target object in which the saliency map was only pointing to one of the target objects. This could potentially indicate a side effect of lower user attention in reviewing cases with incomplete saliency explanations.

To compare measurements between evaluation approaches, we run a linear regression analysis and find that the segmentation mask scores fit with a slope of $w = 0.313$ and intercept of $b = 0.268$ (Figure 2-b, blue trend line), and the fit for human-attention mask scores has a slope of $w = 0.428$ and intercept of $b = 0.210$ (Figure 2-b, orange trend line). Note that the difference between the two linear regression models' slopes with the ideal slope of 1.0 is higher with the segmentation-mask baseline. To examine the statistically significant difference between the measures, we use ANCOVA with a custom model to test for homogeneity of the regression slopes between the two regression models as well as between the calculated regression model and the ideal of slope 1.0 with zero bias. The homogeneity test fails with a significant difference of $p < 0.001$ between the two regression models and the

ideal line. The analysis indicates the subjective measurement of explanations goodness produces significantly different results from both objective ground truth measures.

4.3 Human Biases in Rating

Next, we explore the human judgment evaluation results to find other possible external or internal factors that could affect participants' subjective ratings. For example, human judgment ratings may include user biases toward visual appearance or completeness of saliency maps resulting in biased ratings. We reviewed and compared the results from user studies for rating Grad-CAM and LIME explanations to identify possible human biases toward the visual appearance of saliency explanations. Also, we reviewed the results to assess possible participants' biases toward model explanation FP and FN error types.

To evaluate the effect of the visual appearance of saliency explanations, we compare participants' ratings of saliency map explanations from LIME [30] technique to Grad-CAM explanations on the same subset of images and the same classifier. We run a new user study to collect participants' subjective ratings of LIME explanations. The saliency explanations from the LIME technique (Figure 3-(Bottom)) are visually more chunky and pixelated (mainly due to the use of superpixels in LIME's pipeline) compared to smooth class activation maps from Grad-CAM technique (Figure 3-(Top)).

Figure 4-(a) shows two linear regression models to compare participants' ratings of the two explanation techniques. We find the slope of $w = -0.428$ and intercept of $b = 0.789$ for the user ratings on samples with LIME saliency map (Figure 4-(a) green trend line) and slope of $w = -0.607$ and intercept of $b = 0.947$ for samples with Grad-CAM saliency map (Figure 4-(a), yellow trend line). We would have expected to see the similar regression slopes between the two groups if the users were evaluating both saliency map explanation types similarly. However, the test for homogeneity between the two regression slopes shows a significant difference ($p < 0.001$) between the two model error types. This indicates that users rated saliency maps from the two techniques differently (an indication of possible bias) inconsistent with the ground truth evaluation score (Figure 4-(a), y-axis) for each set of samples.

We then analyze participants' rating behavior with respect to different explanation error types. We first divided the samples for the test set into two groups with high FP explanation errors (when the model is looking at background pixels) and high FN explanation errors (when the model is missing foreground pixels). Using linear regression models, we find the slope of $w = -0.121$ and intercept of $b = 0.265$ for the samples with FP explanation error score (Figure 4-(b) yellow trend line) and slope of $w = -0.306$ and intercept of $b = 0.525$ for samples with high FN explanations error score (Figure 4-(b), green trend line). We would have expected to see the similar regression slopes between the two groups if the users were evaluating both saliency error types similarly. However, the test for homogeneity between the two regression slopes shows a significant difference ($p < 0.001$) between the two explanation error types. This indicates that users pay less attention to FP explanation errors and in turn, are more critical for FN explanation errors. Looking at image samples from the user study, there are several examples in which the target object was on a smaller scale and the model saliency map included more background pixels. Therefore, results revealed user biases on different explanations' error types in their rating of explanation.

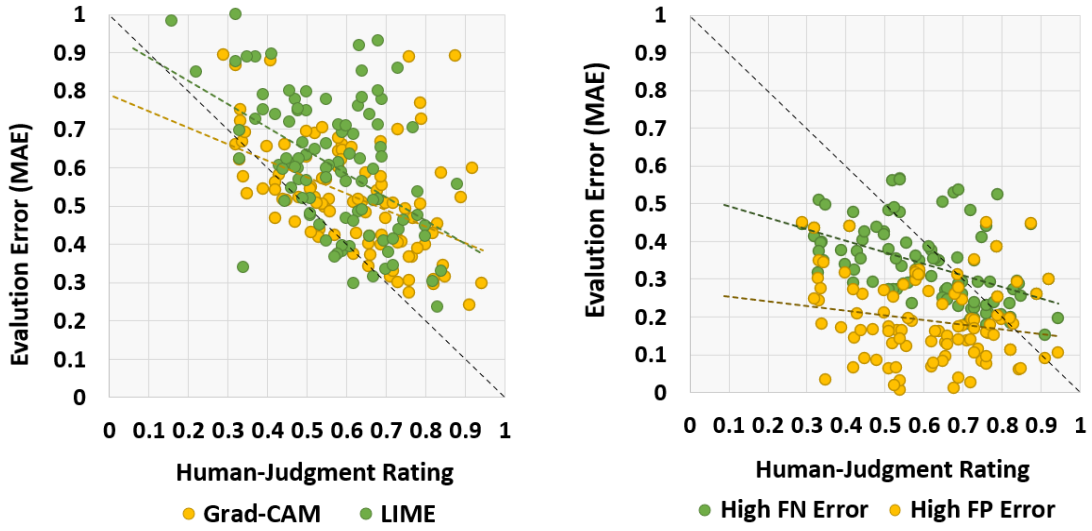
5 DISCUSSION

In this section, we review and discuss the evaluation experiments and open problems around model explanation evaluation. The evaluation experiment results showed that the human-attention benchmark has allowed for a higher level of granularity in the evaluation of saliency maps and also reflects how human attention to certain features is different from the single-layer object's segmentation map. As compared to the human judgment rating evaluations, we also observed signs of participants' bias in their ratings.

5.1 Implications of Results

We ran human-subject experiments to understand the relationship between the subjective and objective evaluations of saliency explanations by inquiring two research questions. With respect to **RQ1**, although the evaluation results from the three methods had positive correlations, statistical testing revealed significant differences among them. The difference in scores was mainly due to the clear non-uniform distribution of feature importance in human attention masks while the segmentation mask weights are uniformly distributed for all features (e.g., pixels, words). Thus, it appears that users consider certain areas of an image or words in a paragraph as being key components of a "good" explanation and adjust their ratings according to that component's inclusion in the explanation. Although the results suggest that both the segmentation mask and the human-attention baselines generate highly correlated explanation errors, there are statistically significant differences. Specifically, the human-attention baseline may better reflect human interests and allow for a more accurate evaluation of a model explanation's correctness (i.e., via FP error) and completeness (i.e., via FN error) with respect to human reasoning and attention to features. For example, in annotations of living things, users were more likely to select facial features as important while the segmentation mask captured any pixels related to the living thing with a uniformly weighted single-layer mask. This is reflected in the human-judgement evaluation results, where participants' ratings of explanations were higher for the human-attention baseline rather than the segmentation mask baseline. Further, using human-attention as the evaluation baseline could in turn lead to designing types of machine learning explanations that are closer to human rationale and more acceptable to end users. Note that we did not explicitly ask participants to rate their trust in the model or its explanations, but "how good" the model is explaining the target object in images. However, more research is needed to expand the quantitative evaluation results from our human-attention baseline with respect to different user trust factors.

Regarding **RQ2**, our experiments revealed user biases toward explanations' visual appearance and error types in their rating of explanations. Specifically, participants treated model saliency explanations from LIME technique with lower rating compared to the Grad-CAM explanations. Also, participants paid less attention to FP explanation errors (i.e. correctness of explanations) and instead



(a) Subjective and ground truth evaluation scores from LIME and Grad-CAM Explanations.

(b) Subjective and ground truth evaluation scores for samples with high FP and High FN explanation MAE error.

Figure 4: Discrepancies between averaged human judgment rating of saliency explanations and human-attention baseline evaluation. Evaluation scores are not normalized and the black dashed lines shows the ideal regression line with the slope equal to -0.5 and intercept of zero. (a) Participants evaluate saliency explanations from LIME and Grad-CAM differently. (b) Participants evaluate saliency explanations' FP error (model looking at background pixels) differently than FN errors (model not looking at target pixels).

were more critical for FN explanation errors (i.e. completeness of explanations). This result shows the importance of taking user biases into account when interpreting user study results involved with machine learning explanations. Our findings are in accordance with previous research indicating unjustified user trust and understanding of machine learning explanations, e.g., [32] and [18]. However, when comparing to these research, we emphasize that studying and measuring user bias could be dependent to the user's task and level of human expertise. Our study results are bounded to the generic user review and rating task and conducted in a crowdsourced platform, thus inheriting certain limitations of non-expert participants performing under unconstrained settings. A similar example is the study by Buccinca et al. [3], which shows that intermediate measures like user trust and acceptance may not correspond to human-AI performance measures in deployed XAI systems since users ultimately behave differently in decision-making situations.

5.2 Values and Limitations of Quantitative Evaluation of Explanations

Our benchmarks are aimed to provide a quantitative evaluation method for machine learning explanations and bridge the trade-off between objectivity and subjectivity of different evaluation methods. Despite the values of quantitative methods, there exist limitations to the quantitative framing of an explanation's trustworthiness. In the following, we review these values and limitations from two perspectives.

Reproducibility of Results: First, one way to categorize different evaluation measures is by the reproducibility of their results. As implemented in our work, users' subjective ratings of explanations could inform results for the goodness of model generated explanations. Ribeiro et al. [30] presented a case for correction of model explanation in which users reject wrong features and add new features for quantitative evaluation of model explanations. However, although these methods can provide detailed insights, subjective user feedback is not necessarily reusable for new models and interpretable techniques, as different explanations may require new human review. This limitation indeed exists in studies for evaluating XAI systems in different applications and domains [7], including tasks and scenarios concerned with the fairness of the decision-making system. A secondary limitation of creating human-attention benchmark is in the annotation cost for multi-level human explanation masks. Asking multiple users to select areas they feel best explain a classification is time consuming and compiling those annotations can be challenging without an accessible system. However, annotation costs (and the associated imperfections) may be justifiable when compared to repeated novel rounds of user evaluation, as the iterative process of design and evaluation for machine-learning based systems typically requires multiple rounds of training and testing. Therefore, our human attention benchmark can potentially reduce evaluation costs over design cycles by providing a baseline to evaluate new implementations.

Ground Truth Objectivity: Although objective evaluations that utilize an element of ground truth provide quantitative and reproducible results, they lack the inherent guidance in human feedback that can provide a finer-grained evaluation on different aspects of explanation goodness and user trust. Thus, one should consider these limitations when interpreting evaluation results and drawing

conclusions from quantitative evaluation of machine learning explanations. Additionally, studying and measuring user bias could be sensitive to the data domain, user task, and user expertise. In this regard, our proposed benchmark and study results are bounded to the generic data collection task of user review and rating in a crowdsourced platform. Therefore, a limitation of our work is that user annotations of data and acceptance ratings may not be transferable to other proxy tasks or user types, as noted by others [3].

6 CONCLUSION AND FUTURE WORK

We present a new model-explanation evaluation benchmark for multiple datasets in image and text domains. Our benchmark is designed for quantitative evaluation of saliency map explanations based on human attention. This human-grounded benchmark enables fast, replicable, and objective execution of evaluation experiments for saliency explanations. We studied the relationships and trade-offs between two different human-grounded evaluation approaches (i.e., single-layer annotation mask and human subjective feedback) to present the efficiency of the proposed human attention baseline. Our study results indicated the difference between threshold-agnostic evaluation with a human-attention baseline as compared to previous methods with binary ground truth masks and labels. Our experiments also revealed user biases on different explanations' visual appearance and error types in the subjective rating of explanations.

In our future work, we plan to study annotators' behavior on objects of different nature and learn general patterns in human attention. This could potentially help to standardize annotators' perception of explanation when performing the annotation task. Lastly, we are interested in examining the use case of the human-attention benchmark for tuning models to improve prediction rationale and its effects on explanation quality.

ACKNOWLEDGMENTS

This research is based on work supported by the DARPA XAI program under Grant #N66001-17-2-4031 and NSF award #1900767.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 9505–9515.

[2] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. *arXiv preprint arXiv:2002.00772* (2020).

[3] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. *arXiv preprint arXiv:2001.08298* (2020).

[4] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. 2020. Evaluating Weakly Supervised Object Localization Methods Right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3133–3142.

[5] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017), 90–100.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. IEEE.

[7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[8] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. 2018. Towards explanation of dnn-based prediction with guided feature inversion. In *KDD*.

[9] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* 111, 1 (Jan. 2015), 98–136.

[10] David Gunning. [n.d.]. Explainable Artificial Intelligence (XAI). ([n. d.]), 36. DARPA report.

[11] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. Trust in Automation. *IEEE Intelligent Systems* 28, 1 (Jan. 2013), 84–88. https://doi.org/10.1109/MIS.2013.24

[12] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. *arXiv:1812.04608 [cs]* (Feb. 2019). http://arxiv.org/abs/1812.04608 arXiv: 1812.04608.

[13] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*. 9734–9745.

[14] T. Huang, K. Cheng, and Y. Chuang. 2009. A collaborative benchmark for region of interest detection algorithms. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 296–303. https://doi.org/10.1109/CVPR.2009.5206765 journalAbbreviation: 2009 IEEE Conference on Computer Vision and Pattern Recognition.

[15] Adam J. Johs, Denise E. Agosto, and Rosina O. Weber. 2020. Qualitative Investigation in Explainable Artificial Intelligence: A Bit More Insight from Social Science. *arXiv e-prints* 2011 (Nov. 2020), arXiv:2011.07130. http://adsabs.harvard.edu/abs/2020arXiv201107130J

[16] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[17] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 59–67.

[18] Himabindu Lakkaraju and Osbert Bastani. 2019. " How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. *arXiv preprint arXiv:1911.06473* (2019).

[19] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *ICML*. 331–339.

[20] Tao Lei, Regina Barzilay, and Tommi S Jaakkola. 2016. Rationalizing Neural Predictions. In *EMNLP*.

[21] Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded Evaluations of Explanation Methods for Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5198–5208.

[22] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *CVPR*.

[23] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015

[24] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[25] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2019. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv preprint arXiv:1811.11839* (2019).

[26] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.

[27] Ahmed Osman, Leila Arras, and Wojciech Samek. 2020. Towards ground truth evaluation of visual explanations. *arXiv preprint arXiv:2003.07258* (2020).

[28] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019* (2019).

[29] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*. ACM.

[31] Philipp Schmidt and Felix Biessmann. 2019. Quantifying Interpretability and Trust in Machine Learning Systems. *arXiv preprint arXiv:1901.08558* (2019).

[32] Johannes Schneider, Joshua Handali, Michalis Vlachos, and Christian Meske. 2020. Deceptive AI Explanations: Creation and Detection. *arXiv preprint arXiv:2001.07641* (2020).

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 618–626.

[34] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM*

Transactions on Interactive Intelligent Systems 10, 4 (Dec. 2020), 1–31. https://doi.org/10.1145/3419764

[35] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[36] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[37] Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. 2018. Beyond sparsity: Tree regularization of deep models for interpretability. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[38] Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*. 260–267.

[39] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.