



Is this AI trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust

Cheng Chen
School of Communications
Elon University
Elon, North Carolina, USA
cchen8@elon.edu

S. Shyam Sundar
Media Effects Research Laboratory
Pennsylvania State University
University Park, Pennsylvania, USA
sss12@psu.edu

ABSTRACT

To promote data transparency, frameworks such as *CrowdWorkSheets* encourage documentation of annotation practices on the interfaces of AI systems, but we do not know how they affect user experience. Will the quality of labeling affect perceived credibility of training data? Does the source of annotation matter? Will a credible dataset persuade users to trust a system even if it shows racial biases in its predictions? To find out, we conducted a user study ($N = 430$) with a prototype of a classification system, using a 2 (labeling quality: high vs. low) \times 4 (source: others-as-source vs. self-as-source cue vs. self-as-source voluntary action, vs. self-as-source forced action) \times 3 (AI performance: none vs. biased vs. unbiased) experiment. We found that high-quality labeling leads to higher perceived training data credibility, which in turn enhances users' trust in AI, but not when the system shows bias. Practical implications for explainable and ethical AI interfaces are discussed.

CCS CONCEPTS

• Human-centered computing \rightarrow Empirical studies in HCI.

KEYWORDS

training data credibility, data labeling quality, labeling source, trust in AI, algorithmic bias

ACM Reference Format:

Cheng Chen and S. Shyam Sundar. 2023. Is this AI trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3544548.3580805>

1 INTRODUCTION

Artificial intelligence (AI) is embedded in almost every aspect of our lives and has been increasingly used in high-stakes domains, ranging from health [28] and employment [35] to finance [42] and justice [20]. However, AI systems are often not transparent enough to explain why a certain decision was made. This “black box” issue,

coupled with increasing reliance on AI, has raised serious concerns over the biases that an AI system may introduce into its predictions.

As a concept in machine learning, bias has multiple meanings. We define algorithmic bias as an unintended consequence of machine learning, in which the model favors the dominant group in society over a minority group of individuals, and thereby violates the principle of group fairness, i.e., equal opportunity of positive classification outcomes regardless of race and other factors [56].

While algorithmic biases can be introduced in almost every stage of machine learning, researchers have argued that racial and other types of biases in algorithms are primarily caused by the nature of training data [4, 11, 12]. Prior research has shown that if the machine learning model is trained on non-representative training data, AI fairness and accuracy are at high risk, with potential to undermine users' trust in AI. This issue raises an important question for designers of AI interfaces: how to communicate training data credibility to end users, so that they can evaluate for themselves whether there are any biases embedded in the AI system, and accordingly calibrate their trust toward it?

Previous studies have explored several ways to convey training data credibility to lay users. Anik and Bunt [2] found that showing training data demographics can help users identify biases in machine learning systems. Similarly, Chen [9] pointed out that displaying racial diversity in either training data or labelers' backgrounds before actual AI interaction is effective in shaping users' expectations and trust in AI. Given that most supervised machine learning needs to be trained on human-labeled data and the proposal of having *CrowdWorkSheets* to document the labeling practice [13], this study explores whether showing labeling quality, i.e., accuracy of the labeled data, would lead to higher perceived training data credibility, and how the perception of training data credibility may influence different aspects of trust in AI, namely cognitive trust, affective trust, and behavioral trust.

Additionally, since labeling is often undertaken by crowd workers, a related question is whether the crowd as a source of training data is seen as credible by users. Do they see themselves as being similar to crowd workers, and if so, would they perceive the training data to be more credible, just as if they themselves participated?

Another important question pertains to the extent to which user perception of training data credibility can help shape trust in AI when its performance is clearly biased. Will biases be seen as exceptions rather than the rule, or will they undermine the credibility of the training data?

In this study, we are guided by two broad research questions: (1) *How will labeling quality (independent variable) influence the perceived training data credibility (mediating variable) and further*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '23, April 23–28, 2023, Hamburg, Germany

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9421-5/23/04...\$15.00
<https://doi.org/10.1145/3544548.3580805>

affect users' trust in AI (dependent variable)? And, (2) How will labeling source and AI performance moderate this mediation model, if at all?

We conducted a user study ($N = 430$) to address these questions. We designed a prototype of an ostensibly AI-based facial expression classification system. We first showed participants the source of the labeled data with variation in who provides the labels, a crowd worker or the user. After this, we presented participants with a snapshot of the labeled data, displaying high or low quality of labeling, after which we measured perceived training data credibility. We then randomly assigned participants to one of three AI-performance conditions: biased, unbiased, or no performance. Users' cognitive, affective, and behavioral trust in AI were assessed as potential outcomes.

By examining the mediation effect of labeling quality on trust in AI through the perceived training data credibility and the moderating roles of labeling source and AI performance, this study contributes to explainable AI (XAI) research in the following ways: first, by showing labeling quality through a snapshot of the labeled data, we propose a novel way to communicate training data credibility, which we demonstrate has positive effects on AI trust. Second, the positive mediation effect of labeling quality on AI trust shows the value of early communication in shaping user perceptions and trust in AI, which broadens current practices of XAI research that is dominated by post-hoc explanations (see examples [6, 16, 27, 31, 48, 53, 59]). Third, we find a nuanced interaction effect between AI performance and perceived training data credibility on AI trust, which specifies the conditions under which perceived training data credibility can positively shape users' trust in AI.

2 RELATED WORK

2.1 Explainable AI

The concern over the black-box issue of AI systems and the need for transparency has given rise to the field of XAI. The idea behind XAI is to clarify the reasoning behind machine learning models and to be understandable to people [3, 30, 44]. Consistent with this goal, practitioners have provided global, local, and counter-example explanations to help people understand algorithmic decisions after the fact.

While post-hoc explanations are promising to reshape users' trust in AI, scholars have argued that explanations after AI performance is not a good approach to combat algorithmic bias because harms have already been inflicted by the deployment of AI [22, 37]. Instead, a better approach is to communicate earlier about training data information of a machine learning model, so that users can assess for themselves if there are any potential sources of bias before AI interaction. This approach is echoed by the recently proposed documentation paradigm in XAI community, which advocates detailed reporting of training data and performance characteristics of machine learning models at the outset of model development [5, 19, 22, 41, 45].

Previous studies have shown that early communication of training data information can help users identify bias in machine learning models [2] and adjust their expectations and trust in AI [9]. This line of research is focused on displaying training data information to users and examining its effect on user perceptions and trust in

AI. However, few have investigated how communicating training data information may influence training data credibility. We argue that it is important to take into account user perceptions of training data credibility because that can shed light on the theoretical mechanisms underlying the positive effect of showing training data information on user trust in AI.

2.2 Training Data Credibility

Based on the construct of credibility, which is defined as believability or trustworthiness of a source [34], we conceptualize training data credibility as the degree of trustworthiness attributed to the training data by end users. Training data credibility could be shaped by many factors in training data provenance and preparation, such as the gender and racial composition of the raw training data [2, 9] and labelers' backgrounds [9]. Given that supervised machine learning is most often trained on human-labeled data and scholars have advocated for having *CrowdWorkSheets* at the beginning of model development [13], this study tests how revealing the labeling practice, especially the quality of labeling, may affect user perceptions of training data credibility.

2.3 Labeling Quality

There are many ways to describe the quality of labeling. We focus on the accuracy of labeling, namely the extent to which the labels match the raw training data. Taking an AI-based facial expression classification system as an example, if all happy faces are labeled as happy and all sad faces are labeled as sad, the labeling quality is high as the labels accurately reflect the facial images. By contrast, if the labels and training data are mixed, such that some happy faces are labeled as sad and some sad faces are labeled as happy, the labeling quality is said to be low. Given that accuracy is a straightforward way to evaluate AI performance and prior studies have shown the positive effect of AI accuracy on user perceptions and trust in AI [43], we predict that showing accurate labeling, i.e., a high-quality labeling practice, will result in higher perceived training data credibility.

2.4 The Moderating Role of Labeling Source

Considering that the labeling task for AI systems is often outsourced to crowd workers, who provide meaningful tags for machines to learn, an important question is the extent to which the crowd-sourcing nature of labeling would influence perceived training data credibility.

Huang and Sundar [24] argued that the evaluation of credibility in the context of crowd-sourcing depends on whether users are reminded that they themselves are part of the crowd. Their evaluation differs based on whether the self or the crowd is more salient in their mind, which dictates the kind of cognitive path or thought process they follow. If users go through the self-as-source path, they perceive the content to be more trustworthy and credible. By contrast, if users follow the others-as-source path, they tend to have a lower perception of content credibility because they do not trust "the crowd."

Notably, there is a difference in how one becomes the source of the labeled data — perceptually or behaviorally. The theory of

interactive media effects (TIME; [51]) and its extension to human-AI interaction (HAI-TIME; [50]) differentiate the perceptual effect from the action effect on AI trust by proposing two routes, namely the cue route and the action route. If users see the presence of the labeling opportunity without actually participating in labeling, they are following the cue route, which may have perceptual effects on AI trust, driven by cognitive heuristics, such as the ownness heuristic (i.e., I had a hand in the training data, therefore the AI system must be good) and the control heuristic (i.e., I felt in charge of data labeling, so the training of the model must be good). If users do participate in data labeling, they are said to be following the action route. Their heightened involvement is likely to reshape their trust in AI, probably due to an enhanced sense of agency [50]. As a result, when individuals feel involved in a task and responsible for the outcome, they tend to engage in self-serving bias [36], which may generate more favorable attitudes towards the labeling task, thus leading to higher perceived credibility of the training data.

Based on this rationale, we hypothesize that the self-as-source cue (compared to others-as-source) will trigger the ownness heuristic and the control heuristic, which in turn will enhance perceived training data credibility. In addition, we predict that the act of participating in the labeling process will enhance one's sense of agency, which will increase perceived training data credibility to a greater extent than simply being exposed to the self-as-source cue on the interface.

Considering the positive effect of self being the source, we also expect that labeling source will moderate the effect of labeling quality on perceived training data credibility. That is, the effect of labeling quality on perceived training data credibility will be stronger when users see themselves as the source (cue) compared to when they perceive others as the source. This self-as-source effect will be even stronger when people actually serve as the source by participating in data labeling, i.e., self-as-source (action).

2.5 Trust in AI

Training data is fundamental for machine learning. If users perceive the training data to be credible, they are likely to form higher trust in the AI system. Drawing on the definition of trust in automation [29], we define trust in AI as the extent to which users believe the AI system will take their welfare into account under conditions of uncertainty and vulnerability. Considering that trust is a multi-faceted construct, we focus on three aspects of AI trust, i.e., cognitive trust, affective trust, and behavioral trust, and discuss how each of them may be related to perceived training data credibility.

Cognitive trust is formed when the AI system shows its competence and reliability [26]. Given that training data drives the process of machine learning, if the training data is perceived to be credible, it is likely that the AI performance will be evaluated as competent and reliable, thus contributing to cognitive trust of AI. Hence, we expect a positive relationship between perceived training data credibility and cognitive trust in AI.

Different from cognitive trust, which is knowledge-driven, affective trust is motivated by emotion as it emphasizes the emotional bond between the user and the system [26]. But, affective trust will probably not be affected by perceived training data credibility, as

credibility is more about evaluation of AI competence and reliability rather than the warmth and caring conveyed by the system.

As an outcome of cognitive and affective trust, behavioral trust refers to the willingness to take actions based on the judgment or information provided by the AI system [1]. If the training data is considered reliable, there is very low risk in relying on the AI system for decision making. Therefore, we predict that those who perceive training data to be credible are likely to take the advice provided by the AI system.

Together, we expect perceived training data credibility to be related to cognitive and behavioral trust in AI but not affective trust in AI.

2.6 The Moderating Role of Racial Bias in AI Performance

Training data credibility is not the only factor that determines trust in AI. In fact, AI performance, especially biased performance, has a greater influence on user experience and trust [17, 57, 58]. Based on the notion of group fairness [56], we define racial bias in AI performance as the extent to which the AI system favors the dominant group over minorities in its predictions. Considering the promising role of training data credibility in shaping trust in AI and the negative effect of racial bias in AI performance upon AI trust, a pertinent question is the extent to which perceived training data credibility would interact with racial bias in AI performance in influencing users' trust in AI.

There are two competing theoretical frameworks that can predict the nature of this interaction effect. One is expectancy violations theory [7], which states that disappointment will result when a positive expectation is violated by an undesirable experience, and this violation will result in negative evaluative outcomes [8, 23], such as lower trust in AI in the current study. The other perspective comes from priming theory [54], which posits that the first stimulus can influence the processing of subsequent stimuli. It means that when primed that the training data is credible, users are likely to maintain their trust in AI even though they may encounter a biased AI performance. Given that these two explanations are both theoretically plausible, we predict that AI performance may moderate the relationship between perceived training data credibility and trust in AI.

2.7 Hypothesis and Research Question

In sum, we propose a mediation model in which labeling quality would affect perceived training data credibility, which in turn would influence trust in AI. We also propose labeling source and AI performance as two moderators playing a role at different stages of the mediation model, as shown in Figure 1. To examine the mediation model as well as the moderation effects, we propose the following research question and hypothesis:

RQ1: How will labeling source and AI performance moderate the indirect effect of labeling quality on AI trust through perceived training data credibility?

H1: Labeling quality will influence perceived training data credibility, which in turn will influence trust in AI.

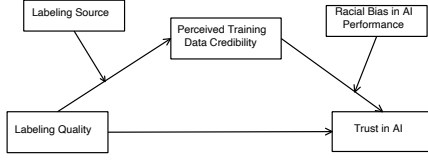


Figure 1: Study Model

3 METHOD

We addressed the hypothesis and research question through a 2 (labeling quality: high vs. low) \times 4 [labeling source: others-as-source vs. self-as-source (cue) vs. self-as-source (voluntary action) vs. self-as-source (forced action)] \times 3 (AI performance: none vs. unbiased performance vs. racially biased performance) between-subjects online experiment. The study was approved by the Institutional Review Board at a large public university and was pre-registered with the Open Science Foundation before accessing the data¹.

3.1 Participants

We recruited a total of 459 participants through CloudResearch [32] in early July 2022. Considering that the study took about 10 minutes to complete based on a pilot test, we paid participants \$1.5 in exchange for their work. The sample size was determined via an *a priori* power analysis to achieve a power of .95, an error lower than .05, and a medium effect size using the *F* test family [18]. After excluding incomplete and inattentive responses, we were left with 430 participants. It is noteworthy that the final sample size is still higher than the sample size requirement for the power of .80 ($n = 372$), which means that we have sufficient power for data analysis.




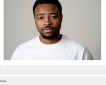


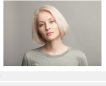
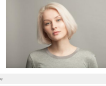

Of the 430 participants, 54.7% ($n = 235$) were females, 44.7% ($n = 192$) were males, and .2% ($n = 1$) were other/non-binary. Also, .5% ($n = 2$) preferred not to reveal their gender. Participant age ranged from 20 to 76 ($M = 44.05$, $SD = 13$). The sample was predominantly White (79.5%), followed by Asian (8.6%), Black or African American (7.7%), Hispanic, Latino or Spanish origin (5.8%), American Indian or Alaska Native (1.6%), Native Hawaiian or other Pacific Islander (.5%), and other races or origins (.9%). The median education level was a bachelor's degree and the median family income was in the range of \$50K to \$75K a year.

3.2 Procedure and Stimuli

Upon consenting, participants were asked to provide their demographics. They were then invited to interact with a prototype of an Emotion Reader AI website, which was introduced as a facial expression classification system, designed to help detect facial expressions through social media images. We told our participants that the AI system had been trained on an open emotion dataset, which had grown to nearly 10,000 facial images. They were told that these images were labeled by volunteers, who provided one of the following tags for each image—joy, sadness, anger, fear, surprise, disgust, and neutral. The algorithms then figured out the patterns by learning from the labeled data. So far, they were told, more than 500 people had participated in data labeling.

¹For more information about the pre-registration, please click here.

Table 1: Manipulation of Labeling Source

Self-as-Source (Cue)	Self-as-Source (Voluntary action)	Self-as-Source (Forced Action)
<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>	<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>	<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>
<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>	<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>	<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>
<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>	<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>	<p>You Can Make a Difference</p> <p>What is the most subtle emotion shown in the image?</p>  <p>OK</p> <p>Cancel</p> <p>Next</p> <p>Previous</p> <p>Done</p> <p>Break the Emotion Reader. It is in the participant phase. The function of data labeling is not available for you.</p>

3.2.1 Manipulation of Labeling Source. After the introduction, we randomly assigned participants to one labeling source condition. They were either (1) seeing how the labeling was done by a crowd worker through a video (others-as-source), (2) seeing the opportunity of data labeling but were not able to use the feature (self-as-source [cue]), or (3) invited to label the data and required to complete the task to proceed (self-as-source [forced action]). To increase ecological validity of the study, we also had a voluntary action condition, in which participants were invited to label the data but had the choice to skip the task. This condition mimics real-life experience as some people may not participate in data labeling even though they are provided the opportunity.

To reinforce the manipulation of self-as-source, we told participants in the voluntary and forced action conditions that their work would be uploaded to the AI system in about an hour. By contrast, we simply thanked participants for “checking the labeling practice” in the others-as-source and self-as-source (cue) conditions. Notably, all participants saw the same set of facial images in all conditions, and the facial expressions in all three images were considered neutral, as shown in Table 1. Sample size across conditions is balanced. We had 107 participants in the others-as-source condition, 106 in the self-as-source cue condition, 107 in the voluntary action condition, and 110 in the forced action condition.

3.2.2 Manipulation of Labeling Quality. Following the source manipulation, we randomly assigned participants to see one level of the ‘labeling quality’ variable: high vs. low. Given that labeling quality was operationalized based on the accuracy of labeling, we had the high-quality condition being 100% accurate, in that all happy faces were labeled as happy, and all sad faces were annotated as

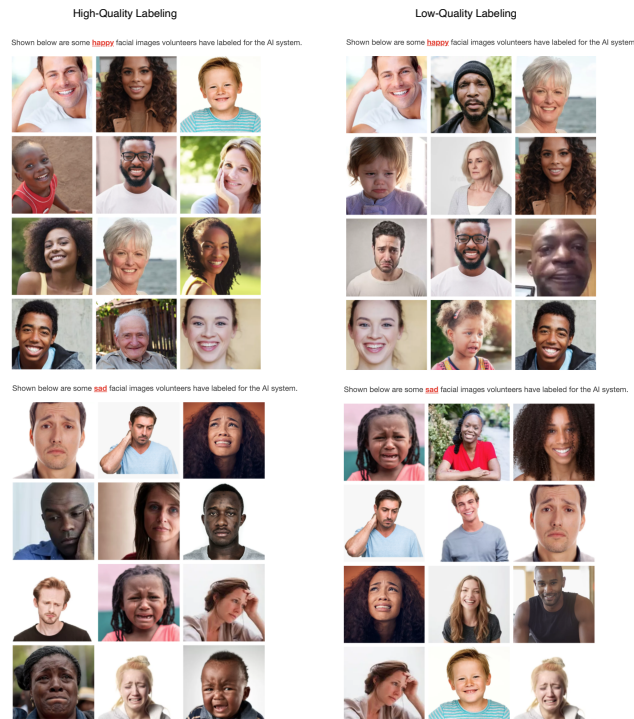


Figure 2: Manipulation of Labeling Quality

sad. By contrast, the low-quality labeling condition was only 50% accurate, which means that half the images were misclassified for both happy and sad facial expressions. The stimuli of labeling quality are presented in Figure 2. The sample size is balanced across high- ($n = 219$) vs. low-quality ($n = 211$) labeling conditions. After this, we measured users' perception of training data credibility.

3.2.3 Manipulation of AI Performance. We manipulated AI performance by randomly assigning participants to one of three conditions: no performance, biased performance, and unbiased performance. In the latter two, we asked participants to view the examples of AI performance, which involves the classification of two Black and two White subject images. Given that racial bias in AI performance was operationalized as the extent to which the AI system favors the dominant group in its predictions, we had the AI system classifying all White subject images with 100% accuracy whereas it classified all Black subject images with 0% accuracy in the "biased performance" condition. The disparity of accuracy showed strong evidence of racial bias in AI performance, as evident in Figure 3. In the "unbiased performance" condition, the AI system performed equally well for both White and Black subject images, i.e., 100% accurate for both White and Black faces. We did not display any examples of AI performance in the "no performance" condition. Again, randomization yielded balanced cell sizes: we had 146 participants in the no performance condition, 144 in the biased performance condition, and 140 in the unbiased performance condition.

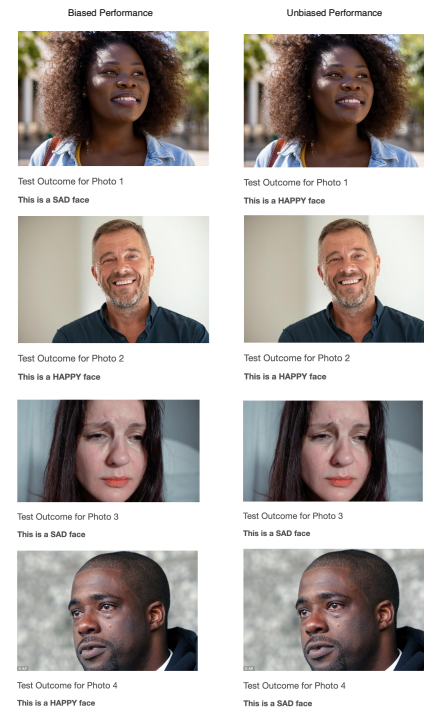


Figure 3: Manipulation of Racial Bias in AI Performance

After they were exposed to the manipulation of AI performance, we measured the dependent variable of interest, i.e., AI trust. We debriefed the participants by the end of the questionnaire by informing participants that there was no AI involved in facial expression classifications. We explained the reason for this deception and also provided resources if participants felt uncomfortable after the study.

3.3 Measures

All items were measured on a 7-point Likert scale, ranging from 1 = strongly disagree to 7 = strongly agree.

3.3.1 Trust in AI. Given that trust is a multi-dimensional construct, we measured cognitive trust (12 items; [25]), affective trust (4 items; [26, 33]), and behavioral trust (3 items; [47]) separately. However, an exploratory factor analysis showed that the 12-item scale of cognitive trust resulted in two distinct factors based on the valence of wording. Thus, we followed the rule of thumb of 60/40 to remove cross-loaded items. That is, we only kept items that showed a factor loading greater than .6 on the primary factor and less than .4 on all other factors. At the end of this procedure, we were left with five negatively worded items. We reverse-coded the items and then averaged them to form an index of cognitive trust, such that the higher value means higher cognitive trust in AI. The indices of cognitive trust ($M = 4.47$, $SD = 1.58$, Cronbach's $\alpha = .92$), as well as affective trust ($M = 2.59$, $SD = 1.37$, Cronbach's $\alpha = .93$) and behavioral trust ($M = 3.06$, $SD = 1.74$, Cronbach's $\alpha = .95$), showed good reliability.

3.3.2 Training Data Credibility. We asked participants the extent to which they agree that the training data of the Emotion Reader AI system is (1) credible, (2) trustworthy, (3) reliable, and (4) dependable. We created an index of perceived training data credibility by averaging the four items ($M = 4.06$, $SD = 2.01$, Cronbach's alpha = .99).

4 RESULTS

4.1 Manipulation Check

After showing the stimuli of the labeling source, we asked participants which condition they encountered in the interaction with the Emotion Reader AI. Four response options were provided: (1) The system showed me how a crowd worker labeled the training data, (2) The system invited me to label the training data, but I can skip it, (3) The system showed me how labeling looks, but the feature was not ready for use, and (4) The system requested me to label the training data, and I have to do it to proceed. A Chi-square test revealed that most participants successfully identified the labeling source condition to which they were assigned: $\chi^2(9) = 747.82$, $p < .001$. However, a majority of participants (96.26%) assigned to the self-as-source voluntary action condition thought that they had to complete the labeling task to proceed even though they were told that the labeling task was voluntary. As a result, all participants in the voluntary action condition labeled the data for the AI system. Given that all manipulations of labeling source were ontologically valid and the purpose of the study is to examine how intrinsic features of the AI medium, i.e., labeling source, affect user perception of training data credibility, we used the manipulated variable in further analysis, following the recommendation from [40].

Regarding the manipulation effectiveness of labeling quality, we asked participants the extent to which they agreed that the classification of happy and sad facial images was (1) accurate and (2) free of error. Results from a one-tailed independent sample t -test showed that people perceived higher accuracy and less error in the high-quality labeling condition ($M = 6.10$, $SD = .94$) than the low quality labeling condition ($M = 1.77$, $SD = 1.41$), $t(363.54) = -30.62$, $p < .001$, $d = -3.64$. Thus, the manipulation of labeling quality was deemed successful.

We used three items, rated on a 7-point scale, to examine the manipulation effectiveness of racial bias in AI performance: (1) The Emotion Reader AI favored White people, (2) The Emotion Reader AI disfavored Black people, and (3) The Emotion Reader AI was racially biased. We created an index by averaging the three items, which is reliable ($M = 2.91$, $SD = 1.90$, Cronbach's alpha = .98). A one-tailed independent sample t -test revealed that participants perceived the AI to be more racially biased in the biased performance condition ($M = 4.07$, $SD = 1.72$) compared to the unbiased performance condition ($M = 1.70$, $SD = .83$), $t(207.11) = 14.84$, $p < .001$, $d = 1.75$. Thus, the manipulation of racial bias in AI performance was also successful.

4.2 Hypothesis Testing

RQ1 asked how would labeling source and AI performance moderate the indirect effect of labeling quality of AI trust through perceived training data credibility. Given that there are two moderators, with one (i.e., labeling source) conditionally influencing

the first stage of mediation and the other (i.e., racial bias in AI performance) conditionally influencing the second stage of mediation, we used Model 21 in SPSS PROCESS Macro [21] to answer this question. Furthermore, We used 5,000 bootstrap resamples and 95% percentile bootstrap confidence intervals (CIs) in each analysis. It is important to note that the effect is said to be significantly different from zero if the upper and lower CIs do not contain a zero between them.

As shown in Table 2, the indirect effect of labeling quality on cognitive trust through perceived training data credibility was true only when the AI showed no performance or unbiased performance, regardless of the labeling source. In other words, if there was a biased result, perceived training data credibility did not assure users' cognitive trust in AI. We present the interaction effect in Figure 5.

The same moderated mediation model was applied to analyze the conditional indirect effect on affective and behavioral trust in AI. As shown in Table 3, labeling quality increased perceived training data credibility, which positively predicted affective trust in AI regardless of labeling source and AI performance. We also found a similar indirect effect on behavioral trust in AI, without significant moderation by either labeling source or AI performance, as presented in Table 4.²

H1 proposed that labeling quality (independent variable) would influence perceived training data credibility (mediating variable), which in turn would be associated with users' trust in AI (dependent variable). Given that this hypothesis states a simple mediation model, in which the effect of the independent variable on the dependent variable is attributable in part to the mediating variable, we used Model 4 in SPSS PROCESS Macro [21] to test this hypothesis. In this mediation analysis, we used 5,000 bootstrap resamples and 95% percentile bootstrap confidence intervals (CIs). Results showed that the proposed mediation effect is statistically significant for all three aspects of AI trust: for cognitive trust, $B = .66$, $SE = .23$, 95% CI: [.22, 1.11], for affective trust, $B = 1.46$, $SE = .20$, 95% CI: [1.09, 1.88], for behavioral trust, $B = 1.97$, $SE = .25$, 95% CI: [1.49, 2.47]. We present the mediation models in Figure 4.

5 DISCUSSION

This study finds that showing labeling quality can shape users' perception of training data credibility, which further influences their trust in AI. This indirect effect is true regardless of who is perceived as the source of labeling, but it is conditional upon AI performance being unbiased or unknown. If an AI shows biased performance, training data credibility does not help salvage user trust.

5.1 Importance of Displaying Labeling Quality on AI Interface

By displaying labeling quality on the interface of AI, this study finds an effective way to communicate training data credibility to end users, which extends the line of research on user perception

²Given that labeling source does not matter when fostering accurate credibility judgment of training data from high vs. low quality of labeling, we did not drill down further to report the main effect of labeling source on perceived training data credibility and trust in AI, although we proposed several hypotheses in our pre-registration.

Table 2: Conditional Effects of Labeling Source and AI Performance on the Indirect Effect of Labeling Quality on Cognitive Trust via Perceived Training Data Credibility

Labeling Source	AI Performance	B	SE	95% CI
Others-as-Source	No	1.35	.27	[.83, 1.90]
Others-as-Source	Biased	.04	.22	[-.40, .48]
Others-as-Source	Unbiased	.66	.21	[-.26, 1.08]
Self-as-Source (Cue)	No	1.76	.34	[1.11, 2.44]
Self-as-Source (Cue)	Biased	.05	.29	[-.53, .63]
Self-as-Source (Cue)	Unbiased	.87	.27	[.34, 1.38]
Voluntary Action	No	1.61	.32	[.98, 2.25]
Voluntary Action	Biased	.05	.27	[-.48, .57]
Voluntary Action	Unbiased	.79	.24	[.31, 1.28]
Forced Action	No	1.40	.29	[.95, 1.99]
Forced Action	Biased	.04	.23	[-.41, .50]
Forced Action	Unbiased	.69	.22	[.27, 1.12]

Table 3: Conditional Effects of Labeling Source and AI Performance on the Indirect Effect of Labeling Quality on Affective Trust via Perceived Training Data Credibility

Labeling Source	AI Performance	B	SE	95% CI
Others-as-Source	No	1.60	.23	[1.17, 2.09]
Others-as-Source	Biased	1.08	.22	[-.69, 1.53]
Others-as-Source	Unbiased	1.29	.23	[-.86, 1.78]
Self-as-Source (Cue)	No	2.09	.27	[1.57, 2.65]
Self-as-Source (Cue)	Biased	1.40	.26	[-.92, 1.95]
Self-as-Source (Cue)	Unbiased	1.69	.27	[1.19, 2.21]
Voluntary Action	No	1.91	.27	[1.39, 2.45]
Voluntary Action	Biased	1.28	.26	[-.80, 1.84]
Voluntary Action	Unbiased	1.54	.26	[1.04, 2.08]
Forced Action	No	1.67	.26	[1.18, 2.20]
Forced Action	Biased	1.12	.25	[-.68, 1.64]
Forced Action	Unbiased	1.35	.25	[-.88, 1.85]

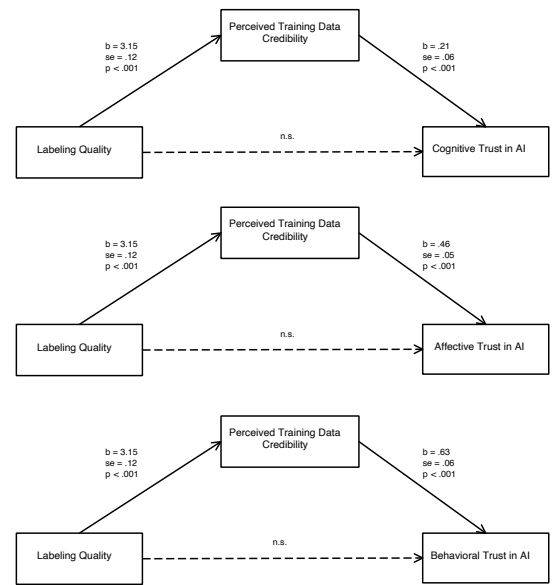
of training data quality. Aside from communicating training data demographics [2] and labelers' racial backgrounds [9], designers can convey the extent to which the labels match the raw training data. The accuracy of labeling can help credibility judgment of the training data, which can further enhance users' trust in AI.

The important role of labeling quality on perceived training data credibility solves the dilemma in a previous study [10], which found that users could not differentiate racially imbalanced from balanced labeled data and perceived them as equally biased compared to an interface showing features underlying facial expression classifications. Findings of this study suggest that the skepticism surrounding labeled data could be mitigated if researchers communicate visually the accuracy of labeling by showing a subset of labeled data on the interface.

Furthermore, displaying labeling quality before users experience the actual performance of the system confirms the value of early communication in shaping user perception and trust in AI. This practice broadens the dominant practice in XAI community, which focuses on providing global (how the AI works), local (how the AI

Table 4: Conditional Effects of Labeling Source and AI Performance on the Indirect Effect of Labeling Quality on Behavioral Trust via Perceived Training Data Credibility

Labeling Source	AI Performance	B	SE	95% CI
Others-as-Source	No	2.32	.27	[1.81, 2.87]
Others-as-Source	Biased	1.39	.25	[-.94, 1.91]
Others-as-Source	Unbiased	1.78	.27	[1.28, 2.31]
Self-as-Source (Cue)	No	3.02	.30	[2.45, 3.66]
Self-as-Source (Cue)	Biased	1.81	.31	[1.23, 2.46]
Self-as-Source (Cue)	Unbiased	2.32	.31	[1.73, 2.93]
Voluntary Action	No	2.76	.31	[2.19, 3.40]
Voluntary Action	Biased	1.65	.30	[1.10, 2.31]
Voluntary Action	Unbiased	2.12	.30	[1.54, 2.71]
Forced Action	No	2.41	.31	[1.82, 3.05]
Forced Action	Biased	1.44	.29	[-.92, 2.07]
Forced Action	Unbiased	1.85	.29	[1.31, 2.47]

**Figure 4: The Indirect Effect of Labeling Quality on AI Trust through Perceived Training Data Credibility**

works for certain individuals), or counter-example explanations (how would the AI work if certain input changes) *after* AI performance (see examples [6, 16, 27, 31, 48, 53, 59]). We argue that, instead of providing only post-hoc explanations, it is equally important to present training data characteristics, such as the labeling quality, prior to AI interaction. This pre-explanation approach can help users assess credibility of the training data for themselves and calibrate their trust in the system accordingly. The positive effect of pre-explanation is echoed by at least one study, which shows

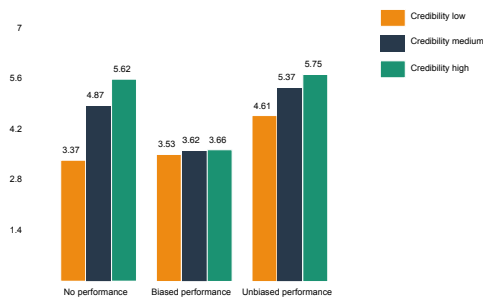


Figure 5: The Interaction Effect of Perceived Training Data Credibility and AI Performance on Cognitive Trust in AI

that displaying racial diversity in training data and labelers' backgrounds via a model card before actual interaction can positively shape users' expectations and trust in AI [9].

To display training data characteristics up front, it requires AI designers and developers to have the training data information handy when designing the interface of AI systems. It would be consonant with the recently proposed documentation approach in XAI community, which calls for releasing information about training data preparation at the outset of model development [19, 22, 37]. Considering that labeling quality can influence AI trust through training data credibility perception, designers should devise creative ways to prepare and present to users a *CrowdWorkSheet* to detail the labeling practice for a machine learning model [13].

5.2 New Concept: Training Data Credibility

We propose a new concept — training data credibility — to measure lay users' perception of the AI system. By asking users the extent to which they perceive the training data to be (1) credible, (2) trustworthy, (3) reliable, and (4) dependable, it specifies how lay users think about the quality of the labeled data instead of the entire AI system. The new measure could be used by not only AI interface designers but also UX researchers to better capture users' perception of training data credibility.

Moreover, findings of the study demonstrate the value of assessing training data credibility perception as it significantly mediates the effect of labeling quality on users' trust in AI. The significant mediating role of perceived training data credibility not only unpacks the theoretical mechanism that drives the perceptual effect on AI trust, but also provides an avenue for designers to adjust users' trust in AI. To increase AI trust, designers should clearly communicate training data credibility to end users, which can be realized by showing labeling quality via a snapshot of the labeled data.

5.3 Labeling Source Does Not Matter When Labeling Quality is Clearly Conveyed

While labeling source was proposed as a moderator on the effect of labeling quality on perceived training data credibility, this study

finds that it does not play a significant role in determining user perception of training data credibility. This finding is not consistent with the line of research on source and credibility [52]. One possible reason is that the manipulation of labeling quality manipulation is too strong, in that it does not leave much room for labeling source to play a role in the model. To better examine the effect of labeling source on perceived training data credibility, future studies could make the labeling quality more ambiguous such that users cannot tell whether the quality of labeling is high or low. This will allow us to better observe how the source of labeling influences perceived training data credibility.

5.4 Negative Influence of Racial Bias in AI Performance on AI Trust

Our data clearly shows that perceived training data credibility does not maintain users' cognitive trust in AI when the system shows bias in its performance. This finding is consistent with previous studies showing the detrimental effect of algorithmic bias on AI trust. People experience a trust breakdown after encountering a biased result [17, 58]. While it is important to have an accurate evaluation of racial bias in AI algorithms from the users' perspective, an obvious follow-up question that AI developers and designers should be asking is: how to involve users to help create a better AI system, such that there is no bias in its performance in the future? Prior studies have shown that simply asking users to indicate their level of agreement with the AI classification outcomes does not help restore users' expectations [46] and trust in AI [9]. Rather, users need to be given more agency [38]. Perhaps future studies may leverage the mutual augmentation approach proposed in the HAI-TIME model [50] to reshape user experience and trust in the AI system. That is, users should be informed about how their effort in correcting the AI performance may lead to an improved AI system and eventually benefit themselves and other users. In doing so, users are likely to maintain their cognitive trust in AI, which is primarily driven by knowledge, understanding, and rationale [26].

Different from the patterns for cognitive trust in AI, perceived training data credibility is positively related to affective and behavioral trust in AI, regardless of AI performance. In other words, when users are primed that the training dataset is credible, they tend to have a stronger emotional connection to the AI system and are more likely to adopt the AI's recommendation to evaluate one's facial expression. This finding corroborates previous research, which showed that expecting the AI system to be fair could lead to higher trust in AI even when users encounter a biased AI performance [9]. In addition, this finding supports the priming effect, which states that the first stimulus influences perceptions of subsequent stimuli [54]. There are several possible explanations for why performance bias did not override the priming effects in our study. First, participants saw only four examples of AI classification, two of which were misclassified, so they may have considered it a one-time mistake. Second, participants may have given more weight to training data characteristics when evaluating AI trust rather than AI performance. Given that AI performance is the outcome of machine learning, which relies on training data to begin with, the important role of training data relative to AI performance may explain the effectiveness of the priming effect.

However, the positive effect of training data credibility on affective and behavioral trust in AI regardless of the nature of AI performance triggers concerns over automation bias, a tendency to overtrust AI even when AI shows obvious errors [39]. To reduce over-reliance on AI for decision making, designers should invite more cognitive engagement with the AI system when it shows a biased performance. For example, designers can afford users an opportunity to take a close look at the criteria used by the system for determining its classification and make changes to the criteria if they deem it necessary. Such an approach, based on "interactive transparency," is known to be effective in calibrating users' trust in AI by increasing perceived disclosure and user agency [38].

5.5 Practical Implications

Findings of the study have significant practical implications for AI developers, designers, and researchers. First, our manipulation of labeling quality was successful. This means when users are forced to examine a snapshot of training data, they are more cognitively engaged, and our data suggest that they are more vigilant when it comes to biases in performance, leading them to question their trust in AI when it produces biased results. A clear practical implication is that presenting users with labeled-data snapshots prior to their interactions is a great way to get them to calibrate their trust and promote thoughtful and responsible use of AI outcomes.

Second, our manipulation of performance bias is very effective despite the use of just two faces of each race (4 total). This means AI users will form a strong opinion about the system's racial bias even if it fails once or twice. Studies have shown that algorithm aversion can set in very quickly [14] and is quite difficult to overcome without enabling some user control over the algorithmic decision [15]. Designers should be aware of the powerful "exemplification effect" [60] of a small sample of AI performance outcomes on users' trust in AI.

Furthermore, labeling source does not matter when the accuracy of labeling is clearly conveyed. It means that it is not necessary for UX designers to motivate users to observe or participate in the labeling practice when interacting with an AI system. How users see themselves in data labeling does not seem to affect training data credibility. Instead, showing labeling quality through a snapshot of labeled data would be a better design solution for fostering accurate credibility judgments.

Notably, it is clear that even when an AI system is not biased in its performance outcomes, poor credibility of training data can undermine user trust (see the columns pertaining to unbiased performance in Figure 5). Thus, a design implication is to deploy interface cues that tout the credibility of the labeled data used for training the AI model rather than simply promote its performance metrics such as accuracy. Coupled with the training data credibility cues, researchers can ask users the extent to which they think the training data is credible, trustworthy, reliable, and dependable, in order to gauge their level of trust in AI. Our mediation model shows that the perceived credibility of the training data is a good indicator to assess users' overall trust in the AI system.

It is important to note that displaying labeling quality is applicable only to supervised machine learning. Given that visuals and labels are key components to conveying labeling quality, this new

design practice may work better for AI systems that are trained on clearly distinct labels, such as cats vs. dogs and men vs. women. It means that AI systems with ambiguous or disputed labels may not benefit from the current design solution. For instance, showing social media posts with a label of hate speech (or not hate speech) may not help communicate training data credibility due to the lack of a clear and universally accepted understanding of hate speech among users. Furthermore, the current design solution may not be ideal for AI systems whose training data is not visual, such as voice samples for a text-to-speech system.

While the novel design is promising in adjusting users' cognitive trust in AI, one negative outcome of showing labeling quality prior to the actual interaction is that users may blindly trust the AI system emotionally and behaviorally even when they see a racially biased outcome by the system, a phenomenon also known as automation bias [39]. Designers need to devise action-based solutions to engage users cognitively to avoid their knee-jerk tendency to over-trust systems that provide disclosure without carefully examining the content of that disclosure.

5.6 Limitations and Future Studies

This study suggests one design solution for communicating labeling quality. By providing a snapshot of training data up front, coupled with the labels, users are able to evaluate for themselves the credibility of the training data and accordingly calibrate their trust in the AI system. To ontologically differentiate high- from low-quality labeling, we have all facial images labeled correctly in the high-quality labeling condition, but only half in the low-quality labeling condition. Considering that most labels are provided by human annotators, there is no guarantee that the labels are 100% accurate for any training dataset in real-life labeling practice. It would be meaningful to explore the accuracy threshold that can influence user perception of training data credibility. Would 90% accuracy, i.e., 90% of the training data matching the label, be enough to contribute to perceived training data credibility? Future studies would do well to examine this question.

Apart from showing the accuracy of labeling, there are many other ways to communicate labeling quality. For example, UX designers can focus on the background of the labelers. According to a recent study, knowing that labelers are from diverse racial backgrounds can increase expectations of AI fairness and accuracy by triggering the representativeness heuristic [55], which further enhances users' trust in AI [9]. Likewise, if the labeling is done by a group of domain experts, it may also increase perceived training-data credibility by triggering the expertise heuristic [49]. If the interface highlights the similarity between the labelers and the user, the triggered similarity heuristic may contribute to the perception of training data credibility as well.

All design solutions have trade-offs. One unknown issue is the extent to which showing labeling quality before the interaction may increase cognitive load and further undermine user experience. To find a balance between the need for explanations and the need for efficiency, designers can make the snapshot of labeled data optional and leave it to the users to decide whether to take a close look at the accuracy of labeling before further interaction. Considering that too much information could be confusing and thereby undermine

user understanding [27], future studies can explore how much to communicate about labeling, so that users will be informed while having a smooth and satisfactory user experience.

6 CONCLUSION

While designers can convince users about training data credibility by showing good examples of quality labeling, they have to be cognizant of its differential effects on different types of trust. Our data show that while labeling quality can cognitively help users calibrate their trust alongside their perceptions of AI performance, it can emotionally and behaviorally lead to overtrust because users do not seem to correct their perceptions even when they encounter a clearly biased performance. In order to promote socially responsible AI, designers ought to make every effort to invoke users' cognitive engagement when making trust decisions upon seeing performance outcomes, rather than letting affective trust and behavioral trust hold sway.

REFERENCES

- [1] Simon Albrecht and Anthony Travaglione. 2003. Trust in public-sector senior management. *International Journal of Human Resource Management* 14, 1 (2003), 76–92. <https://doi.org/10.1080/09585190210158529>
- [2] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–13. <https://doi.org/10.1145/3411764.3445736>
- [3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Benetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [4] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. In *Proceedings of the 35th Conference on Neural Information Processing Systems*. <https://doi.org/10.48550/arXiv.2106.05498>
- [5] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tac1_a_00041
- [6] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–14. <https://doi.org/10.1145/3173574.3173951>
- [7] Judee K Burgoon. 1993. Interpersonal expectations, expectancy violations, and emotional communication. *Journal of language and social psychology* 12, 1-2 (1993), 30–48. <https://doi.org/10.1177/0261927X93121003>
- [8] Judee K Burgoon, Joseph A Bonito, Paul Benjamin Lowry, Sean L Humpherys, Gregory D Moody, James E Gaskin, and Justin Scott Giboney. 2016. Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies* 91 (2016), 24–36. <https://doi.org/10.1016/j.ijhcs.2016.02.002>
- [9] Cheng Chen. 2022. *Communicating racial bias in AI algorithms: Effects of training data diversity and user feedback on AI trust*. Ph.D. Dissertation. Pennsylvania State University, State College, PA.
- [10] Cheng Chen and S. Shyam Sundar. 2021. Combating algorithmic bias: Should AI media show and tell to gain user trust? (2021). Poster presented at the 2021 ICDS Symposium on Fairness of Machine Learning.
- [11] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Proceedings of the 32nd International Conference on Neural Information Processing Systems* 31 (2018), 3543–3554. <https://doi.org/10.5555/3327144.3327272>
- [12] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89. <https://doi.org/10.1145/3376898>
- [13] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 2342–2351. <https://doi.org/10.1145/3531146.3534647>
- [14] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114. <https://doi.org/10.1037/xge0000033>
- [15] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170. <https://doi.org/10.1287/mnsc.2016.2643>
- [16] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, United States, 275–285. <https://doi.org/10.1145/3301275.3302310>
- [17] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful: things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. AAAI Press, Palo Alto, California, USA, 62–71. <https://doi.org/10.1609/icwsm.v11i1.14898>
- [18] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. <https://doi.org/10.3758/bf03193146>
- [19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92. <https://doi.org/10.48550/arXiv.1803.09010>
- [20] Karen Hao. 2019. *AI is sending people to jail—and getting it wrong*. MIT Technology Review. Retrieved December 8, 2022 from <https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/>
- [21] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications, New York, NY, United States.
- [22] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy* 12 (2020), 1. <https://doi.org/10.48550/arXiv.1805.03677>
- [23] Joo-Wha Hong, Ignacio Cruz, and Dmitri Williams. 2021. AI, you can drive my car: How we evaluate human drivers vs. self-driving cars. *Computers in Human Behavior* 125 (2021), 106944. <https://doi.org/10.1016/j.chb.2021.106944>
- [24] Yan Huang and S Shyam Sundar. 2022. Do we trust the crowd? Effects of crowdsourcing on perceived credibility of online health information. *Health Communication* 37, 1 (2022), 93–102. <https://doi.org/10.1080/10410236.2020.1824662>
- [25] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04
- [26] Devon Johnson and Kent Grayson. 2005. Cognitive and affective trust in service relationships. *Journal of Business Research* 58, 4 (2005), 500–507. [https://doi.org/10.1016/S0148-2963\(03\)00140-1](https://doi.org/10.1016/S0148-2963(03)00140-1)
- [27] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 2390–2395. <https://doi.org/10.1145/2858036.2858402>
- [28] Heidi Ledford. 2019. Millions of black people affected by racial bias in health-care algorithms. *Nature* 574, 7780 (2019), 608–610. <https://www.nature.com/articles/d41586-019-03228-6>
- [29] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- [30] Q Vera Liao, Moninder Singh, Yunfeng Zhang, and Rachel Bellamy. 2021. Introduction to explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–3. <https://doi.org/10.1145/3411763.3445016>
- [31] Q Vera Liao and S Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, United States, 1257–1268. <https://doi.org/10.1145/3531146.3533182>
- [32] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49, 2 (2017), 433–442. <https://doi.org/10.3758/s13428-016-0727-z>
- [33] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal* 38, 1 (1995), 24–59. <https://doi.org/10.2307/256727>
- [34] Miriam J Metzger and Andrew J Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics* 59 (2013), 210–220. <https://doi.org/10.1016/j.pragma.2013.07.012>

- [35] David Meyer. 2018. *Amazon reportedly killed an AI recruitment system because it couldn't stop the tool from discriminating against women*. Fortune. Retrieved September 13, 2022 from <https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/>
- [36] Dale T Miller. 1976. Ego involvement and attributions for success and failure. *Journal of Personality and Social Psychology* 34, 5 (1976), 901. <https://doi.org/10.1037/0022-3514.34.5.901>
- [37] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 220–229. <https://doi.org/10.1145/3287560.3287596>
- [38] Maria D Molina and S Shyam Sundar. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication* 27, 4 (2022), zmac010. <https://doi.org/10.1093/jcmc/zmac010>
- [39] Kathleen L Mosier, Linda J Skitka, Susan Heers, and Mark Burdick. 2017. Automation bias: Decision making and performance in high-tech cockpits. In *Decision Making in Aviation*. Routledge, 271–288. https://doi.org/10.1207/s15327108ijap0801_3
- [40] Daniel J O'Keefe. 2003. Message properties, mediating states, and manipulation checks: Claims, evidence, and data analysis in experimental persuasive message effects research. *Communication Theory* 13, 3 (2003), 251–274. <https://doi.org/10.1111/j.1468-2885.2003.tb00292.x>
- [41] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336. <https://doi.org/10.1016/j.patter.2021.100336>
- [42] A Phaneuf. 2020. *Artificial intelligence in financial services: Applications and benefits of AI in finance*. Insider Intelligence. Retrieved December 8, 2022 from <https://www.insiderintelligence.com/insights/ai-in-finance/#:~:text=AI%20is%20particularly%20helpful%20in,underwriting%20and%20reduce%20financial%20risk.>
- [43] Amy Rechkemmer and Ming Yin. 2022. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–14. <https://doi.org/10.1145/3491102.3501967>
- [44] Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer, 5–22. <https://doi.org/10.1145/3491102.3501967>
- [45] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 59–68. <https://doi.org/10.1145/3287560.3287598>
- [46] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–13. <https://doi.org/10.1145/3313831.3376624>
- [47] Hyeonjin Soh, Leonard N Reid, and Karen Whitehill King. 2009. Measuring trust in advertising. *Journal of Advertising* 38, 2 (2009), 83–104. <https://doi.org/10.2753/JOA0091-3367380206>
- [48] Yuan Sun and S Shyam Sundar. 2022. Exploring the effects of interactive dialogue in improving user control for explainable online symptom checkers. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. Association for Computing Machinery, New York, NY, United States, 1–7. <https://doi.org/10.1145/3491101.3519668>
- [49] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative, Cambridge, MA. <https://doi.org/10.1162/dmal.9780262562324.073>
- [50] S Shyam Sundar. 2020. Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAIL). *Journal of Computer-Mediated Communication* 25, 1 (2020), 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- [51] S Shyam Sundar, Haiyan Jia, T Franklin Waddell, and Yan Huang. 2015. Toward a theory of interactive media effects (TIME) four models for explaining how interface features affect user psychology. *The Handbook of the Psychology of Communication Technology* (2015), 47–86. <https://doi.org/10.1002/9781118426456.ch3>
- [52] S Shyam Sundar and Clifford Nass. 2001. Conceptualizing sources in online news. *Journal of Communication* 51, 1 (2001), 52–72. <https://doi.org/10.1111/j.1460-2466.2001.tb02872.x>
- [53] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–17. <https://doi.org/10.1145/3411764.3445101>
- [54] Endel Tulving, Daniel L Schacter, and Heather A Stark. 1982. Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8, 4 (1982), 336. <https://doi.org/10.1037/0278-7393.8.4.336>
- [55] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- [56] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7. <https://doi.org/10.1145/3194770.3194776>
- [57] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–14. <https://doi.org/10.1145/3313831.3376813>
- [58] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–14. <https://doi.org/10.1145/3173574.3174230>
- [59] Wencan Zhang and Brian Y Lim. 2022. Towards relatable explainable AI with the perceptual process. In *CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–24. <https://doi.org/10.1145/3491102.3501826>
- [60] Dolf Zillmann. 2002. Exemplification theory of media influence. In *Media effects*. Routledge, 29–52.