Improving Human-AI Collaboration with Descriptions of AI Behavior

ÁNGEL ALEXANDER CABRERA, Carnegie Mellon University, USA
ADAM PERER, Carnegie Mellon University, USA
JASON I. HONG, Carnegie Mellon University, USA

People work with AI systems to improve their decision making, but often under- or over-rely on AI predictions and perform worse than they would have unassisted. To help people appropriately rely on AI aids, we propose showing them *behavior descriptions*, details of how AI systems perform on subgroups of instances. We tested the efficacy of behavior descriptions through user studies with 225 participants in three distinct domains: fake review detection, satellite image classification, and bird classification. We found that behavior descriptions can increase human-AI accuracy through two mechanisms: helping people identify AI failures and increasing people's reliance on the AI when it is more accurate. These findings highlight the importance of people's mental models in human-AI collaboration and show that informing people of high-level AI behaviors can significantly improve AI-assisted decision making.

CCS Concepts: • **Human-centered computing** → *Empirical studies in HCI*; *HCI theory, concepts and models*; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: machine learning, human-AI collaboration

ACM Reference Format:
Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration with Descriptions of AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 136 (April 2023), 21 pages. https://doi.org/10.1145/3579612

1 INTRODUCTION

Human-AI collaboration is finding real-world use in tasks ranging from diagnosing prostate cancer [12] to screening calls to child welfare hotlines [19, 27]. To effectively work with AI aids, people need to know when to either accept or override an AI's output. People decide when to rely on an AI by using their *mental models* [3, 30], or internal representations, of how the AI tends to behave: when it is most accurate, when it is most likely to fail, etc. A detailed and accurate mental model allows a person to effectively complement an AI system by appropriately relying [37] on its output, while an overly simple or wrong mental model can lead to blind spots and systematic failures [3, 8]. At worst, people can perform worse than they would have unassisted, such as clinicians who made more errors than average when shown incorrect AI predictions [7, 24].

Mental models are an inherently incomplete representation of any system, but numerous factors make it especially challenging to develop adequate mental models of AI systems. First, modern AI systems are often black-box models for which humans cannot see how or why the model made a prediction [54]. Second, black-box models are also often stochastic and can provide different outputs for slightly different inputs without human-understandable reasons [1]. Lastly, people

Authors' addresses: Ángel Alexander Cabrera, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Adam Perer, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Jason I. Hong, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

2573-0142/2023/4-ART136
https://doi.org/10.1145/3579612

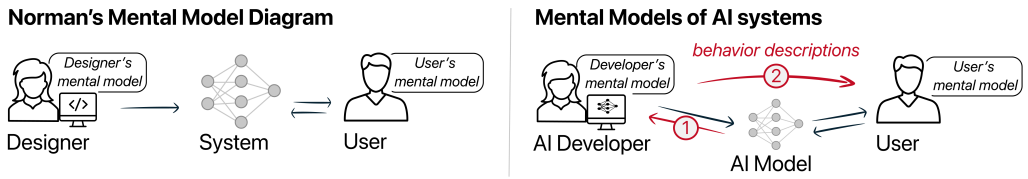

Fig. 1. Don Norman's mental model framework [43] describes how designers use their mental models to implement systems. End-users then interact with the systems and develop their own mental models of how they believe the systems work. While this process is similar for AI models, a key difference is that an AI is not a direct representation of a developer's intent, but a stochastic model learned from data. This means that **(1)** AI developers *themselves* have to make sense of what an AI system has learned through testing and iteration. Subsequently, they can encode these insights as **(2)** *behavior descriptions*, details of how an AI performs on subgroups of instances, that can be shown to end-users to improve human-AI collaboration.

often expect AI models to behave as humans do, which often does not match their actual behavior [40] and makes people unaware of how an AI may fail [25, 65]. These factors make it challenging for people to develop appropriate mental models of AI systems and effectively rely on them to improve their decision making.

To help people collaborate more effectively with AI systems, we propose directly showing end-users insights of AI behavior (Figure 1). We term these insights **behavior descriptions**, details of an AI's performance (metrics, common patterns, potential failures, etc.) on subgroups of instances (subsets of a dataset defined by different features, e.g. "images with low exposure"). Behavior descriptions can take many forms, but should help end-users *appropriately rely* [37] on an AI, using its output when it is most likely correct and overriding it when it is most likely incorrect. The goal of behavior descriptions is similar to that of explainable AI (xAI), but differs in what type of information is provided to end-users. Explainable AI attempts to describe *why* an AI system produced a certain output, while behavior descriptions describe *what* patterns of output a model tends to produce. Thus, xAI and behavior descriptions can be used together to support effective human-AI collaboration.

We hypothesize that behavior descriptions will help people better detect systematic AI failures and improve AI-assisted decision making. Additionally, we hypothesize that people will both trust an AI aid more and find it more helpful when shown behavior descriptions. To test these hypotheses, we conducted human-subject experiments in three distinct domains: fake review detection, satellite image classification, and bird classification. These three domains cover a range of data types, classification tasks, and human and AI accuracies to isolate the effect of behavior descriptions from domain-specific effects.

We found that behavior descriptions can significantly improve the overall accuracy of human-AI teams through two distinct mechanisms. First, for instances with a behavior description, users can directly correct AI failures. Second, people tend to rely more on the AI system for instances without behavior descriptions when behavior descriptions are shown for other underperforming subgroups. We additionally found that showing behavior descriptions had no significant impact on people's qualitative judgements, such as trust and helpfulness, of the AI. Despite the potential benefits of behavior descriptions, their effects are not universal and depend both on how obvious AI failures are to a person and on a person's ability to correct the output once they know the AI is wrong.

In summary, this work introduces **behavior descriptions**, details shown to people working with AI systems of how an AI performs (metrics, common patterns, potential failures, etc.) for specific

subgroups of instances. We show how behavior descriptions can improve the performance of human-AI collaboration in three human-subject experiments. These results indicate that knowing the mental models of end-users in human-AI collaboration is essential to understand how a human-AI team will perform and which interventions and decision aids will be most successful.

2 BACKGROUND AND RELATED WORK

We review three main areas of work related to both creating and applying behavior descriptions. First, we explore existing research on understanding and improving human-AI collaboration. Second, we discuss methods for improving end-user AI understanding using explainable and interpretable AI. Lastly, we describe tools and visualizations for analyzing AI behavior that can be used to create behavior descriptions.

2.1 Human Factors and AI Aids

When a person interacts with an AI system, they develop a *mental model* of how it behaves - when it performs well, when it fails, or when it has quirky results [30]. Mental models let people effectively work with AIs by helping them decide whether to rely on, modify, or override an AI's output. Therefore, it is important for people to have adequate mental models of AI systems so they can appropriately rely on an AI and override it when it is likely to fail [37]. There are various methods for encouraging *appropriate reliance* of AI systems, often called *trust calibration* [57] techniques.

Studies have explored what factors influence people's mental models of AI systems. Kulesza et al. [31] found that people with more complete mental models were able to collaborate more effectively with a recommendation system. Bansal et al. [3] focused on attributes of AI systems and found that systems with *parsimonious* (simple), and *non-stochastic* (predictable) error boundaries were the easiest for humans to work with. Other factors such as stated and perceived accuracy [29], confidence values [66], and model personas [28] can also influence people's mental models and their reliance on AI. Behavior descriptions can complement these existing findings on human-AI collaboration, further helping people better collaborate with AI aids.

Recent methods have explored improving people's mental models, including tutorials explaining a task [33], tuning a model to better match human expectations [41], or adaptive trust calibration [45]. Some methods for improving mental models, such as adaptive trust calibration, use model details such as calibrated confidence scores that can be used in conjunction with behavior descriptions. The technique most similar to our work is Mozannar et al. [42]'s exemplar-based teaching of model behavior. Their method learns nearest neighborhood regions around model failures to help validate people's mental models. Although similar, behavior descriptions are *semantic*, high-level insights of model behavior discovered and validated by developers. Example-based training can fill gaps that were not previously identified by behavior descriptions.

Like model tutorials and trust calibration, behavior descriptions aim to improve human-AI collaboration by updating people's mental models of AI behavior. Behavior descriptions provide additional information to end-users about *when* to rely on an AI, further enriching their mental models and complementing these existing techniques.

2.2 Explainable and Interpretable AI

AI systems can be designed and explained to help people understand model outputs. For example, interpretable model architectures allow users to inspect how a model makes predictions, while explanations can surface factors that affect model behavior.

One approach to improving end-user model understanding is to design inherently interpretable, or "glass-box", models. Glass-box models can be easier to debug and trace, improving people's ability to understand AI predictions [54]. Despite their benefits, glass-box models can be similarly

difficult to collaborate with, for example, a study by Poursabzi-Sangdeh et al. [48] found that more "clear" models can actually lead to information overload and hinder people's ability to detect AI errors. Additionally, glass-box models may not work for more complex domains or data types such as images and videos.

Instead, another approach is model-agnostic, post-hoc explanations of model outputs. LIME [50] is one such method, which fits linear models to a neighborhood of instances to highlight which features most influenced a model's prediction. Instead of using all input features to explain a prediction, follow-up techniques such as Anchors [51] and MUSE [35] learn sets of rules on features that are sufficient to explain predictions for a subset of instances. Behavior descriptions also show details about a model for subsets of instances, but instead of approximating *why* a model produces certain outputs, focuses on describing *what* the pattern of outputs is.

Despite the insights post-hoc model explanations can provide end-users, they have been found to have a small or even detrimental effect on the performance of AI-assisted decision making [17, 34]. Explanations have been shown to potentially increase people's reliance on an AI, whether it is right or wrong, leading to worse team performance than a human or AI alone [5, 26, 63, 66]. By telling users how a model performs, behavior descriptions specifically target people's mental models and inform users when to rely on or override an AI's prediction.

Behavior descriptions can be combined with existing techniques to improve human-AI collaboration. For example, behavior descriptions can still be used with interpretable models or shown in conjunction with black-box explanations.

2.3 Behavioral Analysis of AI Systems

There are numerous tools and techniques that help developers discover and fix model failures, especially those with real-world effects such as safety concerns and biases. Behavioral analysis enables developers to go beyond aggregate metrics, such as accuracy, to discover specific and important behaviors [11, 49].

Interactive tools, visualizations, and algorithmic techniques have shown promise in helping developers discover AI errors. The tools most relevant to this work are focused on exploring model performance on a subset of data. These include systems such as FairVis [10], a visual analytics system for discovering intersectional biases, AnchorViz [15], a visualization for semantically exploring datasets, and Slice Finder [18], a method for automatically surfacing subsets of data with large model loss. There are also tools tailored to specific domains, for example, Errudite [61], a visual analytics system for discovering and testing behaviors in NLP models. More recently, crowd-based systems such as Pandora [44], Deblinder [9], and P-BTM [38] have shown how crowdsourcing can augment behavioral analysis tools by surfacing previously unknown model failures. Lastly, there are algorithmic testing methods for discovering AI behaviors. A common technique is *metamorphic testing* [21], which verifies the relationship between changes to an input and expected changes in an output. Checklist [52] is one such testing tool for NLP models that creates sets of modified instances to test if language models understand basic logic and grammar.

These model debugging techniques can be used to create behavior descriptions, statements of model performance on subgroups of data. Additionally, analyses can be rerun on updated models to update behavior descriptions and directly inform end-users how a model has changed [4].

3 BEHAVIOR DESCRIPTIONS

We define **behavior descriptions** as details of how an AI performs (metrics, common patterns, potential failures, etc.) for a subgroup of instances. Behavior descriptions should be semantically meaningful and human-understandable but can vary significantly between datasets and tasks. For domains like binary classification, e.g. spam detection, they can be simple statements of accuracy

like *our system incorrectly flags marketing emails as spam 53% of the time.* In more complex domains like image captioning, they can describe specific behaviors like *our system often describes mountain climbing as skiing* or *our system can produce run-on sentences.* The goal of behavior descriptions is to help end-users develop better mental models of an AI and thus *appropriately rely* on the model — using the AI output when it is more accurate and overriding the AI when it is more likely to fail.

In addition to what behavior descriptions contain, *how* they are presented to end-users can vary and impact their effectiveness. For example, *when* behavior descriptions are shown can vary, from only showing them for extreme edge cases to every instance. Similarly, behavior descriptions may only be shown during training or for the first few uses of an AI aid. Due to the broad diversity of potential behavior descriptions, we do not attempt to enumerate all possible forms, and focus instead on testing some core assumptions of their efficacy.

3.1 Principles for Effective Behavior Descriptions

To design the behavior descriptions used in this work, we drew from existing studies of human-AI collaboration. From these studies, we derived three properties behavior descriptions should have to maximize their effectiveness. These are not the only principles, but an initial set used to design the behavior descriptions used in this study.

The first property comes directly from the goal of behavior descriptions, to help users appropriately rely on AI output. Therefore, behavior descriptions should provide information that helps end-users decide both *whether* they should rely on an AI and, if the AI is wrong, *how* they should override it. We summarize this principle as **actionable** behavior descriptions that provide end users with concrete information they can act on when using an AI aid.

The second principle comes from studies of AI error boundaries and explainable AI. Bansal et al. [3]'s study of people's mental models of AI systems found that models with errors that were *parsimonious* (e.g. simple to define) led to more effective mental models. Separately, Poursabzi-Sangdeh et al. [48] found that showing end-users more details about an AI reduced their ability to detect AI errors due to information overload. Thus, behavior descriptions should aim to be **simple**, short, and easy to interpret and remember.

The third and final principle comes from findings on alert fatigue and cognitive load. When alerts or messages are shown continuously, people can suffer from *alert fatigue* and begin to ignore the messages [14]. Additionally, showing people more information increases their cognitive load, which can lead to decreased learning and performance [55, 58]. To avoid these pitfalls, behavior descriptions should be limited and focus on the subgroups of instances with the highest impact. Thus, the third principle is to aim for **significant** behavior descriptions, either common behaviors or those with the most serious consequences.

In summary, the three design principles that we followed to create the behavior descriptions in this work are the following:

(1) **Actionable**, suggesting both whether and how a person should override an AI output.
(2) **Simple**, aiming to be as parsimonious and easy to remember as possible.
(3) **Significant**, limited to behaviors that are common and/or have serious consequences.

3.2 Why Not Just Fix AI Failures?

A key question surrounding the utility of behavior descriptions is why developers would not directly fix the systematic model failures or behaviors they discover. Modern AI systems are often stochastic, black-box models like neural networks that cannot be deterministically "fixed" or "updated". ML practitioners interviewed by Holstein et al. [22] would often try to fix one problem that would cause the model to start failing in other unrelated ways. In another empirical study, Hopkins and

Booth [23] found that practitioners would avoid or limit model updates due to concerns about breaking their models and introducing more failures. Challenges to fixing known model failures are also present in research. This is most apparent with natural language processing models, which are approaching human aptitude in many domains such as question answering and programming [6]. Despite the growing capability of NLP models, many state-of-the-art systems still encode serious biases and harms [13] and fail basic logical tests [52] that developers have been aware of for years, but have not been able to fix.

Fixing model failures can also require significant amounts of new training data, which tends to be expensive and time consuming. Additionally, the specific instances needed to improve certain model failures can be difficult to get, such as globally diverse images or accents [32]. Behavior descriptions can be an important intermediate solution for model issues; end-users can effectively work with an imperfect AI that developers are working to improve and fix.

Lastly, behavior descriptions can be helpful when models are updated. Model updates can be incompatible with end-users' existing mental models and violate their expectations of how an AI behaves, leading to new failures [4]. Updated behavior descriptions can be deployed with a new model to directly update end-user's mental models and avoid decreased performance.

4 EXPERIMENTAL DESIGN

To directly test how behavior descriptions impact human-AI collaboration, we conducted a set of human-subject experiments across three different tasks. The tasks range across dimensions such as human accuracy, AI accuracy, and data type to reduce the chance that our results are confounded by domain-specific differences.

4.1 Experimental Setup

To test the effect of behavior descriptions, we conducted experiments across three different classification tasks. All three tasks shared the same core setup and only varied in the type of classification task (binary or multiclass) and what participants were asked to label. For each task, we tested three between-subjects conditions to isolate the effect of behavior descriptions:

- **No AI:** Participants were asked to classify instances without any assistance.
- **AI:** Participants were asked to classify instances with the help of an AI they were told was 90% accurate.
- **AI + Behavior Descriptions (BD):** Participants were asked to classify instances with the help of an AI they were told was 90% accurate. Additionally, for instances that were part of a behavior description group (10/30 instances), participants were shown a behavior description stating the AI accuracy for that type of instance (see Section 4.1.1 for details about behavior description groups).

Participants in every condition and task were first shown a consent form, introduced to the task, and shown example instances and labels. Those in the condition with the AI were also shown a screen before labeling that introduced the AI and stated its overall accuracy of 90%. The participants were then shown and asked to label 30 instances (see Figure 3 for an example UI), 20 instances from the overall dataset, and 5 instances each from two subsets of the data, *behavior description groups*, where the AI performance was significantly worse than for the overall model (see Section 4.1.1 for details). After labeling the 30 instances, participants completed a short questionnaire with Likert scale questions, open-ended text responses, and an attention check question.

The experiments were conducted on Amazon Mechanical Turk (AMT) with participants from the United States. We selected participants who had completed more than 1,000 tasks and had an approval rating of more than 98% to ensure high-quality responses. Participants that failed the

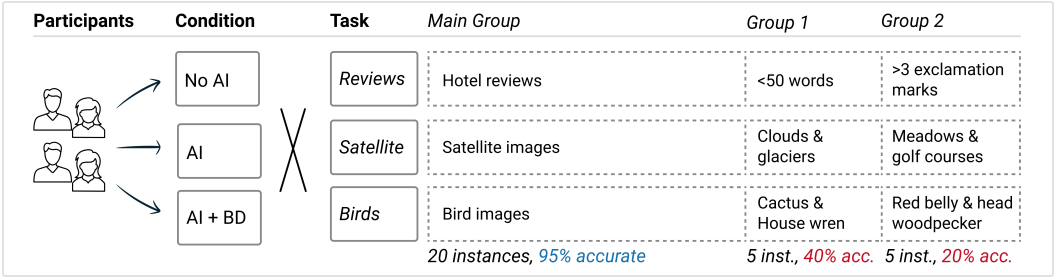

Fig. 2. **Experimental setup**. Each participant was randomly assigned to a condition and dataset (3x3 between-subjects study, 25 participants per condition, 225 participants total). For each dataset, participants saw 30 instances, 20 instances from the whole dataset with an AI accuracy of 95%, and 5 instances each from two subsets of the dataset with an AI accuracy of 40% and 20% respectively (simulating subgroups behavior descriptions would be useful for). In the *AI + Behavior Description* condition, participants were shown behavior descriptions for instances in group 1 and 2. While the instances shown to participants were randomly chosen from a larger subset of data, each participant saw the same number of AI errors to ensure they observed the same AI accuracies.

attention check, a question asking which step they were currently on, were still paid, but were excluded from the results and analysis. Although we considered providing bonuses as an incentive for accurate responses, we found that the incentive to have the task approved was sufficient to get good results without a bonus. We confirmed this by finding similar average accuracy on the control condition for the reviews task, 56%, to that reported by Lai and Tan [34] on the same task using a bonus, 51%.

The study was approved by an Institutional Review Board (IRB) process. We had a total of 225 participants, 25 per task/condition pair. The number of participants per condition was chosen using a power analysis on the mean and standard error of the accuracy in the initial usability studies for the interface. Of the 225 participants, we removed 13 for failing the attention check. Additionally, we removed three other participants, across two conditions, who had an accuracy of less than 10% (the next highest accuracy being 35%), the same as guessing randomly. We paid participants $2 for the task, which lasted 15 minutes for an hourly compensation of $8 an hour.

4.1.1 Behavior descriptions and wizard-of-oz AI. For this work, we used behavior descriptions stating the model accuracy for subgroups of the data for which the model performs significantly worse than average [10, 18, 44, 52, 62]. Depending on whether the task was binary or multiclass classification, the behavior descriptions resembled text sentences such as "the model is 20% accurate for this type of instance" or "the model confuses these two classes 80% of the time". These types of behavior descriptions are straightforward to create, calculating accuracy on a subset of data, and are actionable for end-users, informing them of how likely it is that they need to override the AI.

In order to control the distribution of instances and AI accuracy each participant saw, we used a wizard-of-oz AI system, mock AI outputs, with a fixed observed accuracy. Of the 30 instances each participant labeled, 20 were randomly chosen from the overall dataset, and the final 10 split into two groups of 5 instances randomly selected from two subsets of the dataset which we term *behavior description groups*. The AI accuracy for instances in the behavior descriptions groups was significantly lower than the overall model to simulate the types of subgroups behavior descriptions would be used for. Lastly, while we fixed the distribution of correct/incorrect model outputs per behavior description group, the instances each participant saw were randomized, both sampled from a larger set of images and randomly ordered.

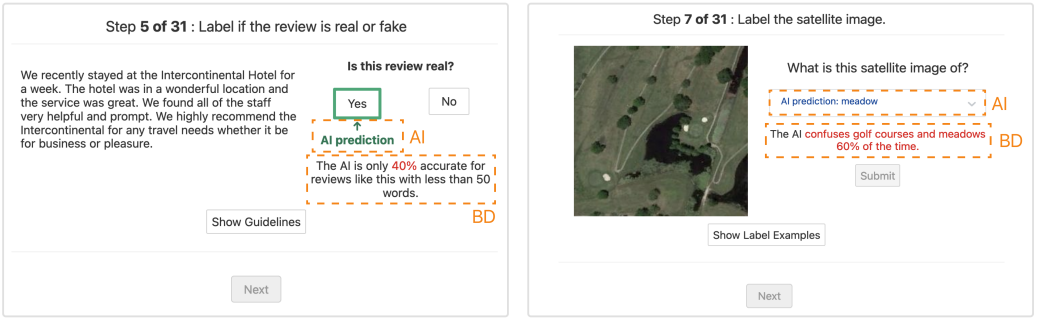

Fig. 3. UI screenshots for the fake reviews (left) and satellite image classification (right) tasks. Each participant labeled 30 instances, distributed according to the instance groups described in Figure 2 and Section 4.1.2. Both screenshots are shown on a labeling step for the *AI + Behavior Description (BD)* condition and on instances that are part of a behavior description group. In the *AI* condition participants are not shown the additional text for instances in a BD group, and in the *No AI* condition participants are not shown the AI output. The bird classification task used the same format as the satellite classification task shown.

The accuracy breakdown for the subgroups was the following: 95% accuracy (1/20 misclassified) for the *main group* of 20 instances, 40% accuracy (3/5 misclassified) for the first behavior description group, *group 1* and 20% accuracy (4/5 misclassified) for the second group, *group 2*. The behavior description groups for each task are detailed in Figure 2 and Section 4.1.2. The purpose of the behavior description groups is to have two concrete subsets representing the type of instances for which a behavior description would be used - these are the instances for which we show behavior descriptions in the AI + BD condition.

The accuracy breakdown above gives an actual AI accuracy of 73.33%, not the 90% accuracy stated to the participants, since the task would have been too long to have both an actual 90% overall accuracy and significantly low accuracies for the two behavior description groups. We wanted to simulate a situation where an AI is reasonably accurate, e.g. > 90%, so that a human would want to work with it. Existing work has explored the impact of stated versus observed accuracy and found that there is a small decrease in agreement with the AI the lower the observed accuracy [64]. Since each condition had the same stated vs. observed accuracy discrepancy (90% vs. 73.33% respectively), it should not impact our relative findings on the efficacy of behavior descriptions.

4.1.2 Classification tasks. We chose three distinct tasks for the study to ensure that our findings are not tied to a specific dataset or task. The three tasks vary by data type (text/image), task type (binary/multiclass classification), and human accuracy (human better/worse than AI). For each task description below, we also detail what types of instances make up *group 1* and *group 2*, the subsets of instances (behavior description groups) for which the AI is less accurate and for which behavior descriptions are shown in the BD condition (see Figure 2). The tasks are the following:

Fake Review Detection. The dataset of deceptive reviews from Ott et al. [46, 47] has 800 truthful reviews and 800 deceptive reviews for hotels in Chicago. The truthful reviews were collected from online sites such as TripAdvisor, while false reviews were collected from Amazon Mechanical Turk workers. The objective of the task is to determine whether the review is "truthful", written by someone who actually stayed at the hotel, or "deceptive", written by someone who has not. We chose this task since it has been used in previous studies of human-AI collaboration to test the effect of explanations [34] and tutorials [33]. BD groups were chosen from research on common failures in language models:

Group 1 - Reviews with less than 50 words [52].
Group 2 - Reviews with more than 3 exclamation marks [56].

Satellite Image Classification. The satellite images come from the NWPU-RESISC45 dataset [16], which has 31,500 satellite images across 45 classes. The task for the dataset is multiclass classification, labeling each square satellite image with a semantic class. We selected a subset of 10 classes for the task in order to show participants at least a couple instances per class. This task was inspired by real-world human-AI systems for labeling and segmenting satellite images [39]. The BD groups were chosen from areas of high error in the confusion matrices from the original paper [16]:
Group 1 - Cloud and glacier images
Group 2 - Meadow and golf course images

Bird Classification. The bird images came from the Caltech-UCSD Birds 200 dataset [60], made up of 6,033 images of 200 bird species. As in the satellite image task, we chose a subset of 10 classes from the dataset for multiclass classification. The task was inspired by numerous apps and products for classifying bird species [59]. The BD groups were chosen from birds in the same family, classes that are the most similar and difficult to distinguish:
Group 1 - Cactus Wrens and House Wrens
Group 2 - Red Bellied Woodpeckers and Red Headed Woodpeckers

4.2 Hypotheses

From the primary goal of behavior descriptions, helping end-users appropriately rely on an AI, we formulated the following hypotheses of how we expect behavior descriptions to affect human-AI collaboration. The hypotheses focus both on quantitative measures of performance and qualitative opinions from participants. Our first hypothesis is that behavior descriptions will improve the overall accuracy of human-AI teams. We hypothesize that behavior descriptions will help end-users more appropriately rely on the AI [37], leading to improved performance.

H1. Showing participants behavior descriptions (BD) results in higher overall accuracy than just showing the AI prediction or no AI.

We hypothesize that this improved performance will primarily come from participants overriding the systematic failures identified by behavior descriptions. By providing actionable descriptions of when the model is most likely to be wrong, for instances in BD groups, we hypothesize that the participants will be able to better identify and correct errors when shown behavior descriptions.

H2. The higher accuracy from showing BDs is due to a higher accuracy on instances that are part of BD groups.

Lastly, we have a set of hypotheses on how we expect people's perception of the AI to change when they are shown behavior descriptions. We hypothesize that participants will find the AI to be *more helpful*, will be *more likely to want to use the AI*, and will *trust the AI more* when they are shown behavior descriptions. These hypotheses come from Yin et al. [64]'s study on accuracy and trust in AI systems, which found that observed accuracy significantly affected trust and reliance on AI systems.

H3a. Participants shown BDs trust the AI more than when just shown the AI output.
H3b. Participants shown BDs find the AI more helpful than when just shown the AI output.
H3c. Participants shown BDs are more likely to want to use the AI in the future than when just shown the AI output.

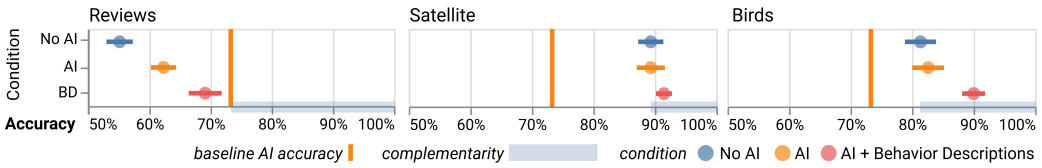

Fig. 4. **Average participant accuracy by task and condition.** The vertical orange bar indicates the AI accuracy, what would be the participant's accuracy if they picked the AI response every time. The blue shaded area indicates *complementarity*, the region where the human+AI accuracy is higher than either the human or AI alone. We find that behavior descriptions led to higher accuracy in the reviews and birds tasks, with complementarity in the birds task (red point in rightmost chart). The error bars represent standard error.

5 RESULTS

To assess the significance of different conditions on participant accuracy, we used ANOVA tests with Tukey post-hoc tests to correct for multiple comparisons. For the Likert scale questions, we used Mann-Whitney U tests with Bonferroni corrections. Lastly, we used linear models to test for learning effects, also using a Bonferroni correction for multiple comparisons. We used a p value of 0.05 as the cutoff for significance.

5.1 Overall Accuracy

To test **H1** we can compare the average human-AI team accuracy for each task across the three conditions. Overall, we found that the *AI* and *AI + behavior descriptions (BD)* interventions have a different effect on team performance in each task (Figure 4).

For the reviews task, there was a significant difference in accuracy across the three conditions ($F_{2,64} = 9.18,\ p < 0.001$). The only significant pairwise difference was between the *No AI* and *AI + BD* conditions ($p < 0.001,\ 95\%\ C.I. = [0.06, 0.22]$). While the AI by itself did not significantly improve the accuracy of the participants, AI supplemented with behavior descriptions led to significantly higher team accuracy, supporting **H1**. Despite the increased performance, there was no human-AI complementarity – higher accuracy than either the human or AI alone – as participant accuracy at every condition was lower than the baseline AI accuracy.

In the satellite classification task there was no significant difference between conditions ($F_{2,67} = 0.63,\ p = 0.534$). The baseline human accuracy without AI support was the highest across tasks, around 90%, so there was a smaller margin to improve the accuracy of the participants using an AI with a significantly worse accuracy.

Lastly, in the birds classification task, there was a significant difference in participant accuracy ($F_{2,65} = 3.98,\ p = 0.023$). As in the reviews task, the only pairwise difference was between the *No AI* and *AI + BD* conditions ($p = 0.048,\ 95\%\ C.I. = [0.00, 0.16]$), showing how behavior descriptions can lead to significant increases in participant accuracy using an AI and supporting **H1**. This increased performance also led to complementary human-AI accuracy, higher than that of both the AI or human alone.

In sum, these results **partially support H1**, with behavior descriptions leading to significantly higher accuracy in two of the three tasks. This suggests that while behavior descriptions will not universally improve the accuracy of human-AI teams, they can lead to significant improvements in certain tasks and domains.

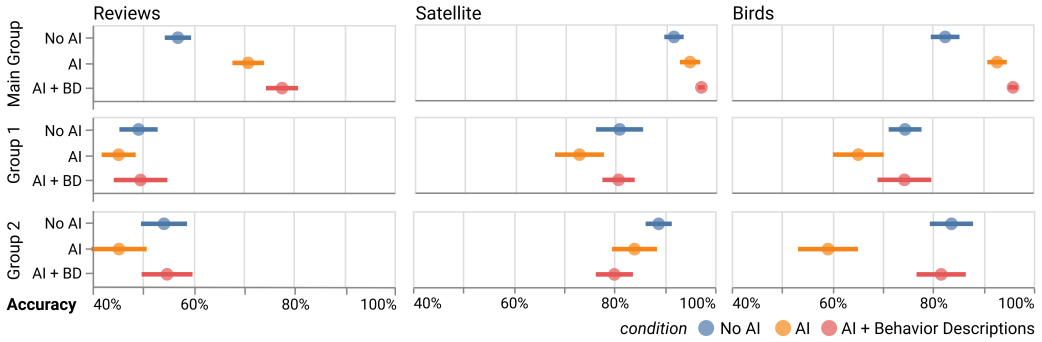

Fig. 5. **Average team accuracy by task, condition and instance group.** We further break down accuracy by instance type (see Section 4.1.1): the main group (20 instances), and two behavior description groups (5 instances each). The average human-AI accuracy across the three groups of instances gives us an idea of *how* behavior descriptions improve the performance of human-AI teams. We find that participants relied more on the AI when shown BDs in every task. Participant performance on the different behavior description groups was mixed, from no effect to significant improvement in group 2 birds (bottom right). These results highlight the two effects of behavior descriptions, increasing human reliance on a more accurate AI and overriding systematic AI errors. The error bars represent standard error.

5.2 Accuracy by Behavior Description Group

To better understand the ways in which behavior descriptions impact performance, we can look at participant accuracy across both conditions and instances in the three different behavior description groups described in Section 4.1.1 (Figure 5). This allows us to directly test **H2** by seeing if participants in the *AI + BD* condition perform significantly better on the two subsets of instances that have behavior descriptions. The results can be seen in Figure 5. We found that this is partially true, as the increased accuracy of the *AI + BD* condition was due both to higher accuracy on the behavior description groups *and* instances in the main group.

In the reviews task, there was only a significant difference between conditions for instances in the normal group ($F_{2,64} = 11.82,\ p < 0.001$), with no differences between conditions for either of the behavior description groups. The *AI* ($p = 0.007,\ 95\%\ C.I. = [0.03, 0.24]$) and *AI + BD* ($p < 0.001,\ 95\%\ C.I. = [0.10, 0.31]$) conditions were significantly more accurate than the *No AI* condition for instances in the main group. These results do not support **H2**, as the higher overall accuracy of the *AI + BD* condition was primarily due to greater reliance on the AI for instances in the main group, not higher accuracy on instances with behavior descriptions.

Despite there being no difference in overall accuracy between conditions for the satellite task, there were differences when looking at the behavior description groups. As with the reviews task, there was a significant difference between conditions for instances in the main group ($F_{2,67} = 5.34,\ p = 0.007$). Participants in both the *AI* ($p = 0.026,\ 95\%\ C.I. = [0.00, 0.09]$ and *AI + BD* ($p = 0.007,\ 95\%\ C.I. = [0.01, 0.10]$ conditions had a higher accuracy than participants in the *No AI* condition for instances in the main group. This is the same result we found in the reviews task, where there were significant accuracy differences for instances in the main group. Despite this difference, the increased accuracy did not translate to a higher overall accuracy for the *AI* and *AI + BD* conditions. Since we did not find any difference between conditions for the two BD groups, these findings do not support **H2**.

Lastly, we found significant differences between conditions for multiple behavior description groups in the bird classification task. As with the other two tasks, there is a significant difference

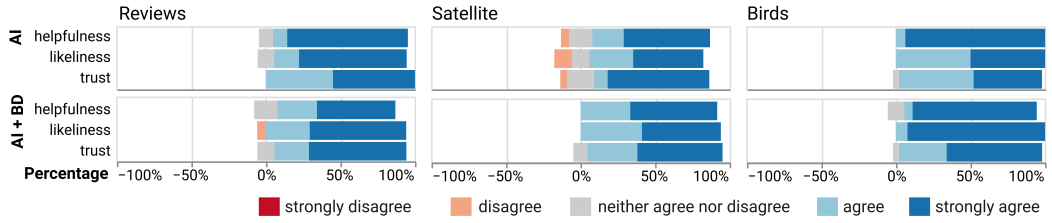

Fig. 6. **Likert-scale responses on perception of AI.** The diverging stacked bar chart centered around the neutral response shows that participants across all conditions and subjective measures overwhelmingly viewed the AI favorably. There were no significant differences in user's perception of the AI when they were give behavior descriptions.

in accuracy between conditions for instances in the main group ($F_{2,65} = 12.22,\ p < 0.001$). Both *AI* and *AI + BD* have a significantly higher accuracy than *No AI*, but there is no significance between *AI* and *AI + BD*. Although there is no significance for instances in group 1, we do see a difference between conditions for instances in group 2 ($F_{2,65} = 6.81,\ p = 0.002$). Both the *No AI* and *AI + BD* conditions both have a higher accuracy than just *AI*. This shows that while participants were able to distinguish between the two types of woodpeckers in group 2 without the AI or when informed about the AI failures, participants trusted the AI and performed significantly worse when just shown the AI prediction. These results support **H2**, as the increased accuracy on group 2 led to higher accuracy in the AI + BD condition.

In summary, these results **partially support H2**. Depending on the task, the higher accuracy in the *AI + BD* condition was due to a higher accuracy on instances that were part of BD groups *and* and a greater reliance on the AI for instances in the main group.

5.3 Qualitative Results

After labeling the 30 instances, participants were shown a questionnaire page asking about their opinions and feelings of the AI aid using Likert scale questions (Figure 6). Since these questions were directly related to the AI output, they were only shown to participants in the *AI* and *AI + BD* conditions. The questions were the following: (1) How **helpful** was the AI for this task? (2) How **likely** are you to use this AI again in a future task? (3) How much do you **trust** the AI? We did not find significant differences between the *AI* and *AI + BD* conditions for any of the Likert scale questions across tasks. These results reject **H3a, H3b, H3c**, and indicate that behavior descriptions do not significantly impact participant perceptions of AI despite differences in how participants use AI predictions.

5.4 Additional Findings

In addition to the accuracy metrics and Likert scale responses, we collected and analyzed additional measurements and qualitative responses to further unpack the dynamics of behavior descriptions. Although these are post hoc, exploratory results for which we did not have hypotheses, they can serve as inspiration for further, more formal studies.

One such factor was the time it took participants to label each instance, which can potentially surface interesting insights when compared between conditions and behavior description groups. Unfortunately, the time per instance had high variance and was inconsistent, with numerous outliers. This is likely due to the way AMT workers complete tasks, as they often take breaks or search for new HITs while working on a task [36]. Future studies could incentivize quick responses to gather more accurate time data and measure the speed of participant responses across conditions.

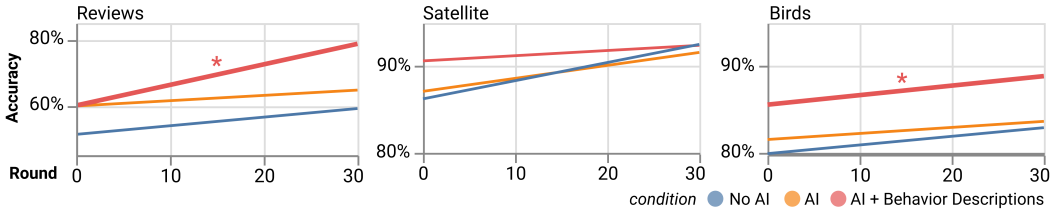

Fig. 7. **Learning curves by task and condition.** We fit linear models of accuracy on round number to measure learning effects. We found that the two conditions in which BDs significantly improved performance also had significant learning effects, the *AI + BD* conditions in the reviews and birds tasks (denoted by *).

We also tracked in which round each instance was labeled, $(n/30)$, to detect any learning effects (Figure 7). To measure learning effects, we fit a linear model of average participant accuracy for each condition and each step. We found that for the two domains in which behavior descriptions were effective, reviews and birds, there was also a significant learning effect for the *AI + BD* condition (reviews/AI + BD: $\beta = 0.0062, p = 0.020$; birds/AI + BD: $\beta = 0.0031, p = 0.035$).

The endpoints of the learning curves also show some interesting patterns. For the reviews task, participants in the *AI* and *AI + BD* conditions started at similar accuracies and only over time did participants in the *AI + BD* condition learn to effectively work with the AI and make use of the behavior descriptions. In contrast, in the birds classification task, participants in the *AI + BD* condition were consistently better than the *AI* condition and improved at a similar rate over time. Interestingly, for the satellite domain, the *AI + BD* condition learning curve ends at a point similar to the *No AI* condition but starts significantly higher. This suggests that, while over time participants learned the correct satellite labels or when the AI tended to fail, the behavior descriptions helped bootstrap the learning process.

Lastly, we collected qualitative free response answers from participants about patterns of AI behavior they noticed and general comments they had about the task. As expected, participants in both the *AI* and *AI + BD* noticed that the AI failed in the behavior description groups. While more participants described failures in the *AI + BD* condition than the *AI* condition, the comments were inconsistent and did not show many significant differences between conditions. We found an interesting pattern from comments in the birds task, where participants in the *AI* condition described more general but correct patterns of AI failure. Specifically, a participant found that *"the AI is good at predicting the main class of birds, but might get the sub class incorrect,"* which was the common failure reason between the two BD groups.

6 DISCUSSION

These results show that directly informing people about AI behaviors can significantly improve human-AI collaboration. We hope these initial insights spark future work on understanding people's mental models and developing new types of behavior-based decision aids.

6.1 Effectiveness of Behavior Descriptions

Overall, the results generally supported our primary hypothesis that behavior descriptions can significantly improve the accuracy of participants working with AI aids. Surprisingly, however, we found that the improved accuracy came from two complementary effects: people overriding systematic AI failures *and* relying more often on the AI for instances without behavior descriptions.

The first effect, helping participants fix systematic AI failures, was the initial goal of behavior descriptions, but we found that it was inconsistent and varied significantly between conditions and

behavior description groups. For example, in the reviews task, there was no significant accuracy difference for instances in either behavior description group. This is likely due to the behavior descriptions in the reviews domain not being sufficiently *actionable* and the participants not knowing whether or not to override the AI when they knew it was likely wrong. On the contrary, there *was* a significant difference in accuracy for instances in group 2 of the bird classification task, where participants in the BD condition were able to fix AI failures that most participants with just the AI did not notice. Thus, while we did find some support for this effect, participants fixing AI failures in BD groups was often not the main driver of increased overall accuracy.

In contrast, the more consistent effect was participants in the *AI + BD* condition relying on the AI more often for instances not in the BD groups (main group) compared to the *AI* condition. This increased reliance on the AI when it was more accurate, on instances in the main group, contributed to the higher overall accuracy in the *AI + BD* conditions for the reviews and birds tasks. This was an unexpected effect which was a factor in the higher overall accuracy for participants in the *AI + BD* condition. Although unexpected, the effect is corroborated by studies on trust in AI systems that found that low reliability, AI output that violates people's expectations over time, decreases trust and reliance in an AI [20]. By isolating common AI errors into well-defined, predictable subgroups, the AI appears more reliable to people and can increase their trust and reliance.

Although two of the three domains showed an overall increase in accuracy with behavior descriptions, there was no significant increase in the satellite classification task. We believe that this is due to the high baseline accuracy of humans performing the task, with approximately 90% accuracy, which left little room for improvement with a significantly less accurate AI. The high accuracy of the participants also allowed them to detect and fix the AI errors in the BD groups without any prompting in the *AI* condition. In summary, while behavior descriptions can improve performance, they are not the panacea to human-AI collaboration — AI aids must still provide value and complementarity to human decision makers.

6.2 Learning and Behavior Descriptions

We found that participants improved more quickly when using behavior descriptions, learning to effectively complement the AI. The primary pattern we noticed was the significant learning rate in the *AI + BD* condition for the reviews and birds task, as participants quickly learned when they should override the AI. This could be an important property for AI systems that are updated often, as new behavior descriptions could be used to directly update end-users mental models and avoid failures from incompatible updates [4].

A secondary effect we found was a higher initial accuracy in the *AI + BD* condition for the satellite and birds task. While it took time for participants in the *AI* condition to notice how a model tended to fail, the participants with behavior descriptions were aware of the failures right from the start. Even if people's accuracy converges over time, as in the reviews task, behavior descriptions can speed up end-user onboarding and improve early-stage performance.

6.3 Authoring Behavior Descriptions

The AI debugging techniques described in Section 2.3 can be used to create behavior descriptions, but tools designed specifically for creating BDs could provide important benefits. For example, behavior descriptions do not *have* to be AI failures, but could highlight consistent output patterns or areas where the model performs much better than humans. Bespoke tools could also optimize for creating behavior descriptions with effective properties such as those described in Section 3.1. Tools customized to create behavior descriptions could optimize for these different properties.

The types of people who create behavior descriptions can also be much broader than AI/ML developers. With the right tools, stakeholders ranging from quality assurance engineers to domain-specific teams could discover and deploy their own behavior descriptions. In the future, techniques from crowdsourcing could be used to harness end-user's own insights for generating behavior descriptions. This could be, for example, prompting end-users to report consistent failures and patterns that could then be aggregated and voted on [9]. The insights could be processed to automatically generate up-to-date behavior descriptions.

6.4 Understanding and Improving Mental Models of AI

These experiments also implicitly tested the more general effect of human mental models on how they collaborate with AIs. We found that when humans have more detailed mental models of how an AI performs, they are more likely to rely on the AI in general. This result is interesting regardless of the use of behavior descriptions, as more experienced end-users will likely develop better mental models of AI aids and gradually change how they rely on the AI.

Developing better techniques for quantifying end-user's mental models of AI systems can help researchers design effective decision aids such as behavior descriptions. This work could take inspiration from studies in HCI and psychology on *cognitive modeling*, mathematical models of human cognition [2]. Cognitive models have been used in education by simulating how people learn math to dynamically teach students [53]. For human-AI collaboration, cognitive models of how people learn AI behavior could inform the design of decision aids such as behavior descriptions.

Our results can also guide the design of machine learning models that can more directly provide behavior descriptions. For example, sparse decision trees could be used to directly generate behavior descriptions to show to end-users. New algorithmic methods that are more amenable to finding clusters of high error could lead to better human-AI collaborative systems.

7 LIMITATIONS AND FUTURE WORK

The participants in our study were Amazon Mechanical Turk workers with limited domain expertise in the tasks they completed. Domain experts and professionals, e.g., doctors or lawyers, have deeper expertise in their fields and may develop different mental models of AI systems they work with. For example, they may notice AI failures more often or be less influenced by the information provided by behavior descriptions. Future studies can explore the dynamics of using BDs with domain experts.

Although we selected domains that reflect potential real-world examples of human-AI collaboration, our experiment was a controlled study in a simulated setting. The impact of behavior descriptions may vary when applied to real-world situations with consequential outcomes [66], such as a radiologist classifying tumors. When classification errors have a much higher cost, people may update their mental models differently or act more conservatively when shown BDs.

Our study used one specific type of behavior description, subgroup accuracy. There are countless variations of BDs that can be explored further, such as highlighting subgroups with high accuracy, describing what features of an instance are correlated with failure, or suggesting alternative labels. Future experiments could test these variations to disentangle which features of behavior descriptions are the most effective in improving people's performance.

Lastly, our study focused on the relatively simple domain of classification. Modern AI systems are used in much more complex tasks such as image captioning, human pose estimation, and even image generation. The behavior descriptions for these domains will likely look significantly different from those tested in this work. For example, BDs for a captioning model might focus on grammatical issues and object references rather than statistical metrics. The impact of behavior

descriptions will likely vary significantly in these domains, and specific studies could explore both their effect and optimal designs.

8 CONCLUSION

We introduce **behavior descriptions**, details shown to people in human-AI teams of how a model performs for subgroups of instances. In a series of user studies with 225 participants in 3 distinct domains, we find that behavior descriptions can significantly improve human-AI team performance by helping people both correct AI failures and rely on the AI when it is more accurate. These results highlight the importance of people's mental models of AI systems and show that methods directly improving mental models can improve people's performance when using AI aids. This work opens the door to designing behavior-based AI aids and better understanding how humans represent, develop, and update mental models of AI systems.

ACKNOWLEDGMENTS

This material is based upon work supported by an Amazon grant, a National Science Foundation grant under No. IIS-2040942, and the National Science Foundation Graduate Research Fellowship Program under grant No. DGE-1745016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Amazon or the National Science Foundation.

REFERENCES

[1] Naveed Akhtar and Ajmal Mian. 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access 6 (2018), 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385

[2] John R. Anderson, Michael Matessa, and Christian Lebiere. 1997. ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention. Human–Computer Interaction 12, 4 (Dec. 1997), 439–462. https://doi.org/10.1207/s15327051hci1204_5

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7, 1 (2019), 19.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. Proceedings of the AAAI Conference on Artificial Intelligence 33 (July 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[5] Gagan Bansal, Tongshuang Wu, and Joyce Zhou. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA, 1–26. https://doi.org/10.1145/3411764.3445717

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, Vol. 33. Curran Associates, Inc., 1877–1901. ISSN: 10495258.

[7] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015 (2015), 160–169. https://doi.org/10.1109/ICHI.2015.26 Publisher: IEEE ISBN: 9781467395489.

[8] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287

[9] Ángel Alexander Cabrera, Abraham J. Druck, Jason I. Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (Oct. 2021), 1–22. https://doi.org/10.1145/3479569

[10] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST). 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

[11] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Rob DeLine, Adam Perer, and Steven M. Drucker. 2022. What Did My AI Learn? How Data Scientists Make Sense of Model Behavior. ACM Transactions on Computer-Human Interaction (June 2022), 3542921. https://doi.org/10.1145/3542921

[12] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proc. ACM Hum.-Comput. Interact. 3, CSCW (Nov. 2019). https://doi.org/10.1145/3359206

[13] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. Science 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230

[14] Jared J. Cash. 2009. Alert Fatigue. American Journal of Health-System Pharmacy 66, 23 (Dec. 2009), 2098–2101. https://doi.org/10.2146/ajhp090181

[15] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating Classifier Error Discovery through Interactive Semantic Data Exploration. In 23rd International Conference on Intelligent User Interfaces. ACM, Tokyo Japan, 269–280. https://doi.org/10.1145/3172944.3172950

[16] Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proc. IEEE 105, 10 (Oct. 2017), 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998

[17] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. arXiv:2007.12248 [cs, stat] (July 2020). http://arxiv.org/abs/2007.12248 arXiv: 2007.12248.

[18] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Slice Finder: Automated Data Slicing for Model Validation. In 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, Macao, Macao, 1550–1553. https://doi.org/10.1109/ICDE.2019.00139

[19] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376638

[20] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals 14, 2 (2020), 627–660. https://doi.org/10.5465/annals.2018.0057

[21] Pinjia He, Clara Meister, and Zhendong Su. 2020. Structure-Invariant Testing for Machine Translation. In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. ACM, Seoul South Korea, 961–973. https://doi.org/10.1145/3377811.3380339

[22] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, 1–16. https://doi.org/10.1145/3290605.3300830

[23] Aspen Hopkins and Serena Booth. 2021. Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. In AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, Inc, 134–145. https://doi.org/10.1145/3461702.3462527

[24] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry 11, 1 (June 2021), 108. https://doi.org/10.1038/s41398-021-01224-x

[25] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. 2015. My Data Just Goes Everywhere: User Mental Models of the Internet and Implications for Privacy and Security. In Eleventh Symposium On Usable Privacy and Security (SOUPS 2015). USENIX Association, Ottawa, 39–52. https://www.usenix.org/conference/soups2015/proceedings/presentation/kang

[26] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. CHI Conference on Human Factors in Computing Systems Proceedings (2020), 1–14. https://doi.org/10.1145/3313831.3376219 ISBN: 9781450367080.

[27] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In CHI Conference on Human Factors in Computing Systems. ACM, New Orleans LA USA, 1–18. https://doi.org/10.1145/3491102.3517439

[28] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. Proceedings of the ACM on Human-Computer Interaction 4, CSCW2 (2020). https://doi.org/10.1145/3415234

[29] 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. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300641

[30] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In Proceedings of the SIGCHI Conference on Human Factors in

Computing Systems. ACM, Austin Texas USA, 1–10. https://doi.org/10.1145/2207676.2207678

[31] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng Keen Wong. 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC (2013), 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. International Journal of Computer Vision 128, 7 (2020), 1956–1981. https://doi.org/10.1007/s11263-020-01316-z

[33] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376873

[34] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, Atlanta GA USA, 29–38. https://doi.org/10.1145/3287560.3287590

[35] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. ACM, Honolulu HI USA, 131–138. https://doi.org/10.1145/3306618.3314229

[36] Laura Lascau, Sandy J. J. Gould, Anna L. Cox, Elizaveta Karmannaya, and Duncan P. Brumby. 2019. Monotasking or Multitasking: Designing for Crowdworkers' Preferences. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300649

[37] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 46, 1 (Jan. 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[38] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. 2020. Towards Hybrid Human-AI Workflows for Unknown Unknown Detection. In Proceedings of The Web Conference 2020. ACM, Taipei Taiwan, 2432–2442. https://doi.org/10.1145/3366423.3380306

[39] Tomaz Logar, Joseph Bullock, Edoardo Nemni, Lars Bromley, John A. Quinn, and Miguel Luengo-Oroz. 2020. PulseSatellite: A tool using human-AI feedback loops for satellite image analysis in humanitarian contexts. arXiv:2001.10685 [cs, eess] (Jan. 2020). http://arxiv.org/abs/2001.10685 arXiv: 2001.10685.

[40] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, San Jose California USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[41] Jonathan Martinez, Kobi Gal, Ece Kamar, and Levi H. S. Lelis. 2020. Personalization in Human-AI Teams: Improving the Compatibility-Accuracy Tradeoff. arXiv:2004.02289 [cs] (Aug. 2020). http://arxiv.org/abs/2004.02289

[42] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2021. Teaching Humans When To Defer to a Classifier via Exemplars. arXiv:2111.11297 [cs] (Dec. 2021). http://arxiv.org/abs/2111.11297 arXiv: 2111.11297.

[43] Donald A Norman. 2014. Some Observations on Mental Models. In Mental Models (0 ed.), Dedre Gentner and Albert L. Stevens (Eds.). Psychology Press, 15–22. https://www.taylorfrancis.com/books/9781317769408/chapters/10.4324/9781315802725-5

[44] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 6. 10.

[45] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. PLoS ONE 15, 2 (2020), 1–20. https://doi.org/10.1371/journal.pone.0229132

[46] Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative Deceptive Opinion Spam. 9–14.

[47] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11). Association for Computational Linguistics, USA, 309–319.

[48] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, Yokohama Japan, 1–52. https://doi.org/10.1145/3411764.3445315

[49] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. 2019. Machine Behaviour. Nature 568, 7753 (April 2019), 477–486. https://doi.org/10.1038/s41586-019-1138-y

[50] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and

Data Mining. ACM, San Francisco California USA, 1135–1144. https://doi.org/10.1145/2939672.2939778
[51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. Proceedings of the AAAI Conference on Artificial Intelligence 32, 1 (April 2018). https://ojs.aaai.org/index.php/AAAI/article/view/11491
[52] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of {NLP} Models with {C}heck{L}ist. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442
[53] Steven Ritter, John R. Anderson, Kenneth R. Koedinger, and Albert Corbett. 2007. Cognitive Tutor: Applied research in mathematics education. Psychonomic Bulletin & Review 14, 2 (April 2007), 249–255. https://doi.org/10.3758/BF03194060
[54] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 5 (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x
[55] John Sweller. 2010. Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. Educational Psychology Review 22, 2 (June 2010), 123–138. https://doi.org/10.1007/s10648-010-9128-5
[56] Phoey Lee Teh, Paul Rayson, Irina Pak, and Scott Piao. 2015. Sentiment Analysis Tools Should Take Account of the Number of Exclamation Marks!!!. In Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services. ACM, Brussels Belgium, 1–6. https://doi.org/10.1145/2837185.2837216
[57] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. Patterns 1, 4 (2020), 100049. https://doi.org/10.1016/j.patter.2020.100049 Publisher: Elsevier Inc.
[58] Tamara Van Gog, Liesbeth Kester, and Fred Paas. 2011. Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. Applied Cognitive Psychology 25, 4 (July 2011), 584–587. https://doi.org/10.1002/acp.1726
[59] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Boston, MA, USA, 595–604. https://doi.org/10.1109/CVPR.2015.7298658
[60] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001. California Institute of Technology.
[61] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. {E}rrudite: Scalable, Reproducible, and Testable Error Analysis. Proceedings of the 57th Conference of the Association for Computational Linguistics (2019), 747–763. https://www.aclweb.org/anthology/P19-1073
[62] Tongshuang Wu, Daniel S. Weld, and Jeffrey Heer. 2019. Local Decision Pitfalls in Interactive Machine Learning: An Investigation into Feature Selection in Sentiment Analysis. ACM Transactions on Computer-Human Interaction 26, 4 (July 2019), 1–27. https://doi.org/10.1145/3319616
[63] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In Proceedings of the 25th International Conference on Intelligent User Interfaces. ACM, Cagliari Italy, 189–201. https://doi.org/10.1145/3377325.3377480
[64] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300509
[65] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. Proceedings of the ACM on Human-Computer Interaction 4, CSCW3 (Jan. 2021), 1–25. https://doi.org/10.1145/3432945
[66] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, Barcelona Spain, 295–305. https://doi.org/10.1145/3351095.3372852

A ADDITIONAL STUDY DETAILS

To improve the reproducibility of this work, we include additional details about the study here. Table 1 shows the final count of participants per condition after removing those who did not pass the attention check. Table 2 shows the specific language used for each behavior description in each task. Lastly, Figure 8 shows the specific instructions shown to participants before labeling for each task and the subsets of labels used in the two multiclass tasks.

Table 1. The number of participants per condition after removing failed attention checks.

	Reviews	Satellite	Birds
No AI	24	23	22
AI	23	24	23
AI + BD	23	25	25

Table 2. The specific behavior descriptions used for each behavior description group.

	Reviews	Satellite	Birds
Group 1	The AI is only 40% accurate for reviews like this with less than 50 words.	The AI confuses golf courses and meadows 60% of the time.	The AI confuses Cactus Wrens with House Wrens 60% of the time.
Group 2	The AI is only 20% accurate for reviews like this with more than 3 exclamation marks.	The AI confuses glaciers and clouds 80% of the time.	The AI confuses Read Headed Woodpecker with Red Bellied Woodpecker 80% of the time.

Received July 2022; revised October 2022; accepted January 2023

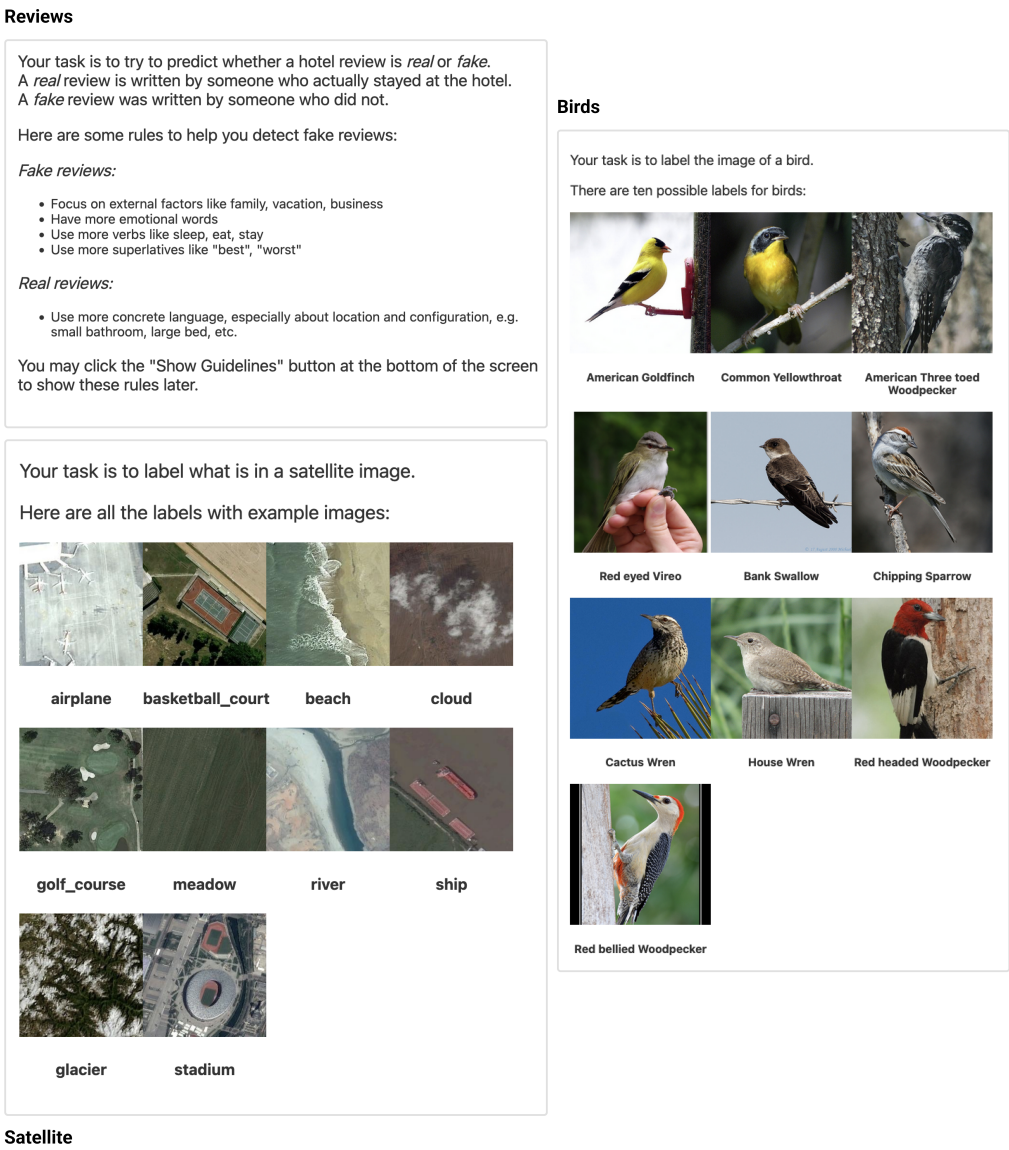

Fig. 8. The instructions shown to participants for each task before labeling.