On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection

Vivian Lai
University of Colorado Boulder
vivian.lai@colorado.edu

Chenhao Tan
University of Colorado Boulder
chenhao.tan@colorado.edu

ABSTRACT

Humans are the final decision makers in critical tasks that involve ethical and legal concerns, ranging from recidivism prediction, to medical diagnosis, to fighting against fake news. Although machine learning models can sometimes achieve impressive performance in these tasks, these tasks are *not* amenable to full automation. To realize the potential of machine learning for improving human decisions, it is important to understand how assistance from machine learning models affects human performance and human agency.

In this paper, we use deception detection as a testbed and investigate how we can harness explanations and predictions of machine learning models to improve human performance while retaining human agency. We propose a spectrum between full human agency and full automation, and develop varying levels of machine assistance along the spectrum that gradually increase the influence of machine predictions. We find that without showing predicted labels, explanations alone slightly improve human performance in the end task. In comparison, human performance is greatly improved by showing predicted labels (>20% relative improvement) and can be further improved by explicitly suggesting strong machine performance. Interestingly, when predicted labels are shown, explanations of machine predictions induce a similar level of accuracy as an explicit statement of strong machine performance. Our results demonstrate a tradeoff between human performance and human agency and show that explanations of machine predictions can moderate this tradeoff.

CCS CONCEPTS

• **Applied computing → Law, social and behavioral sciences**.

KEYWORDS

human agency, human performance, explanations, predictions

ACM Reference Format:
Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3287560.3287590

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287590

1 INTRODUCTION

Machine learning has achieved impressive success in a wide variety of tasks. For instance, neural networks have surpassed human-level performance in *ImageNet classification* (95.06% vs. 94.9%) [29]; Kleinberg et al. [36] demonstrate that in bail decisions, machine predictions of recidivism can reduce jail rates by 41.9% with no increase in crime rates, compared to human judges; Ott et al. [60] show that linear classifiers can achieve ∼90% accuracy in detecting deceptive reviews while humans perform no better than chance. As a result of these achievements, machine learning holds promise for addressing important societal challenges.

However, it is important to recognize different roles that machine learning can play in different tasks in the context of human decision making. In tasks such as object recognition, human performance can be considered as the upper bound, and machine learning models are designed to emulate the human ability to recognize objects in an image. A high accuracy in such tasks presents great opportunities for large-scale automation and consequently improving our society's efficiency. In contrast, efficiency is a lesser concern in tasks such as bail decisions. In fact, full automation is often not desired in these tasks due to ethical and legal concerns. These tasks are *challenging* for humans and for machines, but with vast amounts of data, machines can sometimes identify patterns that are *unsalient, unknown, or counterintuitive* to humans. If the patterns embedded in the machine learning models can be elucidated for humans, they can provide valuable support when humans make decisions.

The goal of our work is to investigate best practices for integrating machine learning into human decision making. We propose a spectrum between full human agency, where humans make decisions entirely on their own, and full automation, where machines make decisions without human intervention (see Figure 1 for an illustration). We then develop varying levels of machine assistance along the spectrum using explanations and predictions of machine learning models. We build on recent developments in interpretable machine learning that provide useful frameworks for generating explanations of machine predictions [34, 35, 45, 50, 64, 65]. Instead of using these explanations to help users debug machine learning models, we incorporate the explanations as assistance for humans to improve *human* performance while retaining human agency in the decision making process. Accordingly, we directly evaluate human performance in the end task through user studies.

In this work, we focus on a constrained form of decision making where humans make individual predictions. Specifically, we ask humans to decide whether a hotel review is genuine or deceptive based on the text. This prediction problem allows us to focus on the *integration* of machine learning into human predictions. In comparison, prior work in decision theory and decision support systems focuses on modeling preferences and utilities as well as building

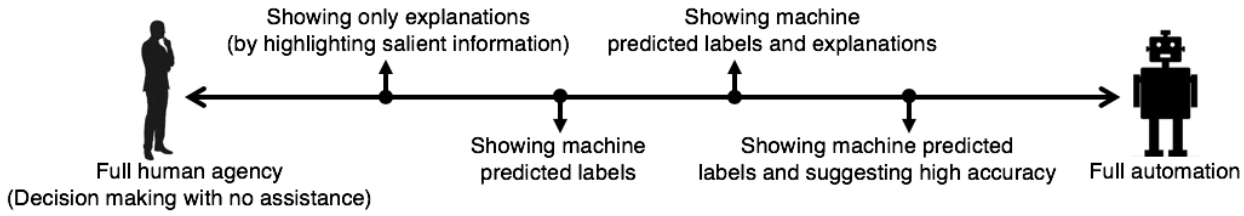


Figure 1: A spectrum between full human agency and full automation illustrating how machine learning can be integrated in human decision making. The detailed explanation of each method is in Section 3.

knowledge databases and representations to reason about complex decisions [5, 31, 33, 55, 67]. Moreover, since many policy decisions can be formulated as prediction problems [37], understanding human predictions with assistance from machine learning models constitutes an important step towards empowering humans with machine learning in critical challenging tasks.

Deception detection as a testbed. In this work, we use deception detection as our testbed for three reasons. First, deceptive information is prevalent on the Internet. For instance, Ott et al. [58] find that deceptive reviews are a growing problem on multiple platforms such as TripAdvisor and Yelp. Fake news has also received significant attention recently [43, 74] and might have influenced the outcome of the U.S. presidential election in 2016 [3]. Enhancing humans' ability in detecting deception can potentially alleviate these issues.

Second, deception detection is a challenging task for humans and has been extensively studied [1, 2, 22, 24, 60]. It is promising that machines show preliminary success in prior work. For example, machines are able to achieve an accuracy of ~90% in distinguishing genuine reviews from deceptive ones, while human performance is no better than chance [60]. Machines can identify unsalient and counterintuitive signals, e.g., deceptive reviews are less specific about spatial configurations and tend to include less sensorial and concrete language. It is worth noting that we should take the high machine accuracy with a grain of salt in the general domain because deception detection is a complex problem.[1] The task introduced by Ott et al. [60] nevertheless provides an ideal sandbox to understand human predictions with assistance from machine learning models.

Third, full automation is not desired in critical tasks such as deception detection because of ethical and legal concerns. The government should not have the authority to automatically block information from individuals, e.g., in the context of "fake news". Furthermore, full automation may not comply with legal requirements. For instance, in the case of recidivism prediction, the Wisconsin Supreme Court ruled that "judges be made aware of the limitations of risk assessment tools" and "a COMPAS risk assessment should not be used to determine the severity of a sentence or whether an offender is incarcerated" [47, 71]. Similarly, the trial judge is required to act as a gatekeeper regarding the evidence from a polygraph (lie detector) [70]. Therefore, it is crucial to retain human agency and understand human predictions with assistance from machine learning models.

[1]For instance, one can argue that it is impossible to fully address the issue of deception in online reviews only based on textual information as an adversarial user can copy another user's review, which becomes a deceptive review but with exactly the same text as a genuine one.

Organization and Highlights. We start by reviewing related work to provide the necessary background for our study (Section 2). Our focus in this work is on investigating human predictions with assistance from machine learning models in the context of deceptive review detection. To explore the spectrum between full human agency and full automation in Figure 1, we develop varying levels of assistance from machine learning models (Section 3). For example, the following three levels of machine assistance gradually increase the influence of machine predictions: 1) showing only explanations of machine predictions *without* revealing predicted labels; 2) showing predicted labels without revealing high machine accuracy; 3) showing predicted labels with an explicit statement of strong machine accuracy.

In Section 4, we investigate human performance under different experimental setups along the spectrum. We show that explanations alone slightly improve human performance, while showing predicted labels achieves great improvement (~21% relative improvement in human accuracy). However, this improvement is still moderate compared to "full" priming with an explicit statement of machine accuracy (~46% relative improvement in human accuracy). Our findings suggest that there exists a tradeoff between human performance and human agency. Interestingly, when predicted labels are shown, explanations of machine predictions can achieve a similar effect as an explicit statement of machine accuracy. We also find that humans tend to trust correct machine predictions more than incorrect ones, indicating that they can somewhat identify when machines are correct.

We further examine the effect of statements of machine accuracy by varying the accuracy numbers (Section 5). Surprisingly, we find that our participants are not sensitive to statements of machine accuracy and are more likely to trust machine predictions with an accuracy statement than without, even if the accuracy statement suggests poor machine performance. These observations echo with prior work on numeracy and suggest that it is difficult for humans to interpret and act on numbers [6, 62, 63, 69]. We also find that frequency explanations (e.g., 5 out of 10 for explaining 50%) can help humans calibrate the accuracy numbers. Note that we do not recommend these presentations on the spectrum because they present untruthful information.

We discuss the limitations of our work and provide concluding thoughts regarding future directions of investigating best practices for integrating artificial intelligence into human decision making in Section 6.

2 RELATED WORK

We summarize related work in two areas to put our work in context: interpretable machine learning and deception and misinformation.

Interpretable machine learning. Machine learning models remain as black boxes despite wide adoption. Blindly following machine predictions may lead to dire repercussions, especially in scenarios such as medical diagnosis and justice systems [9, 36, 73]. Therefore, improving their transparency and interpretability has attracted broad interest [34, 35, 45, 50, 64, 65], dating back to early work on recommendation systems [13, 30]. In the case of general automation, researchers have also studied issues of appropriate reliance and trust [8, 18, 44, 61, 76].

There are two major approaches to providing explanations of machine learning models: example-based and feature-based. For example, an example-based explanation framework is MMD-critic proposed by Kim et al. [34], which selects both prototypes and criticisms. Ribeiro et al. [64] propose a feature-based approach, LIME, that fits a sparse linear model to approximate non-linear models locally. Similarly, Lundberg and Lee [50] present a unified framework that assigns each feature an importance value for a particular prediction.

We would like to emphasize two unique aspects of our work: task difficulty and interpretability evaluation. First, compared to categorizing text into topics and object recognition, deception detection is a challenging task for humans and it remains an open question whether humans can leverage help from machine learning models in such settings. Second, we directly measure human performance in the end task. In comparison, prior work in interpretable machine learning aims to help humans understand how machine learning models work and/or debug them, the evaluation is thus mostly based on either the understanding of the models or the improvement in machine performance. Concurrently, several recent studies have also examined how explanations relate to human performance [10, 23]. Our work also resonates with the seminal work on mixed-initiative user interfaces [31] and intelligence augmentation [4]. In addition, our work is connected to cognitive studies on understanding effective explanations beyond the context of machine learning [48, 49].

Deception and misinformation. Deception is a widely studied phenomenon in many disciplines [75]. In psychology, deception is defined as an act that is intended to foster in another person a belief or understanding which the deceiver considers false [41]. To detect deception, researchers have examined the role of behavioral, emotional, and linguistic cues [17, 19, 39, 42, 54, 75].

Since people are increasingly relying on online reviews to make purchase decisions [11, 72, 78, 81], machine learning methods have been used to detect deception in online reviews [22, 24, 32, 60, 77, 79]. An important challenge in detecting deception in online reviews is to obtain the groundtruth labels of reviews. Ott et al. [60] create the first sizable dataset in deception detection by asking Amazon Mechanical Turkers to write deceptive reviews. Deceptive reviews can also be seen as an instance of spamming and online fraud [2, 16, 27, 56].

More recently, the issue of misinformation and fake news has drawn much attention from both the public and the research community [21, 43]. Most relevant to our work is Zhang et al. [80], which explores varying types of credibility annotations specifically designed for news articles. In addition, Nyhan and Reifler [57] demonstrate the "backfire" effect, which suggests that corrections of misperceptions may enhance people's false beliefs, and Vosoughi et al. [74] show that fake news is more innovative and spreads faster than real news.

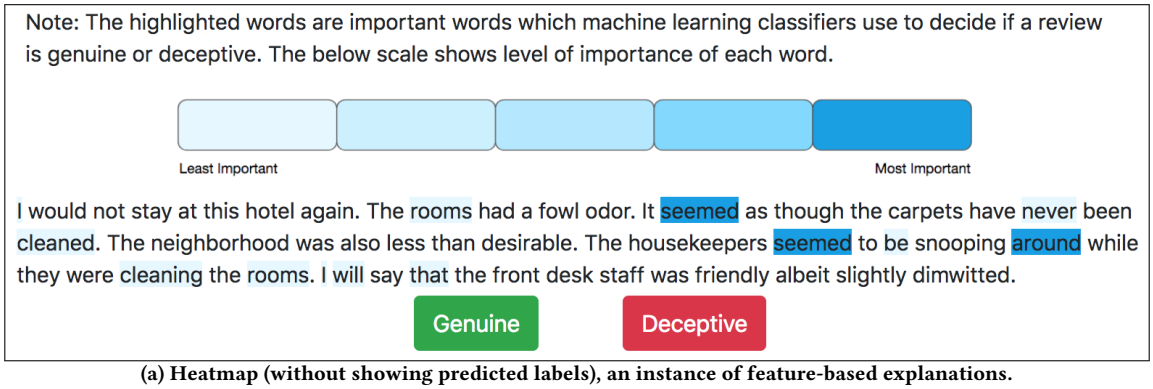
It is worth noting that deception detection is a broad and complex issue. For instance, fake news can be hard to define and may not be easily separated into two classes. Moreover, detecting fake news is different from detecting deceptive reviews as the former task requires other skills such as fact checking. It is important to note that our focus in this work is on investigating *how humans interact with assistance from machine learning algorithms in decision making*. We thus adopt the task of distinguishing genuine reviews from deceptive ones based on textual information in Ott et al. [60] as a sandbox. Our results on this constrained deception detection task can potentially contribute valuable insights to future solutions of the broader issue of deception detection.

3 EXPERIMENTAL SETUP AND HYPOTHESES

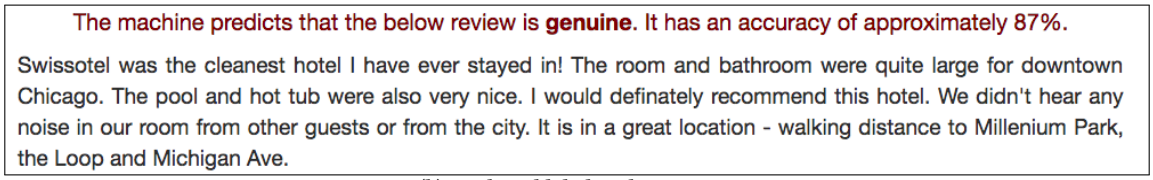
Our goal is to understand whether machine predictions and their explanations can improve human performance in challenging tasks, such as deception detection, and how humans interpret assistance from machine learning models. In this section, we first present our task setup and then develop varying levels of machine assistance along the spectrum introduced in Figure 1. We finally formulate our hypotheses and define our evaluation metrics.

Experimental setup. We employ the deception detection task developed by Ott et al. [60] and evaluate human performance in this task with varying levels of machine assistance. The dataset in Ott et al. [60] includes 800 genuine and 800 deceptive hotel reviews for 20 hotels in Chicago. The genuine reviews were extracted from TripAdvisor and the deceptive ones were written by turkers. We use 80% of the reviews as training data and the remaining 20% as the heldout test set. Since the machine performance with linear SVM in Ott et al. [60] already surpasses humans (~50%) by a wide margin and linear classifiers are generally considered more interpretable, we follow Ott et al. [59] and use linear SVM with bag-of-words features as our machine learning model. The linear SVM classifier achieves an accuracy of 87% on the heldout test set.

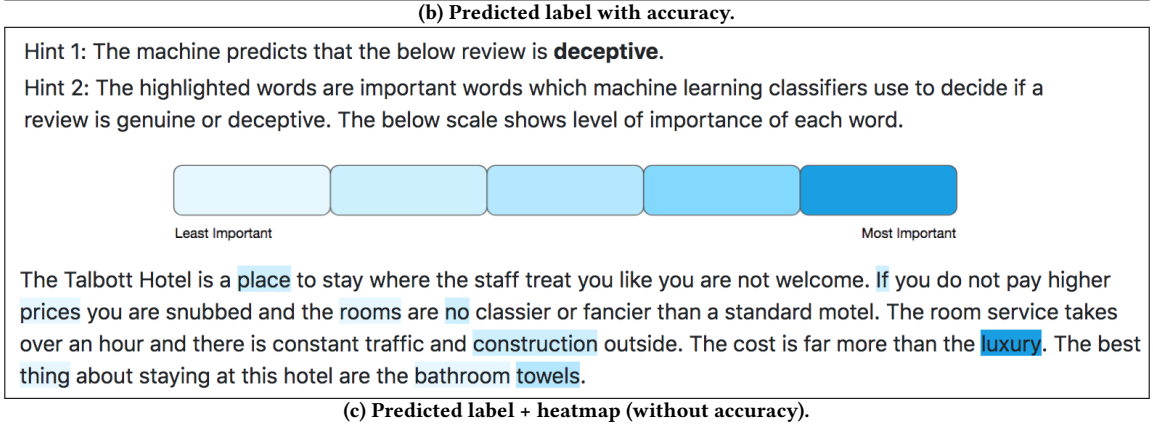
Our main task in this paper is to evaluate human performance with assistance from machine learning models. To do that, we conduct a user study on Amazon Mechanical Turk. Turkers are recruited to determine whether a review in the heldout test set is genuine or deceptive. In other words, humans are asked to perform the same task as the *machine* on the test set. We follow a between-subject design: each turker is assigned a level of machine assistance along the spectrum (Figure 1) and labels 20 reviews after going through three training examples and correctly answering an attention-check question. To incentivize turkers to perform at their best, we provide 40% bonus for each correct prediction in addition to the 5 cent base rate for a review. We also solicit our participants' estimation of their own performance and basic demographic information such as gender and education background through an exit survey. We only allow a turker to participate in the study once to guarantee sample independence across experimental



(a) Heatmap (without showing predicted labels), an instance of feature-based explanations.



(b) Predicted label with accuracy.



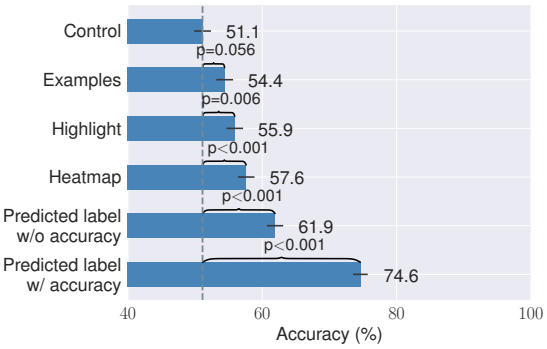
(c) Predicted label + heatmap (without accuracy).

Figure 2: Example interfaces with varying levels of machine assistance. Figure 2a only presents feature-based explanations of machine predictions in the form of *heatmap*. Figure 2b shows both the predicted label and an explicit statement about machine accuracy (87%). Figure 2c shows the predicted label with heatmap, but does not present machine accuracy. We crop the "Genuine" and "Deceptive" buttons in Figure 2b and 2c to save space.

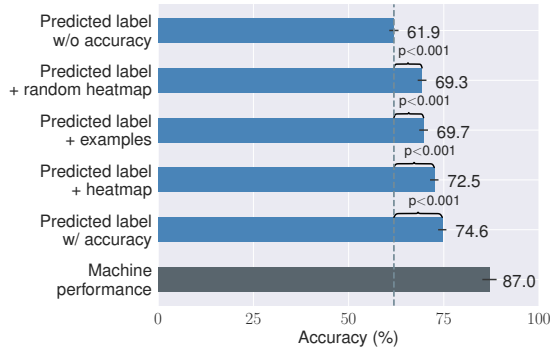
setups. Given that there are 320 test reviews and that we collect five turker predictions for each review, each experimental setup has a total of 80 turkers. Refer to the appendix for more details regarding our user study and survey questions.

Varying levels of machine assistance. Humans are the main agents in our experiments and make final decisions; machines only provide assistance, which can be ignored if humans deem it useless. An ideal outcome is that human performance can be improved with minimal information from machine learning models so that humans retain their agency in the decision making process. To examine how humans perform under different levels of influence from machine learning models, we consider the following presentations along the spectrum in Figure 1 (we only show three interfaces in Figure 2 for space reasons; see the appendix for more).

- **Control.** Humans are only presented a review. This setup contains no information from machine learning models and humans have full agency.
- **Feature-based explanations.** Since our machine learning model is linear, we present two versions of feature-based explanations by highlighting words based on absolute values of weight coefficients. First, we highlight the top 10 words in each review with the same color (*highlight*). Second, we use *heatmap* to show gradual changes in weight coefficients among the top 10 words. The most heavily-weighted words are highlighted in the darkest shade of blue. Soft-highlighting (heatmap) has been shown to improve visual search on targeted areas for humans [40]. Note that we do not indicate the sign of features to avoid revealing predicted labels. Humans may pay extra attention to the highlighted words and accordingly make decisions on their own. Figure 2a shows an example interface for *heatmap*.
- **Example-based explanations.** This method (*examples*) is inspired by example-based interpretable machine learning [34]. Humans are presented two additional reviews from the training data, one deceptive and one genuine that are most similar to the review under consideration. This setup resonates with nearest



(a) Human accuracy with varying levels of machine assistance.



(b) Human accuracy with predicted labels (and other information).

Figure 3: Human accuracy with varying levels of assistance. In Figure 3a, *control* provides no assistance; *examples, highlight,* and *heatmap* present explanations of machine predictions alone; *predicted label w/o accuracy* shows predicted labels; *predicted label w/ accuracy* shows predicted labels and reports machine accuracy that suggests strong machine performance. It is clear that showing predicted labels is crucial for improving human accuracy. Adding an explicit statement of machine accuracy further improves human accuracy. Figure 3b further investigates the combinations of predicted labels and their explanations, and presents *machine performance* as a benchmark. Intriguingly, we find that adding explanations achieves a similar effect as adding an explicit statement of machine accuracy. All p-values are computed by conducting t-test between the corresponding setup and the first experimental setup in the figure ("control" in Figure 3a and "predicted label w/o accuracy" in Figure 3b).

neighbor classifiers. Humans can potentially make better decisions in this setup than in *control* by comparing the similarity between reviews.

- **Predicted label without accuracy.** The above two approaches only show explanations of machine predictions, but do not reveal any information about predicted labels. The next level of priming presents the predicted label. If humans fully follow machine predictions, they will perform much better than chance and likely lead to an upper bound in this deception detection task for humans. However, humans may not trust the machine due to algorithm aversion [15].
- **Predicted label with accuracy.** We may further influence human decisions by explicitly suggesting that machines perform well in this task with 87% accuracy. Figure 2b shows an example for *predicted label with accuracy*. Note that such strong recommendations may not be desired due to ethical and legal concerns (see our discussion in the introduction).
- **Combinations.** Finally, we combine feature (example)-based explanations and predicted labels. Note that we do not show machine performance to avoid strong priming. Figure 2c shows an example of *predicted label + heatmap* without information about machine performance.

Hypotheses. We formulate the following hypotheses regarding *how well humans can perform with machine assistance* and *how often humans trust machine predictions when predicted labels are available.*

- *Hypothesis 1a.* Feature-based explanations and example-based explanations improve human performance over *control.*
- *Hypothesis 1b. Heatmap* is more effective than *highlight* as gradual changes in weight coefficients can be useful, as shown in Kneusel and Mozer [40] for visual search. Feature-based explanations are more effective than example-based explanations since the latter requires a greater cognitive load, i.e., reading two more reviews.
- *Hypothesis 2. Showing predicted labels* significantly improves human performance compared to feature (example)-based explanations alone. Assuming that humans trust the machine and follow its prediction, showing predicted labels can likely improve human performance because the machine accuracy is 87%. However, showing predicted labels reduces human agency, so it is important to understand the size of the performance gap and make informed design choices.
- *Hypothesis 3.* By combining predicted labels and feature (example)-based explanations, the trust that humans place on machine predictions increases, as it has been shown that concrete details can influence the level of trust in general automation [44].

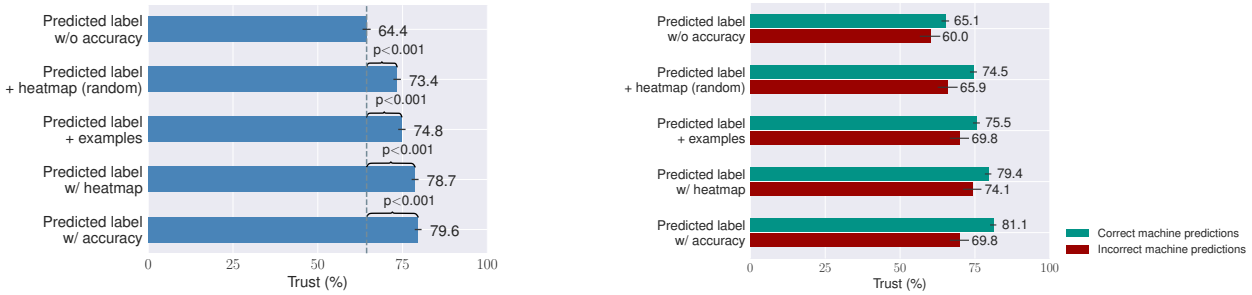
We evaluate the above hypotheses using two metrics, accuracy and trust. *Accuracy* is defined as the percentage of correctly predicted instances by humans; *trust* is defined as the percentage of instances for which humans follow the machine prediction. Note that we can only compute trust when predicted labels are available.

4 RESULTS

In this section, we investigate how varying levels of assistance from machine learning models along the spectrum in Figure 1 affect human predictions. We first discuss aggregate human performance using human accuracy and trust. Our results show that in this challenging task, explanations alone slightly improve human performance, while showing predicted labels can significantly improve human performance. When predicted labels are shown, we examine the level of trust that humans place on machine predictions. Our results suggest that humans can somewhat differentiate correct machine predictions from incorrect ones. Finally, we present individual differences among our participants based on information collected in the exit survey. Our dataset and demonstration are available at https://deception.machineintheloop.com/.

4.1 Human Accuracy

We first present human accuracy measured by the percentage of correctly predicted instances by humans. Our results suggest that



(a) Trust in machine predictions. **(b) Trust in correct and incorrect machine predictions.**

Figure 4: The trust that humans place on machine predictions. Figure 4a shows that adding feature-based explanations (heatmap) can effectively increase the trust level compared to *predicted label w/o accuracy*. p-value in Figure 4a is computed by conducting t-test between the corresponding setup and *predicted label w/o accuracy*. Figure 4b breaks down the trust based on whether machine predictions are correct or incorrect and shows that humans trust correct machine predictions more than the incorrect ones in all the five experimental setups, although the differences are only statistically significant in two setups.

showing predicted labels is crucial for improving human performance. Featured-based explanations coupled with predicted labels are able to induce similar human performance as an explicit statement of strong machine accuracy. As such, adding feature-based explanations to predicted labels may be more ideal than suggesting strong machine performance as the priming is weaker and may facilitate a higher level of human agency in decision making.

Explanations alone slightly improve human performance (Figure 3a). As Figure 3a shows, human performance in *control* is no better than chance (51.1%). This finding is consistent with Ott et al. [60] and decades of research on deception detection [7]. Explanations alone slightly improve human performance over control, and the differences are statistically significant for *highlight* and *heatmap*, not for *examples*. However, the best explanations, *heatmap*, is not statistically significantly different from *highlight* ($p = 0.335$) or *examples* ($p = 0.069$). As a result, our findings partially supports *Hypothesis 1a* and rejects *Hypothesis 1b*.

These findings suggest that it is difficult for humans to understand explanations on their own. This is plausible for example-based explanations since it requires extra cognitive burden and estimating text similarity is a nontrivial task for humans. For feature-based explanations, it seems that the improvement is driven by the small number of training reviews that we provide to explain the task. First-person singular pronouns provide a good example: one of the training reviews is deceptive and highlight many occurrences of the word, "my". A participant said, *"I tried to match the pattern from the example. In the example. the review with the most "My's" and "I's" were deceptive"*. In other words, the improvement in *heatmap* and *highlight* may not happen at all without the training reviews, which indicates the difficulty of interpreting these feature-based explanations and the importance of explaining the explanations. One possible direction is to develop automatic tutorials to teach the intuitions behind important features, which is related to machine teaching [51, 68, 82].

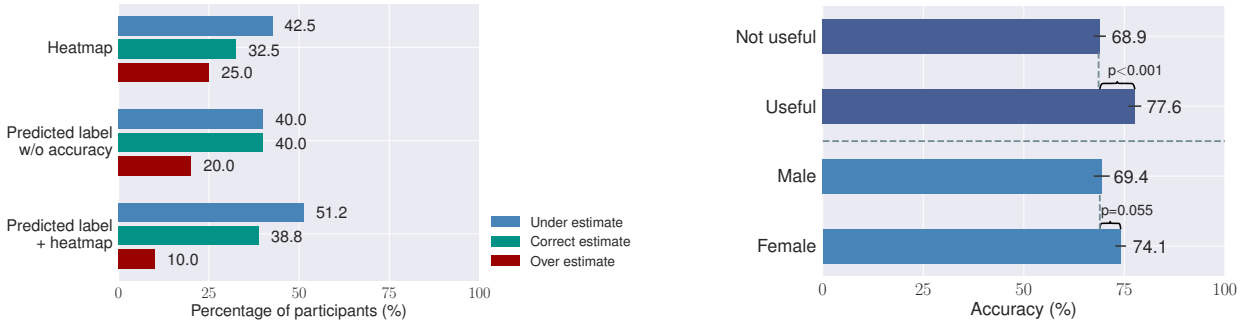
Showing predicted labels significantly improves human performance (Figure 3a and 3b). As Figure 3a shows, showing predicted labels drastically improves human performance (61.9% for *predicted label w/o accuracy*, a 21% relative improvement over *control*; the difference with *heatmap* is statistically significant ($p < 0.001$)). By presenting machine accuracy as shown in Figure 2b, the performance is further improved to 74.6% (*predicted label w/ accuracy* in Figure 3a, a 46% relative improvement over *control*).

These results are consistent with *Hypothesis 2*. The big performance gap between showing predicted labels and showing feature (example)-based explanations alone suggests that when humans interact with machine learning models, it makes a significant difference whether predicted labels are shown. However, this observation also echoes with concerns about humans overly relying on machines [44].

To further understand human performance with predicted labels, we examine all experimental setups with predicted labels in Figure 3b. Although showing predicted labels seems necessary for achieving sizable human performance improvement, the effect of presenting machine accuracy can be moderated by showing feature (example)-based explanations. We find that *predicted label + examples* and *predicted label + heatmap* outperform *predicted label w/o accuracy* (69.7% and 72.5% vs. 61.9%), without presenting the machine accuracy. In this case, we observe that heatmap is more effective than examples, and leads to comparable human performance with *predicted label w/ accuracy*. There is still a gap between the best human performance (*predicted label w/ accuracy*) and *machine performance* (74.6% vs. 87.0%). These observations suggest that humans do not necessarily trust machine predictions.

4.2 Trust

We further examine the levels of trust that humans place on machine predictions when predicted labels are available. Since machine performance surpasses human performance in *control* by a wide margin in this task, higher levels of trust are correlated with higher levels of accuracy in our experiments. However, these two metrics capture different dimensions of human predictions because trust is tied to machine predictions. This becomes clear when we break down human trust by whether machine predictions are correct or not. We find that humans tend to trust correct machine predictions more than incorrect ones, which suggests that humans can somewhat effectively identify cases where machines are wrong. It is important to emphasize that our focus is on understanding how



(a) Human estimation of their own performance. **(b) Gender and hint usefulness in *predicted label + heatmap*.**

Figure 5: Heterogeneity findings among participants in our study. Figure 5a shows performance estimation by participants in three different experimental setups. Figure 5b presents the performance of participants in *predicted label + heatmap* group by two variables, hint usefulness and gender.

human trust varies along the spectrum rather than manipulating the trust of humans in machines.

Feature (example)-based explanations increase the trust that humans place on machine predictions (Figure 4a). We further introduce random heatmap by randomly highlighting an equal number of words as in *heatmap* to examine whether humans are influenced by any explanations including random ones.

Our results are consistent with *Hypothesis 3*: both feature-based and example-based explanations increase the trust of humans in machine predictions. In fact, *predicted label + heatmap* leads to a similar level of trust as *predicted label w/ accuracy*, although the latter explicitly tells humans that the machine learning model "has an accuracy of approximately 87%". In other words, when predicted labels are shown, heatmap can nudge humans in decision making without making strong statements of machine accuracy. Interestingly, random heatmap also increases the trust level significantly, suggesting that even irrelevant details can increase the trust of humans in machine predictions. The fact that heatmap is significantly more effective than random heatmap (78.7% vs. 73.4%, $p < 0.001$) indicates that humans can interpret valuable information in weight coefficients beyond "the placebo effect".

Humans tend to trust machine predictions more when machine predictions are correct. (Figure 4b). We next examine whether humans trust machine predictions more when machine predictions are correct than when they are incorrect. Figure 4b shows that in all the five experimental setups with predicted labels, our participants trust correct machine predictions more than incorrect ones. However, the difference is statistically significant only in *predicted label w/ accuracy* ($p < 0.001$) and *predicted label w/ heatmap (random)* ($p = 0.015$). These results suggest that humans can somewhat differentiate correct machine predictions from incorrect ones. Further evidence is required to fully understanding the reasons why humans (don't) trust (in)correct machine predictions. Such understandings can improve both machine learning models and their presentations to support human decision making.

4.3 Heterogeneity in Human Perception and Performance

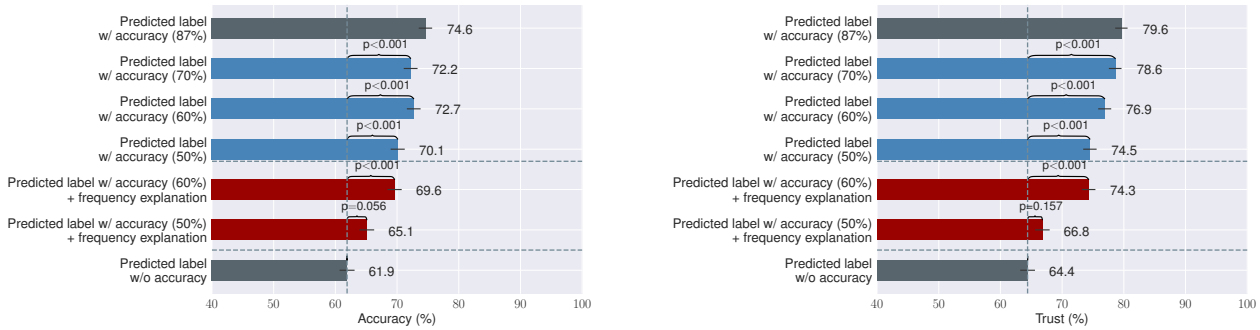
We finally discuss the heterogeneity between participants in our study. Here we focus on the participants' estimation of their own performance and gender differences. Refer to the appendix for additional comparisons.

Human estimation of their own performance (Figure 5a). We ask participants to estimate their own performance in our exit survey. Our results are not exactly aligned with the previous finding that humans tend to overestimate their capacity of detecting lying [20]. In fact, ~42% of the participants correctly predicted their performance. Among the remaining, ~18% overestimated their performance, while ~40% underestimated their performance. Figure 5a shows the breakdown for three experimental setups. In general, it seems difficult for humans to estimate their performance. One participant who overestimated his performance (estimated 11-15 but got 10 correct) said, *"I enjoyed this hit. When I was a young man, I was a manager in the hotel business and got to read a lot of comment cards from guests. I hope that I was pretty accurate in my answers"*. Another participant who underestimated his performance (estimated 6-10 but got 15 correct) said, *"It was difficult to determine if they were genuine or deceptive. I don't feel certain on any of my choices"*.

Heterogeneity in performance across individuals (Figure 5b). We have so far focused on average human performance comparisons between different experimental setups. It is important to recognize that the performance of individuals can vary. Exit survey responses allow us to study such heterogeneity. We focus on two properties in the interest of space. Refer to the appendix for a complete discussion of heterogeneity between individuals.

First, individuals who find the hints useful outperform those who find the hints not useful. The difference between these two groups in Figure 5b (*predicted label + heatmap*) is statistically significant. This observation resonates with our analysis regarding the trust of humans in machine predictions and holds in 5 out of 8 experimental setups (this question was not asked in *control*), although the differences are only statistically significant in three setups.[2] Second, we find that females generally outperform males. This observation holds in 8 out of 9 experimental setups, but none of the differences is statistically significant. Our results contribute to mixed observations regarding gender differences in deception detection [14, 46, 52, 53].

[2]The low number of statistically significant differences is expected, because human performance is low unless we show predicted labels.



(a) Human accuracy. **(b) Trust.**

Figure 6: Human accuracy and trust given varying statements of machine accuracy. Figure 6a and Figure 6b show that human accuracy and trust generally decline with statements of decreasing machine accuracy despite the fact that machine predictions remain unchanged. Note that the decline of human trust with statements of decreasing accuracy is small. Only by adding frequency explanations, human accuracy and trust become closer to not showing any indication of machine accuracy, i.e., *predicted label w/o accuracy*.

5 VARYING STATEMENTS OF MACHINE ACCURACY

Given the strong influence of predicted labels and machine accuracy, a natural question to ask is how human judgment changes if we vary the statement of machine accuracy. For example, instead of the true accuracy of 87%, we could claim that the machine has an accuracy of 60%. It is important to emphasize that since these statements of accuracy are not true, we do not recommend this approach as part of our spectrum in Figure 1 and thus put these results in a separate section. However, we think that it is valuable to understand how varying statements of accuracy might influence human predictions.

Although human accuracy and trust generally decline with statements that suggest lower accuracy, statements of machine accuracy improve human trust in machine predictions even when the claimed accuracy is only 50%. To understand human accuracy with varying statements of machine accuracy, we use *predicted label w/o accuracy* and *predicted label w/ accuracy (87%)* as benchmarks. In Figure 6a and Figure 6b, human accuracy and trust with varying statements of machine accuracy all fall between these two benchmarks as expected. Here we focus on the blue bars filled with forward slashes that correspond to simple statements of machine accuracy, "The machine predicts that the below review is deceptive. It has an accuracy of approximately x%" ($x = 70, 60, 50$). As the claimed accuracy declines from 87% to 50%, human accuracy and trust decrease, with the exception of human accuracy from 70% to 60%. However, the decline in human trust and accuracy is fairly small. For instance, *predicted label w/ accuracy (50%)* still outperforms *predicted label w/o accuracy* significantly. The results are surprising and counterintuitive since one should put less trust in a machine that has only an accuracy of 50% as compared to a machine that boasts 87%. Our findings suggest that any indication of machine accuracy, be it high or low, improves human trust in the machine. This observation echoes prior work on numeracy that suggests that average humans and even doctors struggle with interpreting and acting on numbers [6, 62, 63, 69]. Therefore, it is crucial that we develop a better *empirical* understanding of how humans interact with explanations and predictions of machine learning models in decision making before using these machine learning models in the loop of human decision making.

Frequency explanations can help humans interpret and act on statements of machine accuracy. To further investigate human interaction with varying statements of machine accuracy, we add frequency explanations to the statement with accuracy 50% and 60%. Specifically, we show participants "The machine predicts that the below review is *deceptive*. It has an accuracy of approximately 50%, which means that it is correct 5 out of 10 times." instead of "The machine predicts that the below review is *deceptive*. It has an accuracy of approximately 50%." The results are shown with the red bars filled with stars in Figure 6a and Figure 6b. We find that frequency explanations reduce the trust that humans place on machine predictions. For instance, human accuracy in *predicted label w/ accuracy (50%) + frequency explanation* is ~7% lower (p=0.003) than in *predicted label w/ accuracy (50%)*. Similarly, human trust in *predicted label w/ accuracy (50%) + frequency explanation* is ~10% lower (p<0.001) than in *predicted label w/ accuracy (50%)*. Furthermore, the differences in human accuracy and trust are not statistically significant between *predicted label w/ accuracy (50%) + frequency explanation* and *predicted label w/o accuracy*. These observations suggest that frequency explanations can help humans interpret statements of machine accuracy, in which case a statement of 50% accuracy with frequency explanation is almost the same as not showing machine accuracy. Our frequency explanations are also known as frequent format and have been shown to be more effective for conveying uncertainty than stating the probability [25, 26, 66].

6 CONCLUDING DISCUSSION

In this paper, we conduct the first empirical study to investigate whether machine predictions and their explanations can improve human performance in challenging tasks such as deception detection. We propose a spectrum between full human agency and full automation, and design machine assistance with varying levels of priming along the spectrum. We find that explanations alone slighlty improve human performance, while showing predicted labels significantly improves human performance. Adding an explicit statement of strong machine performance can further improve

human performance. Our results demonstrate a tradeoff between human performance and human agency, and explaining machine predictions may moderate this tradeoff.

We find interesting results regarding the trust that humans place on machine predictions. On the one hand, humans tend to trust correct machine predictions more than incorrect ones, which indicates that it is possible to improve human decision making while retaining human agency. On the other hand, we show that human trust can be easily enhanced by adding random heatmap as explanations or statements of low accuracies that do not justify trusting machine predictions. In other words, additional details including irrelevant ones can improve the trust that humans place on machine predictions. These findings highlight the importance of taking caution in using machine learning for supporting decision making and developing methods to improve the transparency of machine learning models and its associated human interpretation.

As machine learning gets employed to support decision making in our society, it is crucial that the machine learning community not only advances machine learning models, but also develops a better understanding of how these machine learning models are used and how humans interact with these models in the process of decision making. Our study takes an initial step towards understanding human predictions with assistance from machine learning models in challenging tasks.

Implications and future directions. Our results show that explanations alone slightly improve human performance. One reason for the limited improvement with explanations alone is that although we provide explanations during the decision making process, we provide limited resources to "teach" these explanations. A possible future direction is to develop tutorials for machine learning models and their explanations to relieve some cognitive burden from humans, e.g., summarizing the model as a list of rules, adding heatmap in examples or providing a sequence of training examples with explanations and sufficient coverage. This direction also connects to the area of machine teaching [51, 68, 82].

Another possible direction to improve the effectiveness of explanations is to provide narratives. Our results suggest that feature-based and example-based explanations provide useful details for machine predictions to improve the trust of humans in machine predictions. It can be useful if we can similarly provide rationales behind feature-based and example-based explanations in the form of narratives. A qualitative understanding of how turkers interpret hints from machine learning models may shed light on the requirements of effective narratives.

Last but not least, it is important to study the ethical concerns of providing assistance from machine learning models in human decision making. Our results demonstrate a clear tradeoff in this space: it is difficult to improve human performance without showing predicted labels, but showing predicted labels, especially alongside machine performance, runs the risk of removing human agency. Human decision makers with assistance from machines further complicate the current discussions on the issue of fairness in algorithmic decision making [12, 28, 38]. As the adoption of machine learning approaches can have broad impacts on our society, such questions require inputs from machine learning researchers, legal scholars, and the entire society.

Limitations. We use Amazon Mechanical Turk to recruit participants, but this may not be a representative sample of the population. However, we would like to emphasize that turkers are likely to provide a better proxy than machine learning experts for understanding how humans interact with assistance from machine learning models in critical challenging tasks. Also, our explanations are derived from a linear SVM classifier and nearest neighbors. It may be even more challenging for humans to interpret explanations of non-linear classifiers.

Another important challenge in understanding how humans interact with machine learning models lies in the difficulty to assess the generalizability of our results. Our formulation of deception detection represents a scenario where machines outperform humans by a wide margin and humans may have developed false beliefs about this task, as most humans have read reviews online. In order to consider a wide range of tasks, e.g., bail decisions and medical diagnosis, we need a framework to compare different tasks. Machine performance and humans' prior intuition are probably important factors that can influence human interpretation of the explanations. However, it remains an open question whether there exists a principled framework to reason about these tasks. At the very least, it is important for our community to go beyond simple visual tasks such as OCR and object recognition, especially for the purpose of improving human performance in decision making.

REFERENCES

[1] Mohamed Abouelenien, Veronica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2014. Deception detection using a multimodal approach. In *Proceedings of ICMI*.

[2] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. 2013. Opinion Fraud Detection in Online Reviews by Network Effects.. In *Proceedings of ICWSM*.

[3] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–236.

[4] W Ross Ashby. 1957. An introduction to cybernetics. (1957).

[5] James O Berger. 2013. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

[6] Donald M Berwick, Harvey V Fineberg, and Milton C Weinstein. 1981. When doctors meet numbers. *The American journal of medicine* 71, 6 (1981), 991–998.

[7] Charles F Bond Jr and Bella M DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review* 10, 3 (2006), 214–234.

[8] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on*. IEEE, 160–169.

[9] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of KDD*.

[10] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *Proceedings of EMNLP*.

[11] Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43, 3 (2006), 345–354.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of KDD*.

[13] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of CHI*.

[14] Bella M DePaulo, Jennifer A Epstein, and Melissa M Wyer. 1993. Sex differences in lying: How women and men deal with the dilemma of deceit. (1993).

[15] 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.

[16] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. 1999. Support vector machines for spam categorization. *IEEE Transactions on Neural networks* 10, 5 (1999), 1048–1054.

[17] Earl F Dulaney. 1982. Changes in language behavior as a function of veracity. *Human Communication Research* 9, 1 (1982), 75–82.

[18] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

[19] Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology* 39, 6 (1980), 1125.

[20] Eitan Elaad. 2003. Effects of feedback on the overestimated capacity to detect lies and the underestimated ability to tell lies. *Applied Cognitive Psychology* 17, 3 (2003), 349–363.

[21] Diane Farsetta and Daniel Price. 2006. Fake TV news: Widespread and undisclosed. *Center for Media and Democracy* 6 (2006).

[22] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of ACL (short papers)*.

[23] Shi Feng and Jordan Boyd-Graber. 2018. What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play. *arXiv preprint arXiv:1810.09648* (2018).

[24] Vanessa Wei Feng and Graeme Hirst. 2013. Detecting deceptive opinions with profile compatibility. In *Proceedings of IJCNLP*.

[25] Gerd Gigerenzer. 1996. The psychology of good judgment: frequency formats and simple algorithms. *Medical decision making* 16, 3 (1996), 273–280.

[26] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review* 102, 4 (1995), 684.

[27] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with trustrank. In *Proceedings of VLDB*.

[28] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of NIPS*.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of ICCV*.

[30] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of CSCW*.

[31] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of CHI*.

[32] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of WSDM*.

[33] Peter GW Keen. 1978. *Decision support systems; an organizational perspective*. Technical Report.

[34] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of NIPS*.

[35] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Proceedings of NIPS*.

[36] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293.

[37] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.

[38] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *Proceedings of ITCS*.

[39] Mark L Knapp, Roderick P Hart, and Harry S Dennis. 1974. An exploration of deception as a communication construct. *Human communication research* 1, 1 (1974), 15–29.

[40] Ronald T Kneusel and Michael C Mozer. 2017. Improving Human-Machine Cooperative Visual Search With Soft Highlighting. *ACM Transactions on Applied Perception (TAP)* 15, 1 (2017), 3.

[41] Robert M Krauss, Valerie Geller, and Christopher Olson. 1976. Modalities and cues in the detection of deception. In *Meeting of the American Psychological Association, Washington, DC*.

[42] Mark L Knapp and Mark E Comaden. 1979. Telling it like it isn't: A review of theory and research on deceptive communications. *Human Communication Research* 5, 3 (1979), 270–285.

[43] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[44] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[45] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *Proceedings of EMNLP*.

[46] Li Li. 2011. Sex differences in deception detection.

[47] Adam Liptak. 2017. Sent to Prison by a Software Program's Secret Algorithms.

[48] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[49] Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology* 55, 3 (2007), 232–257.

[50] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of NIPS*.

[51] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. 2018. Teaching categories to human learners with visual explanations. In *Proceedings of CVPR*.

[52] Samantha Mann, Aldert Vrij, and Ray Bull. 2004. Detecting true lies: police officers' ability to detect suspects' lies. *Journal of applied psychology* 89, 1 (2004), 137.

[53] Steven A McCornack and Malcolm R Parks. 1990. What women know that men don't: Sex differences in determining the truth behind deceptive messages. *Journal of Social and Personal Relationships* 7, 1 (1990), 107–118.

[54] Albert Mehrabian. 1971. *Silent messages*. Vol. 8. Wadsworth Belmont, CA.

[55] Allen Newell and Herbert Alexander Simon. 1972. *Human problem solving*. Vol. 104. Prentice-Hall Englewood Cliffs, NJ.

[56] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of WWW*.

[57] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[58] Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of WWW*.

[59] Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of NAACL*.

[60] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of ACL*.

[61] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[62] Ellen Peters, Daniel Västfjäll, Paul Slovic, CK Mertz, Ketti Mazzocco, and Stephan Dickert. 2006. Numeracy and decision making. *Psychological science* 17, 5 (2006), 407–413.

[63] Valerie F Reyna and Charles J Brainerd. 2008. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and individual differences* 18, 1 (2008), 89–107.

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

[65] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of AAAI*.

[66] Peter Sedlmeier and Gerd Gigerenzer. 2001. Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General* 130, 3 (2001), 380.

[67] Jung P Shim, Merrill Warkentin, James F Courtney, Daniel J Power, Ramesh Sharda, and Christer Carlsson. 2002. Past, present, and future of decision support technology. *Decision support systems* 33, 2 (2002), 111–126.

[68] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. 2014. Near-Optimally Teaching the Crowd to Classify.. In *Proceedings of ICML*.

[69] Paul Slovic and Ellen Peters. 2006. Risk perception and affect. *Current directions in psychological science* 15, 6 (2006), 322–325.

[70] Supreme Court of the United States. 1993. Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579.

[71] Supreme Court of Wisconsin. 2016. State of Wisconsin, Plaintiff-Respondent, v. Eric L. Loomis, Defendant-Appellant.

[72] Michael Trusov, Randolph E Bucklin, and Koen Pauwels. 2009. Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of marketing* 73, 5 (2009), 90–102.

[73] Kush R Varshney. 2016. Engineering safety in machine learning. In *Information Theory and Applications Workshop (ITA), 2016*.

[74] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[75] Aldert Vrij. 2000. *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley.

[76] Christopher D Wickens, Justin G Hollands, Simon Banbury, and Raja Parasuraman. 2015. *Engineering psychology & human performance*. Psychology Press.

[77] Guangyu Wu, Derek Greene, Barry Smyth, and Pádraig Cunningham. 2010. Distortion as a validation criterion in the identification of suspicious reviews. In *Proceedings of the First Workshop on Social Media Analytics*.

[78] Qiang Ye, Rob Law, Bin Gu, and Wei Chen. 2011. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human behavior* 27, 2 (2011), 634–639.

[79] Kyung-Hyan Yoo and Ulrike Gretzel. 2009. Comparison of deceptive and truthful travel reviews. *Information and communication technologies in tourism 2009* (2009), 37–47.

[80] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Proceedings of WWW (Companion)*.

[81] Ziqiong Zhang, Qiang Ye, Rob Law, and Yijun Li. 2010. The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management* 29, 4 (2010), 694–700.

[82] Xiaojin Zhu. 2015. Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education. In *Proceedings of AAAI*.

ACKNOWLEDGMENTS

We would like to thank Elizabeth Bradley, Michael Mozer, Sendhil Mullainathan, Amit Sharma, Adith Swaminathan, and anonymous reviewers for helpful discussions and feedback. This material is based upon work supported by the National Science Foundation under Grant No. 1837986. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

A APPENDIX

A.1 Amazon Mechanical Turk Setup

To ensure quality results, we include several criteria for turkers: 1) the turker is based in the United States so that we assume English fluency; 2) the turker has completed at least 50 HITs (human intelligence tasks); 3) the turker has an approval rate of at least 99%.

Before working on the main task, turkers need to go through a short training session, in which we show three reviews from the training data. We present the correct answer after turkers make their prediction. The interface during training is exactly the same as in the actual experiment. After making predictions for 20 reviews, turkers are required to fill out an exit survey that solicits their estimation of their own performance in this task and basic demographic information including age, gender, education background, and experience with online reviews (screenshots in Figure 15 and Figure 16). If the HIT is approved, the turker is compensated a dollar and bonuses depending on the number of reviews he correctly predicted. For example, if a turker makes 11 correct predictions, he is compensated $0.22 in addition to a dollar. The average duration for finishing our HIT is about 11 minutes (Figure 7 shows the CDF of the duration). Turkers spend the shortest amount of time on average (8.3 minutes) in *predicted labels w/ accuracy* and the longest amount of time on average (14.4 minutes) in *examples*, which is consistent with our expectation about extra cognitive burden from reading two more reviews. To sanity check that participants pay similar attention throughout the study, Figure 8 shows the average accuracy with respect to the order in which reviews show up[3]: there does not exist a downward trend. All results are based on the 9 experimental setups in Section 4 of the main paper and results with varying statements of accuracy are not included.

A.2 Experiment Interfaces

This section shows example interfaces for the other five experimental setups that are not shown in the main paper (*predicted label + heatmap (random)* has the same interface as *predicted label + heatmap* except that words are highlighted randomly).

- Control (Figure 17a).
- Highlight (Figure 17b).
- Examples (Figure 18a).
- Predicted label w/o accuracy (Figure 18b).
- Predicted label + examples (Figure 19).

[3]Thanks to suggestions from anonymous reviewers.

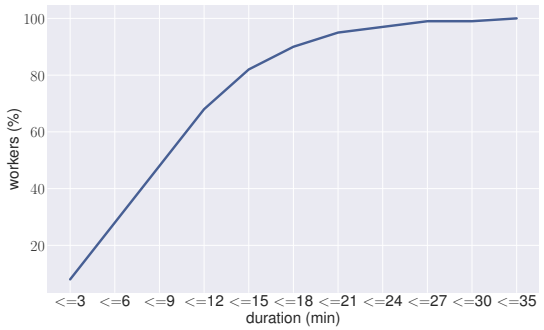


Figure 7: Cumulative distribution of study duration in 9 experimental setups.

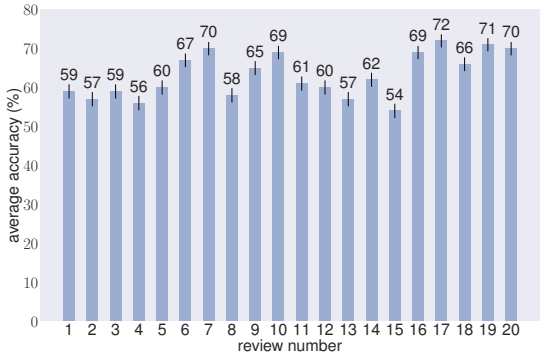


Figure 8: Average accuracy with respect to review ordering in 9 experimental setups.

A.3 Individual Differences

Here we present further results on heterogeneous performance among individuals. We present figures for four experimental setups that are representative of different levels of priming: *heatmap*, *examples*, *predicted label w/o accuracy*, and *predicted label + heatmap*.

Hint usefulness (Figure 9). As discussed in the main paper, human performance is better for participants who find hints useful than those who do not find hints useful in 5 out of 8 experimental setups. *Highlight*, *heatmap* and *predicted label w/o accuracy* are the exceptions. The difference in three setups (*predicted label + heatmap*, *predicted label + heatmap (random)*, *predicted label w/ accuracy*) is statistically significant.

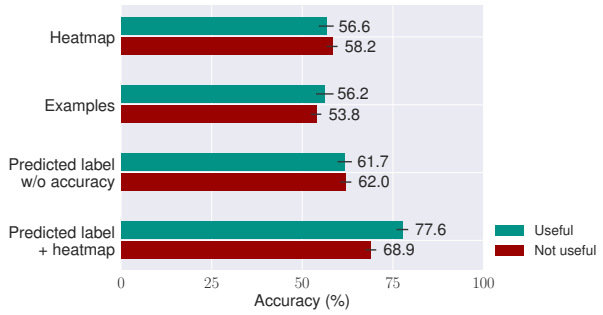


Figure 9: Human accuracy vs. usefulness of hints.

Gender differences (Figure 10). Females generally outperform males, in 8 out of 9 experimental setups. None of the differences is statistically significant.

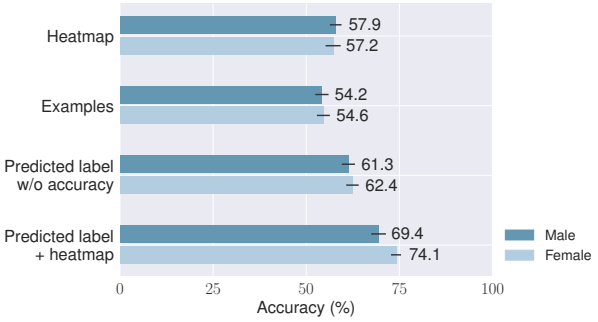


Figure 10: Human accuracy vs. gender.

Review sentiments (Figure 11). One possible hypothesis is that humans perform differently depending on the sentiment of reviews. Indeed, we observe that humans consistently perform better for positive reviews (8 out 9 experimental setups). However, the difference is only statistically significant for *predicted label w/o accuracy*.

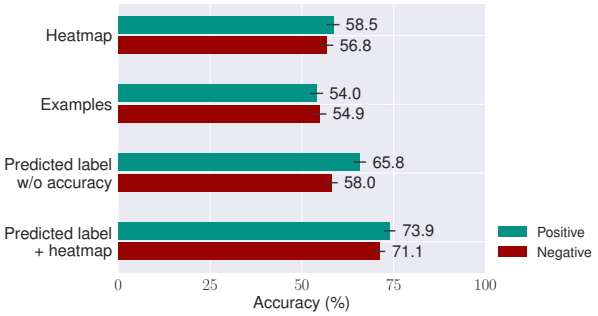


Figure 11: Human accuracy vs. review sentiment.

Education background (Figure 12). There is no clear trend regarding education background, which suggests that education levels do not correlate with the ability to detect deception. For instance, high school graduates perform the best in *predicted label w/o accuracy*, but the worst in *examples*. Since there are five groups, each group is relatively sparse. We thus did not conduct statistical testing for these observations.

Age group (Figure 13). There is no clear trend regarding age groups either. For instance, participants that are 61 & above perform the best in *predicted label w/o accuracy*, but worst in *predicted label + heatmap*. Similarly, since there are five groups and that each group is also relatively sparse, we did not conduct statistical testing for these observations.

Review experience (Figure 14). There is no clear trend regarding experience of writing reviews. With the exception of *control* and *predicted label + heatmap (random)*, the group that reports the best performance is either users who write reviews weekly or users who write reviews frequently. Again, we did not conduct statistical testing for review experience.

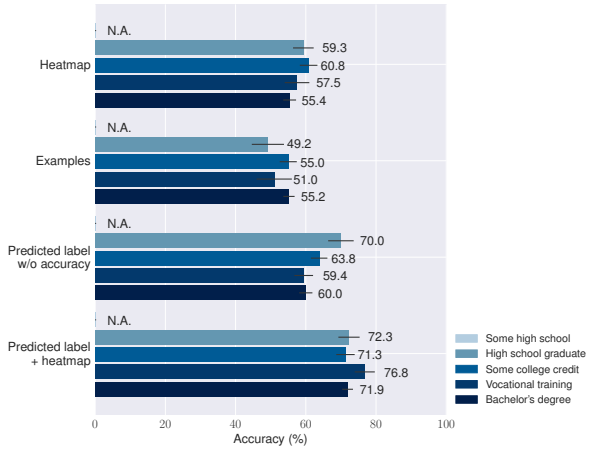


Figure 12: Human accuracy vs. education background.

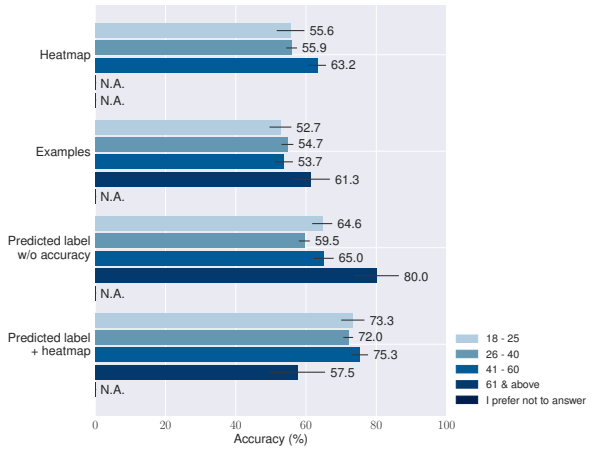


Figure 13: Human accuracy vs. age groups.

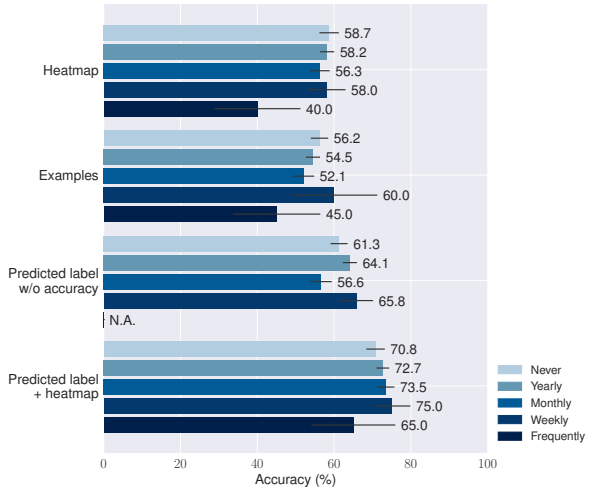


Figure 14: Human accuracy vs. review writing experience.

***1. How many answers do you think that you have answered correctly?**

- 0-5
- 6-10
- 11-15
- 16-20

***2. What is your gender?**

- Female
- Male
- I prefer not to answer

***3. What is your age?**

- 18-25
- 26-40
- 41-60
- 61 and above
- I prefer not to answer

***4. What is the highest degree or level of school you have completed? If currently enrolled, select highest degree received.**

- Some high school, no diploma, and below
- High school graduate, diploma or the equivalent (for example: GED)
- Some college credit, no degree
- Trade/technical/vocational training
- Bachelor's degree, and above
- I prefer not to answer

***5. How often do you write reviews on the Internet?**

- Never
- Yearly
- Monthly
- Weekly
- More frequently than weekly

***6. How often do you make purchase decisions based on online reviews?**

- Never
- Yearly
- Monthly
- Weekly
- More frequently than weekly

7. Please give us your feedback.

Figure 15: Survey questions for control group.

*1. How many answers do you think that you have answered correctly?
- 0-5
- 6-10
- 11-15
- 16-20

*2. What is your gender?
- Female
- Male
- I prefer not to answer

*3. What is your age?
- 18-25
- 26-40
- 41-60
- 61 and above
- I prefer not to answer

*4. What is the highest degree or level of school you have completed? If currently enrolled, select highest degree received.
- Some high school, no diploma, and below
- High school graduate, diploma or the equivalent (for example: GED)
- Some college credit, no degree
- Trade/technical/vocational training
- Bachelor's degree, and above
- I prefer not to answer

*5. How often do you write reviews on the Internet?
- Never
- Yearly
- Monthly
- Weekly
- More frequently than weekly

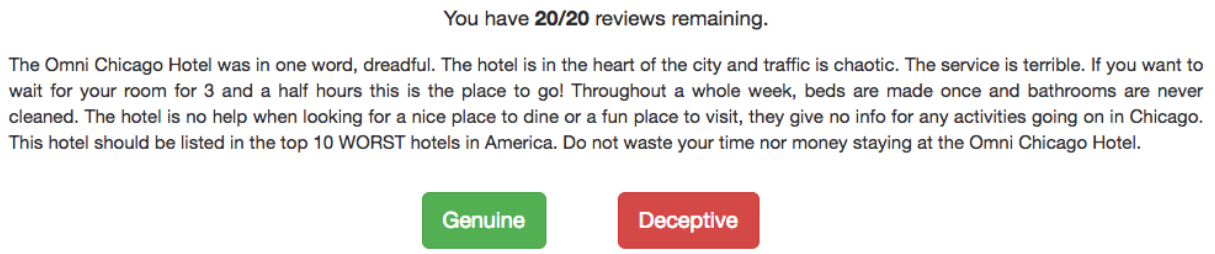
*6. How often do you make purchase decisions based on online reviews?
- Never
- Yearly
- Monthly
- Weekly
- More frequently than weekly

*7a. Did giving you hints (e.g. highlight of words, displaying genuine and deceptive review, machine's prediction) on reviews influence your decision?
- Yes
- No

*7b. Please explain how.

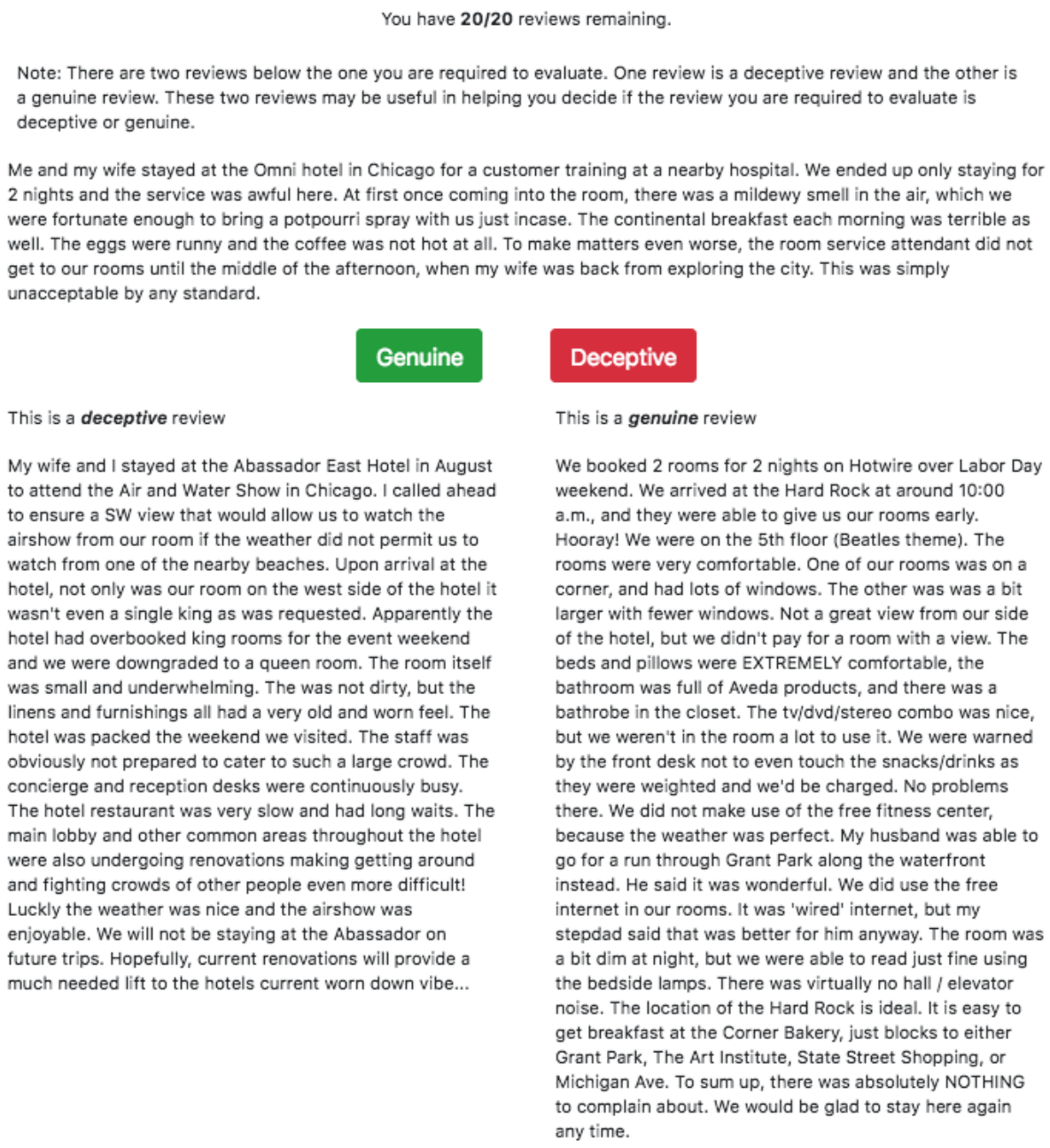
8. Please give us your feedback.

Figure 16: Survey questions for all the other groups.

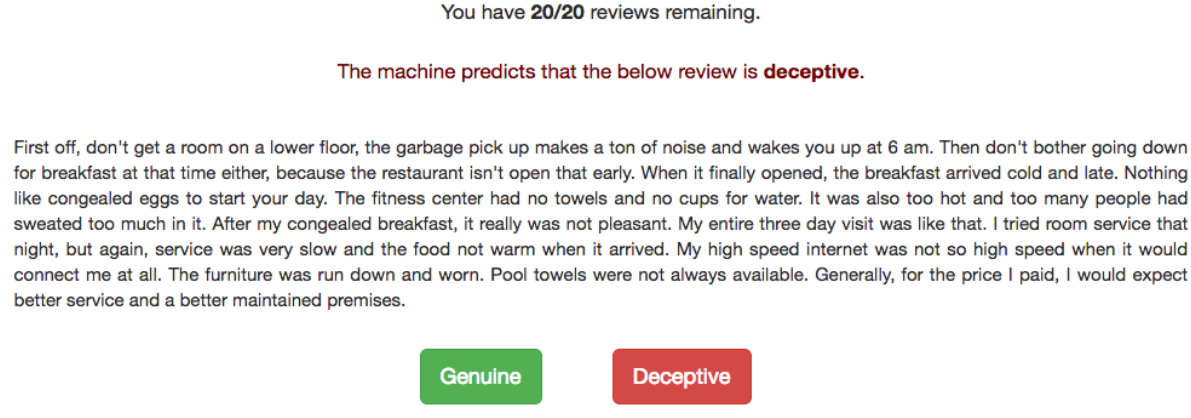


(a) Example interface for *control*.

(b) Example interface for *highlight*.



(a) Example interface for *examples*.



(b) Example interface for *predicted label w/o accuracy*.

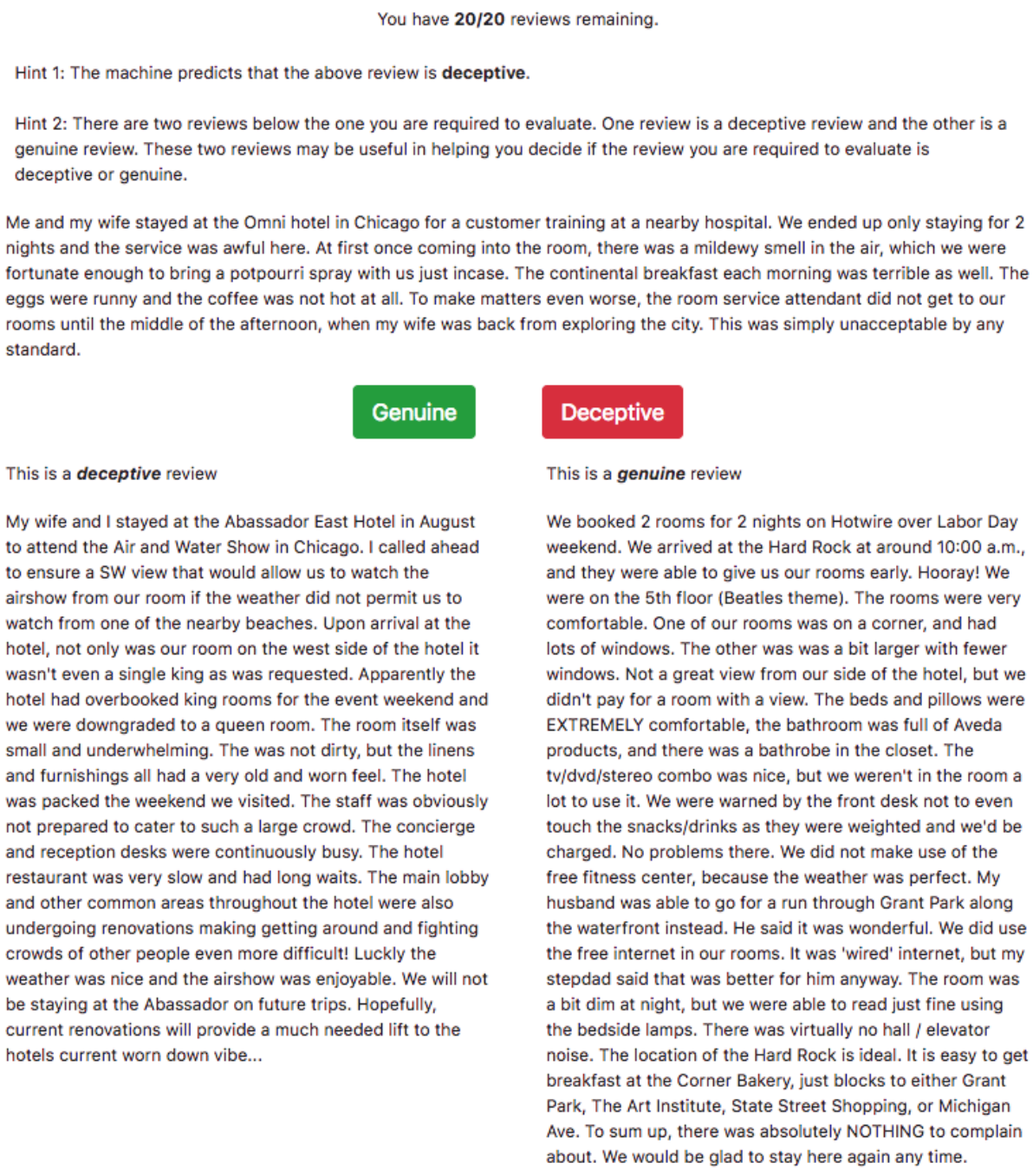


Figure 19: Example interface for *predicted label + examples*.