Empowering Humans in Human-AI Decision Making

by

Vivian Lai

B.S., Singapore Management University, 2015

M.S., University of Colorado Boulder, 2018

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2022

Committee Members:

James H. Martin, Chair

Q. Vera Liao

Tamara Sumner

Chenhao Tan

Tom Yeh

Lai, Vivian (Ph.D., Computer Science)

Empowering Humans in Human-AI Decision Making

Thesis directed by Prof. James H. Martin

Due to recent advances in Artificial Intelligence (AI), AI models are able to surpass human performance in various tasks unprecedentedly and are rapidly integrated into systems that assist humans in making decisions. However, deploying such systems into the real world requires an understanding of the potential risks and challenges we might face. How do we interpret and explain AI models' predictions while being aware of their biases and weaknesses? In this thesis, I discuss my work that empowers humans to make better decisions with AI models through AI-backed interactive systems. I describe (1) how humans make decisions with models (Chapter 2), (2) how explanations differ across models and methods (Chapter 3), (3) how humans learn counterintuitive patterns from models (Chapter 4), and (4) how humans and imperfect models could collaborate effectively (Chapter 5). I conclude by discussing future research perspectives on making human-AI collaborations better and more accessible.

Dedication

To my grandparents.

Acknowledgements

I want to extend my deepest gratitude to my advisor, Chenhao Tan, for believing in me and constantly pushing me to be better over the past six years. He has shown me the disposition of a humble yet brilliant researcher, and I hope to demonstrate the values I learned from him through life and work in the future.

I am fortunate to have collaborated with and mentored by Kori Inkpen, Q. Vera Liao, and Alison Smith-Renner. They taught me skills in research but also demonstrated the female power in a male-dominated industry.

My Ph.D. journey would not be manageable and exciting without my lab mates (whom I also consider my close friends): Han Liu, Joe Chao-Chun Hsu, Rosa Zhou YangQiaoYu , Chacha Chen, Karen Zhou, and Mourad Heddeya. And also, Samuel Carton, Shi Feng, and Jason Shuo Zhang for giving me helpful advice on career, research, and life.

Lastly, I want to thank my parents, who supported me during my graduate studies. Special thanks to my sister, who has never failed to encourage me during difficult times. And my cats, Coal and Oreo, for being my source of strength during the pandemic.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

1.1 Understanding the use of AI

Since the development of algorithms, they have been frequently used by researchers in both academia and industry to solve ongoing problems. As technology advances, complex algorithms no longer exist just on papers and in theories but could be built to realize and solve more complicated problems. The boom of deep learning has led to the birth of frameworks capable of impressive predictive performance in both simple and challenging tasks. However, we do not fully understand the consequences of using AI integrated into our society. There are numerous open questions on the unintended harm and consequences of using AI systems built to help humans.

To understand the complexity of this problem, let us consider the following scenario. A model is trained with given data to help humans with a challenging task, predicting if a particular person's profile would recidivate. Does the human, in this case, a judge, rely on the system's decision entirely? How much trust should the judge place in the system? How can the judge understand why the system is making a particular prediction?

Researchers have explored and developed various frameworks for generating explanations of AI predictions in the field of interpretable machine learning [Kim et al., 2016, Ribeiro et al., 2016, Lundberg and Lee, 2017, Ribeiro et al., 2018a, Lei et al., 2016, Kim et al., 2014]. However, instead of using explanations to debug machine learning models, we use explanations as AI assistance to empower humans in making decisions in the decision making process [Feng and Boyd-Graber, 2018, Green and Chen, 2019b, Cheng et al., 2019]. While there are two extreme options in making a decision, relying on a human's decision or the AI's decision, what are some best practices for integrating AI into the human decision making process? In Chapter 2, we propose a spectrum between full human agency and full automation and investigate how various degrees of AI assistance affect human performance.

While both explanations and AI predictions are used as AI assistance to help humans make decisions, explanations are shown as assistance more often than predicted labels as the latter has stronger priming and tends to sway humans' decisions. Additionally, due to the complex nature of deep learning models, explanations have quickly become the vehicle for explaining AI predictions. Explanations, which are also commonly feature importance, could be derived from the model's built-in mechanism (i.e., coefficients, attention scores) or post-hoc methods that serve as an option to allow AI's predictions to be more predictable. In this thesis, we will also explore how similar feature importance are across different models and methods (Chapter 3).

1.2 Empowering humans with AI

The first step of empowering humans with AI in the decision making process is to understand the best practices of using AI and the differences using explanations retrieved from different models and methods. The second step is the actual task of empowering humans with AI in the decision making process. With AI in the equation, it is difficult to use a single metric to measure and determine the effectiveness of AI. However, one of the more common metrics is complementary performance [Bansal et al., 2019a]. Intuitively, when we include AI as assistance, the ideal situation is that human-AI performance exceeds human alone and AI alone performance.

However, prior work has shown that complementary performance is not within an arm's reach. There are many questions and factors leading to the undesired situation. The second part of this thesis proposes different ways to bridge the gap between expectations and reality. To understand how this could be done, let us consider another hypothetical scenario. To predict if a hotel review is deceptive or genuine, the human is shown "explanations" to help them understand the AI's predictions. Despite being shown "explanations", the human fails to understand why certain words are more *deceptive* or *genuine*. In Chapter 4, we will explore how model-driven tutorials are used to help humans understand the explanations and thereby improving human-AI team performance.

Due to changes in data distribution between the training and test set, models could never be perfect or achieve 100 percent accuracy. As a result, it is pertinent for humans to learn to collaborate with imperfect models. Prior work has also focused on human-AI collaboration in high-stakes scenarios such as the medical and justice domains. In Chapter 5, we will explore how conditional delegation, a new paradigm of human-AI collaboration, is potentially helpful for humans and imperfect AI models to collaborate effectively in low-stakes decisions.

1.3 Organization and Contributions

In order to empower humans with AI in the decision making process, we have conducted a series of studies and they are organized into four chapters. Each chapter is summarized by a short phrase that is representative of it.

- **Chapter 2: A Spectrum Between Human Agency & Human Performance.** Due to recent advance in technology, AI has shown impressive predictive capability and is rapidly integrated in our society, especially in societal critical tasks. To better understand how we can capitalize the potential of AI through harnessing explanations and its predictions, we propose a spectrum between full human agency and full automation. Given the rampant fake news spread throughout the internet, it is generally useful for humans to learn and discern between genuineness and deceptiveness. Using deception detection as a testbed, we show that explanations alone slightly improve human performance but showing predicted labels can improve human performance drastically. The results demonstrate a tradeoff between human agency and human performance.

- **Chapter 3: Many Faces of Feature Importance.** While deep learning models' predictive ability are becoming more impressive and powerful, at the same time their inherent complexity has caused humans difficulty and sometimes confusion in understanding their predictions. In an attempt to understand models' predictions better, feature importance is commonly used as "explanations" to explain the respective model's prediction. However, as feature importance could be derived from different models and methods (i.e., built-in and post-hoc), how similar or consistent are they when compared across? Using text classification as a testbed, traditional models are more similar to each other than with deep learning models. When in a debate or discussion, humans often agree on perspectives with similar reasons. However, unlike in models, we find that important features do not always resemble each other better when two models agree on the predicted label than when they disagree.

- **Chapter 4: Model-driven Tutorials and Simple Explanations.** Although explanations and AI predictions can help improve human performance, elusive complementary performance is not achieved in challenging tasks. Complementary performance Bansal et al. [2019a] is commonly used as a measure in literature to determine if human-AI collaboration is performing up to expectations. It is achieved when human and AI performance exceeds both human alone and AI alone performance. To understand how complementary performance could be achieved, we explore model-driven tutorials to help humans understand counterintuitive patterns hidden in training data. We find that tutorials help to improve human performance and while deep learning models have shown impressive performance on tasks, explanations from simple models are preferred by and more useful to humans.

- **Chapter 5: Conditional Delegation as an Alternative Paradigm.** Prior work has explored how humans can make decisions with explanations and AI predictions [Zhang et al., 2020, Green and Chen, 2019b, Lai and Tan, 2019b]. While explanations and AI predictions are helpful for domains where decisions have dire consequences, they are not helpful to other domains where the large number of decisions is the key challenge. Using content moderation as a testbed, we investigate how humans can work effectively with imperfect models. We propose conditional delegation as an alternative paradigm for humans and imperfect AI where they work together to find trustworthy regions of the model. Our results demonstrate that conditional delegation is promising as it is able to achieve complementary performance. Humans working on conditional delegation are also more engaged and explanations improve task efficiency.

- **Chapter 6.** I conclude my thesis with thoughts on future work.

Chapter 2

A Spectrum Between Human Agency & Human Performance

2.1 Overview

In this chapter, we 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 AI assistance along the spectrum that gradually increase the influence of machine predictions. 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. The results demonstrate a tradeoff between human performance and human agency and show that explanations of machine predictions can moderate this tradeoff.

Most of the contents in this chapter are published in Lai and Tan [2019a]. This is joint work with Chenhao Tan.

2.2 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%) [He et al., 2015]; Kleinberg et al. [2017a] demonstrate that in bail decisions, machine

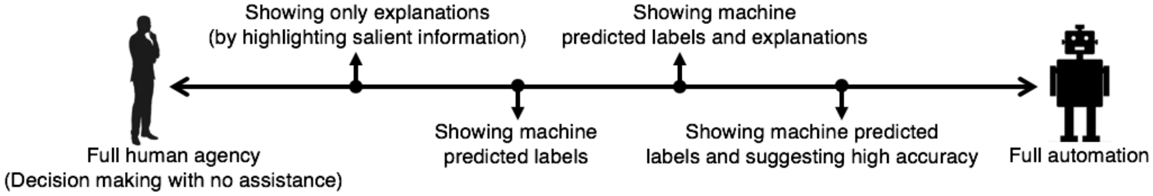

Showing only explanations
(by highlighting salient information)
Showing machine
predicted labels and explanations
Full human agency
(Decision making with no assistance)
Showing machine
predicted labels
Showing machine predicted
labels and suggesting high accuracy
Full automation

Figure 2.1: 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 §2.4.

predictions of recidivism can reduce jail rates by 41.9% with no increase in crime rates, compared to human judges; Ott et al. [2011] 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 **not salient, 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.

This chapter investigates the 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 2.1 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 [Kim et al., 2016, Ribeiro et al., 2016, Lundberg and Lee, 2017, Ribeiro et al., 2018a, Lei et al., 2016, Kim et al., 2014]. 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 knowledge databases and representations to reason about complex decisions [Keen, 1978, Berger, 2013, Newell and Simon, 1972, Horvitz, 1999, Shim et al., 2002]. Moreover, since many policy decisions can be formulated as prediction problems [Kleinberg et al., 2015], 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. [2012] find that deceptive reviews are a growing problem on multiple platforms such as TripAdvisor and Yelp. Fake news has also received significant attention recently [Lazer et al., 2018, Vosoughi et al., 2018] and might have influenced the outcome of the U.S. presidential election in 2016 Allcott and Gentzkow [2017]. 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 [Feng et al., 2012, Feng and Hirst, 2013, Ott et al., 2011, Abouelenien et al., 2014, Akoglu et al., 2013]. 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 [Ott et al., 2011]. Machines can identify not salient 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. [2011] 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" [Liptak, 2017, Supreme Court of Wisconsin, 2016]. Similarly, the trial judge is required to act as a gatekeeper regarding the evidence from a polygraph (lie detector) [Supreme Court of the United States, 1993]. Therefore, it is crucial to retain human agency and understand human predictions with assistance from machine learning models.

2.3 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 [Caruana et al., 2015, Kleinberg et al., 2017a, Varshney, 2016]. Therefore, improving their transparency and interpretability has attracted broad interest [Kim et al., 2016, Ribeiro et al., 2016, Lundberg and Lee, 2017, Ribeiro et al., 2018a, Lei et al., 2016, Kim et al., 2014], dating back to early work on recommendation systems Herlocker

[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.

et al. [2000], Cosley et al. [2003]. In the case of general automation, researchers have also studied issues of appropriate reliance and trust [Lee and See, 2004, Wickens et al., 2015, Parasuraman and Riley, 1997, Bussone et al., 2015, Dzindolet et al., 2003].

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. [2016], which selects both prototypes and criticisms. Ribeiro et al. [2016] propose a feature-based approach, LIME, that fits a sparse linear model to approximate non-linear models locally. Similarly, Lundberg and Lee [2017] 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 [Chandrasekaran et al., 2018, Feng and Boyd-Graber, 2018]. Our work also resonates with the seminal work on mixed-initiative user interfaces [Horvitz, 1999] and intelligence augmentation [Ashby, 1957]. In addition, our work is connected to cognitive studies on understanding effective explanations beyond the context of machine learning [Lombrozo, 2007, 2006].

Deception and misinformation. Deception is a widely studied phenomenon in many disciplines [Vrij, 2000]. 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 [Krauss et al., 1976]. To detect deception, researchers have examined the role of behavioral, emotional, and linguistic cues [Ekman et al., 1980, Dulaney, 1982, Knapp et al., 1974, Mehrabian, 1971, L Knapp and Comaden, 1979, Vrij, 2000].

Since people are increasingly relying on online reviews to make purchase decisions [Ye et al., 2011, Zhang et al., 2010, Chevalier and Mayzlin, 2006, Trusov et al., 2009], machine learning methods have been used to detect deception in online reviews [Jindal and Liu, 2008, Wu et al., 2010, Yoo and Gretzel, 2009, Ott et al., 2011, Feng et al., 2012, Feng and Hirst, 2013]. An important challenge in detecting deception in online reviews is to obtain the groundtruth labels of reviews. Ott et al. [2011] 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 [Drucker et al., 1999, Gyöngyi et al., 2004, Ntoulas et al., 2006, Akoglu et al., 2013].

More recently, the issue of misinformation and fake news has drawn much attention from both the public and the research community Farsetta and Price [2006], Lazer et al. [2018]. Most relevant to our work is Zhang et al. [2018], which explores varying types of credibility annotations specifically designed for news articles. In addition, Nyhan and Reifler [2010] demonstrate the "backfire" effect, which suggests that corrections of misperceptions may enhance people's false beliefs, and Vosoughi et al. [2018] show that fake news is more innovative and spreads faster than real news.

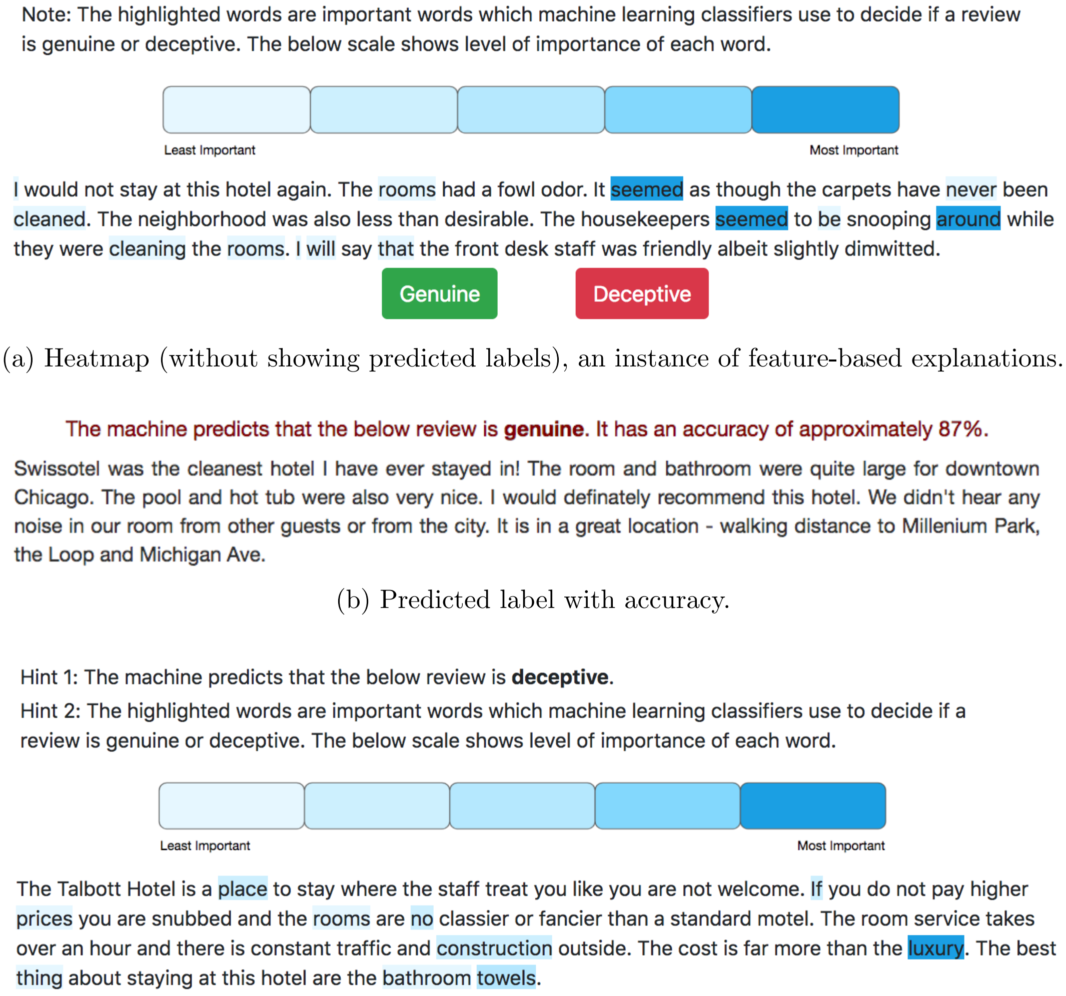
It is worth noting that deception detection is a broad and complex issue. For instance, fake news can be hard to define and may not be easily separated into two classes. Moreover, detecting fake news is different from detecting deceptive reviews as the former task requires other skills such as fact checking. 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. [2011] 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.

2.4 Experiment 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 2.1. We finally formulate our hypotheses and define our evaluation metrics.

Experimental setup. We employ the deception detection task developed by Ott et al. [2011] and evaluate human performance in this task with varying levels of machine assistance. The dataset in Ott et al. [2011] 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. [2011] already surpasses humans (∼50%) by a wide margin and linear classifiers are generally considered more interpretable, we follow Ott et al. [2013] 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 2.1) 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 setups. Given that there



Note: The highlighted words are important words which machine learning classifiers use to decide if a review is genuine or deceptive. The below scale shows level of importance of each word.

I would not stay at this hotel again. The rooms had a fowl odor. It seemed as though the carpets have never been cleaned. The neighborhood was also less than desirable. The housekeepers seemed to be snooping around while they were cleaning the rooms. I will say that the front desk staff was friendly albeit slightly dimwitted.

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

The machine predicts that the below review is **genuine**. It has an accuracy of approximately 87%.

Swissotel was the cleanest hotel I have ever stayed in! The room and bathroom were quite large for downtown Chicago. The pool and hot tub were also very nice. I would definately recommend this hotel. We didn't hear any noise in our room from other guests or from the city. It is in a great location - walking distance to Millenium Park, the Loop and Michigan Ave.

(b) Predicted label with accuracy.

Hint 1: The machine predicts that the below review is **deceptive**.

Hint 2: The highlighted words are important words which machine learning classifiers use to decide if a review is genuine or deceptive. The below scale shows level of importance of each word.

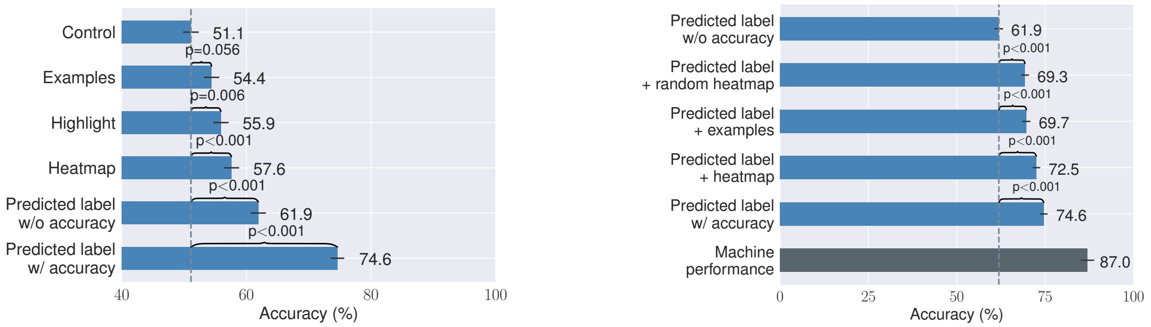
The Talbott Hotel is a place to stay where the staff treat you like you are not welcome. If you do not pay higher prices you are snubbed and the rooms are no classier or fancier than a standard motel. The room service takes over an hour and there is constant traffic and construction outside. The cost is far more than the luxury. The best thing about staying at this hotel are the bathroom towels.

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

Figure 2.2: Example interfaces with varying levels of machine assistance. Figure 2.2a only presents feature-based explanations of machine predictions in the form of **heatmap**. Figure 2.2b shows both the predicted label and an explicit statement about machine accuracy (87%). Figure 2.2c shows the predicted label with heatmap, but does not present machine accuracy. We crop the “Genuine” and “Deceptive” buttons in Figure 2.2b and 2.2c to save space.

are 320 test reviews and that we collect five turker predictions for each review, each experimental setup has a total of 80 turkers.

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



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

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

Figure 2.3: Human accuracy with varying levels of assistance. In Figure 2.3a, **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 2.3b 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 2.3a and "predicted label w/o accuracy" in Figure 2.3b).

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 2.1 (we only show three interfaces in Figure 2.2 for space reasons; see §2.8 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 [Kneusel and Mozer, 2017]. 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 2.2a shows an example interface for **heatmap**.

- **Example-based explanations.** This method (**examples**) is inspired by example-based interpretable machine learning [Kim et al., 2016]. 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 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 [Dietvorst et al., 2015].

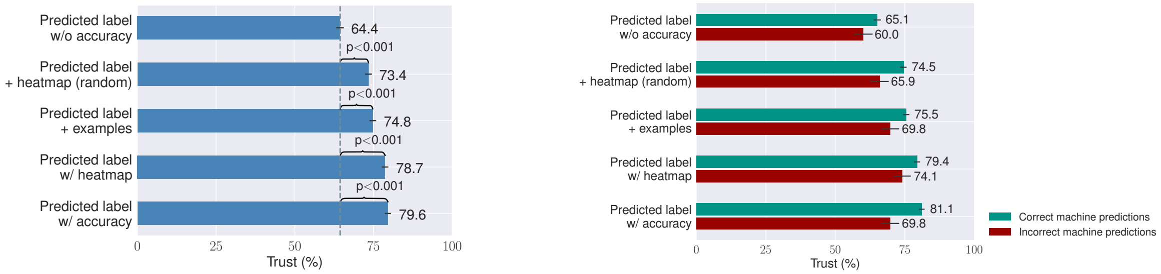
- **Predicted label with accuracy.** We may further influence human decisions by explicitly suggesting that machines perform well in this task with 87% accuracy. Figure 2.2b shows an example for **predicted label with accuracy**. Note that such strong recommendations may not be desired due to ethical 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 2.2c 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 [2017] 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 [Lee and See, 2004].

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.



(a) Trust in machine predictions.

(b) Trust in correct and incorrect machine predictions.

Figure 2.4: The trust that humans place on machine predictions. Figure 2.4a shows that adding feature-based explanations (heatmap) can effectively increase the trust level compared to **predicted label w/o accuracy**. p-value in Figure 2.4a is computed by conducting t-test between the corresponding setup and **predicted label w/o accuracy**. Figure 2.4b 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.

2.5 Results

In this section, we investigate how varying levels of assistance from machine learning models along the spectrum in Figure 2.1 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://machineintheloop.com/deception`.

2.5.1 Human Accuracy

We first present human accuracy measured by the percentage of correctly predicted instances by humans. Our results suggest that 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 2.3a). As Figure 2.3a shows, human performance in **control** is no better than chance (51.1%). This finding is consistent with Ott et al. [2011] and decades of research on deception detection [Bond Jr and DePaulo, 2006]. 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 [Zhu, 2015, Mac Aodha et al., 2018, Singla et al., 2014].

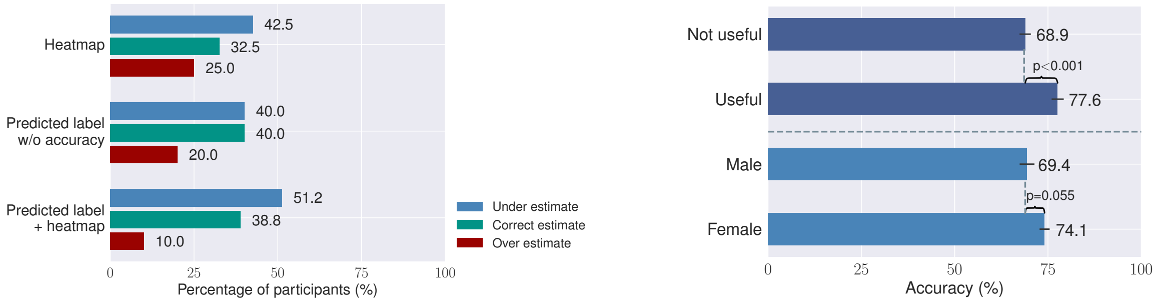
Showing predicted labels significantly improves human performance (Figure 2.3a and 2.3b). As Figure 2.3a shows, showing predicted labels drastically improves human performance (61.9% for **predicted label w/o accuracy**, a 21% relative improvement over **control**; the difference with **heatmap** is statistically significant (p <0.001)). By presenting machine accuracy as shown in Figure 2.2b, the performance is further improved to 74.6% (**predicted label w/ accuracy** in Figure 2.3a, 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 [Lee and See, 2004].

To further understand human performance with predicted labels, we examine all experimental setups with predicted labels in Figure 2.3b. 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.

2.5.2 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



(a) Human estimation of their own performance.

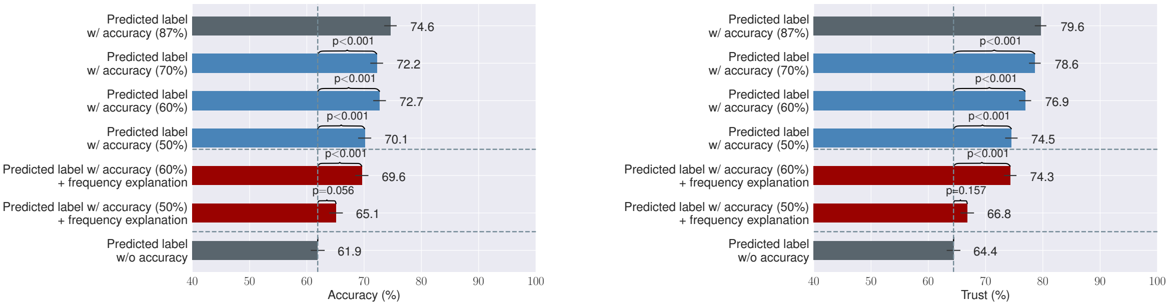
(b) Gender and hint usefulness in **predicted label + heatmap**.

Figure 2.5: Heterogeneity findings among participants in our study. Figure 2.5a shows performance estimation by participants in three different experimental setups. Figure 2.5b presents the performance of participants in **predicted label + heatmap** group by two variables, hint usefulness and gender.

effectively identify cases where machines are wrong. It is important to emphasize that our focus is on understanding how 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 2.4a). 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



(a) Human accuracy.

(b) Trust.

Figure 2.6: Human accuracy and trust given varying statements of machine accuracy. Figure 2.6a and Figure 2.6b 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**.

coefficients beyond "the placebo effect".

Humans tend to trust machine predictions more when machine predictions are correct. (Figure 2.4b). We next examine whether humans trust machine predictions more when machine predictions are correct than when they are incorrect. Figure 2.4b 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.

2.5.3 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.

Human estimation of their own performance (Figure 2.5a). 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 Elaad [2003]. In fact, ~42% of the participants correctly predicted their performance. Among the remaining, ~18% overestimated their performance, while ~40% underestimated their performance. Figure 2.5a 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 2.5b). 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 2.5b (**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 [Mann et al., 2004, DePaulo et al., 1993, McCornack and Parks, 1990, Li, 2011].

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

2.6 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 2.1 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 2.6a and Figure 2.6b, 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 [Reyna and Brainerd, 2008, Berwick et al., 1981, Slovic and Peters, 2006, Peters et al., 2006]. 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 2.6a and Figure 2.6b. 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 [Sedlmeier and Gigerenzer, 2001, Gigerenzer, 1996, Gigerenzer and Hoffrage, 1995].

2.7 Conclusion

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 slightly 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 [Zhu, 2015, Mac Aodha et al., 2018, Singla et al., 2014].

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 [Kleinberg et al., 2017b, Hardt et al., 2016, Corbett-Davies et al., 2017]. 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.

2.8 Appendix

2.8.1 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 2.15 and Figure 2.16). 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 2.7 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 2.8 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

[3] Thanks to suggestions from anonymous reviewers.

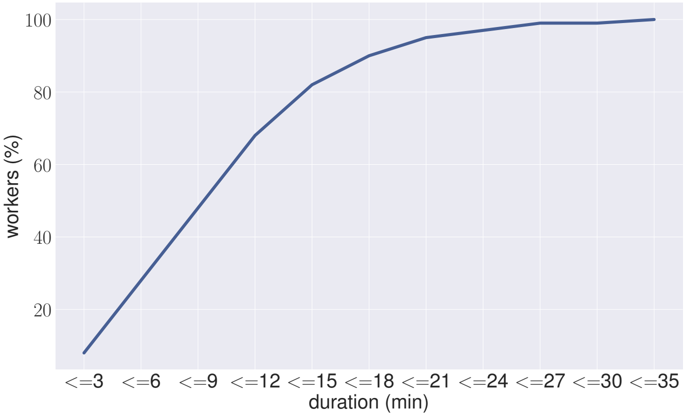


Figure 2.7: Cumulative distribution of study duration in 9 experimental setups.

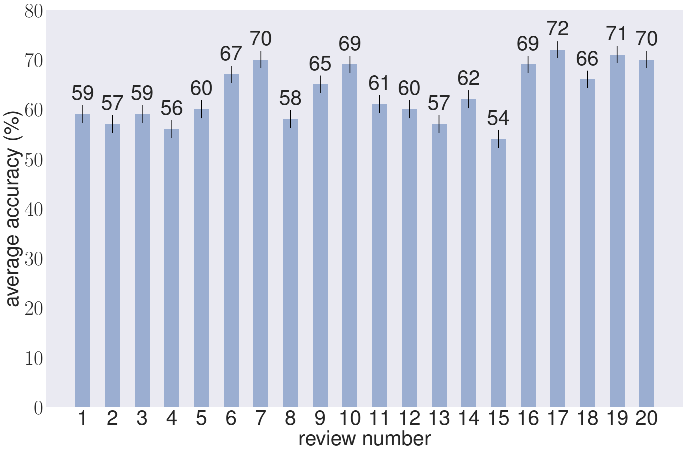


Figure 2.8: Average accuracy with respect to review ordering in 9 experimental setups.

are not included.

2.8.2 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 2.17a).
- Highlight (Figure 2.17b).

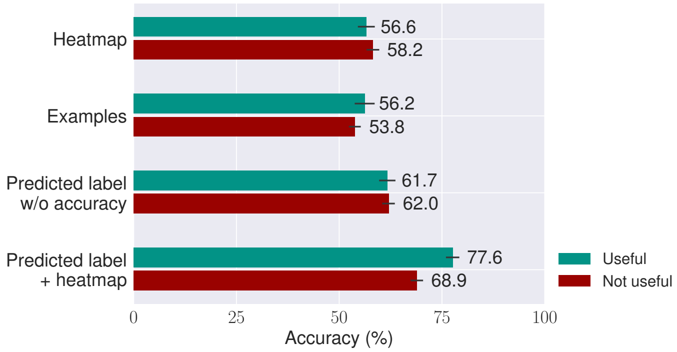


Figure 2.9: Human accuracy vs. usefulness of hints.

- Examples (Figure 2.18a).
- Predicted label w/o accuracy (Figure 2.18b).
- Predicted label + examples (Figure 2.19).

2.8.3 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 2.9). 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.

Gender differences (Figure 2.10). Females generally outperform males, in 8 out of 9 experimental setups. None of the differences is statistically significant.

Review sentiments (Figure 2.11). 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

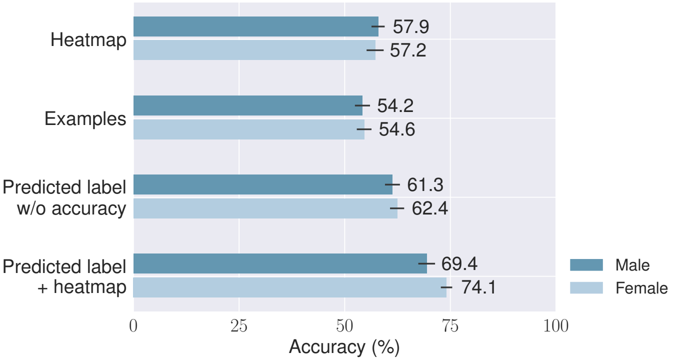


Figure 2.10: Human accuracy vs. gender.

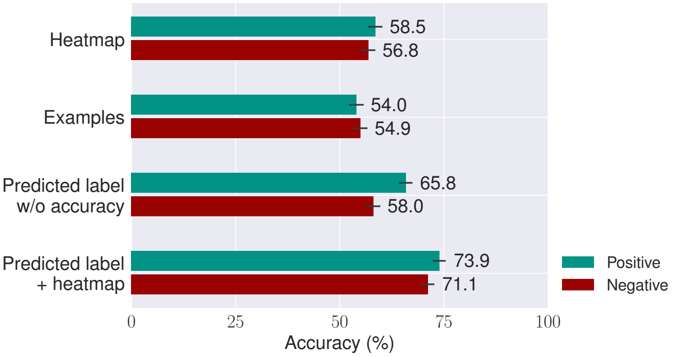


Figure 2.11: Human accuracy vs. review sentiment.

significant for **predicted label w/o accuracy**.

Education background (Figure 2.12). 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 accurcay**, 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 2.13). 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 2.14). There is no clear trend regarding experience of writing

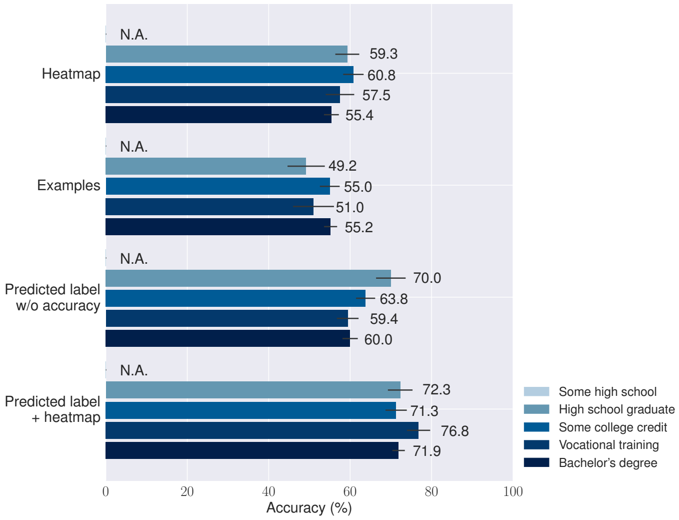


Figure 2.12: Human accuracy vs. education background.

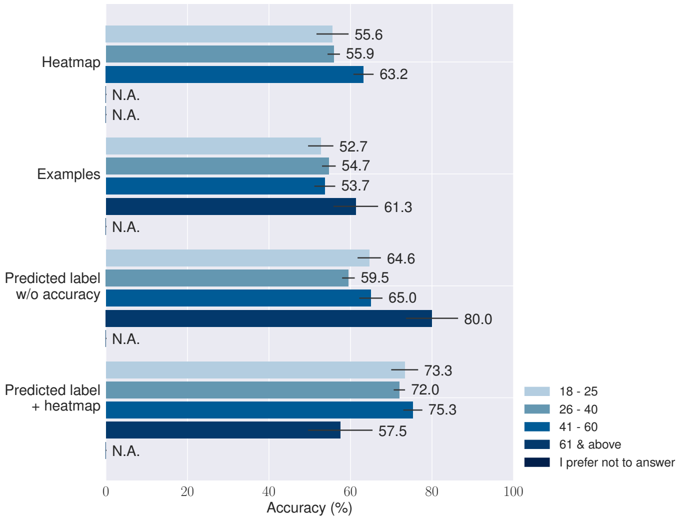


Figure 2.13: Human accuracy vs. age groups.

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.

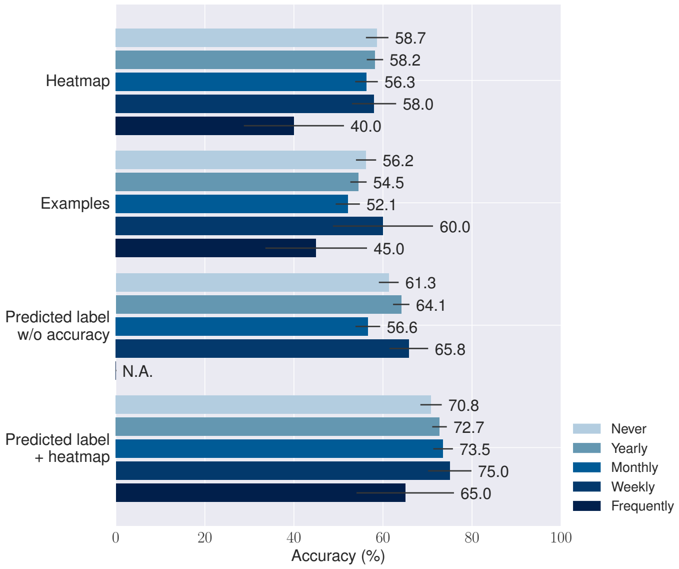
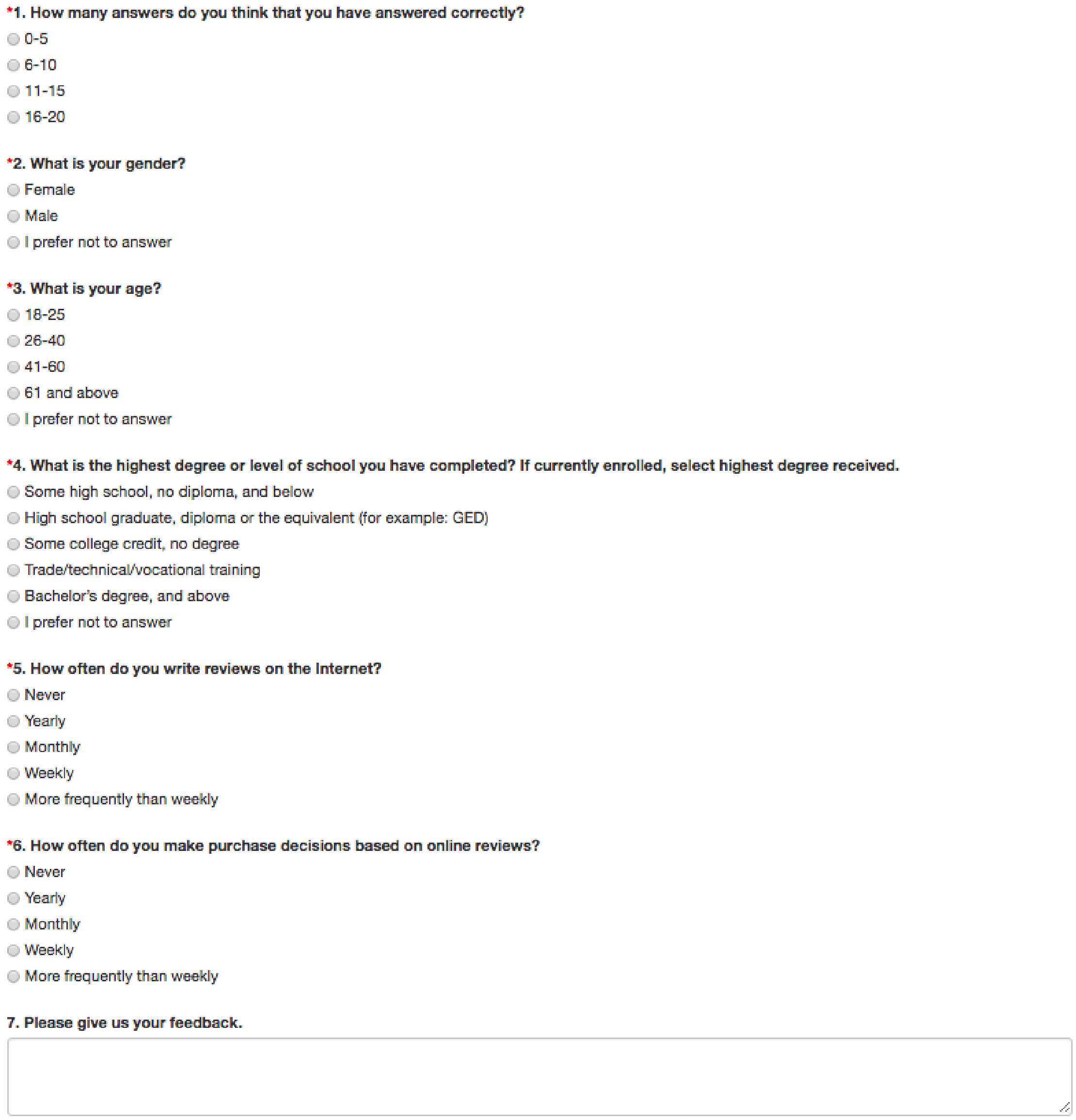


Figure 2.14: Human accuracy vs. review writing experience.



*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?

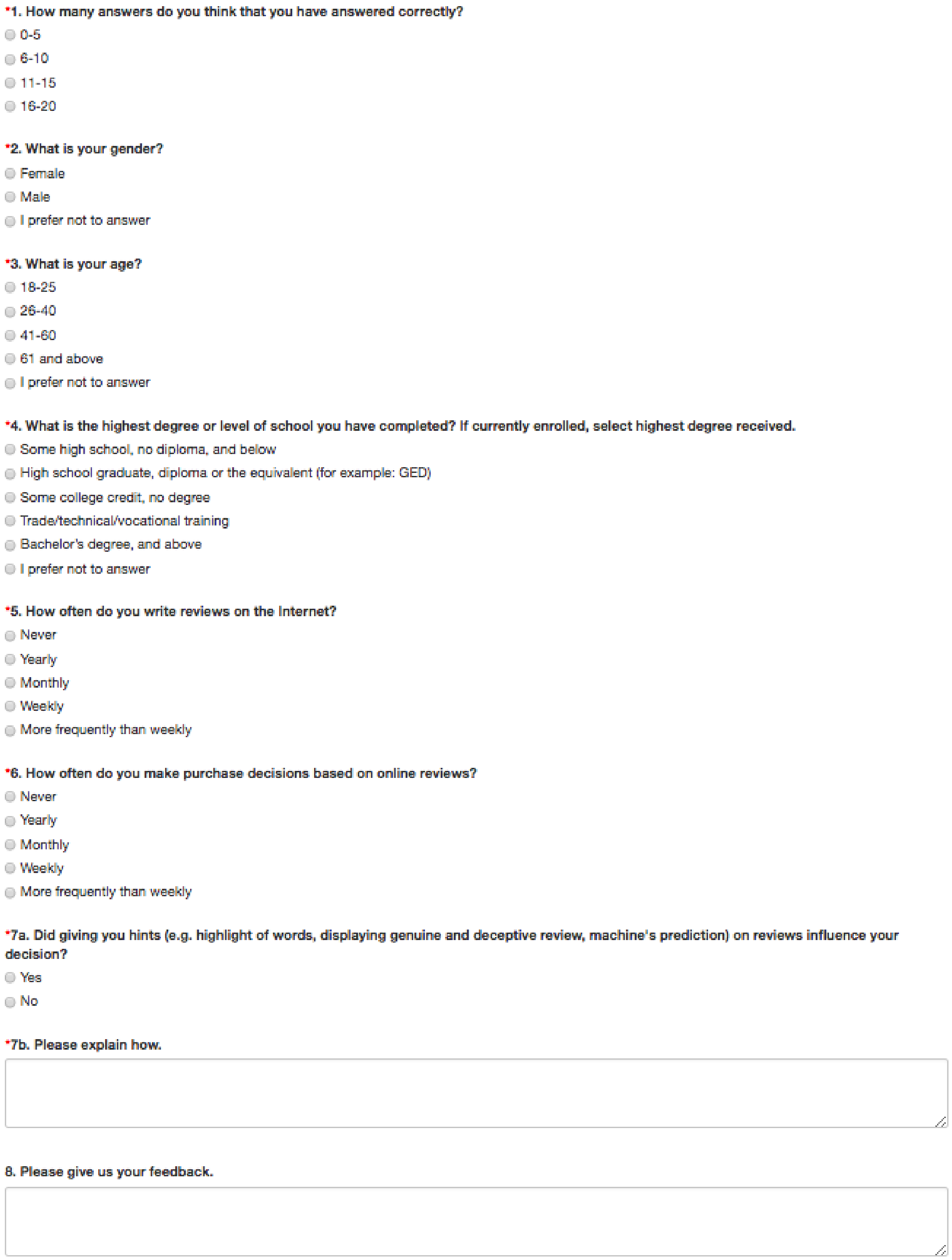
○ Never
○ Yearly
○ Monthly
○ Weekly
○ More frequently than weekly

*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 2.15: 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

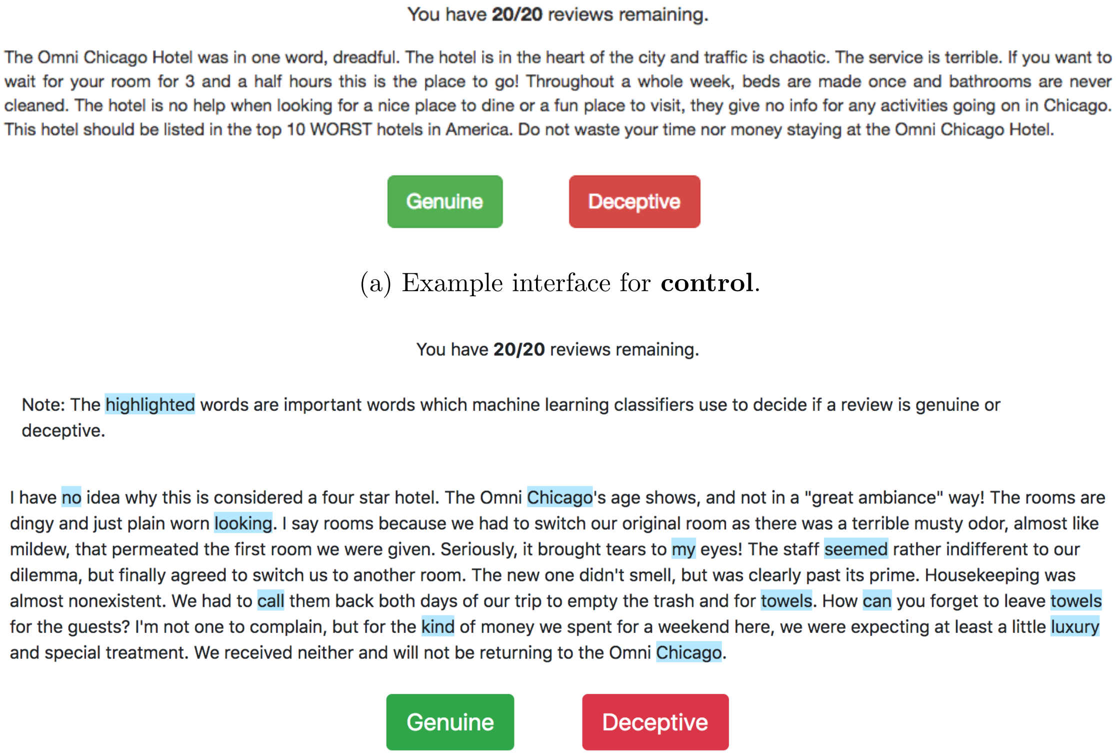
***7a. Did giving you hints (e.g. highlight of words, displaying genuine and deceptive review, machine's prediction) on reviews influence your decision?**

○ Yes
○ No

***7b. Please explain how.**

8. Please give us your feedback.

Figure 2.16: Survey questions for all the other groups.



You have **20/20** reviews remaining.

The Omni Chicago Hotel was in one word, dreadful. The hotel is in the heart of the city and traffic is chaotic. The service is terrible. If you want to wait for your room for 3 and a half hours this is the place to go! Throughout a whole week, beds are made once and bathrooms are never cleaned. The hotel is no help when looking for a nice place to dine or a fun place to visit, they give no info for any activities going on in Chicago. This hotel should be listed in the top 10 WORST hotels in America. Do not waste your time nor money staying at the Omni Chicago Hotel.

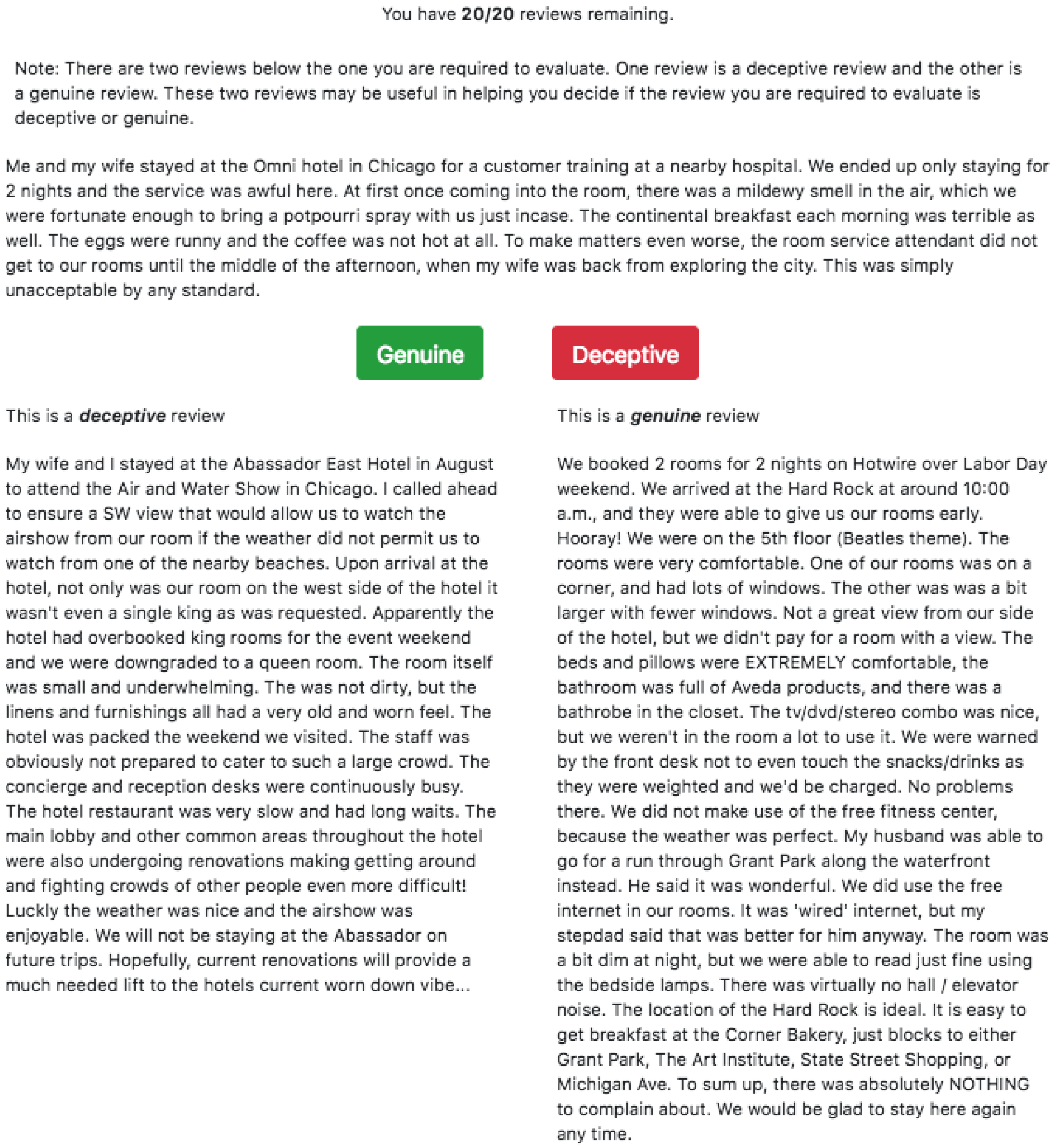
(a) Example interface for **control**.

You have **20/20** reviews remaining.

Note: The highlighted words are important words which machine learning classifiers use to decide if a review is genuine or deceptive.

I have no idea why this is considered a four star hotel. The Omni Chicago's age shows, and not in a "great ambiance" way! The rooms are dingy and just plain worn looking. I say rooms because we had to switch our original room as there was a terrible musty odor, almost like mildew, that permeated the first room we were given. Seriously, it brought tears to my eyes! The staff seemed rather indifferent to our dilemma, but finally agreed to switch us to another room. The new one didn't smell, but was clearly past its prime. Housekeeping was almost nonexistent. We had to call them back both days of our trip to empty the trash and for towels. How can you forget to leave towels for the guests? I'm not one to complain, but for the kind of money we spent for a weekend here, we were expecting at least a little luxury and special treatment. We received neither and will not be returning to the Omni Chicago.

(b) Example interface for **highlight**.



You have **20/20** reviews remaining.

Note: There are two reviews below the one you are required to evaluate. One review is a deceptive review and the other is a genuine review. These two reviews may be useful in helping you decide if the review you are required to evaluate is deceptive or genuine.

Me and my wife stayed at the Omni hotel in Chicago for a customer training at a nearby hospital. We ended up only staying for 2 nights and the service was awful here. At first once coming into the room, there was a mildewy smell in the air, which we were fortunate enough to bring a potpourri spray with us just incase. The continental breakfast each morning was terrible as well. The eggs were runny and the coffee was not hot at all. To make matters even worse, the room service attendant did not get to our rooms until the middle of the afternoon, when my wife was back from exploring the city. This was simply unacceptable by any standard.

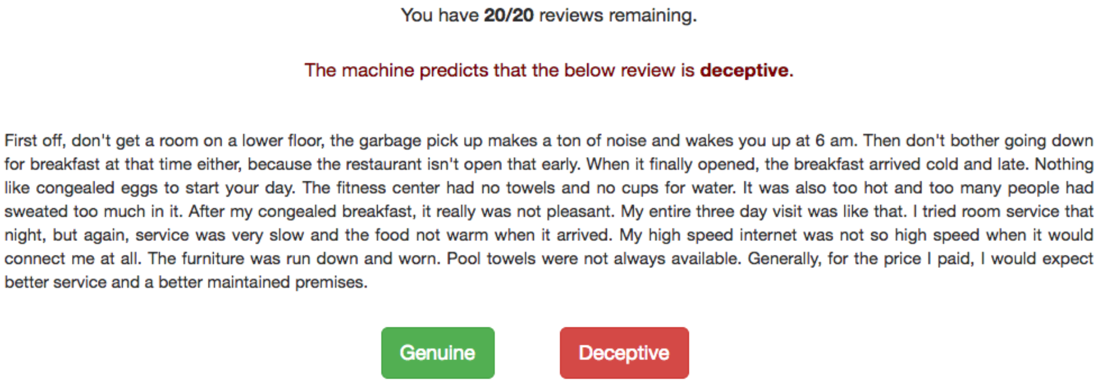
This is a ***deceptive*** review

My wife and I stayed at the Abassador East Hotel in August to attend the Air and Water Show in Chicago. I called ahead to ensure a SW view that would allow us to watch the airshow from our room if the weather did not permit us to watch from one of the nearby beaches. Upon arrival at the hotel, not only was our room on the west side of the hotel it wasn't even a single king as was requested. Apparently the hotel had overbooked king rooms for the event weekend and we were downgraded to a queen room. The room itself was small and underwhelming. The was not dirty, but the linens and furnishings all had a very old and worn feel. The hotel was packed the weekend we visited. The staff was obviously not prepared to cater to such a large crowd. The concierge and reception desks were continuously busy. The hotel restaurant was very slow and had long waits. The main lobby and other common areas throughout the hotel were also undergoing renovations making getting around and fighting crowds of other people even more difficult! Luckly the weather was nice and the airshow was enjoyable. We will not be staying at the Abassador on future trips. Hopefully, current renovations will provide a much needed lift to the hotels current worn down vibe...

This is a ***genuine*** review

We booked 2 rooms for 2 nights on Hotwire over Labor Day weekend. We arrived at the Hard Rock at around 10:00 a.m., and they were able to give us our rooms early. Hooray! We were on the 5th floor (Beatles theme). The rooms were very comfortable. One of our rooms was on a corner, and had lots of windows. The other was was a bit larger with fewer windows. Not a great view from our side of the hotel, but we didn't pay for a room with a view. The beds and pillows were EXTREMELY comfortable, the bathroom was full of Aveda products, and there was a bathrobe in the closet. The tv/dvd/stereo combo was nice, but we weren't in the room a lot to use it. We were warned by the front desk not to even touch the snacks/drinks as they were weighted and we'd be charged. No problems there. We did not make use of the free fitness center, because the weather was perfect. My husband was able to go for a run through Grant Park along the waterfront instead. He said it was wonderful. We did use the free internet in our rooms. It was 'wired' internet, but my stepdad said that was better for him anyway. The room was a bit dim at night, but we were able to read just fine using the bedside lamps. There was virtually no hall / elevator noise. The location of the Hard Rock is ideal. It is easy to get breakfast at the Corner Bakery, just blocks to either Grant Park, The Art Institute, State Street Shopping, or Michigan Ave. To sum up, there was absolutely NOTHING to complain about. We would be glad to stay here again any time.

(a) Example interface for **examples**.

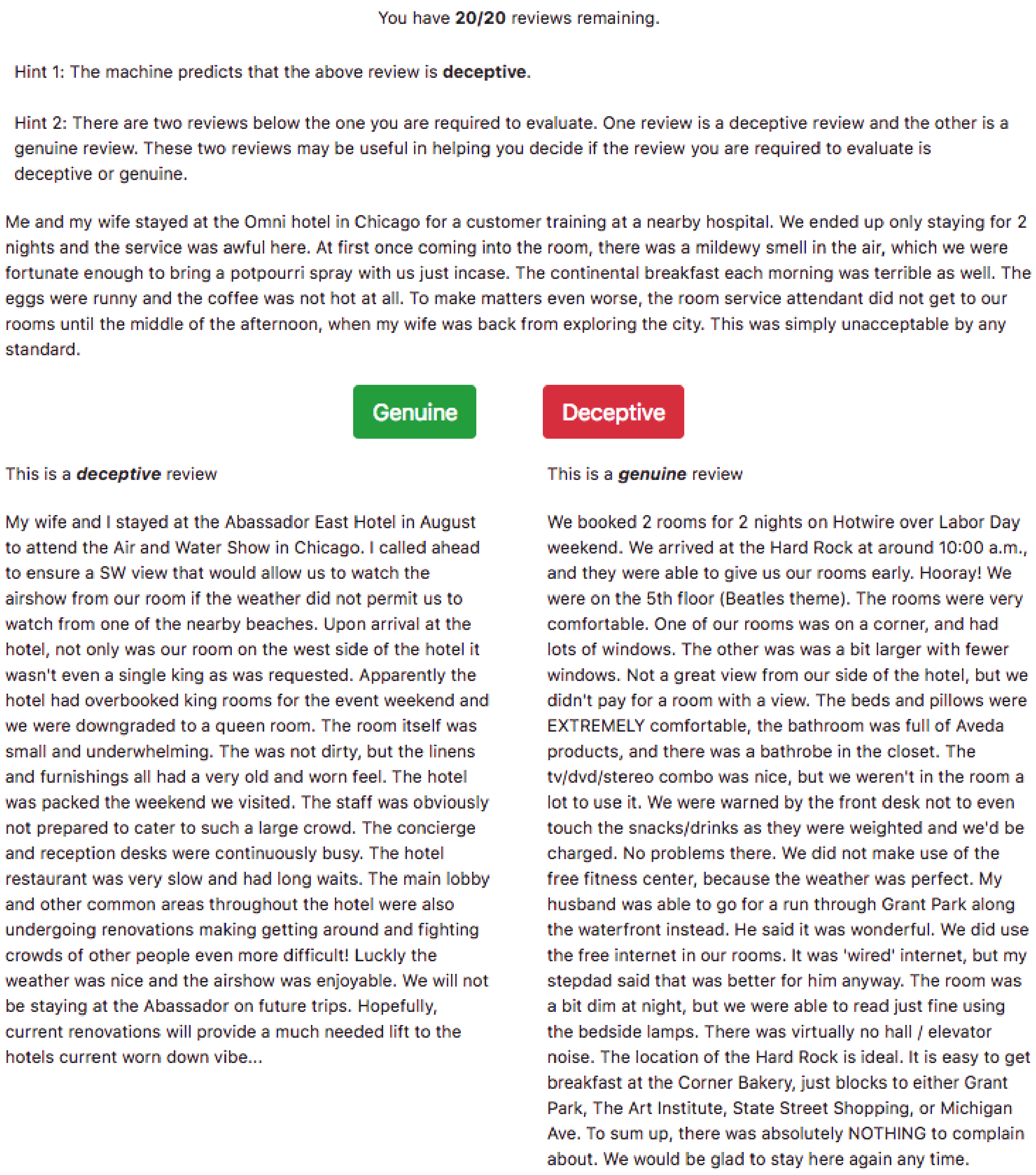


You have **20/20** reviews remaining.

The machine predicts that the below review is **deceptive**.

First off, don't get a room on a lower floor, the garbage pick up makes a ton of noise and wakes you up at 6 am. Then don't bother going down for breakfast at that time either, because the restaurant isn't open that early. When it finally opened, the breakfast arrived cold and late. Nothing like congealed eggs to start your day. The fitness center had no towels and no cups for water. It was also too hot and too many people had sweated too much in it. After my congealed breakfast, it really was not pleasant. My entire three day visit was like that. I tried room service that night, but again, service was very slow and the food not warm when it arrived. My high speed internet was not so high speed when it would connect me at all. The furniture was run down and worn. Pool towels were not always available. Generally, for the price I paid, I would expect better service and a better maintained premises.

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



You have **20/20** reviews remaining.

Hint 1: The machine predicts that the above review is **deceptive**.

Hint 2: There are two reviews below the one you are required to evaluate. One review is a deceptive review and the other is a genuine review. These two reviews may be useful in helping you decide if the review you are required to evaluate is deceptive or genuine.

Me and my wife stayed at the Omni hotel in Chicago for a customer training at a nearby hospital. We ended up only staying for 2 nights and the service was awful here. At first once coming into the room, there was a mildewy smell in the air, which we were fortunate enough to bring a potpourri spray with us just incase. The continental breakfast each morning was terrible as well. The eggs were runny and the coffee was not hot at all. To make matters even worse, the room service attendant did not get to our rooms until the middle of the afternoon, when my wife was back from exploring the city. This was simply unacceptable by any standard.

This is a ***deceptive*** review

My wife and I stayed at the Abassador East Hotel in August to attend the Air and Water Show in Chicago. I called ahead to ensure a SW view that would allow us to watch the airshow from our room if the weather did not permit us to watch from one of the nearby beaches. Upon arrival at the hotel, not only was our room on the west side of the hotel it wasn't even a single king as was requested. Apparently the hotel had overbooked king rooms for the event weekend and we were downgraded to a queen room. The room itself was small and underwhelming. The was not dirty, but the linens and furnishings all had a very old and worn feel. The hotel was packed the weekend we visited. The staff was obviously not prepared to cater to such a large crowd. The concierge and reception desks were continuously busy. The hotel restaurant was very slow and had long waits. The main lobby and other common areas throughout the hotel were also undergoing renovations making getting around and fighting crowds of other people even more difficult! Luckly the weather was nice and the airshow was enjoyable. We will not be staying at the Abassador on future trips. Hopefully, current renovations will provide a much needed lift to the hotels current worn down vibe...

This is a ***genuine*** review

We booked 2 rooms for 2 nights on Hotwire over Labor Day weekend. We arrived at the Hard Rock at around 10:00 a.m., and they were able to give us our rooms early. Hooray! We were on the 5th floor (Beatles theme). The rooms were very comfortable. One of our rooms was on a corner, and had lots of windows. The other was was a bit larger with fewer windows. Not a great view from our side of the hotel, but we didn't pay for a room with a view. The beds and pillows were EXTREMELY comfortable, the bathroom was full of Aveda products, and there was a bathrobe in the closet. The tv/dvd/stereo combo was nice, but we weren't in the room a lot to use it. We were warned by the front desk not to even touch the snacks/drinks as they were weighted and we'd be charged. No problems there. We did not make use of the free fitness center, because the weather was perfect. My husband was able to go for a run through Grant Park along the waterfront instead. He said it was wonderful. We did use the free internet in our rooms. It was 'wired' internet, but my stepdad said that was better for him anyway. The room was a bit dim at night, but we were able to read just fine using the bedside lamps. There was virtually no hall / elevator noise. The location of the Hard Rock is ideal. It is easy to get breakfast at the Corner Bakery, just blocks to either Grant Park, The Art Institute, State Street Shopping, or Michigan Ave. To sum up, there was absolutely NOTHING to complain about. We would be glad to stay here again any time.

Figure 2.19: Example interface for **predicted label + examples**.

Chapter 3

Many Faces of Feature Importance

3.1 Overview

Feature importance is commonly used to explain machine predictions. While feature importance can be derived from a machine learning model with a variety of methods, the consistency of feature importance via different methods remains understudied. In this work, we systematically compare feature importance from built-in mechanisms in a model such as attention values and post-hoc methods that approximate model behavior such as LIME. Using text classification as a testbed, we find that 1) no matter which method we use, important features from traditional models such as SVM and XGBoost are more similar with each other, than with deep learning models; 2) post-hoc methods tend to generate more similar important features for two models than built-in methods. We further demonstrate how such similarity varies across instances. Notably, important features do **not** always resemble each other better when two models agree on the predicted label than when they disagree.

3.2 Introduction

As machine learning models are adopted in societally important tasks such as recidivism prediction and loan approval, explaining machine predictions has become increasingly important [Doshi-Velez and Kim, 2017, Lipton, 2016]. Explanations can potentially improve the trustworthiness of algorithmic decisions for decision makers, facilitate model developers in debugging, and even allow regulators to identify biased algorithms.

One of favorite places to eat on the King W side, simple and relatively quick. I typically always get the chicken burrito and the small is enough for me for dinner. Ingredients are always fresh and watch out for the hot sauce cause it's skull scratching hot. Seating is limited so be prepared to take your burrito outside or you can even eat at Metro Hall Park.

methodsmodels	SVM (ℓ_2)	XGBoost	LSTM with attention	BERT
built-in	sauce, seating, park, prepared, even, always, can, fresh, quick, favorite	is, can, quick, fresh, at, to, always, even, favorite, and	me, be, relatively, enough, always, fresh, ingredients, prepared, quick, favorite	., ingredients, relatively, quick, places, enough, dinner, typically, me, i
LIME	the, dinner, be, quick, and, even, you, always, fresh, favorite	you, to, fresh, quick, at, can, even, always, and, favorite	dinner, ingredients, typically, fresh, places, cause, quick, and, favorite, always	one, watch, to, enough, limited, cause, and, fresh, hot, favorite

Table 3.1: 10 most important features (separated by comma) identified by different methods for different models for the given review. In the interest of space, we only show built-in and LIME here.

A popular approach to explaining machine predictions is to identify important features for a particular prediction [Luong et al., 2015b, Ribeiro et al., 2016, Lundberg and Lee, 2017]. Typically, these explanations assign a value to each feature (usually a word in NLP), and thus enable visualizations such as highlighting top k features.

In general, there are two classes of methods: 1) built-in feature importance that is embedded in the machine learning model such as coefficients in linear models and attention values in attention mechanisms; 2) post-hoc feature importance through credit assignment based on the model such as LIME. It is well recognized that robust evaluation of feature importance is challenging [Jain and Wallace, 2019a, Nguyen, 2018], which is further complicated by different use cases of explanations (e.g., for decision makers vs. for developers). Throughout this work, we refer to machine learning models that learn from data as **models** and methods to obtain local explanations (i.e., feature importance in this work) for a prediction by a model as **methods**.

While prior research tends to focus on the internals of models in designing and evaluating methods of explanations, e.g., how well explanations reflect the original model [Ribeiro et al., 2016], we view feature importance itself as a subject of study, and aim to provide a systematic characterization of important features obtained via different methods for different models. This view is particularly important when explanations are used to support decision making because they are the only exposure to the model for decision makers. It would be desirable that explanations are consistent across different instances. In comparison, debugging represents a distinct use case where developers often know the mechanism of the model beyond explanations. Our view also connects to studying explanation as a **product** in cognitive studies of explanations [Lombrozo, 2012], and is complementary to the model-centric perspective.

Given a wide variety of models and methods to generate feature importance, there are basic open questions such as how similar important features are between models and methods, how important features distribute across instances, and what linguistic properties important features tend to have. We use text classification as a testbed to answer these questions. We consider built-in importance from both traditional models such as linear SVM and neural models with attention mechanisms, as well as post-hoc importance based on LIME and SHAP. Table 3.1 shows important features for a Yelp review in sentiment classification. Although most approaches consider "fresh" and "favorite" important, there exists significant variation.

We use three text classification tasks to characterize the overall similarity between important features. Our analysis reveals the following insights:

- (Comparison between approaches) Deep learning models generate more different important features from traditional models such as SVM and XGBoost. Post-hoc methods tend to reduce the dissimilarity between models by making important features more similar than the built-in method. Finally, different approaches do not generate more similar important features even if we focus on the most important features (e.g., top one feature).
- (Heterogeneity between instances) Similarity between important features is not always greater when two models agree on the predicted label, and longer instances are less likely to share important features.
- (Distributional properties) Deep models generate more diverse important features with

higher entropy, which indicates lower consistency across instances. Post-hoc methods bring the POS distribution closer to background distributions.

In summary, our work systematically compares important features from different methods for different models, and sheds light on how different models/methods induce important features. Our work takes the first step to understand important features as a product and helps inform the adoption of feature importance for different purposes. Our code is available at `https://github.com/BoulderDS/feature-importance`.

3.3 Related work

To provide further background for our work, we summarize current popular approaches to generating and evaluating explanations of machine predictions, with an emphasis on feature importance.

Approaches to generating explanations. A battery of approaches have been recently proposed to explain machine predictions (see Guidotti et al. [2019] for an overview), including example-based approaches that identifies "informative" examples in the training data [e.g., Kim et al., 2016] and rule-based approaches that reduce complex models to simple rules [e.g., Malioutov et al., 2017]. Our work focuses on characterizing properties of feature-based approaches. Feature-based approaches tend to identify important features in an instance and enable visualizations with important features highlighted. We discuss several directly related post-hoc methods here and introduce the built-in methods in §**??**. A popular approach, LIME, fits a sparse linear model to approximate model predictions locally [Ribeiro et al., 2016]; Lundberg and Lee [2017] present a unified framework based on Shapley values, which can be computed with different approximation methods for different models. Gradients are popular for identifying important features in deep learning models since these models are usually differentiable [Shrikumar et al., 2017], for instance, Li et al. [2016] uses gradient-based saliency to compare LSTMs with simple recurrent networks.

Definition and evaluation of explanations. Despite a myriad of studies on approaches to explaining machine predictions, explanation is a rather overloaded term and evaluating explanations is challenging. Doshi-Velez and Kim [2017] lays out three levels of evaluations: functionally-grounded evaluations based on proxy automatic tasks, human-grounded evaluations with laypersons on proxy tasks, and application-grounded based on expert performance in the end task. In text classification, Nguyen [2018] shows that automatic evaluation based on word deletion moderately correlate with human-grounded evaluations that ask crowdworkers to infer machine predictions based on explanations. However, explanations that help humans infer machine predictions may not actually help humans make better decisions/predictions. In fact, recent studies find that feature-based explanations alone have limited improvement on human performance in detecting deceptive reviews and media biases [Lai and Tan, 2019a, Horne et al., 2019].

In another recent debate, Jain and Wallace [2019a] examine attention as an explanation mechanism based on how well attention values correlate with gradient-based feature importance and whether they exclusively lead to the predicted label, and conclude that attention is not explanation. Similarly, Serrano and Smith [2019] show that attention is not a fail-safe indicator for explaining machine predictions based on intermediate representation erasure. However, Wiegreffe and Pinter [2019a] argue that attention can be explanation depending on the definition of explanations (e.g., plausibility and faithfulness).

In comparison, we treat feature importance itself as a subject of study and compare different approaches to obtaining feature importance from a model. Instead of providing a normative judgment with respect to what makes good explanations, our goal is to allow decision makers or model developers to make informed decisions based on properties of important features using different models and methods.

3.4 Approach

In this section, we first formalize the problem of obtaining feature importance and then introduce the models and methods that we consider in this work. Our main contribution is to compare important features identified for a particular instance through different methods for different models.

Feature importance. For any instance t and a machine learning model $m : t \to y \in \{0, 1\}$, we use method h to obtain feature importance on an interpretable representation of t, $\mathcal{I}_t^{m,h} \in \mathbb{R}^d$, where d is the dimension of the interpretable representation. In the context of text classification, we use unigrams as the interpretable representation. Note that the machine learning model does not necessarily use the interpretable representation. Next, we introduce the models and methods in this work.

Models (m). We include both recent deep learning models for NLP and popular machine learning models that are not based on neural networks. In addition, we make sure that the chosen models have some built-in mechanism for inducing feature importance and describe the built-in feature importance as we introduce the model.[1]

- Linear SVM with ℓ_2 (or ℓ_1) regularization. Linear SVM has shown strong performance in text categorization [Joachims, 1998]. The absolute value of coefficients in these models is typically considered a measure of feature importance [e.g., Ott et al., 2011]. We also consider ℓ_1 regularization because ℓ_1 regularization is often used to induce sparsity in the model.

- Gradient boosting tree (XGBoost). XGBoost represents an ensembled tree algorithm that shows strong performance in competitions [Chen and Guestrin, 2016]. We use the default option in XGBoost to measure feature importance with the average training loss gained when using a feature for splitting.

- LSTM with attention (often shortened as LSTM in this work). Attention is a commonly used technique in deep learning models for NLP [Bahdanau et al., 2015]. The intuition is to assign a weight to each token before aggregating into the final prediction (or decoding in machine translation). We use the dot product formulation in Luong et al. [2015b]. The

[1] For instance, we do not consider LSTM as a model here due to the lack of commonly-accepted built-in mechanisms.

weight on each token has been commonly used to visualize the importance of each token. To compare with the previous bag-of-words models, we use the average weight of each type (unique token) in this work to measure feature importance.

- BERT. BERT represents an example architecture based on Transformers, which could show different behavior from LSTM-style recurrent networks [Devlin et al., 2019, Vaswani et al., 2017, Wolf, 2019]. It also achieves state-of-the-art performance in many NLP tasks. Similar to LSTM with attention, we use the average attention values of 12 heads used by the first token at the final layer (the representation passed to fully connected layers) to measure feature importance for BERT.[2] Since BERT uses a subword tokenizer, for each word, we aggregate the attention on related subparts. BERT also requires special processing due to the length constraint; please refer to the supplementary material for details. As a result, we focus on presenting LSTM with attention in the main paper for ease of understanding.

Methods (h). For each model, in addition to the built-in feature importance that we described above, we consider the following two popular methods for extracting post-hoc feature importance (see the supplementary material for details of using the post-hoc methods).

- LIME [Ribeiro et al., 2016]. LIME generates post-hoc explanations by fitting a local sparse linear model to approximate model predictions. As a result, the explanations are sparse.
- SHAP [Lundberg and Lee, 2017]. SHAP unifies several interpretations of feature importance through Shapley values. The main intuition is to account the importance of a feature by examining the change in prediction outcomes for all the combinations of other features. Lundberg and Lee [2017] propose various approaches to approximate the computation for different classes of models (including gradient-based methods for deep models).

Note that feature importances obtained via all approaches are all local, because the top

[2] We also tried to use the max of 12 heads and previous layers, and the average of the final layer is more similar to SVM (ℓ_2) than the average of first layer. Results are in the supplementary material. Vig [2019] show that attention in BERT tends to be on first words, neighboring words, and even separators. The complex choices for BERT further motivate our work to view feature importance as a subject of study.

features are conditioned on an instance (i.e., words present in an instance) even for the built-in method for SVM and XGBoost.

Comparing feature importance. Given $\mathcal{I}_t^{m,h}$ and $\mathcal{I}_t^{m',h'}$, we use Jaccard similarity based on the top k features with the greatest absolute feature importance, $\frac{|\,\mathrm{TopK}(\mathcal{I}_t^{m,h}) \cap \mathrm{TopK}(\mathcal{I}_t^{m',h'})|}{|\,\mathrm{TopK}(\mathcal{I}_t^{m,h}) \cup \mathrm{TopK}(\mathcal{I}_t^{m',h'})|}$, as our main similarity metric for two reasons. First, the most typical way to use feature importance for interpretation purposes is to show the most important features [Lai and Tan, 2019a, Ribeiro et al., 2016, Horne et al., 2019]. Second, some models and methods inherently generate sparse feature importance, so most feature importance values are 0.

It is useful to discuss the implication of similarity before we proceed. On the one hand, it is possible that different models/methods identify the same set of important features (high similarity) and the performance difference in prediction is due to how different models weigh these important features. If this were true, the choice of model/method would have mattered little for visualizing important features. On the other hand, a low similarity poses challenges for choosing which model/method to use for displaying important features. In that case, this work aims to develop an understanding of how the similarity varies depending on models and methods, instances, and features. We leave it to future work for examining the impact on human interaction with feature importance. Low similarity may enable model developers to understand the differences between models, but may lead to challenges for decision makers to get a consistent picture of what the model relies on.

3.5 Experimental Setup and Hypotheses

Our goal is to characterize the similarities and differences between feature importances obtained with different methods and different models. In this section, we first present our experimental setup and then formulate our hypotheses.

Experimental setup. We consider the following three text classification tasks in this work. We choose to focus on classification because classification is the most common scenario used for examining feature importance and the associated human interpretation [Jain and Wallace, 2019a].

Model	Yelp	SST	Deception
SVM (ℓ_2)	92.3	80.8	86.3
SVM (ℓ_1)	91.5	79.2	84.4
XGBoost	88.8	75.9	83.4
LSTM w/ attention	93.9	82.6	88.4
BERT	95.5	92.2	90.9

Table 3.2: Accuracy on the test set.

- Yelp [Yelp, 2019]. We set up a binary classification task to predict whether a review is positive (rating ≥ 4) or negative (rating ≤ 2). As the original dataset is huge, we subsample 12,000 reviews for this work.
- SST [Socher et al., 2013]. It is a sentence-level sentiment classification task and represents a common benchmark. We only consider the binary setup here.
- Deception detection [Ott et al., 2013, 2011]. This dataset was created by extracting genuine reviews from TripAdvisor and collecting deceptive reviews using Turkers. It is relatively small with 1,200 reviews and represents a distinct task from sentiment classification.

For all the tasks, we use 20% of the dataset as the test set. For SVM and XGBoost, we use cross validation on the other 80% to tune hyperparameters. For LSTM with attention and BERT, we use 10% of the dataset as a validation set, and choose the best hyperparameters based on the validation performance. We use spaCy to tokenize and obtain part-of-speech tags for all the datasets [Honnibal and Montani, 2017]. Table 4.1 shows the accuracy on the test set and our results are comparable to prior work. Not surprisingly, BERT achieves the best performance in all three tasks. For important features, we use $k \leq 10$ for Yelp and deception detection, and $k \leq 5$ for SST as it is a sentence-level task. See supplementary materials for details of preprocessing, learning, and dataset statistics.

Hypotheses. We aim to examine the following three research questions in this work: 1) How similar are important features between models and methods? 2) What factors relate to the heterogeneity across instances? 3) What words tend to be chosen as important features?

Overall similarity. Here we focus on discussing comparative hypotheses, but we would like to note that it is important to understand to what extent important features are similar across models (i.e., the value of similarity score). First, as deep learning models and XGBoost are nonlinear, we hypothesize that built-in feature importance is more similar between SVM (ℓ_1) and SVM (ℓ_2) than other model pairs (**H1a**). Second, LIME generates more similar important features to SHAP than to built-in feature importance because both LIME and SHAP make additive assumptions, while built-in feature importance is based on drastically different models (**H1b**). It also follows that post-hoc explanations of different models show higher similarity than built-in explanations across models. Third, the similarity with small k is higher (**H1c**) because hopefully, all models and methods agree what the most important features are.

Heterogeneity between instances. Given a pair of (model, method) combinations, our second question is concerned with how instance-level properties affect the similarity in important features between different combinations. We hypothesize that 1) when two models agree on the predicted label, the similarity between important features is greater (**H2a**); 2) longer instances are less likely to share similar important features (**H2b**). 3) instances with higher type-token ratio,[3] which might be more complex, are less likely to share similar important features (**H2c**).

Distribution of important features. Finally, we examine what words tend to be chosen as important features. This question certainly depends on the nature of the task, but we would like to understand how consistent different models and methods are. We hypothesize that 1) deep learning models generate more diverse important features (**H3a**); 2) adjectives are more important in sentiment classification, while pronouns are more important in deception detection as shown in prior work (**H3b**).

3.6 Similarity between Instance-level Feature Importance

We start by examining the overall similarity between different models using different methods. In a nutshell, we compute the average Jaccard similarity of top k features for each pair of (m, h)

[3] Type-token ratio is defined as the number of unique tokens divided by the number of tokens.

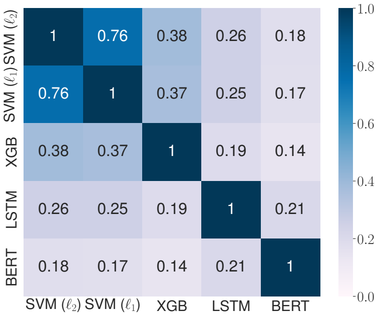


Figure 3.1: Jaccard similarity between the top 10 features of different models based on built-in feature importance on Yelp. The similarity is the greatest between SVM (ℓ_2) and SVM (ℓ_1), while LSTM with attention and BERT pay attention to quite different features from other models.

and (m', h'). To facilitate effective comparisons, we first fix the method and compare the similarity of different models, and then fix the model and compare the similarity of different methods. Figure 3.1 shows the similarity between different models using the built-in feature importance for the top 10 features in Yelp ($k = 10$). Consistent with **H1a**, SVM (ℓ_2) and SVM (ℓ_1) are very similar to each other, and LSTM with attention and BERT clearly lead to quite different top 10 features from the other models. As the number of important features (k) can be useful for evaluating the overall trend, we thus focus on line plots as in Figure 3.2 in the rest of the paper. This heatmap visualization represents a snapshot for $k = 10$ using the built-in method. Also, we only include SVM (ℓ_2) in the main paper for ease of visualization and sometimes refer to it in the rest of the paper as SVM.

No matter which method we use, important features from SVM and XGBoost are more similar with each other, than with deep learning models (Figure 3.2). First, we compare the similarity of feature importance between different models using the same method. Using the built-in method (first row in Figure 3.2), the solid line (SVM x XGBoost) is always above the other lines, usually by a significant margin, suggesting that deep learning models such as LSTM with attention are less similar to traditional models. In fact, the similarity between XGBoost and LSTM with attention is lower than random samples for $k = 1, 2$ in SST. Similar results also hold for BERT (see supplementary materials). Another interesting observation is that post-hoc methods tend to generate greater similarity than built-in methods, especially for LIME (the dashed line (LIME) is always above the solid line (built-in) in the second row of Figure 3.2). This is likely because LIME only depends on the model behavior (i.e., what the model predicts) and does not account for how the model works.

The similarity between important features from different methods tends to be lower for LSTM with attention (Figure 3.3). Second, we compare the similarity of feature importance derived from the same model with different methods. For deep learning models such as LSTM with attention, the similarity between feature importance generated by different methods is the lowest, especially comparing LIME with SHAP. Notably, the results are much more cluttered in deception detection. Contrary to **H1b**, we do not observe that LIME is more similar to SHAP than built-in. The order seems to depend on both the task and the model: even within SST, the similarity between built-in and LIME can rank as third, second, or first. In other words, post-hoc methods generate more similar important features when we compare different models, but that is not the case when we fix the model. It is reassuring that that similarity between any pairs is above random, with a sizable margin in most cases (BERT on SHAP is an exception; see supplementary materials).

Relation with k. As the relative order between different approaches can change with k, we have so far only focused on relatively consistent patterns over k and classification tasks. Contrary to **H1c**, the similarity between most approaches is not drastically greater for small k, which suggests that different approaches may not even agree on the most important features. In fact, there is no consistent trend as k grows: similarity mostly **increases** in SST (while our hypothesis is that it decreases), increases or stays level in Yelp, and shows varying trends in deception detection.

3.7 Heterogeneity between Instances

Given the overall low similarity between different methods/models, we next investigate how the similarity may vary across instances.

The similarity between models is not always greater when two models agree on the

predicted label (Figure 3.4). One hypothesis for the overall low similarity between models is that different models tend to give different predictions therefore they choose different features to support their decisions. However, we find that the similarity between models is not particularly high when they agree on the predicted label, and are sometimes even lower than when they disagree. This is true for LIME in Yelp and for all methods in deception detection. In SST, the similarity when the models agree on the predicted label is generally greater than when they disagree. We show the comparison between SVM (ℓ_2) and LSTM here, and similar results hold for other combinations (see supplementary materials). This observation suggests that feature importance may not connect with the predicted labels: different models agree for different reasons and also disagree for different reasons.

The similarity between models and methods is generally negatively correlated with length but positively correlated with type-token ratio (Figure 3.5). Our results support **H2b**: Spearman correlation between length and similarity is mostly below 0, which indicates that the longer an instance is, the less similar the important features are. The negative correlation becomes stronger as k grows, indicating that length has a stronger effect on similarity when we consider more top features. However, this is not true in the case of LIME and SHAP where the correlation between length and similarity are occasionally above 0 and sometimes even the declining relationship with k does not hold. Our result on type-token ratio is opposite to **H2c**: the greater the type-token ratio, the higher the similarity (see supplementary materials). We believe that the reason is that type-token ratio is strongly negatively correlated with length (the Spearman correlation for Yelp, SST and deception dataset is -0.92, -0.59 and -0.84 respectively). In other words, type-to-token ratio becomes redundant to length and fails to capture text complexity beyond length.

3.8 Distribution of Important Features

Finally, we examine the distribution of important features obtained from different approaches. These results may partly explain our previously observed low similarity in feature importance.

Important features show higher entropy using LSTM with attention and lower entropy with XGBoost (Figure 3.6). As expected from **H3a**, LSTM with attention (the pink lines) are usually at the top (similar results for BERT in the supplementary material). Such a high entropy can contribute to the low similarity between LSTM with attention and other models. However, as the order in similarity between SVM and XGBoost is less stable, entropy cannot be the sole cause.

Distribution of POS tags (Figure 3.7 and Figure 3.8). We further examine the linguistic properties of important words. Consistent with **H3b**, adjectives are more important in sentiment classification than in deception detection. On the contrary to our hypothesis, we found that pronouns do not always play an important role in deception detection. Notably, LSTM with attention puts a strong emphasis on nouns in deception detection. In all cases, determiners are under-represented among important words. With respect to the distance of part-of-speech tag distributions between important features and all words (background), post-hoc methods tend to bring important words closer to the background words, which echoes the previous observation that post-hoc methods tend to increase the similarity between important words (Figure 3.8).

3.9 Conclusion

In this work, we provide the first systematic characterization of feature importance obtained from different approaches. Our results show that different approaches can sometimes lead to very different important features, but there exist some consistent patterns between models and methods. For instance, deep learning models tend to generate diverse important features that are different from traditional models; post-hoc methods lead to more similar important features than built-in methods.

As important features are increasingly adopted for varying use cases (e.g., decision making vs. model debugging), we hope to encourage more work in understanding the space of important features, and how they should be used for different purposes. While we focus on consistent patterns across classification tasks, it is certainly interesting to investigate how properties related to tasks and data affect the findings. Another promising direction is to understand whether more concentrated important features (lower entropy) lead to better human performance in supporting decision making.

3.10 Appendix

3.10.1 Preprocessing and Computational Details

Preprocessing. We used spaCy for tokenization and part-of-speech tagging. All the words are lowercased. Table 3.3 shows basic data statistics.

dataset	average tokens
Yelp	134.6
SST	20.0
Deception	163.7

Table 3.3: Dataset statistics.

Hyperparameter tuning. Hyperparameters for both SVM and XGBoost are tuned using cross validation. The only hyperparameter tuned for SVM includes C. We try a range of Cs from log space -5 to 5. The finalized value of C ranges between 1 and 5. Hyperparameters tuned for XGBoost include learning rate, max depth of tree, gamma, number of estimators and colsample by tree. We lay out the range of values tried in the process of hyperparameter tuning, learning rate: 0.1 to 0.0001, max depth of tree: 3 to 7, gamma: 1 to 10, number of estimators: 1000 to 10000 and colsample by tree: 0.1 to 1.0. Hyperparameters for LSTM with attention are tuned using the validation dataset which comprises 10% of the entire dataset. They include embedding dimension, hidden dimension, learning rate, number of epochs and the type of optimizer. The range of values tried in the process of hyperparameter tuning, hidden dimension: 256 and 512, learning rate: 0.01 to 0.0001, number of epochs: 3 to 20 and type of optimizer: SGD and adam.

BERT fine-tuning. We fine-tuned BERT from a pre-trained BERT model provided by its original release and pytorch implementation Wolf [2019]. We use the same architecture of 12 layers Transformer with 12 attention heads. The hidden dimension of each layer is 768. The vocabulary size is 30522. The initial learning rate we use is $5 * e^{-5}$, and we add an extra ℓ_2 regularization on the parameters that are not bias terms or normalization layer with a coefficient of 0.01. We do early stopping according to the validation set within the first 20 epochs with batch size no larger than 4. The attention weights we consider are the self-attention weights of the first token of each text instance, namely the attention weights from "[CLS]", since according to BERT's design, the first token will generate the sentence representation fed into the classification layer. For the three target tasks, we choose different maximum lengths according to their natural length. For the deception detection task, the maximum sequence length is 300 tokens. For the SST binary classification task, we choose the default 128 tokens as the maximum length and for the yelp review classification task we use 512 tokens.

BERT alignment. Given that BERT tokenizes a text instance with its own tokenizer, we map the important features from BERT tokens to tokenize results from spaCy we used for other models. To be specific, we generate token start-end information as a tuple and call it token spans. We show an example for text instance "It's a good day.":

tokenization 1: [It's], [a], [good], [day], [.]

token spans 1: (0,3),(4,4),(5,8),(9,11),(12,12)

tokenization 2: [It], ['s], [a], [go], [od], [day], [.]

token spans 2: (0,1), (2,3), (4,4), (5,6), (7,8), (9,11), (12,12)

With the span information, we can identify how a token in the first tokenization relates to tokens in the second tokenization and then aggregate all the attention values to the sub-parts. Formally,

$$W^{(1)}_{(i,j)} = \sum_{(s,t) \text{ s.t. } t \geq i, s \leq j} \min(1, \tfrac{t-i+1}{t-s+1}, \tfrac{j-s+1}{t-s+1}) W^{(2)}_{(k,p)}.$$

In other words, for partial span overlapping, we allocate the weight according to the span overlapping ratio. For example: if $\text{span}^{(1)}_i = (2,5)$ and $\text{span}^{(2)}_{k-1} = (2,3), \text{span}^{(2)}_k = (4,6)$, then $W^{(1)}_{(2,5)} = W^{(2)}_{(2,3)} + \frac{2}{3} W^{(2)}_{(4,6)}$. Here $W^{(2)}$ represents the importance weight according to the second tokenization, $W^{(1)}_{(i,j)}$ represents the aligned feature importance for the token that has span (i, j)

in the first tokenization. By definition, $\sum_{(i,j)} W^{(1)}_{(i,j)} = \sum_{(i,j)} W^{(2)}_{(i,j)} = 1$ for attention values.

LIME. We use the LimeTextExplainer and write a wrapper function that returns actual probabilities of the respective model. Since the LinearSVM generates only binary predictions, we return 0.999 and 0.001 instead. We use 1,000 samples for fitting the local classifier.

SHAP. We use a LinearExplainer for linear SVM, a TreeExplainer for XGBoost, and adapt the gradient-based DeepExplainer for our neural models. The main adaptation required for the neural method is to view the embedding lookup layer as a matrix multiplication layer so that the entire network is differentiable on the input token ids.

3.11 Additional Figures

Similarity between BERT layers and SVM (ℓ_2). Important features using the final layer are more similar to that from SVM (ℓ_2) than using the first layer. See Figure 3.9.

Built-in similarity is much lower with deep learning models, and post-hoc methods "smooth" the distance. Similar results are observed in SVM (ℓ_1) and BERT. See Figure 3.10.

Similarity between methods is lower for deep learning models. Similar results are observed in SVM (ℓ_1), XGBoost and BERT. See Figure 3.11.

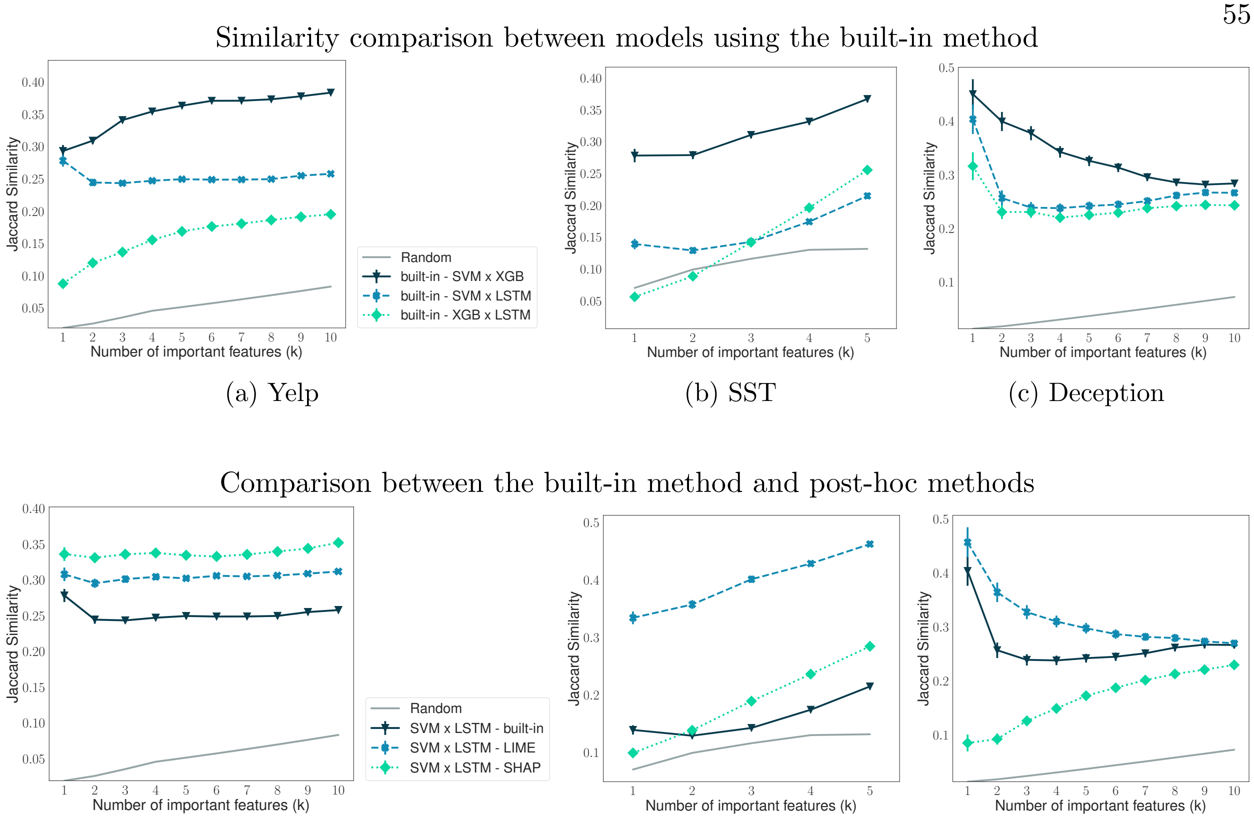
Similarity vs. predicted labels. Similarity is not necessarily higher when predictions agree, it is also not necessarily lower when predictions disagree. See Figure 3.12 and Figure 3.13.

Similarity vs. length. The negative correlation between length and similarity grows stronger as k grows. See Figure 3.14.

Similarity vs. type-token ratio. The positive correlation between type-token ratio and similarity grows stronger as k grows. See Figure 3.15 and Figure 3.16.

Entropy. Deep learning models generate more diverse important features than traditional models. See Figure 3.17.

Jensen-shannon distance between POS. Distance of part-of-speech tag distributions between important features and all words is generally smaller with post-hoc methods for traditional models. See Figure 3.18.



(a) Yelp

(b) SST

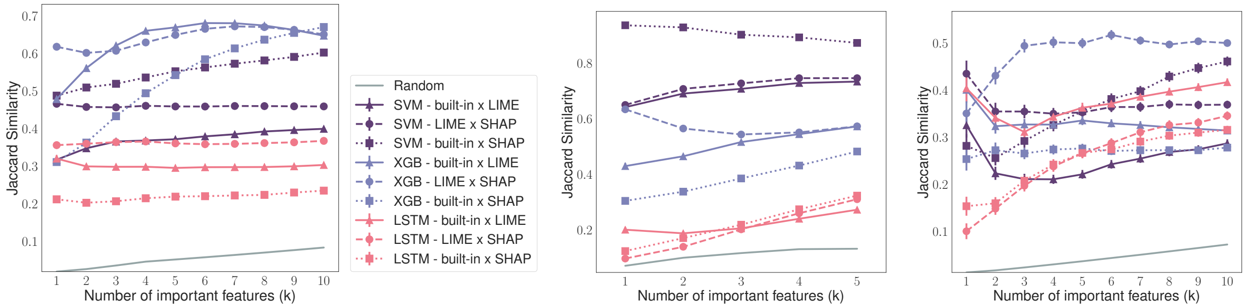
(c) Deception

(d) Yelp

(e) SST

(f) Deception

Figure 3.2: Similarity comparison between models with the same method. x-axis represents the number of important features that we consider, while y-axis shows the Jaccard similarity. **Error bars represent standard error throughout the paper.** The top row compares three pairs of models using the built-in method, while the second row compares three methods on SVM and LSTM with attention (LSTM in figure legends always refers to LSTM with attention in this work). The random line is derived using the average similarity between two random samples of k features from 100 draws.

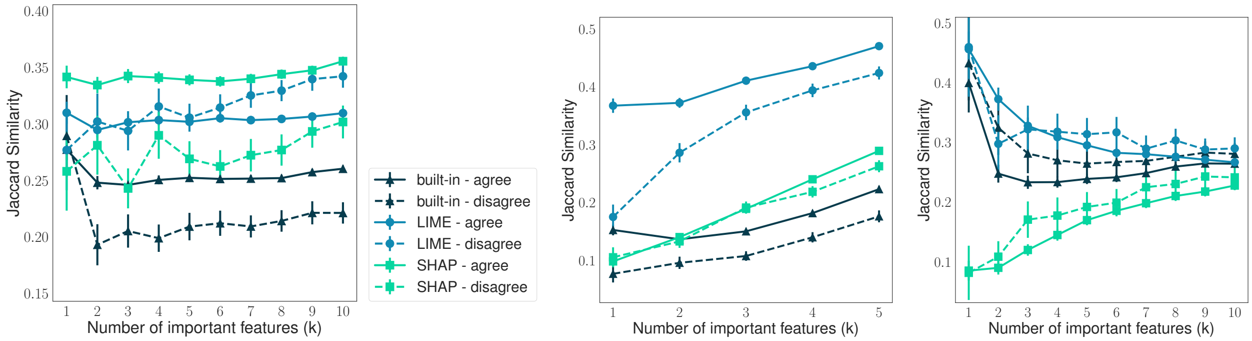


(a) Yelp

(b) SST

(c) Deception

Figure 3.3: Similarity comparison between methods using the same model. The similarity between different methods based on LSTM with attention is generally lower than other methods. Similar results hold for BERT (see the supplementary material).

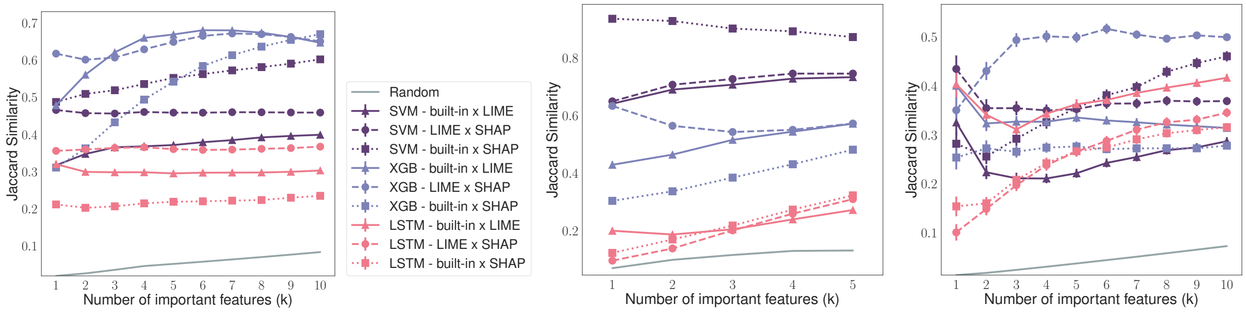


(a) Yelp

(b) SST

(c) Deception

Figure 3.4: Similarity between SVM (ℓ_2) and LSTM with attention with different methods grouped by whether these two models agree on the predicted label. The similarity is not always greater when they agree on the predicted labels than when they disagree.

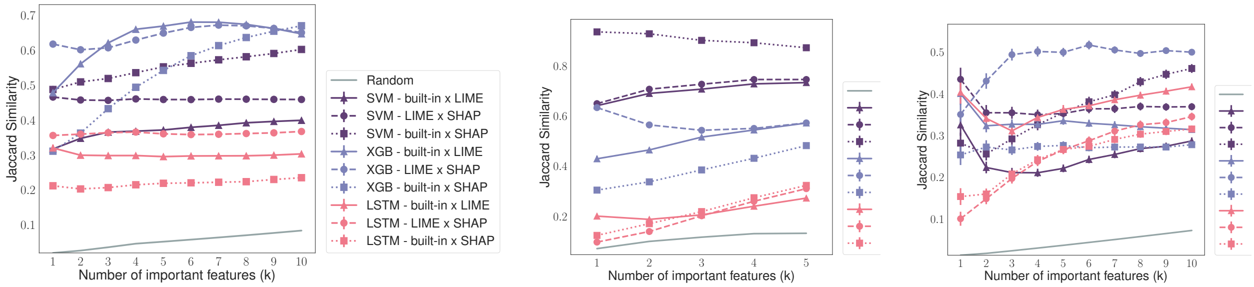


(a) Yelp

(b) SST

(c) Deception

Figure 3.5: In most cases, the similarity between feature importance is negatively correlated with length. Here we only show the comparison between different methods based on the same model. Similar results hold for comparison between different models using the same method. For ease of comparison, the gray line marks the value 0. Generally as k grows, relationship becomes even more negatively correlated.

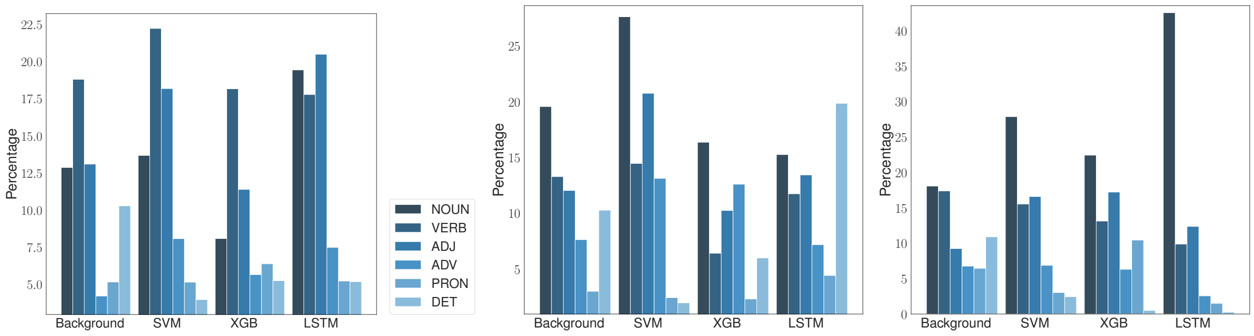


(a) Yelp

(b) SST

(c) Deception

Figure 3.6: The entropy of important features. LSTM with attention generates more diverse important features than SVM and XGBoost.

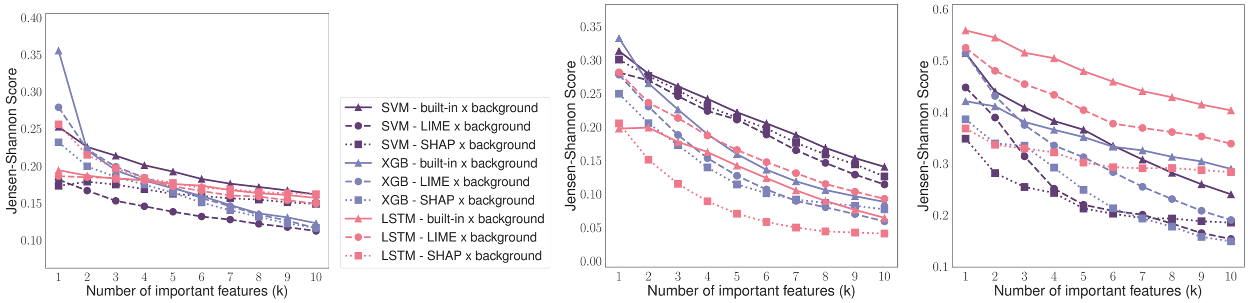


(a) Yelp

(b) SST

(c) Deception

Figure 3.7: Part-of-speech tag distribution with the built-in method. “Background” shows the distribution of all words in the test set. LSTM with attention puts a strong emphasis on nouns in deception detection, but is not necessarily more different from the background than other models.



(a) Yelp

(b) SST

(c) Deception

Figure 3.8: Distance of the part-of-speech tag distributions between important features and all words (background). Distance is generally smaller with post-hoc methods for all models, although some exceptions exist for LSTM with attention and BERT.

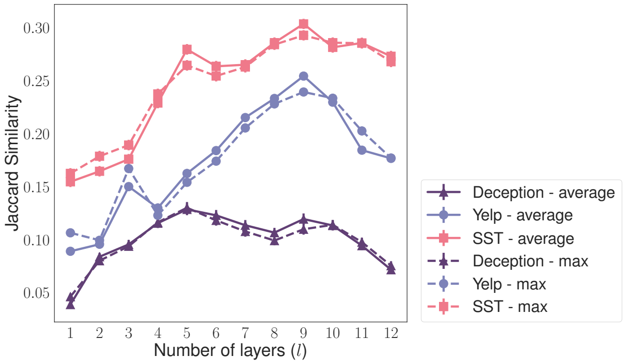
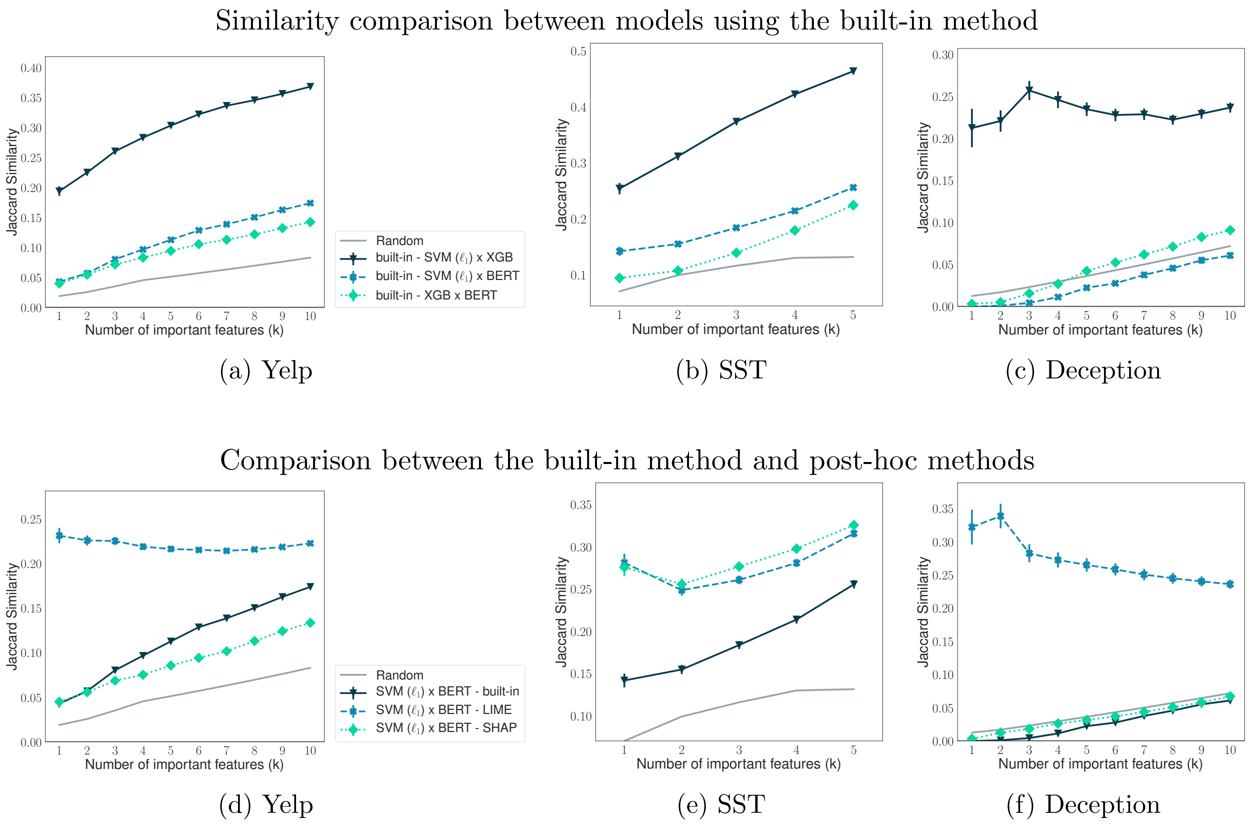


Figure 3.9: Similarity comparison between BERT layers using average or maximum attention heads score ($k = 10$). In general, similarity becomes greater as l increases, but the last layer is not necessarily the greatest. Similarity is slightly higher when average attention heads score is computed.



(a) Yelp

(b) SST

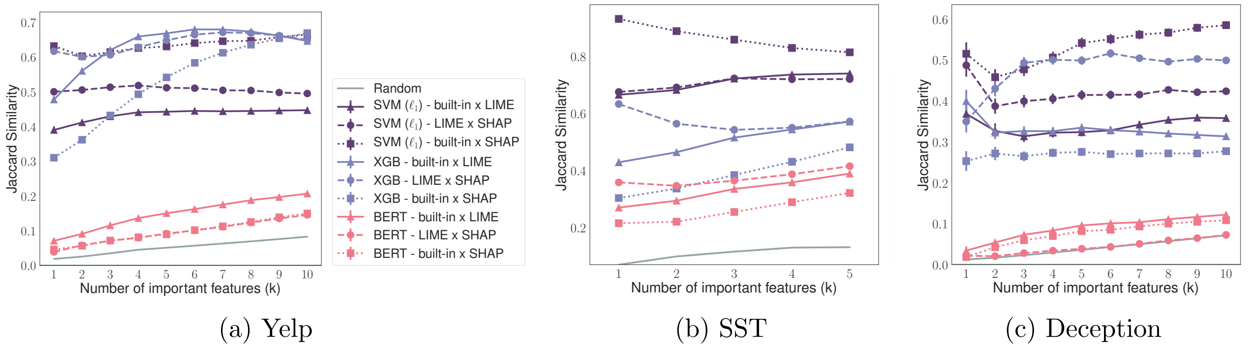
(c) Deception

(d) Yelp

(e) SST

(f) Deception

Figure 3.10: Similarity comparison between models with the same method. x-axis represents the number of important features that we consider, while y-axis shows the Jaccard similarity. **Error bars represent standard error throughout the paper.** The top row compares three pairs of models using the built-in method, while the second row compares three methods on SVM (ℓ_1) and BERT. The random line is derived using the average similarity between two random samples of k features from 100 draws.

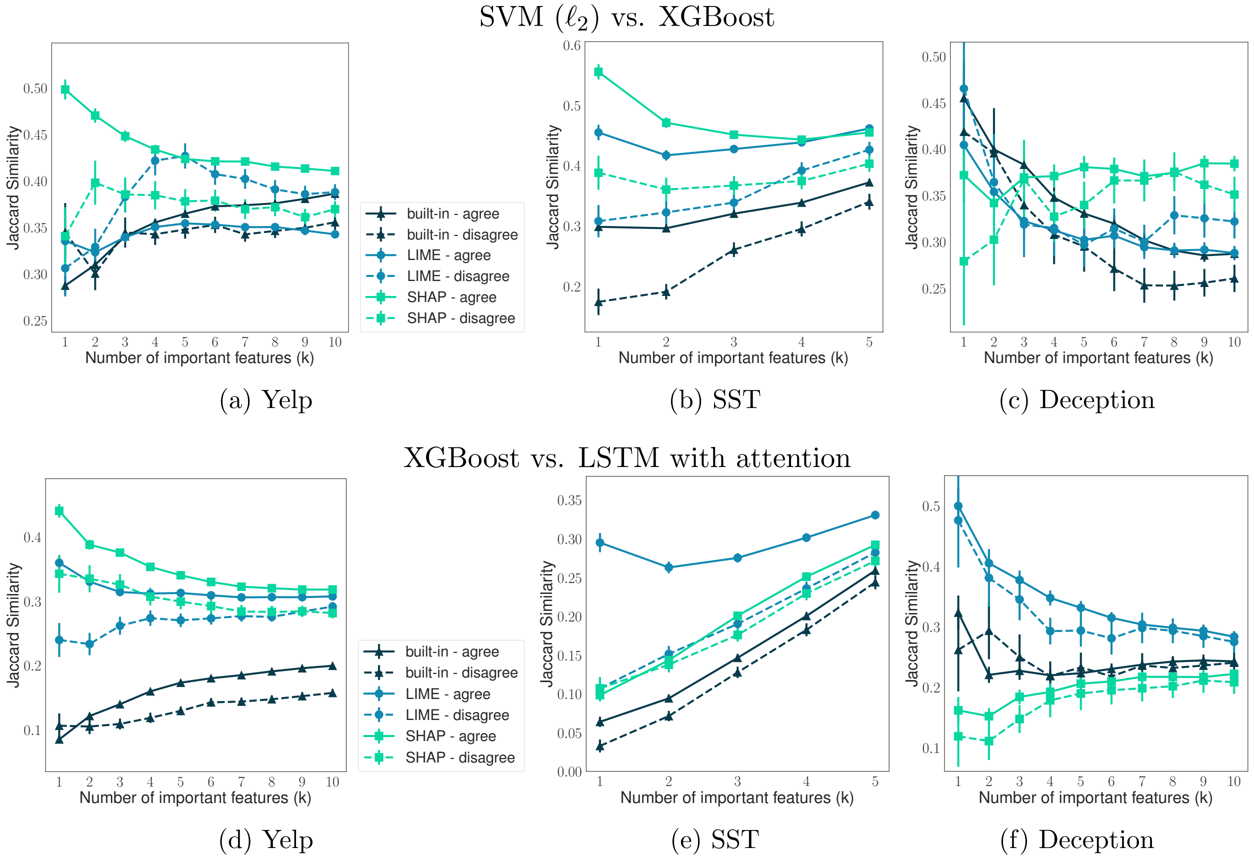


(a) Yelp

(b) SST

(c) Deception

Figure 3.11: Similarity comparison between methods using the same model for SVM (ℓ_1), XGBoost, and BERT. BERT is much closer to random in deception.



SVM (ℓ_2) vs. XGBoost

(a) Yelp

(b) SST

(c) Deception

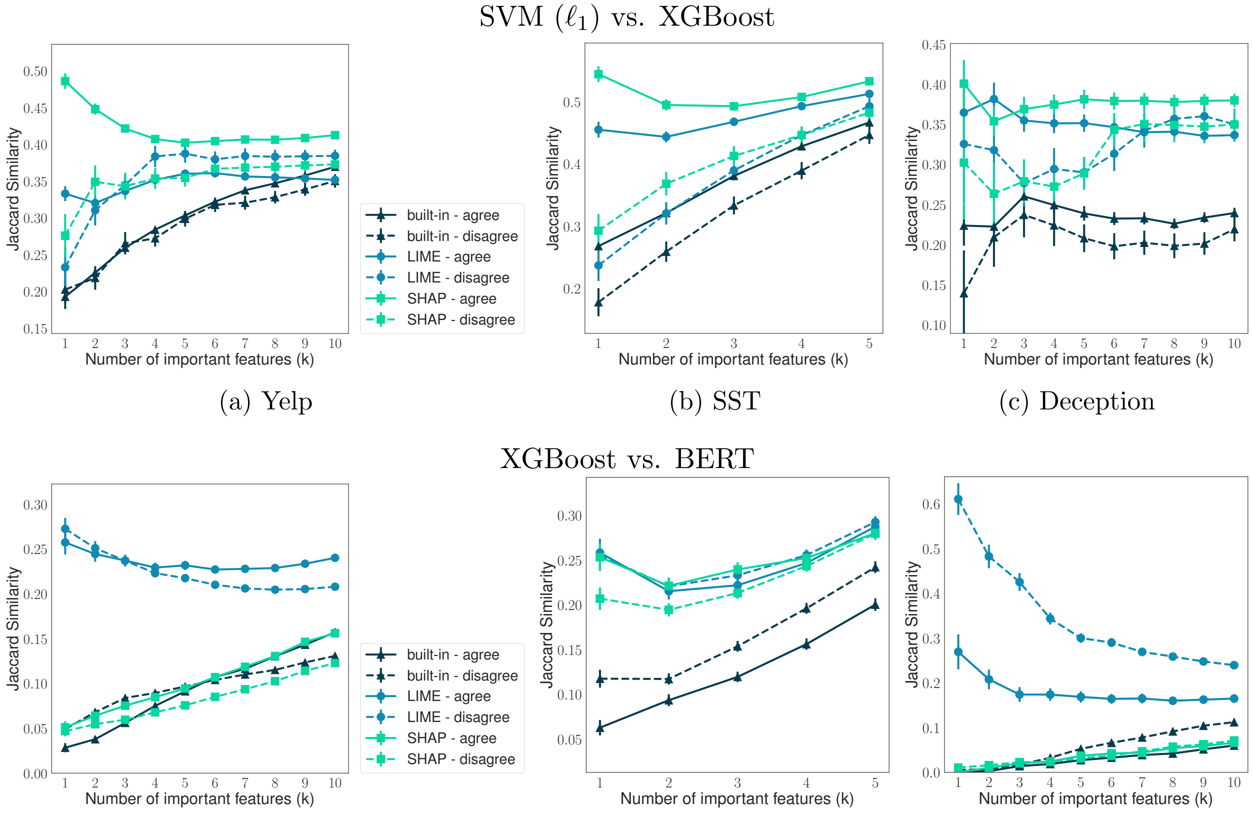
XGBoost vs. LSTM with attention

(d) Yelp

(e) SST

(f) Deception

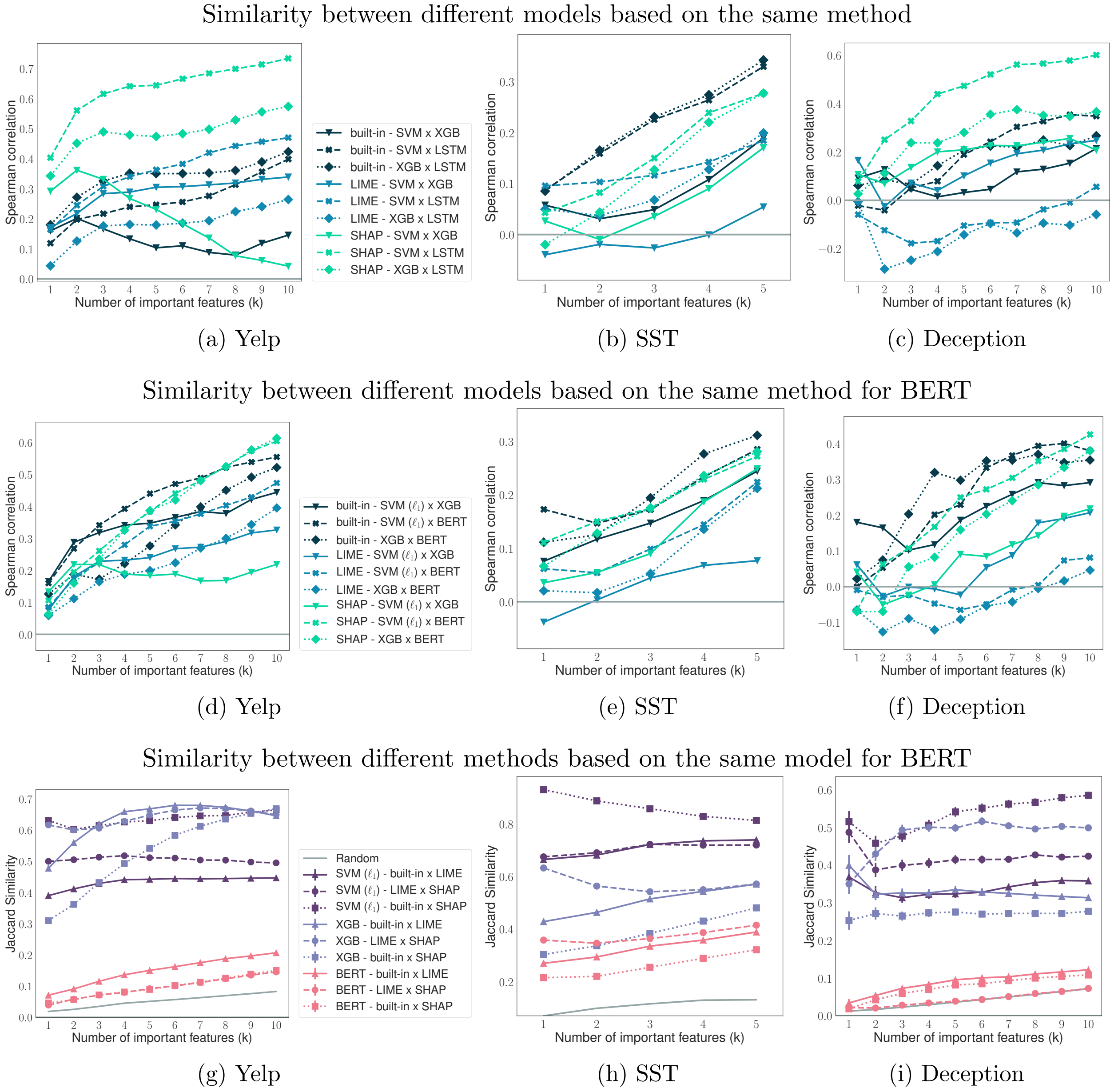
Figure 3.12: Similarity between two models is not necessarily greater when they agree on the predictions, and sometimes, e.g., SVM (ℓ_2) x XGB with LIME method, it is sometimes lower than when they disagree on the predicted labels.



(a) Yelp (b) SST (c) Deception

(d) Yelp (e) SST (f) Deception

Figure 3.13: Similarity between two models is not necessarily greater when they agree on the predictions, and in some scenarios, e.g., SVM (ℓ_1) x XGB with LIME method, XGB x BERT with LIME method, and XGB x BERT with built-in method, they are sometimes lower than when they disagree on the predicted labels.

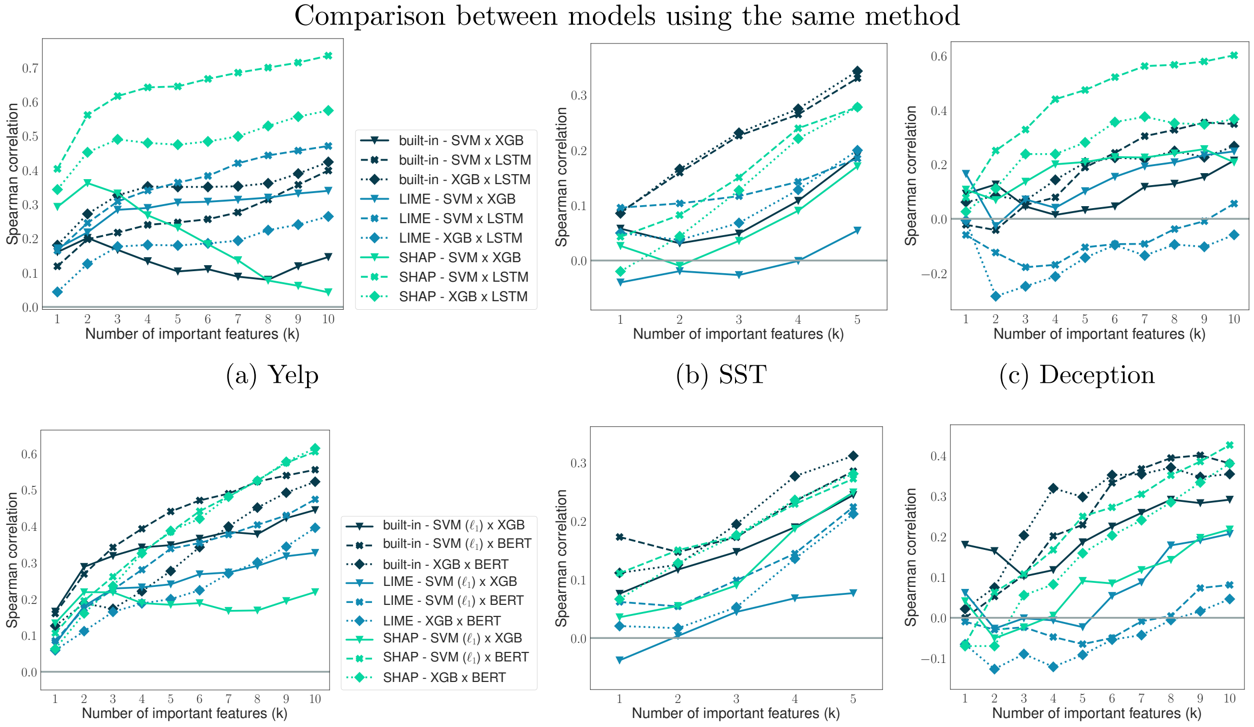


(a) Yelp (b) SST (c) Deception

(d) Yelp (e) SST (f) Deception

(g) Yelp (h) SST (i) Deception

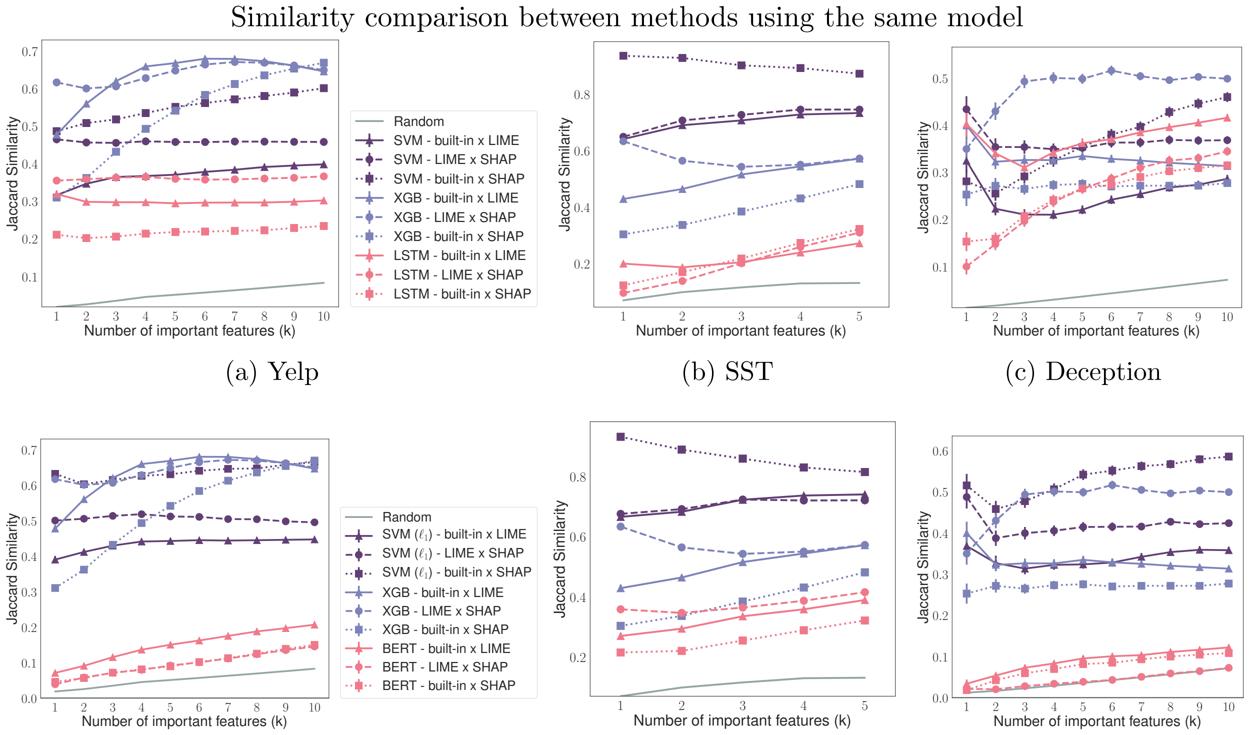
Figure 3.14: Similarity comparison vs. length. The longer the length of an instance, the less similar the important features are. The negative correlation becomes stronger as k grows. In certain scenarios, e.g., XGB - built-in x LIME and XGB - LIME x SHAP, correlation occasionally goes above 0.



(a) Yelp (b) SST (c) Deception

(d) Yelp (e) SST (f) Deception

Figure 3.15: Similarity comparison vs. type-token ratio. The higher the type-token ratio, the more similar the important features are. The positive correlation becomes stronger as k grows. In some cases, e.g., LIME method on deception dataset, correlation becomes weaker as k grows.



(a) Yelp

(b) SST

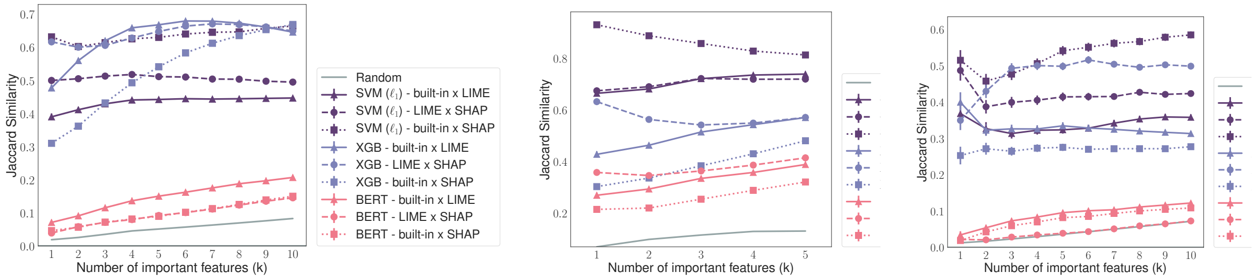
(c) Deception

(d) Yelp

(e) SST

(f) Deception

Figure 3.16: Similarity comparison vs. type-token ratio. The higher the type-token ratio, the more similar the important features are. The positive correlation becomes stronger as k grows. In some cases, e.g., XGB - built-in and LIME and XGB - LIME and SHAP on Yelp dataset, correlation becomes weaker as k grows.

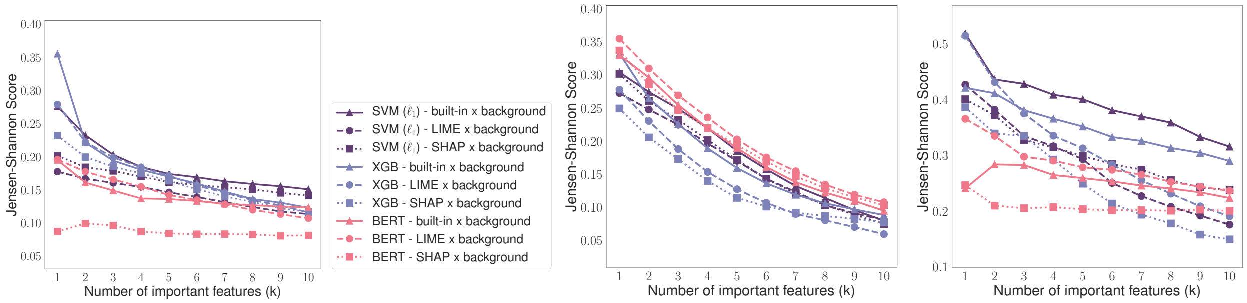


(a) Yelp

(b) SST

(c) Deception

Figure 3.17: The entropy of important features. In general, BERT generates more diverse important features than SVM (ℓ_1) and XGBoost.



(a) Yelp (b) SST (c) Deception

Figure 3.18: Distance of the part-of-speech tag distributions between important features and all words (background). Distance is generally smaller with post-hoc methods for all models, although some exceptions exist for LSTM with attention and BERT.

Chapter 4

Model-driven Tutorials and Simple Explanations

4.1 Overview

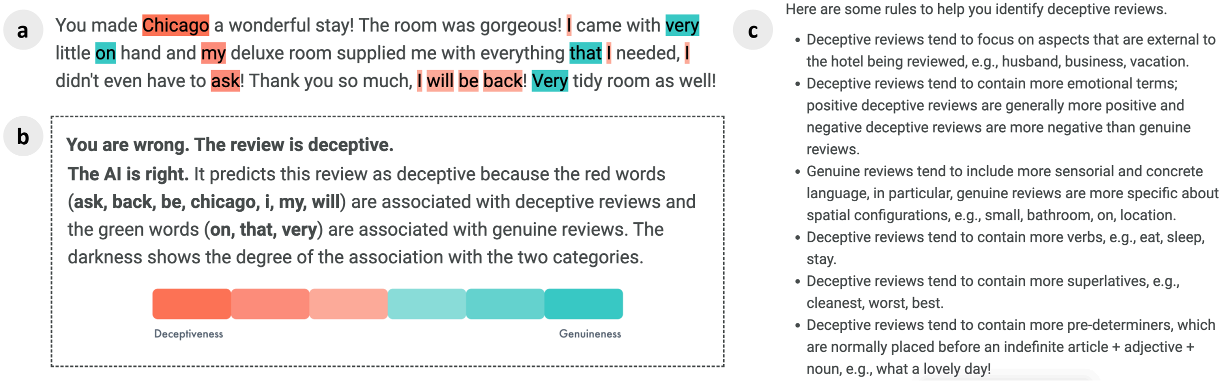
To support human decision making with machine learning models, we often need to elucidate patterns embedded in the models that are not salient, unknown, or counterintuitive to humans. While existing approaches focus on explaining machine predictions with real-time assistance, we explore model-driven tutorials to help humans understand these patterns in a training phase. We consider both tutorials with guidelines from scientific papers, analogous to current practices of science communication, and automatically selected examples from training data with explanations. We use deceptive review detection as a testbed and conduct large-scale, randomized human-subject experiments to examine the effectiveness of such tutorials. We find that tutorials indeed improve human performance, with and without real-time assistance. In particular, although deep learning provides superior predictive performance than simple models, tutorials and explanations from simple models are more useful **to humans**. Our work suggests future directions for human-centered tutorials and explanations towards a synergy between humans and AI.

4.2 Introduction

Interpretable machine learning (ML) has attracted significant interest as ML models are used to support human decision making in societally critical domains such as justice systems and healthcare [Doshi-Velez and Kim, 2017, Guidotti et al., 2019, Lipton, 2016]. In these domains, full automation is often not desired and humans are the final decision makers for legal and ethical reasons. In fact, the Wisconsin Supreme Court ruled that "a COMPAS risk assessment should not be used to determine the severity of a sentence or whether an offender is incarcerated", but does not eliminate the use of ML models if "judges be made aware of the limitations of risk assessment tools" [Liptak, 2017, Supreme Court of Wisconsin, 2016]. Therefore, it is crucial to **enhance** human performance with the assistance of machine learning models, e.g., by explaining the recommended decisions.

However, recent human-subject studies tend to show limited effectiveness of explanations in improving human performance [Bussone et al., 2015, Horne et al., 2019, Lai and Tan, 2019a, Weerts et al., 2019]. For instance, Lai and Tan [2019a] show that explanations alone only slightly improve human performance in deceptive review detection; Weerts et al. [2019] similarly find that explanations do not improve human performance in predicting whether one's income exceeds 50,000 in the Adult dataset. These studies explain a machine prediction by revealing model internals, e.g., via attributing importance weights to features and then visualizing feature importance. We refer to such assistance as real-time assistance because they are provided as humans make individual decisions. To understand such limited effectiveness, we argue that it is useful to distinguish two **distinct** modes in which ML models are being used: **emulating** and **discovering**. In tasks such as object recognition [Deng et al., 2009, He et al., 2015], datasets are crowdsourced because humans are considered the gold standard, and ML models are designed to emulate human intelligence.[1] In contrast, in the discovering mode, datasets are usually collected from observing social processes, e.g., whether a person commits crime on bail for bail decisions [Kleinberg et al., 2017a] and what the writer intention is for deceptive review detection [Abouelenien et al., 2014, Ott et al., 2011]. ML models can thus often identify patterns that are unsalient, unknown, and even counterintuitive to humans, and may even outperform humans in **constrained** datasets [Kleinberg et al., 2017a, Ott et al., 2011, Tan et al., 2014]. Notably, many critical policy decisions such as bail decisions resemble the discovering mode more than the emulating mode because policy decisions are usually

[1] As a corollary, it is usually considered overfitting the dataset when machine learning models outperform humans in these tasks.



You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

You are wrong. The review is deceptive.
The AI is right. It predicts this review as deceptive because the red words (**ask, back, be, chicago, i, my, will**) are associated with deceptive reviews and the green words (**on, that, very**) are associated with genuine reviews. The darkness shows the degree of the association with the two categories.

Here are some rules to help you identify deceptive reviews.

- Deceptive reviews tend to focus on aspects that are external to the hotel being reviewed, e.g., husband, business, vacation.
- Deceptive reviews tend to contain more emotional terms; positive deceptive reviews are generally more positive and negative deceptive reviews are more negative than genuine reviews.
- Genuine reviews tend to include more sensorial and concrete language, in particular, genuine reviews are more specific about spatial configurations, e.g., small, bathroom, on, location.
- Deceptive reviews tend to contain more verbs, e.g., eat, sleep, stay.
- Deceptive reviews tend to contain more superlatives, e.g., cleanest, worst, best.
- Deceptive reviews tend to contain more pre-determiners, which are normally placed before an indefinite article + adjective + noun, e.g., what a lovely day!

Figure 4.1: Illustration of example-driven tutorials and guidelines shown to participants during the training phase: a) top 10 features of the review text are highlighted in green and red (**signed highlights**), where green words are associated with genuine reviews and red words are associated with deceptive reviews; b) participants are presented the actual label, the predicted label, and textual explanations for a review after choosing the label of the review in example-driven tutorials; c) a list of guidelines for identifying deceptive reviews extracted from scientific papers.

challenging (to humans) in nature [Kleinberg et al., 2015].

Studies on how explanations affect human performance tend to employ these challenging tasks for humans (the **discovering** mode for ML models) because humans need **little** assistance to perform tasks in the emulating mode (except for scalability). This observation highlights different roles of explanations in these two modes. In the emulating mode, explanations can help debug and identify biases and robustness issues in the models for future **automation**. In the discovering mode, if the patterns embedded in ML models can be elucidated for humans, they may enhance human knowledge and improve human decision making.[2] Moreover, it might help humans identify spurious patterns in ML models and account for potential mistakes to generalize beyond a **constrained** dataset.

To further illustrate the difficulty of interpreting explanations in the discovering mode, Figure 4.1(a) shows an example from a deceptive review detection task, where the goal is to distinguish deceptive reviews written by people who did not stay at the hotel from genuine ones. "Chicago" is highly associated with deceptive reviews because people are more likely to mention the city name instead of specific places when they imagine their experience. Such a pattern can be hard to

[2] It is worth noting that these two modes represent two ends of a continuum, e.g., emulating experts lead to discoveries for novices.

comprehend for humans, especially when the highlights are shown as real-time assistance without any other information. Instead of throwing people in at the deep end directly with real-time assistance, we propose a novel training phase that can help humans understand the nature of a task and the patterns embedded in a model. This training step is analogous to offline coaching and can be complementary to real-time assistance in explaining machine predictions. We consider two types of model-driven tutorials: 1) guidelines extracted from scientific papers [Li et al., 2014, Ott et al., 2013, 2011] (Figure 4.1(c)), which reflects the current practices of science communication; 2) example-driven tutorials where we select examples from the training data and present them along with explanations in the form of highlights (Figure 4.1(a)&(b)). We also develop a novel algorithm that incorporates spaced repetition to help humans understand the patterns in a machine learning model, and conduct an in-person user study to refine the design of our tutorials.

Our main contribution in this work is to design large-scale, randomized, pre-registered human-subject experiments to investigate whether tutorials provide useful training to humans, using the aforementioned deceptive review detection task as a testbed. We choose this task because 1) deceptive information including fake news is prevalent on the Internet [Allcott and Gentzkow, 2017, Grinberg et al., 2019, Lazer et al., 2018, Ott et al., 2012] and mechanical turkers can provide a reasonable proxy for humans facing this challenge compared to other tasks such as bail decisions and medical diagnosis that require domain expertise; 2) while humans struggle with detecting deception [Bond Jr and DePaulo, 2006], machine learning models are able to learn useful patterns in constrained settings (in particular, ML models achieve an accuracy of above 85% in our deceptive review detection task); 3) full automation might not be desired in this case because the government should not have the authority to automatically block information from individuals, and it is important to **enhance** human ability with a machine in the loop. Specifically, we focus on the following three research questions:

- **RQ1:** Do model-driven tutorials improve human performance without any real-time assistance?
- **RQ2:** How do varying levels of real-time assistance affect human performance **after** train-

ing?

- **RQ3:** How do model complexity and explanation methods affect human performance with/without training?

In all experiments, if training is provided, human subjects first go through a training phase with model-driven tutorials, and then enter the prediction phase to determine whether a review is deceptive or genuine. The prediction phase allows us to evaluate human performance after training.

Our first experiment aims to compare the effectiveness of different model-driven tutorials. Ideally, we would hope that these tutorials can help humans understand the patterns embedded in the ML models well enough that they can perform decently in the prediction phase without any real-time assistance. Our results show that human performance after tutorials are always better than without training, and the differences are statistically significant for two types of tutorials. However, the improvement is relatively limited: human performance reaches ∼60%, while the ML models are above 85%. Meanwhile, there is no statistically significant difference between human performance after any type of tutorial, which suggests that all model-driven tutorials are similarly effective.

One possible reason for the limited improvement of human performance in Experiment 1 is that the patterns might be too complicated for humans to apply in the prediction phase without any real-time assistance. Therefore, our second experiment is designed to understand the effect of tutorials with real-time assistance. Inspired by Lai and Tan [2019a], we develop a spectrum with varying levels of real-time assistance between full human agency and full automation (Figure 4.2). Our results demonstrate that real-time assistance can indeed significantly improve human performance to above 70%. However, compared to Lai and Tan [2019a], the best human performance is not significantly improved.[3] It suggests that given real-time assistance, tutorials are mainly useful in that humans can perform similarly well in the prediction phase with only signed highlights, thus retaining a higher level of human agency.

[3] We only discuss qualitative differences from Lai and Tan [2019a], as these are separate experiments subject to different randomization processes.

Finally, in order to understand how our results generalize to different kinds of models, we would like to examine the effect of model complexity and methods of deriving explanations. Our first two experiments use a linear SVM classifier because linear models are typically deemed interpretable, but deep learning models are increasingly prevalent because of their superior predictive power. While it is well recognized that deep learning models are more complex, it remains an open question how human performance changes with assistance from deep learning models (e.g., BERT) vs. simple models (e.g., linear SVM). Our results show that tutorials and explanations of simple models lead to better human performance than deep learning models, which highlights the tradeoff between **model complexity** and **interpretability**. We also show that for BERT, post-hoc signed explanations from LIME are more effective than built-in explanations derived from attention mechanisms. Moreover, tutorials are effective in improving human performance for both kinds of models compared to without training.

Overall, our results show that model-driven tutorials can somewhat improve human performance with and without real-time assistance, and humans also find these tutorials useful. However, the limited improvement also points to future directions of human-centered interpretable machine learning. We highlight two implications here and present further discussions in the Discussion section. First, it is important to explain beyond the surface patterns and facilitate humans in reasoning about why a feature is important. A strategy is to develop interactive explanations that allow humans to explore the patterns in both the training and the prediction phase. Second, it is useful to bridge the gap between training and generalization in developing tutorials because the model behavior and performance in training data might differ from that on unseen data. The ability to understand this difference is crucial for humans to calibrate trust and generalize beyond the constrained dataset.

4.3 Related work

We start by introducing recent methods for interpretable ML, and then discuss experimental studies on human interaction with explanations and predictions derived from ML models. We end by summarizing related work on deception detection.

4.3.1 Methods for interpretable machine learning

A battery of studies propose various algorithms to explain a machine prediction by uncovering model internals (also known as local explanations) [Guidotti et al., 2019]. Most relevant to our work is feature attribution that assigns an importance weight to each feature [Lei et al., 2016, Lundberg and Lee, 2017, Ribeiro et al., 2016, 2018a]. For instance, Ribeiro et al. [2016] propose LIME that fits a sparse linear model to approximate local machine predictions, and coefficients in this linear model are used as explanations. Lai et al. [2019] compare the built-in and post-hoc explanations methods in text classification and show that different methods lead to very different explanations, in particular, deep learning models lead to explanations with less consistency than simple models such as linear SVM. Other popular approaches include 1) example-based [Kim et al., 2016, 2014, Mothilal et al., 2019, Russell, 2019, Wachter et al., 2017], e.g., counterfactual explanations find alternative examples that would have obtained a different prediction, and 2) rule-based [Andrews et al., 1995, Guidotti et al., 2018a] that summarizes local rules (e.g., via decision trees). Notably, SP-LIME is an algorithm that selects examples to provide a global understanding of the model [Ribeiro et al., 2016], which aligns with our goal of generating tutorials. However, to the best of our knowledge, there have not been any human-subject experiments with such example-driven tutorials.

4.3.2 Human interaction with explanations and models

The importance of human-subject experiments is increasingly recognized in understanding the effectiveness of explanations because they are ultimately used by humans. In addition to studies mentioned in the introduction, researchers have investigated other desiderata of explanations [Binns et al., 2018, Cai et al., 2019b, Green and Chen, 2019a,b, Kunkel et al., 2019, Poursabzi-Sangdeh et al., 2018, Yin et al., 2019]. For instance, Binns et al. [2018] examine perception of justice given multiple styles of explanations and conclude that there is no best approach to explaining algorithmic decisions. Cai et al. [2019b] show that a user-centered design improves human perception of an image-search tool's usefulness, but does not improve human performance. Green and Chen [2019a] find that humans underperformed a risk assessment tool even when presented with its predictions, and exhibited behaviors that could exacerbate biases against minority groups. Yin et al. [2019] examine the effect of stated accuracy and observed accuracy on humans' trust in models, while Kunkel et al. [2019] study the effect of explanations on trust in recommender systems. This line of work on trust also relates to the literature on appropriate reliance with general automation [Lee and See, 2004, Lewandowsky et al., 2000]. Retaining human agency is particularly important in societally critical domains where consequences can be dire. Finally, Bansal et al. [2019a] provide feedback during decision making, which can be seen as a form of continuous learning. Our focus is to understand the effect of offline tutorials, which can be potentially combined with real-time assistance/feedback in practice.[4]

4.3.3 Deception detection

Deception is a ubiquitous phenomenon and has been studied in many disciplines [Vrij, 2000]. 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 [Krauss et al., 1976]. Computer scientists have been developing machine learning models to identify deception in texts, images, and videos [Abouelenien et al., 2014, Feng et al., 2012, Feng and Hirst, 2013, Jindal and Liu, 2008, Ott et al., 2011, Pérez-Rosas et al., 2015, Wu et al., 2010, Yoo and Gretzel, 2009]. An important challenge in studying deception is to obtain groundtruth labels because it is well recognized that humans struggle at detecting deception [Bond Jr and DePaulo, 2006]. Ott et al. [Ott et al., 2011] created the first sizable dataset in deception detection by employing workers on Amazon Mechanical Turk to write imagined experiences in hotels.

As people increasingly rely on information on the Internet (e.g., online reviews for making purchase decisions [Chevalier and Mayzlin, 2006, Trusov et al., 2009, Ye et al., 2011, Zhang et al.,

[4] Although feedback (e.g., true labels) on real decisions such as bail decisions can take a long time to observe.

2010]), deceptive information also becomes prevalent [Caspi and Gorsky, 2006, Ott et al., 2012, Shin et al., 2011]. The issue of misinformation and fake news has also attracted significant attention from both the public and the research community [Farsetta and Price, 2006, Grinberg et al., 2019, Lazer et al., 2018, Vosoughi et al., 2018, Zhang et al., 2018]. Our work employs the deceptive review detection task in Ott et al. [2013, 2011] to investigate the effectiveness of model-driven tutorials. While this task is a constrained case of deception and may differ from intentionally malicious deception, it represents an important issue that people face on a daily basis and can potentially benefit from assistance from ML models.

4.4 Method

In this section, we introduce the preliminaries for our prediction task, machine learning models, and explanation methods. We then develop tutorials to help humans understand the embedded patterns in the models in the training phase. Finally, we present types of real-time assistance in the prediction phase. A demo is available at `https://machineintheloop.com/deception`.

4.4.1 Dataset, models, and explanations

Dataset and prediction task. We employ the deceptive review detection task developed by Ott et al. [Ott et al., 2013, 2011], consisting of 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 who were asked to imagine their experience. We use 80% of the reviews as the training set and the remaining 20% as the test set. We evaluate human performance based on their accuracy on sampled reviews from the test set. The task for both humans and ML models is to determine whether a review is deceptive or genuine based on the text.

Models. We consider a linear SVM classifier with unigram bag-of-words as features, which represents a simple model, and BERT [Devlin et al., 2019], which represents a deep learning model with state-of-the-art performance in many NLP tasks. The hyperparameter for linear SVM was selected via 5-fold cross validation with the training set; BERT was fine-tuned on 70% of the reviews and the other 10% of the reviews in the training set were used as the development set for selecting hyperparameters. Table 4.1 shows their accuracy on the test set.

Model	Accuracy (%)
SVM	86.3
BERT	90.9

Table 4.1: Accuracy of machine learning models on the test set.

Methods of deriving explanations. We explain a machine prediction by highlighting the most important 10 words. For linear SVM, we use the absolute value of coefficients to determine feature importance, and the highlights are signed because coefficients are either positive or negative. For BERT, we consider two methods following Lai et al. [Lai et al., 2019]: 1) BERT attention based on the built-in mechanism of Transformer [Vaswani et al., 2017] (specifically, feature importance is calculated using the average attention values of 12 heads used by the first token at the final layer; these highlights are unsigned because attention values range between 0 and 1); 2) BERT LIME, where feature importance comes from LIME by fitting a sparse linear model to approximate local model predictions (these highlights are signed as they come from coefficients in a linear model).

4.4.2 Tutorial generation

Our main innovation in this work is to introduce a training phase with **model-driven** tutorials before humans interact with ML models. We consider the following two types of tutorials.

Guidelines. We follow the current practice of science communication and summarize findings in scientific papers Ott et al. [2013, 2011], Li et al. [2014] as a list of guidelines. These guidelines are observations derived from the ML model (see "Figure 4.1(c)") and paraphrased by us. A "Next" button is enabled after a 30-second timer.

Example-driven tutorials. Inspired by Ribeiro et al. [Ribeiro et al., 2016], another way to give humans a global sense of a model is to present a sequence of examples along with predicted labels and explanations of predictions. For each example in our tutorial, informed by our in-person user study, we first ask participants to determine the label of the example, and then reveal the actual

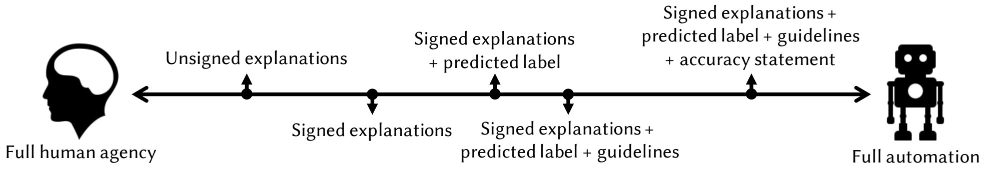


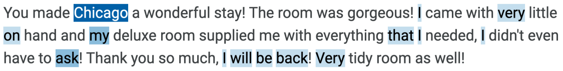
Figure 4.2: An adapted spectrum between full human agency and full automation from Lai and Tan [32]. The order approximates our intuition, but the distance does not reflect linear changes in machine influence. In particular, guidelines do not necessarily increase the influence of predicted labels.

label and the predicted label along with explanations in the form of highlights. The algorithm selects 10 examples that are representative of the patterns that the ML model identifies from the training set.[5] There could be genuine insights as well as spurious patterns. Ideally, these examples allow participants to understand the problem at hand and then apply the patterns, including correcting spurious ones, in the prediction phase. Figure 4.1(a)&(b) presents an example review after the label is chosen and the predicted label and its explanations are shown. A "Continue" button is enabled after a 10-second timer. See the supplementary material for screenshots.

We consider the following algorithms for example selection:

- **Random.** 10 random examples are chosen.
- **SP-LIME.** Ribeiro et al. [Ribeiro et al., 2016] propose SP-LIME to select examples with features that provide great coverage in the training set. To do that, the global importance of each feature is defined as $I_j = \sqrt{\sum_{i=1}^{n} W_{ij}}$, where W_{ij} is the importance of feature j in the i-th instance. Since we only highlight the top 10 features, $W_{ij} = 0$ for any other features. Then, 10 examples are selected to maximize the following objective function: $\text{argmax}_{S,|S|\leq B} \sum_{j=1}^{d} \mathbb{1}(\exists i \in S : W_{ij} > 0) I_j$, where $B = 10$ and d represents the dimension of features. This objective function presents a weighted coverage problem over all features, and is thus submodular. A greedy algorithm provides a solution with a constant-factor approximation guarantee of $1 - 1/e$ to the optimum [Krause and Golovin, 2014].

[5] We chose 10 so that an experiment session finishes within a reasonable amount of time (30 minutes), and all examples happened to be classified correctly by the model (since machine performance is even better on the training set).



You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

Figure 4.3: Unsigned highlights for the example review in Figure 4.1(a).

- **Spaced repetition (SR).** We propose this algorithm to leverage insights from the education literature regarding the effectiveness of spaced repetition (e.g., on long-term retention) [Kang, 2016, Tabibian et al., 2019]. Specifically, we develop the following novel objective function so that users can be exposed to important features repeatedly: $\text{argmax}_{S,|S|\leq B} \sum_{j=1}^{d} U(\{W_{kj}\}_{1\leq k\leq |S|}) I_j$, where $U(\{w_{kj}\}_{1\leq k\leq |S|}) = \mathbb{1}(\max(\{k, W_{kj} > 0\}) - \min(\{k, W_{kj} > 0\}) \geq 3)$. The key difference from SP-LIME is that the weight of a feature is included only if it is repeated in two examples with a gap of at least three.

Finally, we consider the combination of guidelines and examples selected with spaced repetition by first showing the guidelines for 15 seconds, 10 examples selected with spaced repetition, and the guidelines again for 15 seconds.

4.4.3 Real-time assistance

In addition to tutorials in the training phase, we introduce varying levels of real-time assistance in the prediction phase. Inspired by Lai and Tan [Lai and Tan, 2019a], we design six levels of real-time assistance, as illustrated in Figure 4.2.

- **No machine assistance.** Participants are not exposed to any real-time machine assistance.
- **Unsigned highlights.** Top 10 features are highlighted in shades of blue. The darker the color, the more important the feature. See Figure 4.3 for an example.
- **Signed highlights.** Top 10 features are highlighted in shades of green and red: green words are associated with genuine reviews, while red words are associated with deceptive reviews. The darker the color, the more important the feature. See Figure 4.1(a) for an example.[6]

[6] We use an attention check question to make sure that participants can distinguish red from green.

- **Signed highlights + predicted label.** In addition to signed highlights, we display the predicted label.
- **Signed highlights + predicted label + guidelines.** We additionally provide the option of revealing guidelines.
- **Signed highlights + predicted label + guidelines + accuracy statement**. We further add an accuracy statement, "It has an accuracy of approximately 86%", emphasizing the strong performance of the ML model.

These six levels gradually increase the amount of information and prime users towards machine predictions. Ideally, we hope to retain human agency as much as possible while achieving strong human performance.

4.5 In Person User Study

To obtain a qualitative understanding of human interaction with model-driven tutorials, we conduct an in-person semi-structured user study. This user study allows us to gather in-depth insights on how humans learn and apply our tutorials through interviews, as well as feedback on the interface before conducting large-scale, randomized experiments.

4.5.1 Experimental design

We employ a concurrent think-aloud process with participants [Nielsen et al., 2002]. Each participant went through a tutorial and determined the label of 20 reviews from the test set. They were told to verbalize the reason before deciding on the label both in the training and the prediction phase with the following syntax: I think the review is **predicted label** because **reason**. After the prediction phase, we conducted an interview to gather general feedback on tutorials. We manually transcribed the audio recordings after an initial pass with the Google Cloud API.

A total of 16 participants were recruited from mailing lists in our department: 3 were female and 13 were male, ranging between age 20 and 35. All participants were engineering graduate students and most of them studied computer science. Participants were invited to the lab where the study occurred. Either a personal or a provided laptop was used. Participants were compensated between \$15 and \$20 for \$10 every 30 minutes. Four types of tutorials (guidelines, examples selected with SP-LIME, examples selected with SR, guidelines + examples selected with SR) were randomly assigned to participants and each tutorial type had a sample size of 4. Thematic analysis was undertaken to identify common themes in participants' think-aloud processes. Thematic codes were collectively coded by the first two authors.

4.5.2 Results

We summarize the key themes into the following three parts.

Tutorial training and application. 8 out of 8 participants with access to guidelines remembered a couple of "rules" and applied them in the prediction phase. P13 said (the number is randomly assigned), "I believe it is deceptive based on rule No. one and No. three, if I remembered them correctly, it just describes its experience, and does not have a lot of details". 7 out of 12 participants exposed to selected examples adopted pure memorization or pattern-matching during the prediction phase. Participants remembered key deceptive words such as "chicago" to help them decide the review label: P2 said, "My husband is deceptive, I is deceptive, Chicago is deceptive". Some participants were even able to generate similar theories to our guidelines without exposure to it. P14 commented, "The review didn't have anything specific to offer" before deciding that the respective review was deceptive. However, reasoning about the patterns is generally challenging. Quoting from P2, this is mainly because they "can't seem to find a rhyme or reason for those words being genuine or deceptive".

Participants also created theories such as length of review when predicting. P8 remarked, "no one would take that much time to write a review so it won't cross more than 5 lines".

Improvements on tutorials. Participants thought that the guidelines should be available during the prediction phase to better assist them. 4 out of 4 participants felt that they were unable to remember as there were too many guidelines to be memorized. P11 felt that "the tutorial is helpful but it's just hard not being able to reference it" and P9 said that he could "keep checking if it is on the top right corner".

12 out of 12 participants exposed to selected examples expressed confusion about why the features were highlighted as deceptive or genuine but made up their own reasonings for ease of memory. They felt that they would have learned better if some form of explanations were given to justify each feature's indication. P16 remarked that "it would be nice if it can let me know why exactly it thinks the word is deceptive" and P10 commented that on top of the current explanations in selected examples, "more detailed explanation would be helpful" to help understand.

Improvements on the interface. We found that some participants thought that deceptive reviews are written by an AI without reading the instructions, which is false. We thus introduced three additional questions for our large-scale experiments: 1) how are deceptive reviews defined in this study?; 2) identify the color that highlights a word; 3) reiterate the training process and ask user to answer true or false to ensure that the participants know which treatment they are exposed to. We also changed the flow of showing explanations in the training phase: users need to first determine the label for a review before the explanations, the actual label, and the predicted label are shown for at least 10 seconds. Refer to the video and detailed feedback in the supplementary material.

4.6 Experiment 1: Do Tutorials Improve Human Performance without any Real-time Assistance?

As introduced in the Methods section, we hope to build tutorials that can help humans understand the embedded patterns in ML models, which can sometimes be unsalient, unknown, or even counterintuitive to humans. Ideally, humans reflect on these patterns from our tutorials and can apply them in their decision making without any further real-time assistance from ML models. Therefore, we start with RQ1: do tutorials improve human performance without any real-time assistance?

4.6.1 Experimental treatments & hypotheses

We consider the following treatments to examine the effectiveness of various tutorials proposed in the Methods section: 1) guidelines; 2) random examples; 3) examples selected with SP-LIME; 4) examples selected with SR; 5) guidelines + examples selected with SR. All the tutorials and explanations in the tutorials are based on the linear SVM classifier in the Methods section. After a training phase, participants will then decide whether a review is deceptive or genuine based on the text. Note that ML models also rely exclusively on textual information. In addition to these tutorials, we include a control setup where no training was provided to humans.

We hypothesize that 1) training is important for humans to understand this task, since it has been shown that humans struggle with deception detection [Bond Jr and DePaulo, 2006]; 2) it would be easier for participants to understand the patterns embedded in the ML model situated with examples; 3) carefully chosen examples provide more comprehensive coverage and can better familiarize participants with the patterns [Kang, 2016, Tabibian et al., 2019]; 4) guidelines and examples have complementary effects in the training phase. To summarize, our hypotheses in Experiment 1 are as follows:

- (**H1a**) Any tutorial treatment leads to better human performance than the control setup.
- (**H1b**) Examples (including **random examples, examples selected with SP-LIME and SR**) lead to better human performance than **guidelines**.
- (**H1c**) **Selected examples (with SP-LIME or SR)** lead to better human performance than **random examples**.
- (**H1d**) **Examples selected with spaced repetition** lead to better human performance those selected with **SP-LIME**.
- (**H1e**) **Guidelines + examples selected with SR** lead to the best performance.

These five hypotheses were pre-registered on AsPredicted.[7]

[7] The anonymized pre-registration document is available at `https://aspredicted.org/blind.php?x=v8f7zh`. A

4.6.2 Experimental design

To evaluate human performance under different experimental setups, participants were recruited via Amazon Mechanical Turk and filtered to include only individuals residing in the United States, with at least 50 Human Intelligence Tasks (HITs) completed and 99% of HITs approved. Each participant is randomly assigned to one of the six conditions (five types of tutorials + control). We did not allow any repeated participation. We adopted this between-subject design because exposure to any type of tutorial cannot be undone.

In our experiment, each participant finishes the following steps sequentially: 1) reading an explanation of the task and a consent form; 2) answering a few attention-check questions depending on the experimental condition assigned; 3) undergoing a set of tutorials if applicable (training phase); 4) predicting the labels of 20 randomly selected reviews in the test set (prediction phase); 5) completing an exit survey. Participants who failed the attention-check questions are automatically disqualified from the study. Based on feedback from our in-person user study, for each example in the tutorials, a participant first chooses genuine or deceptive without any assistance, and then the answer is revealed and the predicted label and explanations are shown (Figure 4.1(a)&(b)). In the exit survey, participants were asked to report basic demographic information, if the tutorial was helpful (yes or no), and feedback in free responses.[8]

Each participant was compensated $2.50 and an additional $0.05 bonus for each correctly labeled test review. 80 subjects were recruited for each condition so that each review in the test set was labeled five times. In total 480 subjects completed Experiment 1. They were balanced on gender (224 females, 251 males, and 5 preferred not to answer). Refer to the supplementary material for additional information about experiments (e.g., education background, time taken).

To quantify human performance, we measure it by the percentage of correctly labeled instances by humans. In other words, the prediction phase provides an estimate of human accuracy

minor inconsistency is that we did not experiment with "guidelines + examples selected from SP-LIME" as we hypothesized that SR is better.

[8] Feedback from Turkers generally confirmed findings in the in-person user study. See the supplementary material for an analysis.

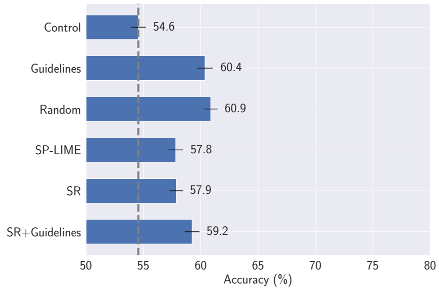


Figure 4.4: Human accuracy without any real-time assistance after different types of tutorials. Error bars represent standard errors. Human accuracy after tutorials is always better than that without any training. Differences are statistically significant between random and control, and guidelines and control based on post-hoc Tukey's HSD test.

through 20 samples. In addition to this objective metric, we also report subject perception of tutorial usefulness reported in the exit surveys.

4.6.3 Results

We first present human accuracy in the prediction phase, an objective measurement of tutorial effectiveness. Our results suggest that tutorials are useful to some extent: all tutorials lead to better human performance (∼60%) than the control setup without any training (Figure 4.4). To formally compare the treatments, we conduct an one-way ANOVA and find a statistically significant effect ($\eta^2 = 0.033$; $p = 7.70e-3$). We further use post-hoc Tukey's HSD test to identify pairs of experimental conditions in which human performance exhibits significant differences. The only statistically significant differences are **guidelines** vs. **control** ($p = 1.75e-2$) and **random** vs. **control** ($p = 7.0e-3$) (the difference between **guidelines+SR** and **control** is borderline significant with $p = 0.10$).

In other words, our experiment results provide partial support to **H1a**, and reject all other hypotheses in Experiment 1. These results suggest that although tutorials provide somewhat useful training, different tutorials are similarly effective. The limited improvement in human performance across all tutorials indicates that the utility of tutorials is small. We hypothesized that it is too challenging for humans to remember all the patterns after a short tutorial (supported by feedback

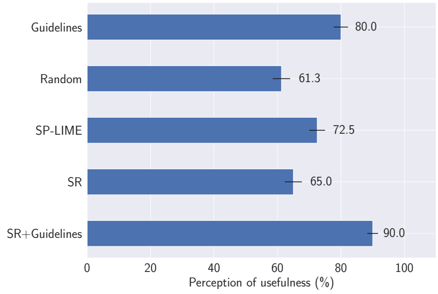


Figure 4.5: Subjective perception of tutorial usefulness. Error bars represent standard errors. Differences are statistically different in the following pairs based on post-hoc Tukey's HSD test: guidelines vs. random, random vs. SR+guidelines, and SR vs. SR+guidelines.

from in-person user study), which motivated Experiment 2 to understand the effect of real-time assistance in conjunction with tutorials. Another contributing factor certainly lies in the design of tutorials, which we will further discuss in the Discussion section.

As for subjective perception of tutorial usefulness, we find that participants generally find our tutorials useful: 73.8% of 400 participants reported that the tutorial was useful (excluding 80 participants in the control setup). Figure 4.5 shows the results by types of tutorials. Among different treatments, participants in **guidelines** and **guidelines + examples selected with SR** find the tutorials most useful, as high as 90% in **guidelines + examples selected with SR**. Formally, post-hoc Tukey's HSD test shows that the differences between the following pairs are statistically different: **guidelines** vs. **random** ($p = 0.048$), **random** vs. **SR+guidelines** ($p < 0.001$), and **SR** vs. **SR+guidelines** ($p = 0.003$). The difference between **SP-LIME** and **SR+guidelines** is borderline significant with $p = 0.078$. These results suggest that tutorials provide strong positive effects in humans' subjective perception.

4.7 Experiment 2: Human Performance with Varying Real-time Assistance after Tutorials

Our second experiment is concerned with human performance with varying levels of real-time assistance after going through the training phase. While Experiment 1 suggests that tutorials provide somewhat useful training, the improvement is limited without any real-time assistance. We hypothesize that human performance could be further improved by introducing real-time assistance. We adapt a spectrum with varying levels of real-time assistance from Lai and Tan [Lai and Tan, 2019a] (Figure 4.2). Moving along the spectrum, the influence of the machine generally becomes greater on the human as more information from the model is presented. For instance, a statement of strong machine performance is likely to bias humans towards machine predictions. Lai and Tan [Lai and Tan, 2019a] find that there exists a tradeoff between human performance and human agency, i.e., as the real-time assistance gives stronger priming along the spectrum, human performance improves and human agency decreases. Explanations such as highlighting important words can moderate this tradeoff **when predicated labels are given**. It remains an open question how this tradeoff unfolds after training.

4.7.1 Experimental treatments & hypotheses

All conditions in Experiment 2 used the **guidelines + selected examples with spaced repetition** tutorial in the training phase because all tutorials are similarly effective and our participants find this one most useful in subjective perception. To examine how humans perform under different levels of real-time assistance from machine learning models, we consider the spectrum in Figure 4.2, inspired by Lai and Tan [Lai and Tan, 2019a].

We hypothesize that 1) real-time assistance results in improved human performance, since it has been shown that highlights and predicted labels improve human performance [Lai and Tan, 2019a]; 2) signed highlights result in better human performance compared to unsigned highlights because signed highlights reveal information about directionality; 3) predicted labels result in better human performance compared to highlights alone; 4) guidelines and signed highlights might moderate the tradeoff between human performance and human agency while achieving the same effect as when an accuracy statement is shown. To summarize, our hypotheses are as follows:

- (**H2a**) Real-time assistance leads to better human performance than no assistance.
- (**H2b**) **Signed highlights** lead to better human performance than **unsigned highlights**.
- (**H2c**) **Predicted label** leads to better human performance than highlights alone.

- (**H2d**) **Signed highlights + predicted label + guidelines + accuracy statement** leads to the best performance.
- (**H2e**) **Signed highlights + predicted label + guidelines** and **Signed highlights + predicted label** perform as well as **Signed highlights + predicted label + guidelines + accuracy statement**.

These five hypotheses were pre-registered on AsPredicted.[9]

4.7.2 Experimental design

We adopted the same experimental design as stated in Experiment 1 except that real-assistance is provided in the prediction phase when applicable. In total 480 subjects completed the experiment (80 participants in each type of real-time assistance). They were balanced on gender (238 females, 237 males, and 5 preferred not to answer). Refer to the supplementary material for additional information about experiments (e.g., education background, time taken).

Human performance is measured by the percentage of correctly predicted instances by humans, which provides an objective measure of human performance with real-time assistance. We also consider the percentage of humans whose performance exceeds machine performance for the corresponding 20 reviews in the prediction phase.[10]

4.7.3 Results

We first present human accuracy in the prediction phase. Our results suggest that real-time assistance is indeed effective: all the levels of real-time assistance except unsigned highlights lead to better human performance than the setup without machine assistance in Figure 4.6. To formally compare the treatments, we conduct an one-way ANOVA and find a statistically significant effect ($\eta^2 = 0.23$; $p = 5.15e - 25$). We further use post-hoc Tukey's HSD test to identify pairs of experimental conditions in which human performance exhibits significant differences. With the exception of **no assistance** vs. **unsigned highlights** ($p = 0.67$), differences in remaining setups

[9] The anonymized pre-registration document is available at `http://aspredicted.org/blind.php?x=fi8kz8`.

[10] We also pre-registered trust as a measure and present the results in the supplementary material for space reasons.

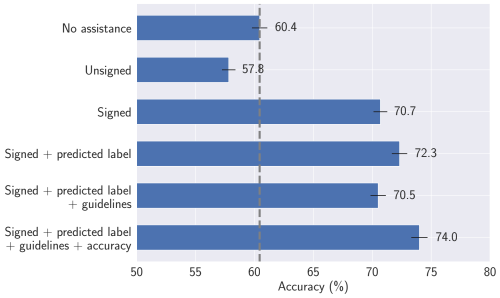


Figure 4.6: Human accuracy with varying levels of real-time assistance after training. Error bars represent standard errors. With the exception of **unsigned highlights**, human accuracy with real-time assistance is better than without real-time assistance. Differences between **no assistance** and any assistance with signed highlights are statistically significant based on post-hoc Tukey's HSD test.

compared to **no assistance** are all statistically significant ($p < 0.001$). Moreover, the difference between **unsigned highlights** and **signed highlights** is significant ($p < 0.001$), demonstrating the effectiveness of signed highlights. Finally, the difference between **signed highlights** and any other real-time assistance with stronger priming (**signed highlights + predicted labels, signed highlights + predicted labels + guidelines, signed highlights + predicted labels + guidelines + accuracy statement**) is not significant.

In summary, our experimental results support **H2a** with the exception of **unsigned highlights**, **H2b**, **H2e**, and reject **H2c** and **H2d** in Experiment 2 (note that **signed highlights + predicted label + guidelines + accuracy statement** indeed leads to the best performance but the difference with other methods is not always statistically significant). These results suggest that **signed highlights** provide sufficient information for improving human performance, and we do not gain much from presenting additional information with stronger priming. While there is significant improvement in human performance with real-time assistance (from ∼60% to ∼70%), the improvement is still limited compared to the machine performance, which is above 85%. This improvement is similar to results reported in Lai and Tan [Lai and Tan, 2019a], which did not use any tutorials other than minimal examples to introduce the task. These observations taken together suggest that the utility of our tutorials mainly lies in that humans can perform well with only signed highlights, a type of real-time assistance with relatively weak priming.

Another ambitious measurement is how frequent humans outperform the ML model. It was rare in Experiment 1 (2 of 480, 0.4%). With effective real-time assistance (i.e., signed highlights included), we find that 26 of 320 (8.1%, 20 times the percentage in Experiment 1) of our participants are able to outperform the ML model. The difference between 8.1% and 0.4% is statistically significant using chi-squared tests ($p < 0.001$). This observation suggests that with the help of tutorial and real-time assistance, there exists hope for a synergy of **humans and AI** outperforming AI alone. We hypothesize that facilitating hypothesis generation is important and present detailed discussions in the Discussion section.

4.8 Experiment 3: The Effect of Model Complexity and Methods of Deriving Explanations

Our experiments so far are based on explanations (coefficients) from a linear SVM classifier. Meanwhile, deep learning models are being widely adopted because of their superior predictive power. However, it is also increasingly recognized that they might be more complex and harder to interpret for humans. Our final experiment investigates how model complexity and methods of deriving explanations relate to human performance and effect of training.

4.8.1 Experimental treatments & hypotheses

Participants are exposed to two different treatments: presence of training and methods of deriving highlights. Where training is present, we use the **selected examples with spaced repetition** tutorial in this experiment. Note that example selection depends on the model and the explanation method (i.e., which features are considered important). In comparison, guidelines are static and are extracted from papers based on linear SVM, so they are not appropriate here. Based on results from Experiment 2, we adopted **signed highlights** as our real-time assistance in the prediction phase when applicable.[11] To summarize, we consider the following setups to examine how humans perform when exposed to training and different methods of deriving explanations: 1)

[11] Since BERT performs better than linear SVM, only showing signed highlights also avoids the potential effect of predicted labels.

no training + SVM coefficients; 2) no training + BERT attention; 3) no training + BERT LIME; 4) training + SVM coefficients; 5) training + BERT attention; 6) training + BERT LIME.

Note that the deep learning model (BERT) leads to both different real-time assistance and examples selected for tutorials because they consider different words important. We can only use unsigned highlights for BERT attention because attention values range between 0 and 1. Refer to the Methods section for details of BERT attention and BERT LIME.

We hypothesize that 1) SVM results in better performance compared to BERT, since it is a common assumption that linear models are more interpretable and it has been shown that SVM results in important features with lower entropy [Lai et al., 2019]; 2) BERT LIME results in better performance compared to BERT attention because signed highlights can reveal more information about the underlying decision; 3) participants would perform better with training than without training. To summarize, our hypotheses in Experiment 3 are as follows:

- (**H3a**) The simple model (**SVM**) leads to better human performance than the deep learning model (**BERT**).
- (**H3b**) **BERT LIME** leads to better human performance than **BERT attention**.
- (**H3c**) **Training** leads to better human performance than **without training**.

These three hypotheses were pre-registered on AsPredicted.[12]

4.8.2 Experimental design

We adopted the same experimental design as in Experiment 1. In total 480 subjects completed the experiment (80 participants in each experimental setup). They were balanced on gender (239 females, 240 males, and 1 preferred not to answer). Refer to the supplementary material for additional information about experiments (e.g., education background, time taken).

To quantify human performance, we measure it by the percentage of correctly predicted instances by humans. In addition to this objective metric, we also report subject perception of tutorial usefulness reported in the exit surveys (note that this is only applicable for the experimental

[12] The anonymized pre-registration document is available at `http://aspredicted.org/blind.php?x=vy794a`.

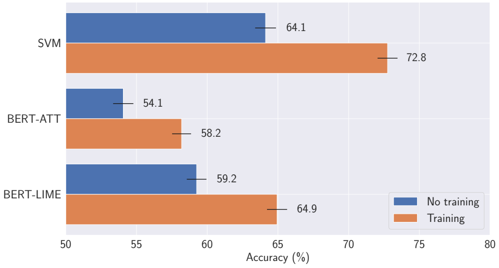


Figure 4.7: Human accuracy grouped by methods of deriving explanations. Error bars represent standard errors. SVM explanations lead to better human performance than explanations based on BERT. Training (second bar from the top in each method) also consistently improves human performance for all explanation methods.

setups with training).

4.8.3 Results

We first present human accuracy in the prediction phase. Our results suggest that methods of deriving explanations make a significant difference (Figure 4.7): 1) human performance is consistently better when important words derived from the linear SVM are highlighted as compared to deep models; 2) BERT LIME leads to better human performance than BERT attention. It also reinforces the point that training leads to better human performance as compared to no training: humans achieve better performance with training with any kind of explanation methods. To formally compare the treatments, we conduct a two-way ANOVA and find a statistically significant effect of tutorials ($\eta^2 = 0.049$; $p = 1.50e-7$) and methods of deriving explanations ($\eta^2 = 0.13$; $p = 4.66e-16$). Differences among all pairs of treatments are also statistically significant using post-hoc Tukey's HSD test ($p < 0.001$).[13]

In other words, our experiment results provide support to all hypotheses in Experiment 3. These results suggest that tutorials are indeed useful in improving human performance, albeit improvement is still limited in the sense that human performance is ∼70% after training with real-time assistance, echoing results in Experiment 2. It also suggests that simple models are preferred to deep learning models when serving as explanations to support human decision making. Between

[13] It is reduced to t-test for the training/no training treatment since the degree of freedom is 1.

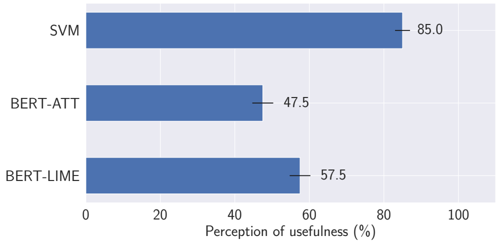


Figure 4.8: Human perception of tutorial usefulness. Error bars represent standard errors. Participants are more likely to find SVM tutorials useful (differences between (SVM, BERT attention) and (SVM, BERT LIME) are statistically significant using post-hoc Tukey's HSD test).

explanations derived from post-hoc and built-in methods from BERT, attention provides the least value for humans, again demonstrating the importance of signed highlights.

The effectiveness of training for simple models is further validated by subjective perception of tutorial usefulness. Figure 4.8 shows that participants are much more likely to find the tutorials derived from SVM explanations useful: 85% of our participants find it useful. The differences between the following pairs are statistically different using post-hoc Tukey's HSD test: **SVM** vs. **BERT attention** ($p < 0.001$) and **SVM** vs. **BERT LIME** ($p < 0.001$). Interestingly, with real-time assistance, humans also find the tutorials more useful compared to the same tutorial in Figure 4.5. These results underscore our findings in Experiment 3 that simple models provide more interpretable tutorials and explanations than deep models.

4.9 Conclusion

In this paper, we conduct the first large-scale, randomized, pre-registered human-subject experiments to investigate whether model-driven tutorials can help humans understand the patterns embedded in ML models and improve human performance. We find that tutorials can indeed improve human performance to some extent, with and without real-time assistance, and humans also find them useful. Moreover, real-time assistance is crucial for further improving human performance in such challenging tasks. Finally, we show that simple models like linear SVM generate more useful tutorials and explanations for humans than complex deep learning models.

Towards human-centered tutorials. Both quantitative results from our randomized experiments and qualitative feedback from in-person user study demonstrate that humans can benefit from model-driven tutorials, which suggests that developing model-driven tutorials is a promising direction for future work in human-centered interpretable machine learning.

However, the improvement in human performance remains limited compared to machine performance in the deceptive review detection task. In order to further advance the synergy between humans and AI, we need to develop human-centered tutorials. Many participants commented that they could not understand why certain words were deceptive or genuine (an example reason could be that imaginative writing does not cover specific details). These results highlight the importance of **facilitating hypothesis generation** in the tutorials. It is insufficient to highlight important features via feature attribution methods, and these tutorials need to also explain why some features are useful. While it is challenging to develop automatic methods that can propose theories about particular features, we might prompt humans to propose theories and evaluate them through the ML model.

Another reason that tutorials had limited improvement in human performance is that the tutorials failed to establish proper trust in machine predictions. It is important to highlight both strengths and caveats of ML models in the tutorials, echoing recent work on understanding trust [Kunkel et al., 2019, Yin et al., 2019]. A challenge lies in how to bridge the gap between training and generalization in tutorials, i.e., model behavior and performance in the tutorials might differ from that in unseen data.

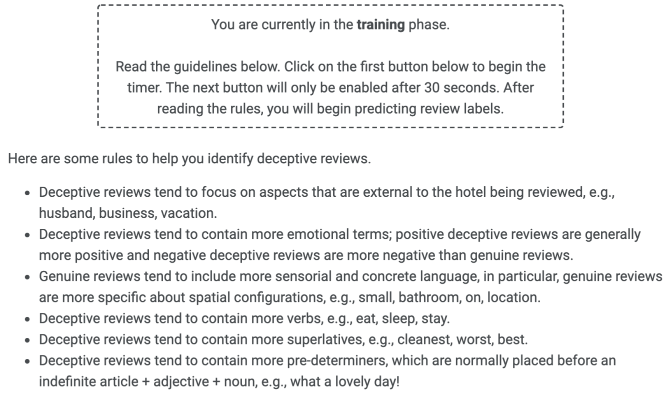
Beyond static explanations. Another important direction is to design interactive explanations beyond static explanations such as simply highlighting important words. Interactive explanations allow humans to experiment with their hypothesis about feature importance. One strategy is to enable humans to inquire about the importance of any word in a review. An alternative strategy is to assess model predictions of counterfactual examples. For instance, humans can remove or add words/sentences in a review, which can help humans understand model behavior in new scenarios.

Choice of tasks. We would like to highlight the importance of task choice in understanding human-AI interaction. Deception detection might simply be too challenging a task for humans, and a short tutorial is insufficient to help humans understand the patterns embedded in ML models. There may also exist significant variation between understanding text and interpreting images, because the former depends on culture and life experience, while the latter relies on basic visual cognition.

We believe that it is important to study human-AI interaction in challenging tasks where human agency is important because the nature of explanations in decision making is distinct from that in debugging. While machines excel at identifying patterns from existing datasets, humans might be able to complement ML models by deriving theories and appropriately correcting machine predictions in unseen data, e.g., spotting mistakes when machines apply patterns ("chicago" becomes a specific comparison point for reviews about a hotel in New York City). So there exists hope for further advancing human performance in these challenging tasks.

Limitation of our samples. Our study is limited by our samples of human subjects. The in-person user study was conducted with university students who tend to have a computer science education, and large-scale, randomized, pre-registered experiments were conducted with Mechanical Turkers from the United States. While our samples are likely to face the challenges of deception on the Internet and would benefit from enhancements in deception detection, they may not be representative of the general population. The effectiveness of model-driven tutorials can also potentially depend on properties of the sample population. In general, we did not find any consistent differences between demographic groups based on age, gender, education background, and review experience (see the supplementary material). It is certainly possible that other demographic information could affect the effectiveness of tutorials. We leave that for future studies.

It is important to point out that our setup employs a random split to obtain training and testing data, which is a standard assumption in supervised machine learning. While humans can ideally improve generalization in this case, humans might be more likely to correct generalization errors in machine learning models when the testing distribution differs from training. In that case, understanding the embedded patterns, especially spotting spurious ones, can help humans generalize these data-driven insights.



You are currently in the **training** phase.

Read the guidelines below. Click on the first button below to begin the timer. The next button will only be enabled after 30 seconds. After reading the rules, you will begin predicting review labels.

Here are some rules to help you identify deceptive reviews.

- Deceptive reviews tend to focus on aspects that are external to the hotel being reviewed, e.g., husband, business, vacation.
- Deceptive reviews tend to contain more emotional terms; positive deceptive reviews are generally more positive and negative deceptive reviews are more negative than genuine reviews.
- Genuine reviews tend to include more sensorial and concrete language, in particular, genuine reviews are more specific about spatial configurations, e.g., small, bathroom, on, location.
- Deceptive reviews tend to contain more verbs, e.g., eat, sleep, stay.
- Deceptive reviews tend to contain more superlatives, e.g., cleanest, worst, best.
- Deceptive reviews tend to contain more pre-determiners, which are normally placed before an indefinite article + adjective + noun, e.g., what a lovely day!

Figure 4.9: Experiment 1 tutorial: guidelines.

In summary, our work highlights the promise of (automatically) building model-driven tutorials to help humans understand the patterns embedded in ML models, especially in challenging tasks. We hope to encourage future work on human-centered tutorials and explanations beyond static real-time assistance towards a synergy between humans and AI.

4.10 Appendix

4.10.1 Experiment Interfaces

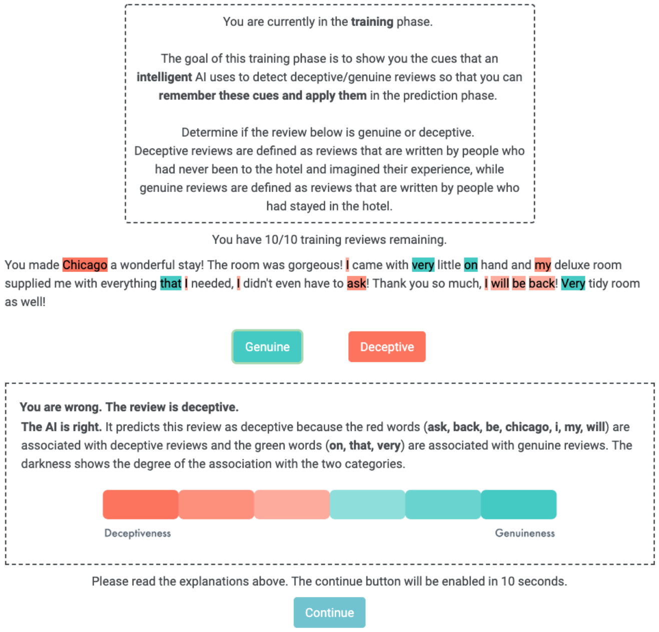
Figure 4.9 - Figure 4.11 shows tutorial interfaces for Experiment 1.

Figure 4.12 - Figure 4.17 shows the prediction phase interfaces for experiment 2.

Figure 4.18 - Figure 4.20 shows examples in different methods deriving explanations for experiment 3.

4.10.1.1 Experiment Details

Among our participants in Experiment 1, 69 were between 18 and 25, 265 were between 26 and 40, 121 were between 41 and 60, 22 were 61 and above, and 3 preferred not to answer. They had a range of education backgrounds, comprising some high school (3), high school graduate (54), some college credit (124), trade/technical/vocational training (42), Bachelor's degree and above (253), and 4 prefered not to answer.



You are currently in the **training** phase.

The goal of this training phase is to show you the cues that an **intelligent** AI uses to detect deceptive/genuine reviews so that you can **remember these cues and apply them** in the prediction phase.

Determine if the review below is genuine or deceptive.
Deceptive reviews are defined as reviews that are written by people who had never been to the hotel and imagined their experience, while genuine reviews are defined as reviews that are written by people who had stayed in the hotel.

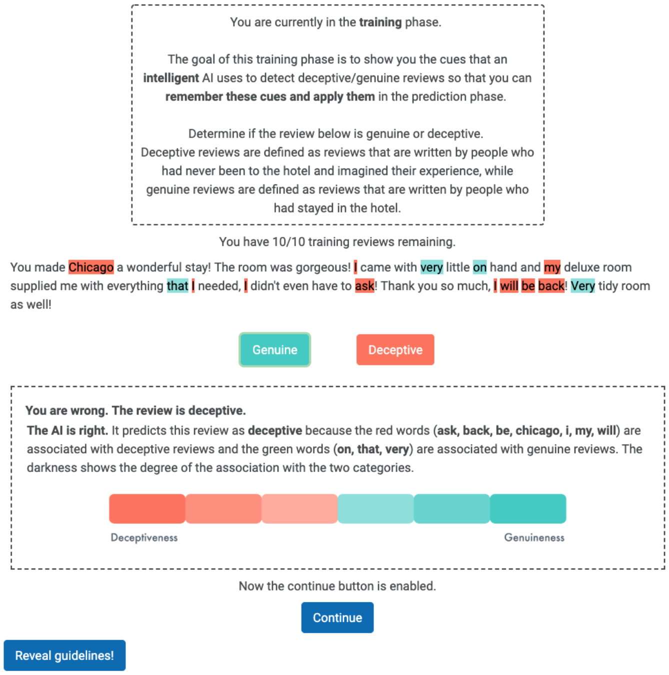
You have 10/10 training reviews remaining.

You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

You are wrong. The review is deceptive.
The AI is right. It predicts this review as deceptive because the red words (**ask, back, be, chicago, i, my, will**) are associated with deceptive reviews and the green words (**on, that, very**) are associated with genuine reviews. The darkness shows the degree of the association with the two categories.

Please read the explanations above. The continue button will be enabled in 10 seconds.

Figure 4.10: Experiment 1 tutorial: selected examples. Selected examples of **random**, **SP-LIME**, and **SR** are captured in video submission.



You are currently in the **training** phase.

The goal of this training phase is to show you the cues that an **intelligent** AI uses to detect deceptive/genuine reviews so that you can **remember these cues and apply them** in the prediction phase.

Determine if the review below is genuine or deceptive.
Deceptive reviews are defined as reviews that are written by people who had never been to the hotel and imagined their experience, while genuine reviews are defined as reviews that are written by people who had stayed in the hotel.

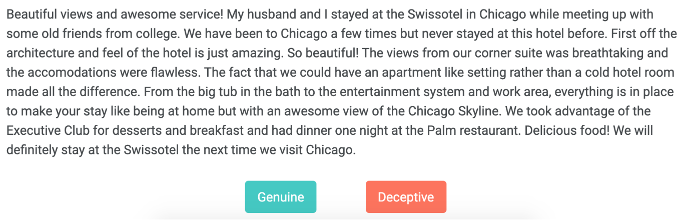
You have 10/10 training reviews remaining.

You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

You are wrong. The review is deceptive.
The AI is right. It predicts this review as **deceptive** because the red words (**ask, back, be, chicago, i, my, will**) are associated with deceptive reviews and the green words (**on, that, very**) are associated with genuine reviews. The darkness shows the degree of the association with the two categories.

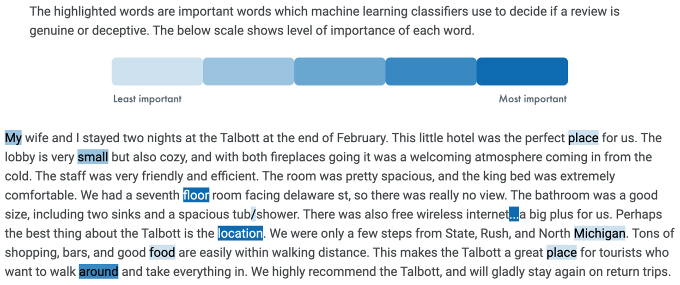
Now the continue button is enabled.

Figure 4.11: Experiment 1 tutorial: selected examples + guidelines. 'Reveal guidelines' shows a list of guidelines as illustrated in Figure 4.9.



Beautiful views and awesome service! My husband and I stayed at the Swissotel in Chicago while meeting up with some old friends from college. We have been to Chicago a few times but never stayed at this hotel before. First off the architecture and feel of the hotel is just amazing. So beautiful! The views from our corner suite was breathtaking and the accomodations were flawless. The fact that we could have an apartment like setting rather than a cold hotel room made all the difference. From the big tub in the bath to the entertainment system and work area, everything is in place to make your stay like being at home but with an awesome view of the Chicago Skyline. We took advantage of the Executive Club for desserts and breakfast and had dinner one night at the Palm restaurant. Delicious food! We will definitely stay at the Swissotel the next time we visit Chicago.

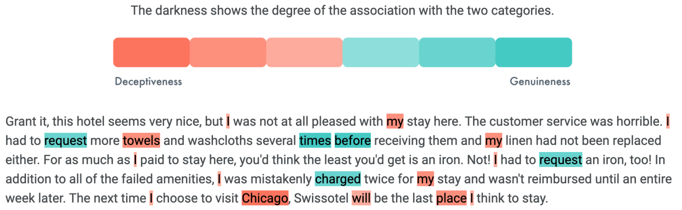
Figure 4.12: Experiment 2 real-time assistance: no assistance.



The highlighted words are important words which machine learning classifiers use to decide if a review is genuine or deceptive. The below scale shows level of importance of each word.

My wife and I stayed two nights at the Talbott at the end of February. This little hotel was the perfect place for us. The lobby is very small but also cozy, and with both fireplaces going it was a welcoming atmosphere coming in from the cold. The staff was very friendly and efficient. The room was pretty spacious, and the king bed was extremely comfortable. We had a seventh floor room facing delaware st, so there was really no view. The bathroom was a good size, including two sinks and a spacious tub/shower. There was also free wireless internet...a big plus for us. Perhaps the best thing about the Talbott is the location. We were only a few steps from State, Rush, and North Michigan. Tons of shopping, bars, and good food are easily within walking distance. This makes the Talbott a great place for tourists who want to walk around and take everything in. We highly recommend the Talbott, and will gladly stay again on return trips.

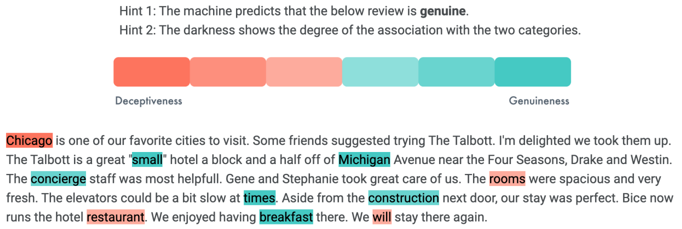
Figure 4.13: Experiment 2 real-time assistance: unsigned highlights.



The darkness shows the degree of the association with the two categories.

Grant it, this hotel seems very nice, but I was not at all pleased with my stay here. The customer service was horrible. I had to request more towels and washcloths several times before receiving them and my linen had not been replaced either. For as much as I paid to stay here, you'd think the least you'd get is an iron. Not! I had to request an iron, too! In addition to all of the failed amenities, I was mistakenly charged twice for my stay and wasn't reimbursed until an entire week later. The next time I choose to visit Chicago, Swissotel will be the last place I think to stay.

Figure 4.14: Experiment 2 real-time assistance: signed highlights.

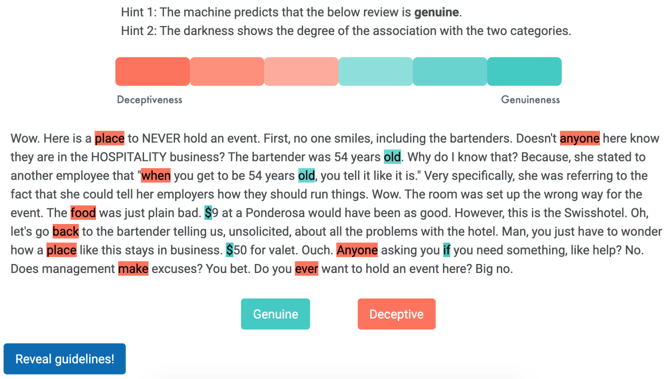


Hint 1: The machine predicts that the below review is **genuine**.
Hint 2: The darkness shows the degree of the association with the two categories.

Chicago is one of our favorite cities to visit. Some friends suggested trying The Talbott. I'm delighted we took them up. The Talbott is a great "small" hotel a block and a half off of Michigan Avenue near the Four Seasons, Drake and Westin. The concierge staff was most helpfull. Gene and Stephanie took great care of us. The rooms were spacious and very fresh. The elevators could be a bit slow at times. Aside from the construction next door, our stay was perfect. Bice now runs the hotel restaurant. We enjoyed having breakfast there. We will stay there again.

Figure 4.15: Experiment 2 real-time assistance: signed highlights + predicted label.

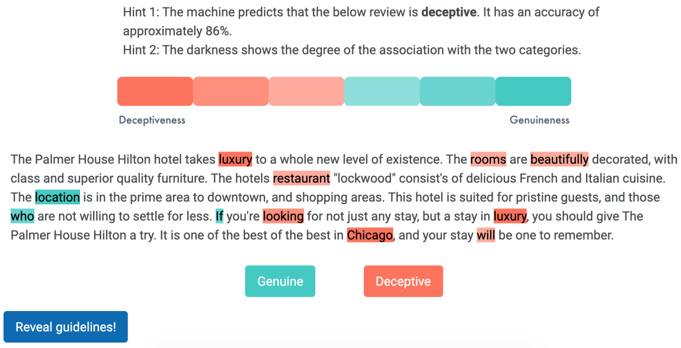
Among our participants in Experiment 2, 64 were between 18 and 25, 270 were between 26 and 40, 116 were between 41 and 60, 26 were 61 and above, and 4 preferred not to answer. They had a range of education backgrounds, comprising some high school (3), high school graduate (44), some college credit (120), trade/technical/vocational training (32), Bachelor's degree and above



Hint 1: The machine predicts that the below review is **genuine**.
Hint 2: The darkness shows the degree of the association with the two categories.

Wow. Here is a place to NEVER hold an event. First, no one smiles, including the bartenders. Doesn't anyone here know they are in the HOSPITALITY business? The bartender was 54 years old. Why do I know that? Because, she stated to another employee that "when you get to be 54 years old, you tell it like it is." Very specifically, she was referring to the fact that she could tell her employers how they should run things. Wow. The room was set up the wrong way for the event. The food was just plain bad. $9 at a Ponderosa would have been as good. However, this is the Swisshotel. Oh, let's go back to the bartender telling us, unsolicited, about all the problems with the hotel. Man, you just have to wonder how a place like this stays in business. $50 for valet. Ouch. Anyone asking you if you need something, like help? No. Does management make excuses? You bet. Do you ever want to hold an event here? Big no.

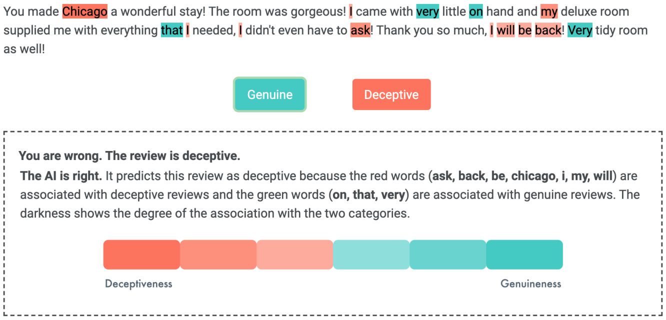
Figure 4.16: Experiment 2 real-time assistance: signed highlights + predicted label + guidelines.



Hint 1: The machine predicts that the below review is **deceptive**. It has an accuracy of approximately 86%.
Hint 2: The darkness shows the degree of the association with the two categories.

The Palmer House Hilton hotel takes luxury to a whole new level of existence. The rooms are beautifully decorated, with class and superior quality furniture. The hotels restaurant "lockwood" consist's of delicious French and Italian cuisine. The location is in the prime area to downtown, and shopping areas. This hotel is suited for pristine guests, and those who are not willing to settle for less. If you're looking for not just any stay, but a stay in luxury, you should give The Palmer House Hilton a try. It is one of the best of the best in Chicago, and your stay will be one to remember.

Figure 4.17: Experiment 2 real-time assistance: signed highlights + predicted label + guidelines + accuracy statement.



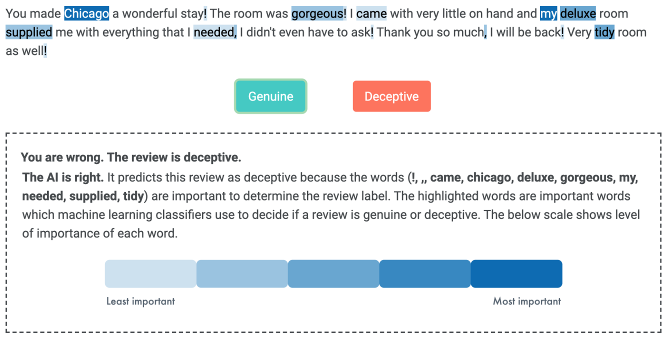
You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

You are wrong. The review is deceptive.
The AI is right. It predicts this review as deceptive because the red words (**ask, back, be, chicago, i, my, will**) are associated with deceptive reviews and the green words (**on, that, very**) are associated with genuine reviews. The darkness shows the degree of the association with the two categories.

Figure 4.18: Experiment 3: top features from SVM are highlighted.

(278), and 3 prefered not to answer.

Among our participants in Experiment 3, 62 were between 18 and 25, 255 were between 26 and 40, 138 were between 41 and 60, 24 were 61 and above, and 1 preferred not to answer. They had a range of educational attainment, comprising some high school (1), high school graduate (51),

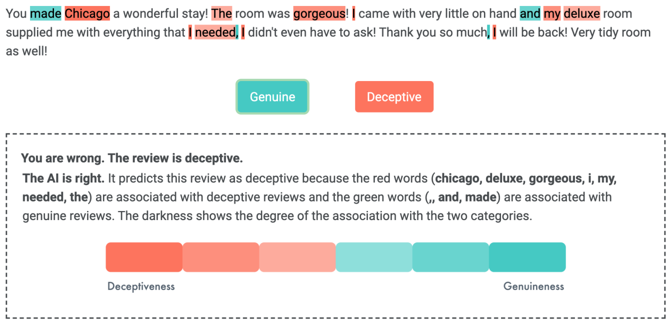


You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

You are wrong. The review is deceptive.

The AI is right. It predicts this review as deceptive because the words (**!, ,, came, chicago, deluxe, gorgeous, my, needed, supplied, tidy**) are important to determine the review label. The highlighted words are important words which machine learning classifiers use to decide if a review is genuine or deceptive. The below scale shows level of importance of each word.

Figure 4.19: Experiment 3: top features from BERT attention are highlighted.



You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

You are wrong. The review is deceptive.

The AI is right. It predicts this review as deceptive because the red words (**chicago, deluxe, gorgeous, i, my, needed, the**) are associated with deceptive reviews and the green words (**,, and, made**) are associated with genuine reviews. The darkness shows the degree of the association with the two categories.

Figure 4.20: Experiment 3: top features from BERT LIME are highlighted.

some college credit (111), trade/technical/vocational training (40), Bachelor's degree and above (274), and 3 preferred not to answer.

We only kept participants that complete the full task and submit a unique survey code. Participants that do not comply with the criteria were not included.

Figure 4.21 - Figure 4.23 show the average time taken in each experiment. We calculated and filtered out outliers from each experiment respectively with an interquartile range. In Figure 4.24 - Figure 4.26 we show the average time taken during prediction phase in each experiment. Outliers were discarded after the same procedures.

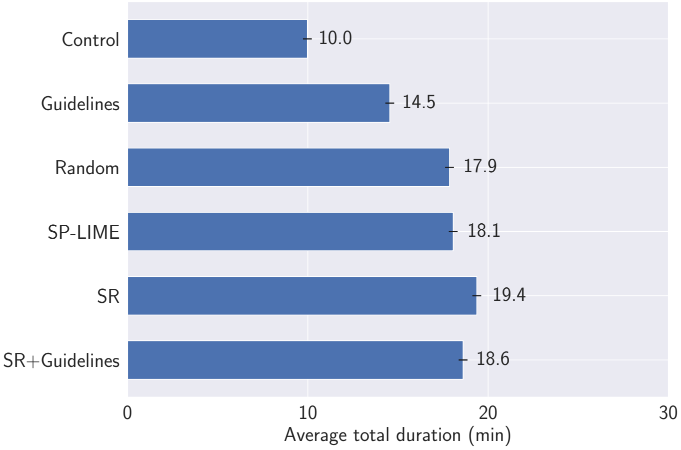


Figure 4.21: Average time taken for each experimental setup in experiment 1.

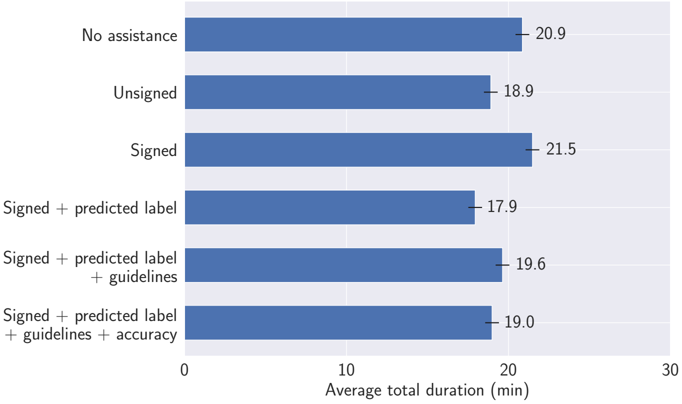


Figure 4.22: Average time taken for each experimental setup in experiment 2.

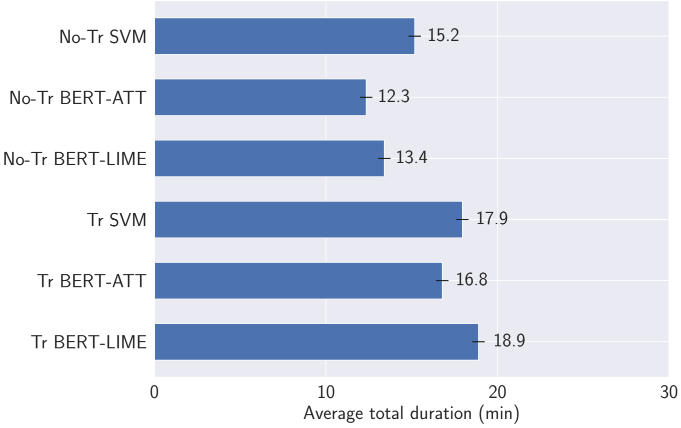


Figure 4.23: Average time taken for each experimental setup in experiment 3.

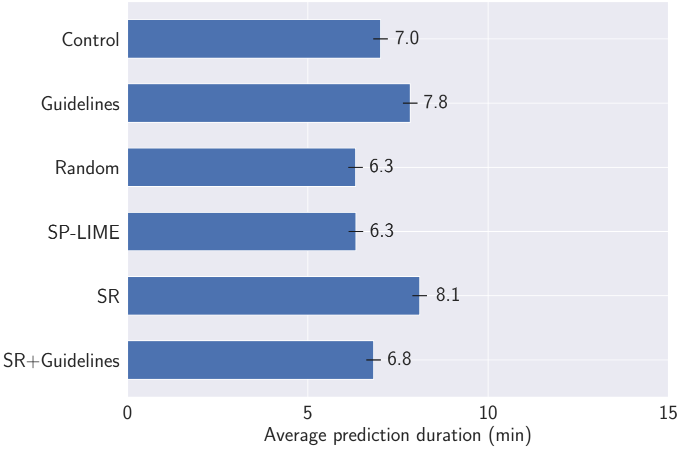


Figure 4.24: Average time taken for the prediction phase in each experimental setup in experiment 1.

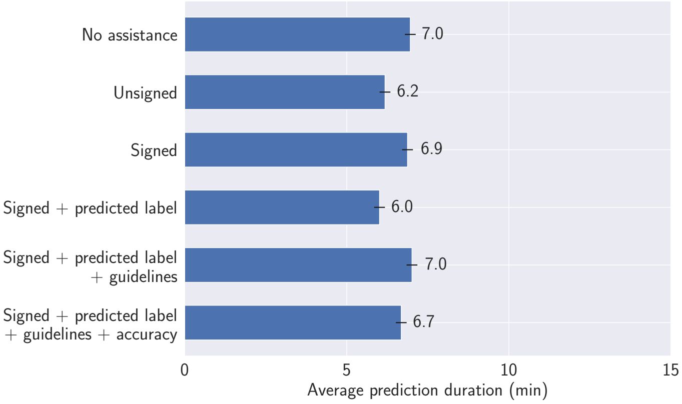


Figure 4.25: Average time taken for the prediction phase in each experimental setup in experiment 2.

4.10.2 Trust Analysis

4.10.2.1 Analysis of Free Responses from Turkers

Free responses from turkers confirmed the findings in the qualitative study. Participants felt that tutorial was useful but could not understand why certain features are deceptive or genuine. One participant commented, "Although I am an English major, the training really helped me to think and consider the nuances of language. I enjoy good writing but I often overlook attempts to manipulate or deceive the reader/audience. I felt this training was very beneficial". Another

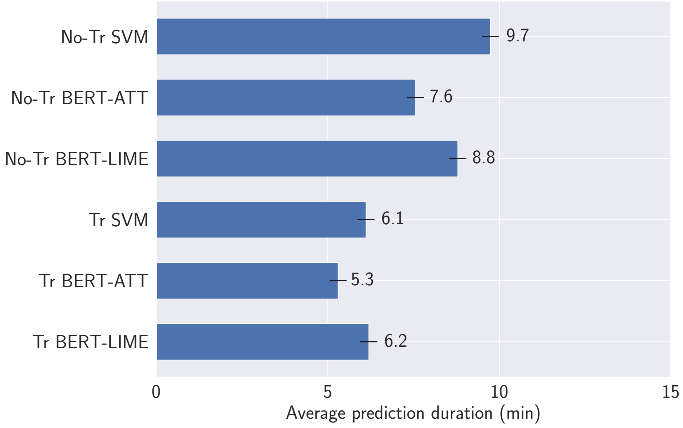


Figure 4.26: Average time taken for the prediction phase in each experimental setup in experiment 3.

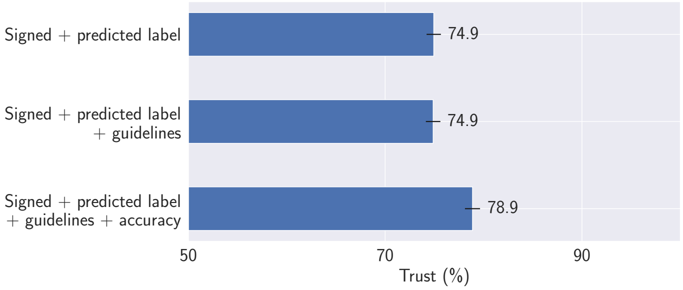


Figure 4.27: Human trust on machine predictions in experiment 2. Differences between all pairs are not statistically significant. These results suggest that guidelines and accuracy statement do not increase human trust in machine learning models significantly.

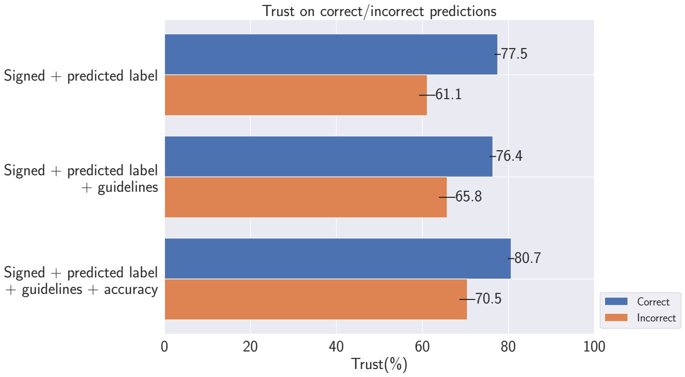


Figure 4.28: Human trust on correct / incorrect machine predictions in experiment 2. Differences between correct predictions and incorrect predictions are statistically significant. These results suggest that human have more trust in correct predictions than incorrect ones.

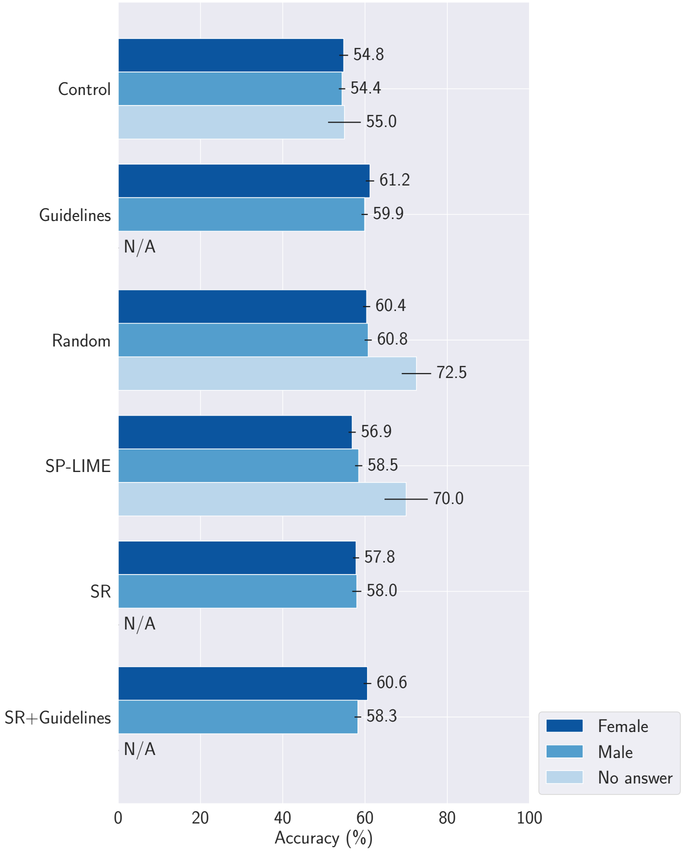


Figure 4.29: Experiment 1: gender. Human accuracy grouped by experimental setups and gender.

participant remarked, “I could not understand why words were chosen for the reason”.

4.10.3 Human Performance Grouped by Demographics

The is no clear trend regarding gender, education background, review writing frequency, and age among experiments.

4.10.4 Attention-check Design

P11 was half way through the session and commented, “I’m trying to think about this from a way of, like, are these reviews being generated by a computer, or are they, like, are all of these reviews from real people, and am I trying to tell if somebody’s, like, lying about the review”. The interviewer then suggested to the participant to read the instructions in the dialogue boxes. P11

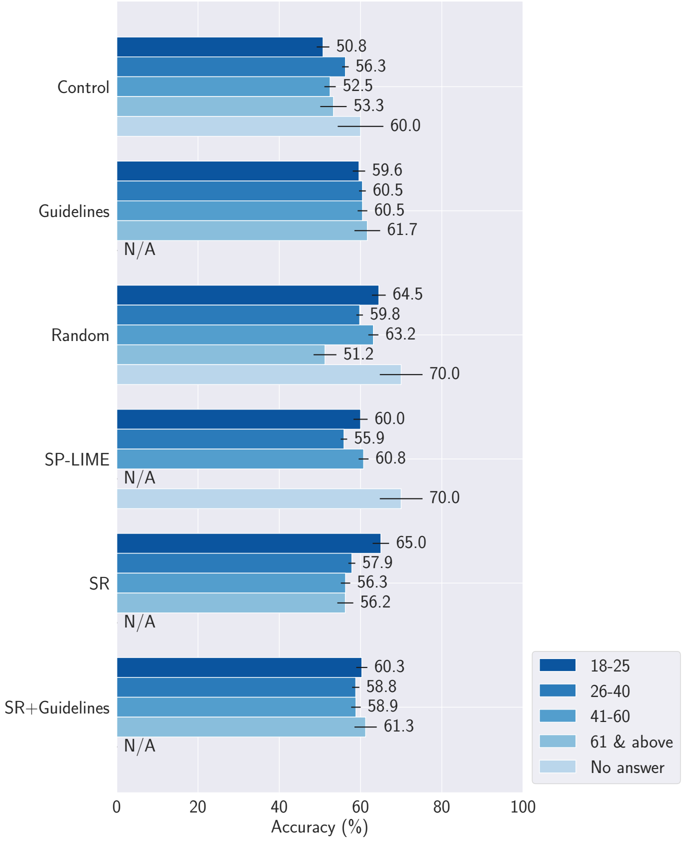


Figure 4.30: Experiment 1: age. Human accuracy grouped by experimental setups and age.

subsequently explained that he “just didn’t notice that because I was just reading the rules and skipped the box”. Similarly, P9 asked the interviewer, “By deceptive review do you mean users typing a review for the sake of tarnishing reputation, or uplifting reputation, or are you referring to computer-generated reviews which are trying to deceive people”. Due to a couple of the above cases, we added additional attention-check questions to ensure that participants are aware of the definition of deceptive reviews. Refer to the outdated and updated attention-check design below.

4.10.5 Exit Survey

Figure 4.43 - Figure 4.45 show exit surveys for experimental setups in Experiment 1.

Figure 4.47 and Figure 4.48 show exit surveys for experimental setups in Experiment 3.

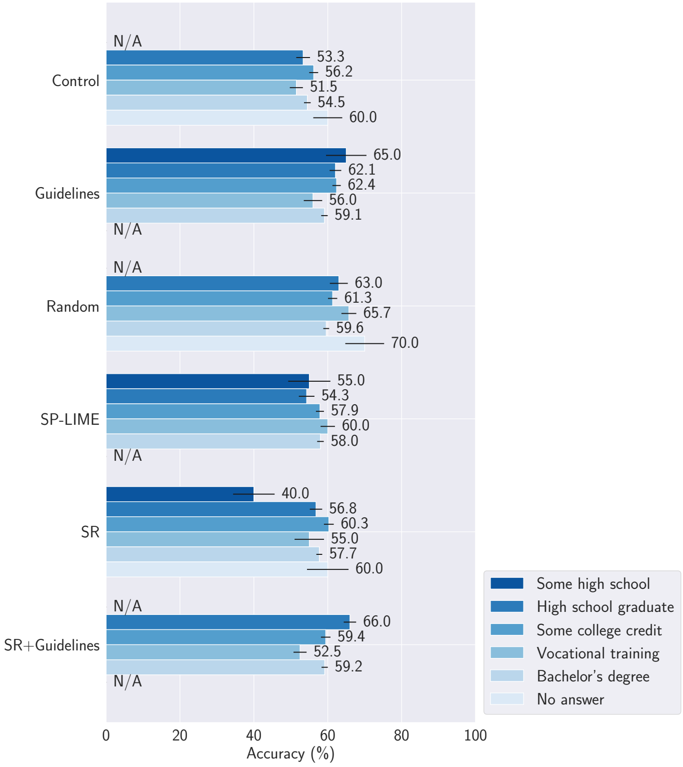


Figure 4.31: Experiment 1: education background. Human accuracy grouped by experimental setups and education background.

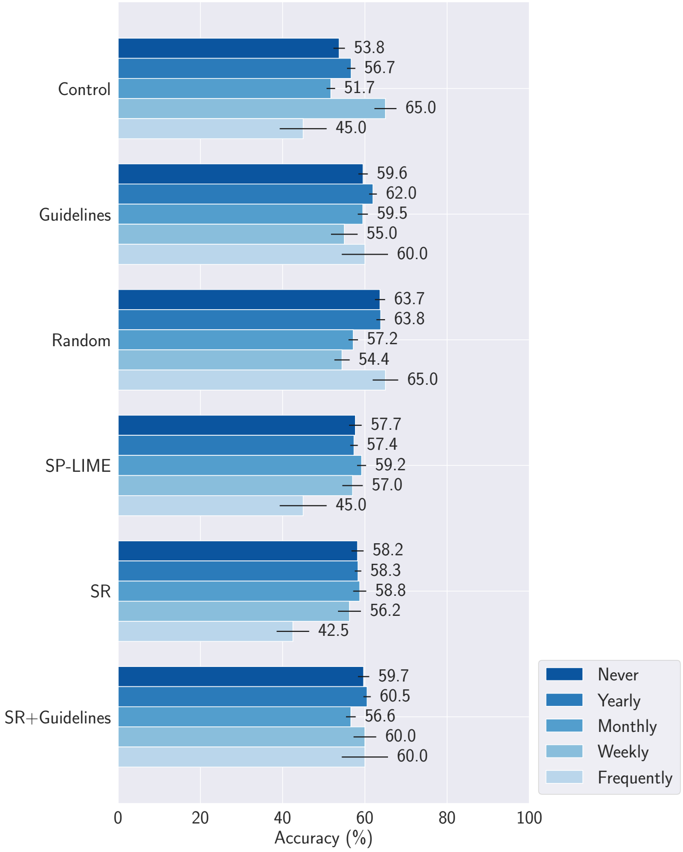


Figure 4.32: Experiment 1: review writing frequency. Human accuracy grouped by experimental setups and review writing frequency.

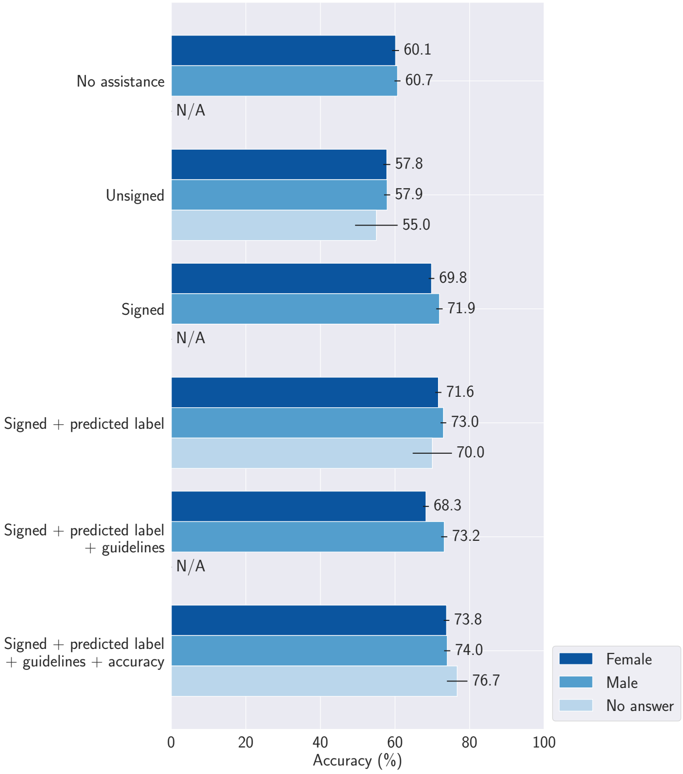


Figure 4.33: Experiment 2: gender. Human accuracy grouped by experimental setups and gender.

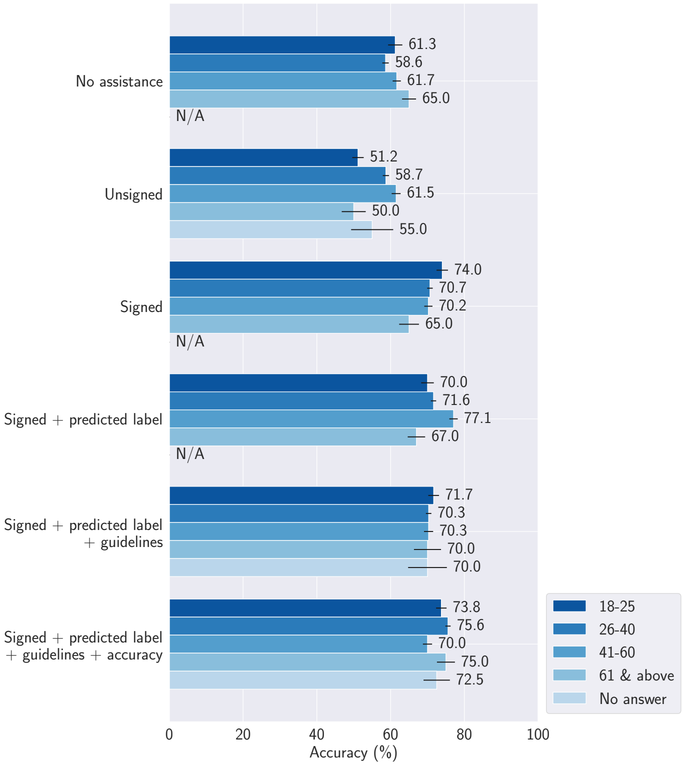


Figure 4.34: Experiment 2: age. Human accuracy grouped by experimental setups and age.

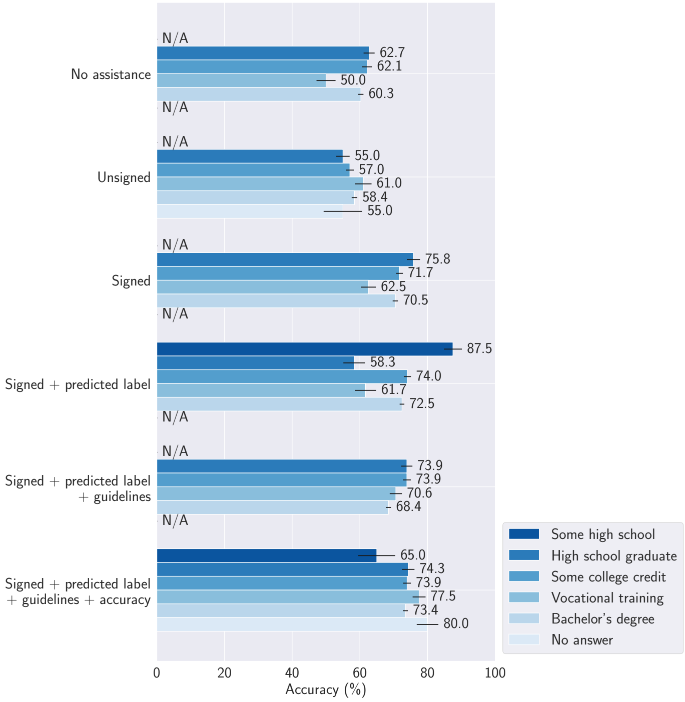


Figure 4.35: Experiment 2: education background. Human accuracy grouped by experimental setups and education background.

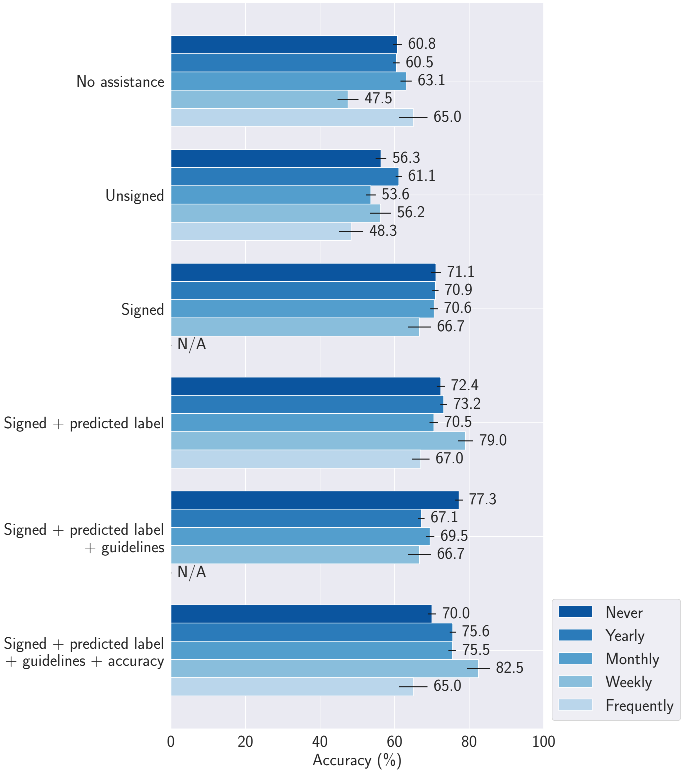


Figure 4.36: Experiment 2: review writing frequency. Human accuracy grouped by experimental setups and review writing frequency.

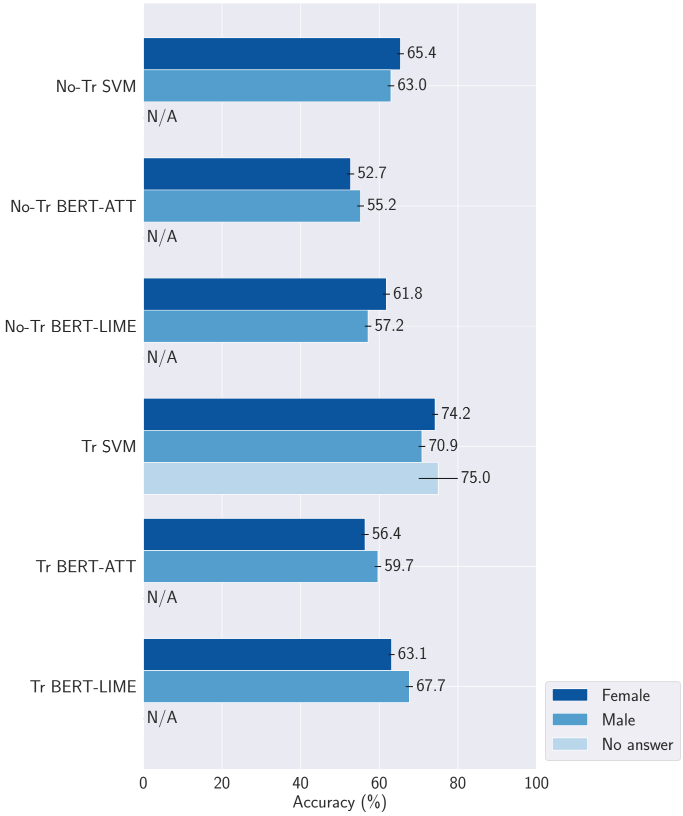


Figure 4.37: Experiment 3: gender. Human accuracy grouped by experimental setups and gender.

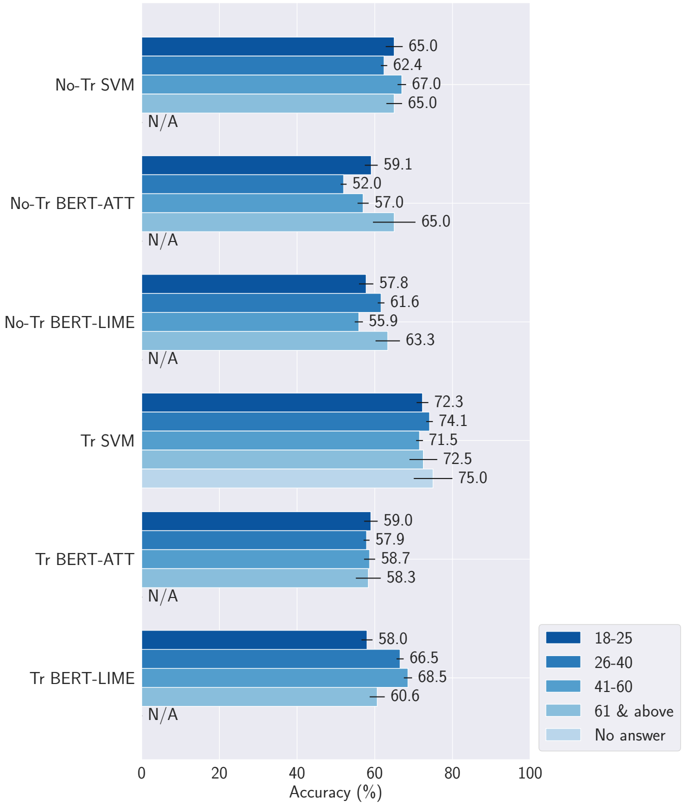


Figure 4.38: Experiment 3: age. Human accuracy grouped by experimental setups and age.

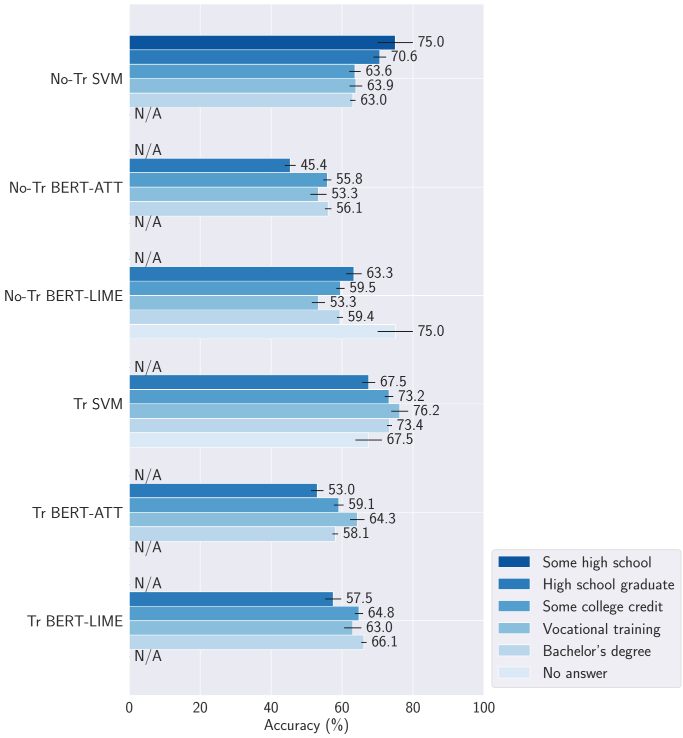


Figure 4.39: Experiment 3: education background. Human accuracy grouped by experimental setups and education background.

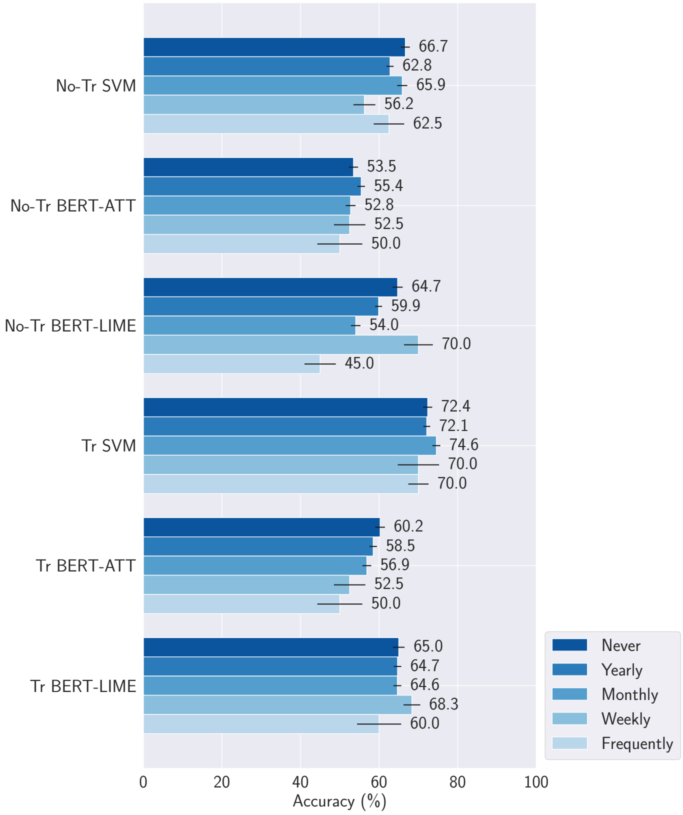
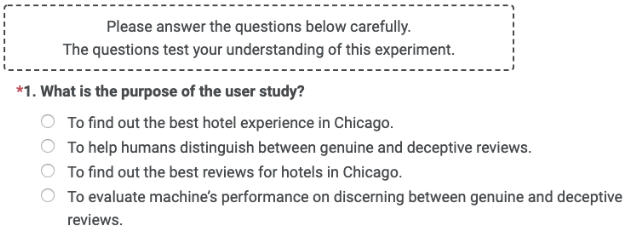


Figure 4.40: Experiment 3: review writing frequency. Human accuracy grouped by experimental setups and review writing frequency.

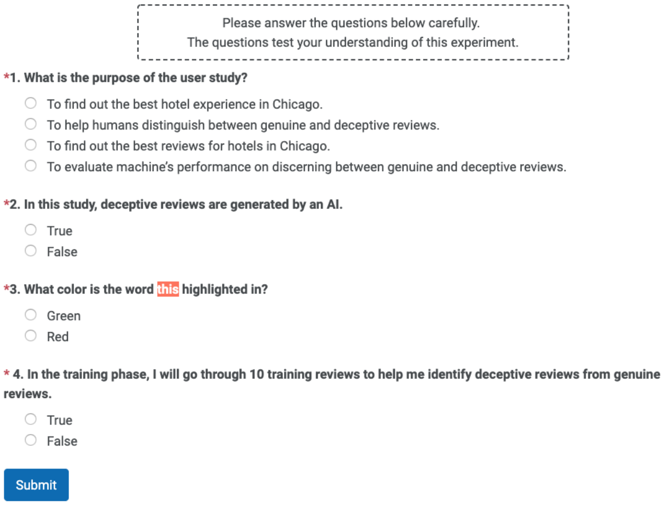


Please answer the questions below carefully.
The questions test your understanding of this experiment.

***1. What is the purpose of the user study?**

- ○ To find out the best hotel experience in Chicago.
- ○ To help humans distinguish between genuine and deceptive reviews.
- ○ To find out the best reviews for hotels in Chicago.
- ○ To evaluate machine's performance on discerning between genuine and deceptive reviews.

Figure 4.41: Outdated attention-check design. The outdated design does not allow participants to confirm on their answers. If they selected the wrong answer, they will be disqualified immediately.



Please answer the questions below carefully.
The questions test your understanding of this experiment.

*1. What is the purpose of the user study?

○ To find out the best hotel experience in Chicago.
○ To help humans distinguish between genuine and deceptive reviews.
○ To find out the best reviews for hotels in Chicago.
○ To evaluate machine's performance on discerning between genuine and deceptive reviews.

*2. In this study, deceptive reviews are generated by an AI.

○ True
○ False

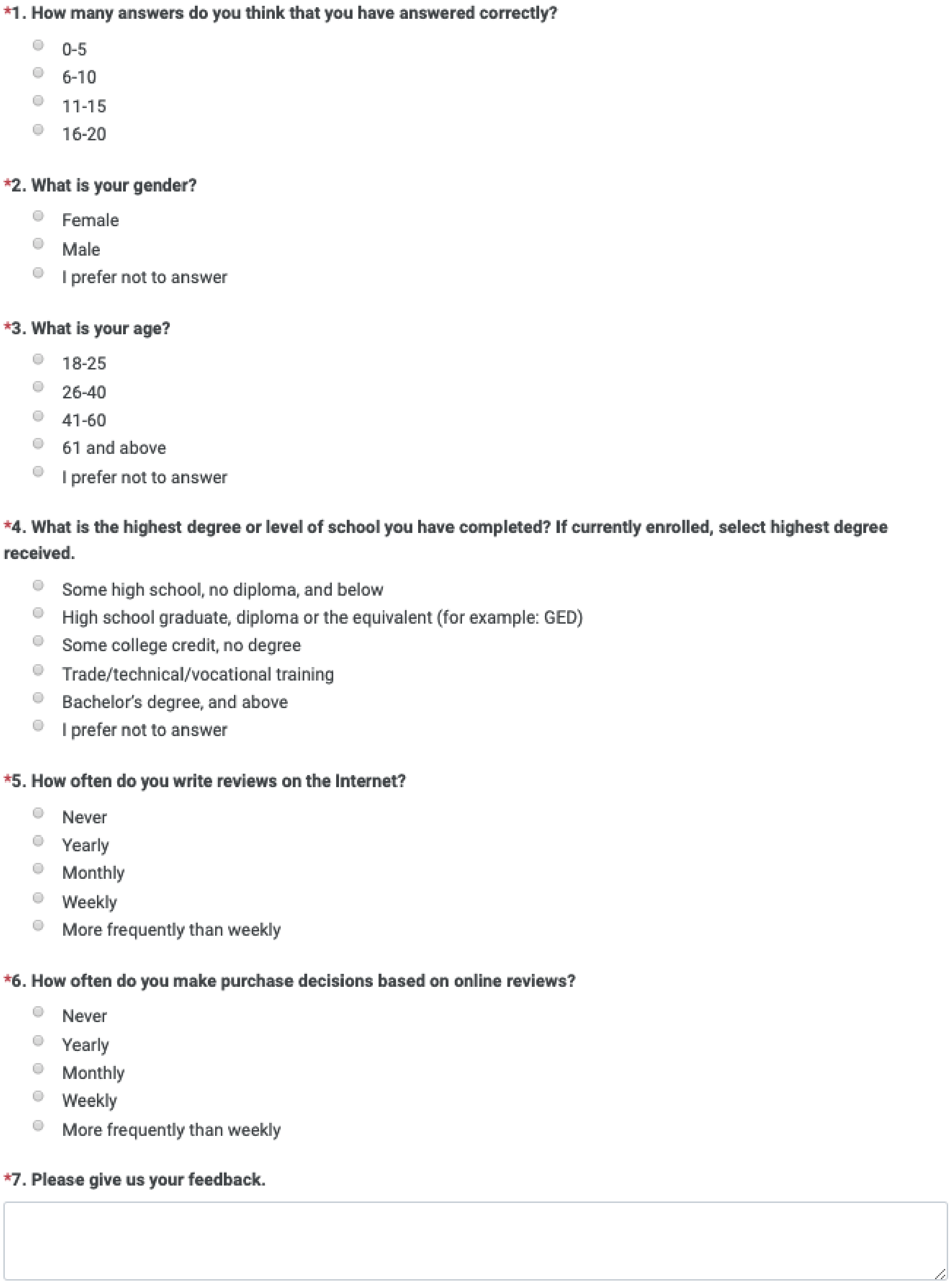
*3. What color is the word this highlighted in?

○ Green
○ Red

* 4. In the training phase, I will go through 10 training reviews to help me identify deceptive reviews from genuine reviews.

○ True
○ False

Figure 4.42: Updated attention-check design. The updated design allows participants to confirm on their answers.



*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?

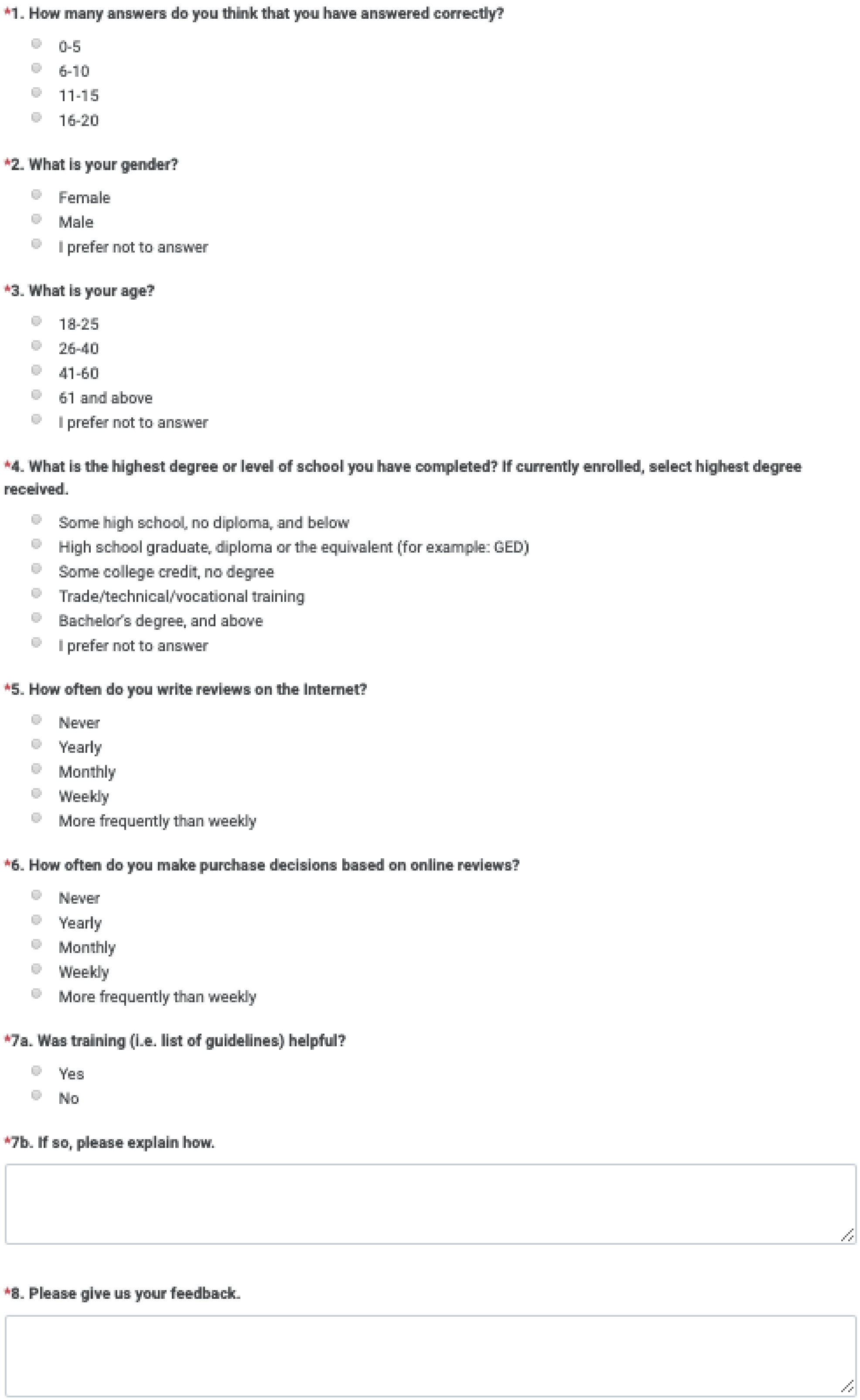
○ Never
○ Yearly
○ Monthly
○ Weekly
○ More frequently than weekly

*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 4.43: Exit survey for control setup in Experiment 1.



*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

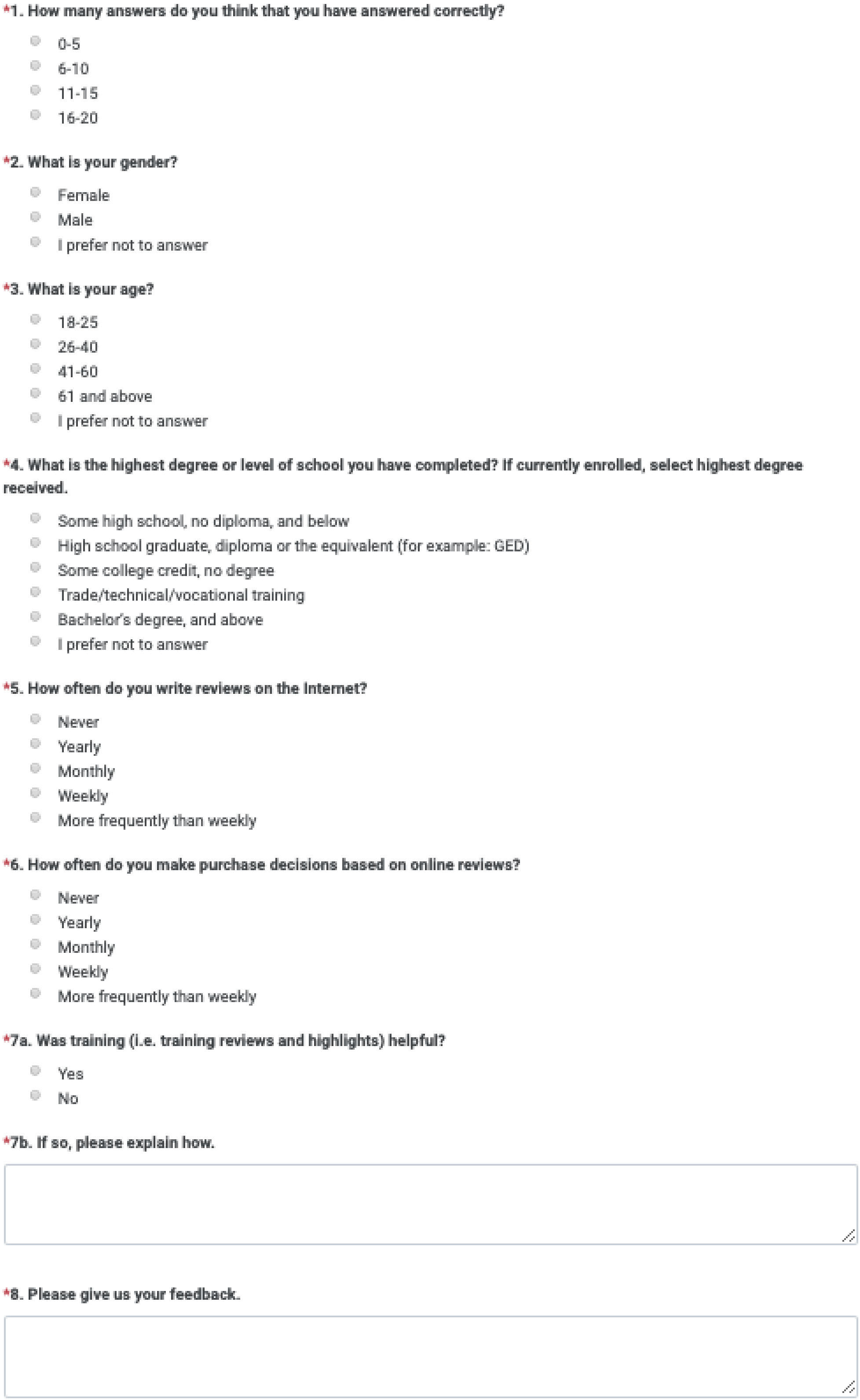
*7a. Was training (i.e. list of guidelines) helpful?

○ Yes
○ No

*7b. If so, please explain how.

*8. Please give us your feedback.

Figure 4.44: Exit survey for guidelines setup in Experiment 1.



*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

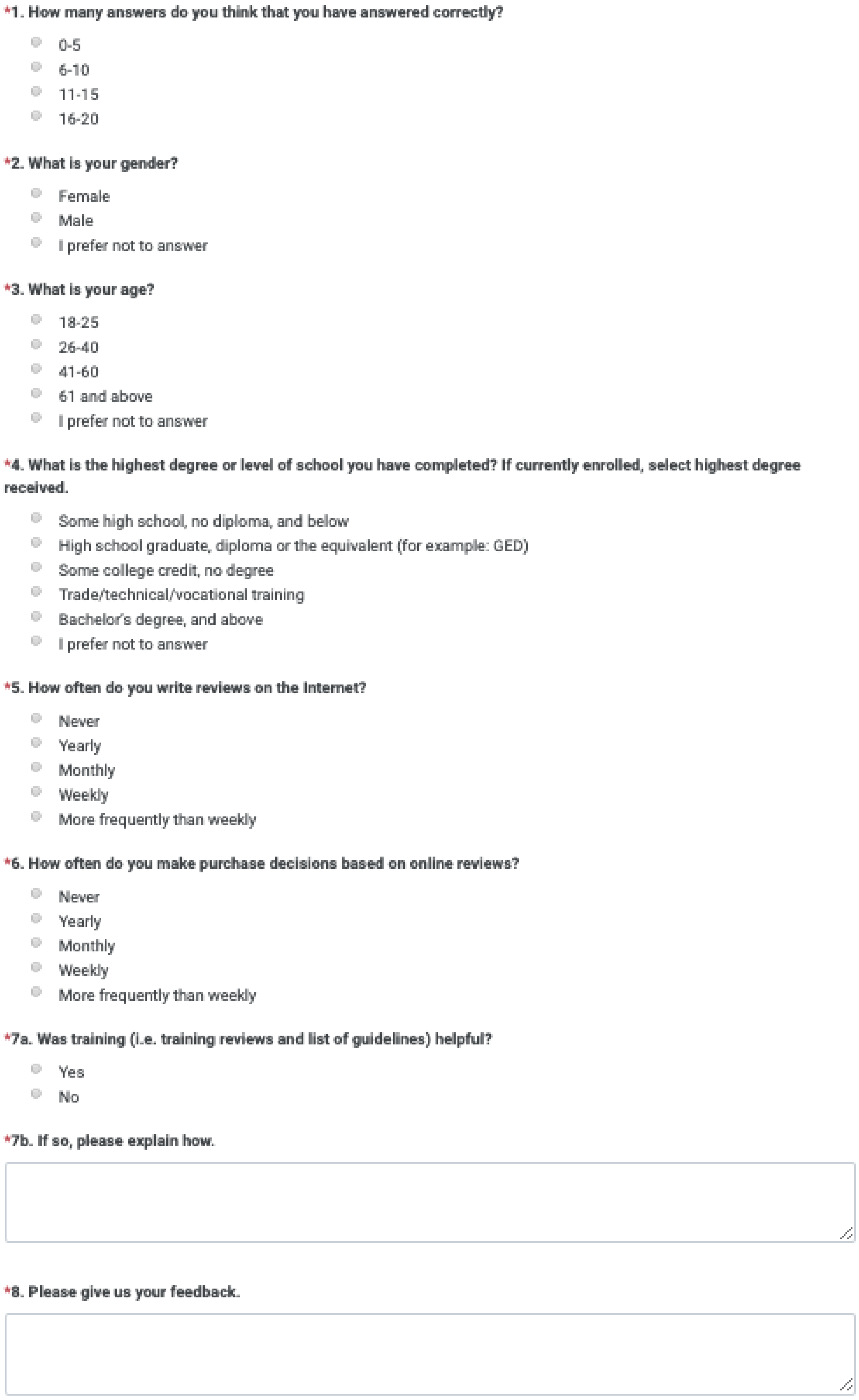
*7a. Was training (i.e. training reviews and highlights) helpful?

○ Yes
○ No

*7b. If so, please explain how.

*8. Please give us your feedback.

Figure 4.45: Exit survey for examples i.e., **random**, **SP-LIME**, and **spaced repetition** in experiment 1. Note that question 7a changes to the following: 'Was training (i.e. training reviews and list of guidelines) useful?' for **SR+guidelines**.



*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

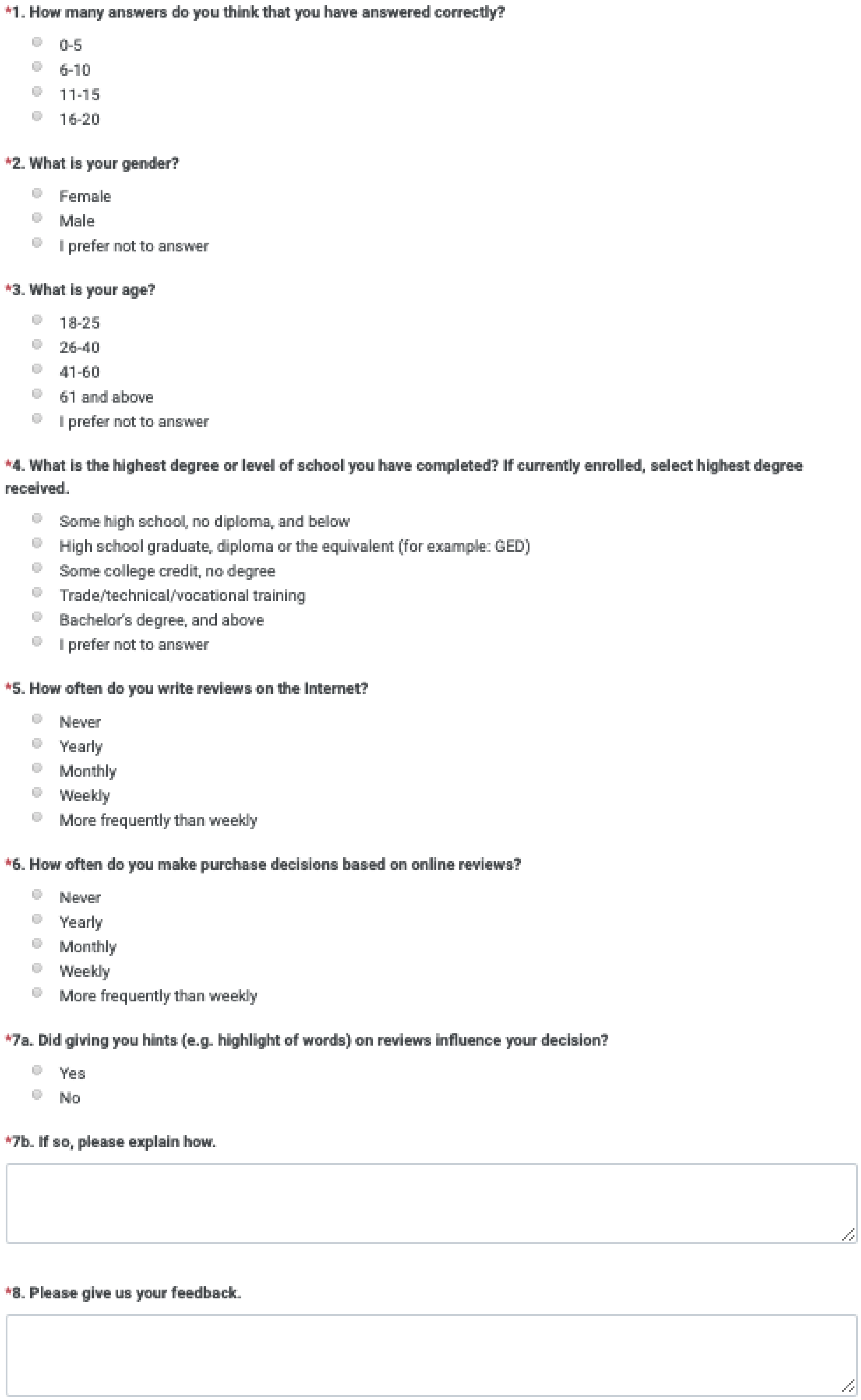
*7a. Was training (i.e. training reviews and list of guidelines) helpful?

○ Yes
○ No

*7b. If so, please explain how.

*8. Please give us your feedback.

Figure 4.46: Exit survey for experimental setup in Experiment 2.



*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

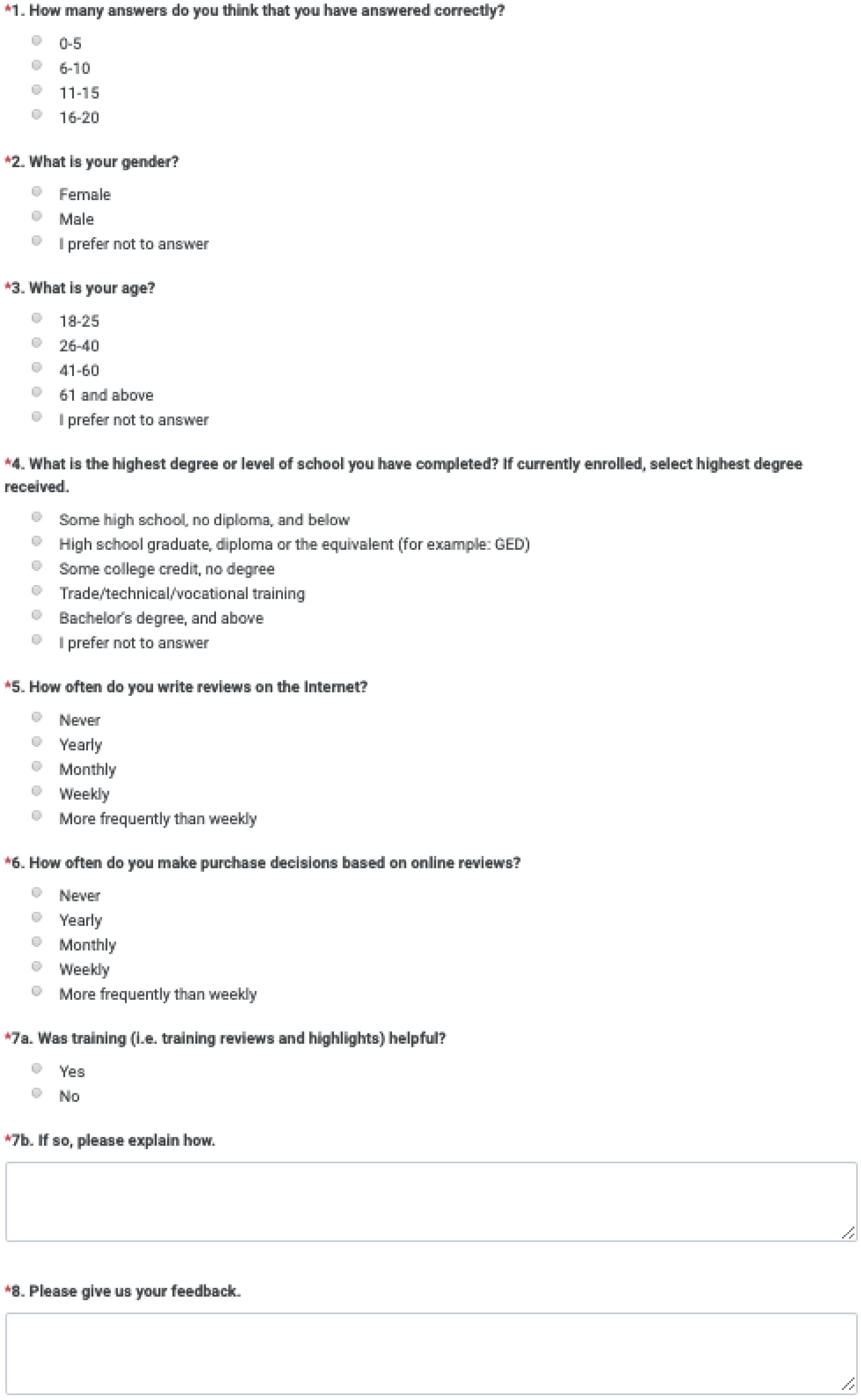
*7a. Did giving you hints (e.g. highlight of words) on reviews influence your decision?

○ Yes
○ No

*7b. If so, please explain how.

*8. Please give us your feedback.

Figure 4.47: Exit survey for non-training experimental setups in Experiment 3.



***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. Was training (i.e. training reviews and highlights) helpful?**

○ Yes
○ No

***7b. If so, please explain how.**

***8. Please give us your feedback.**

Figure 4.48: Exit survey for training experimental setups in Experiment 3.

Chapter 5

Conditional Delegation

5.1 Overview

Despite impressive performance in many benchmark datasets, AI models can still make mistakes, especially among out-of-distribution examples. It remains an open question how such imperfect models can be used effectively in collaboration with humans. Prior work has focused on AI assistance that helps people make individual high-stakes decisions, which is not scalable for a large amount of relatively low-stakes decisions, e.g., moderating social media comments. Instead, we propose conditional delegation as an alternative paradigm for human-AI collaboration where humans create rules to indicate trustworthy regions of a model. Using content moderation as a testbed, we develop novel interfaces to assist humans in creating conditional delegation rules and conduct a randomized experiment with two datasets to simulate in-distribution and out-of-distribution scenarios. Our study demonstrates the promise of conditional delegation in improving model performance and provides insights into design for this novel paradigm, including the effect of AI explanations.

5.2 Introduction

As AI performance grows rapidly and even surpasses humans in benchmark datasets [Kleinberg et al., 2018, He et al., 2015, McKinney et al., 2020, Silver et al., 2018, Brown and Sandholm, 2019], AI models hold great promise for improving human decision making in a wide variety of domains. However, full automation may not be desirable for ethical, legal, and safety reasons, especially in high-stakes domains Cai et al. [2019b], Lubars and Tan [2019], Lai and Tan [2019b], Green and Chen [2019b]. In particular, one well-known problem with the current AI models is **distribution shift**. Namely, AI performance can significantly drop for out-of-distribution examples that are different from the training data (in-distribution examples) [McCoy et al., 2019, Clark et al., 2019, Jia and Liang, 2017, Beede et al., 2020].

Human-AI collaboration is thus critical for effective integration of AI models into human decision making processes Cai et al. [2019c], Wang et al. [2019a], Arous et al. [2020], Ashktorab et al. [2020], Nguyen et al. [2018], Bansal et al. [2019a, 2021], O'Neill et al. [2020]. Many studies have investigated the role of AI in assisting humans in making individual decisions Lai and Tan [2019b], Lai et al. [2020a], Green and Chen [2019b,a], Zhang et al. [2020], Poursabzi-Sangdeh et al. [2021], Carton et al. [2020b], Lin et al. [2020], Weerts et al. [2019], Beede et al. [2020], Wang and Yin [2021], Lundberg et al. [2018b], e.g., predicting whether a person will recidivate in the near future. Such decisions are non-trivial even for human experts (e.g., judges) and AI models can potentially offer insights through their predictions and explanations. This approach is well suited for high-stakes domains, where humans are expected to make the final decision on every case (e.g., judges in bailing decisions). However, human-AI collaboration on every single decision is not scalable and is thus less appropriate for tasks involving a large amount of relatively low-stakes decisions. One such example is content moderation, where moderator decisions on individual comments for further actions (e.g., hiding the content or prompting further review, depending on the community policy) are of limited consequence; instead the key challenge lies in dealing with the massive scale of comments. Such tasks can benefit from a greater level of automation [Gillespie, 2018, Gorwa et al., 2020, Chandrasekharan et al., 2019].

In this work, we propose an alternative paradigm of human-AI collaboration — conditional delegation. Figure 5.1(A) illustrates a general form of conditional delegation. Human and AI work together to identify trustworthy regions of AI before deployment, i.e., model decisions are reliable or trustworthy for examples within these regions. Once deployed, the AI model only affects decisions for instances in the trustworthy regions. For the rest, another set of actions can be taken such as manual review or employing a different model since the given AI's decisions on them cannot be trusted. This approach employs a greater level of automation than human-AI collaboration on every single decision and provides human with active control on when to use an AI model and in what ways.

We use content moderation as a testbed. Figure 5.1(B) shows one possible instantiation in this context. Trustworthy regions can be operationalized with a collection of keyword-based rules created by human-AI collaboration before deployment. For example, after inspecting AI predictions on comments with the word "retard", the human may decide that AI works well on them and set "retard" as a conditional delegation rule. Once deployed, comments that fall within these trustworthy regions, i.e., **containing any keywords** specified by human, if **predicted toxic**, can be reliably reported for final actions, such as being hidden or sent for further review, depending on the community policy.

Notably, the task for humans to create *conditional delegation rules* share some similarity with what many social media moderators are already doing by writing manual automation rules to deal with the massive amount of comments (Figure 5.1(C)). For example, moderators on Reddit use a tool called AutoModerator, with which they manually customize a rule-based system to automatically identify comments for deleting or reporting for further review Jhaver et al. [2019a], Chandrasekharan et al. [2019]. This approach, however, misses out the benefit of AI especially since rigid rules often do not work on informal languages such as social media posts (e.g., containing swear words without being toxic). Without significantly altering content moderators' workflow, conditional delegation offers a promising approach to utilize AI, even if the model is not optimized for the community-specific content and should not be blindly trusted to work alone for every comment (Figure 5.1(D)).

In this instantiation, a key difference from individual human-AI decision making lies in the success criteria: while the quality of individual decisions (e.g., accuracy) is often the target in individual decision making, *precision* and *coverage* are critical for conditional delegation because moderation actions will only happen on comments that are predicted toxic.[1] Precision ensures

[1] Depending on the workflow, avoiding false negatives could be important in other instantiations.

that AI behavior is indeed trustworthy in the delegation mode and avoids unnecessary actions, whether it is mistaken deletion or extra work for further review. Coverage warrants that the AI model can identify as many toxic comments as possible to alleviate the scalability issues. In the context of content moderation, recall (identifying all toxic comments) is often less of a priority given the limited time for content moderators, who are often volunteers, to deal with a massive amount of incoming comments. This is reflected in the current workflow using the manual rule-based approach (Figure 5.1(C)), where comments falling outside the rules are ignored without taking an action. We assume the same workflow in our study and only focus on the precision and coverage related metrics for comments within the scope of keywords rules.

In this study, our **primary** interest is to investigate whether humans can effectively identify trustworthy regions for conditional delegation to improve the model precision with a good coverage, compared to the current manual rule-based approach (Figure 5.1(C)) and the model working alone (Figure 5.1(D)). Furthermore, we explore the effectiveness in two different AI scenarios: using an AI trained on the community specific data (*in-distribution*), and one trained on different data (*out-of-distribution*). The out-of-distribution model would perform much worse, but conditional delegation offers a potential means to improve through human-AI collaboration.

Our second set of contribution is to inform design of interfaces that support people to create high-quality conditional delegation rules. When given an AI model, content moderators often do not have labeled comments to quantify model performance. It would be helpful for them to observe model behaviors on their own data of interest to identify good delegation rules (i.e., trustworthy regions). To facilitate the creation of keyword-based rules, we develop an interface that allows participants who act as moderators to perform keywords search and observe model behavior on the search results. We provide and study the effects of several delegation support features, including predicted labels, local explanations that show the rationales behind predictions, and global explanations that provide an overview of the model.

To summarize, we ask the following research questions:

RQ1. Can users create keyword-based rules for conditional delegation that improves model precision, so that these rules correspond to trustworthy regions?

RQ2. How do the performance of conditional delegation and user experiences (such as engagement and subjective perceptions) vary between in-distribution and out-of-distribution AI?

RQ3. What are the effects of delegation support features on performance and user experiences, including showing prediction labels, local explanations, and global explanations?

Through a randomized experiment with 240 mechanical turkers, we show that even crowdworkers are able to create high-quality rules that lead to higher precision with conditional delegation than the model working alone. Especially when applied to an in-distribution AI, which already outperforms the manual rule-based approach for content moderation, conditional delegation further enhances the performance, leading to "complementary performance" (i.e., human+AI ¿ AI and human+AI ¿ human) [Bansal et al., 2021]. For out-of-distribution AI used in this study, conditional delegation improves the model performance but does not suffice in compensating for the performance disadvantage of AI to outperform the manual rule-based approach. We also found that model explanations can improve efficiency in identifying delegation conditions and, with weak evidence, improve user experiences.

Overall, our work provides a new perspective to the emerging area of human-AI collaboration. Our core contribution is to demonstrate that conditional delegation is a promising alternative paradigm that allows users to control when to trust or distrust AI. We also contribute a set of interface features to assist people in creating conditional delegation rules and an empirical understanding of their effects. The diverging performance of in-distribution and out-of-distribution highlights the importance of considering the effect of distribution shift when conducting empirical studies of human-AI collaboration to inform the generalizablity of results, echoing recent findings in other studies [Chiang and Yin, 2021, Liu et al., 2021].

5.3 Related work

5.3.1 Human-AI Collaboration

Terms like "human-AI collaboration" Cai et al. [2019c], Wang et al. [2019a], Arous et al. [2020], Ashktorab et al. [2020], "human-AI partnership" Nguyen et al. [2018], "human-AI teaming" Bansal et al. [2019a, 2021], O'Neill et al. [2020] have emerged in various literature studying the use of AI systems. They reflect a shift of perspective away from complete automation by AI. Fostering effective human-AI collaboration is not only critical for safety reasons, especially in high-stakes domains Cai et al. [2019b], but also necessary to harnessing the complementarity of human and AI intelligence to achieve optimal outcome Bansal et al. [2019b], Wilder et al. [2020], reduce computational complexity Holzinger [2016], and enable novel technologies that are beyond the current capabilities of AI Wang et al. [2019a], Cranshaw et al. [2017].

Many forms of human-AI collaboration have been explored. The term "human-in-the-loop" is used broadly, but often refers to interactive training paradigm where the AI receives input from the human to improve its performance. For example, the field of interactive Machine Learning Holzinger [2016], Fails and Olsen Jr [2003], Amershi et al. [2014], Dudley and Kristensson [2018], at the intersection of ML and HCI, develops systems that allow end users to guide model behavior. This kind of paradigm allows humans to directly impact the working of AI, and requires using algorithms that can incorporate human input to update the model, which can be technically challenging or infeasible in practice.

Another rich area to study human-AI collaboration is AI-assisted Zhang et al. [2020], Wang and Yin [2021], Buçinca et al. [2021] or "machine/algorithm-in-the-loop" decision-making Green and Chen [2019b], Lai and Tan [2019b]. In this paradigm, AI performs an assistive role by providing a prediction or recommendation, while the human decision maker makes the final call and may choose to accept or reject the AI recommendation. Several studies explored the questions of whether and how to achieve **complementary performance**, i.e., the collaborative decision outcome outperforming human or AI alone Zhang et al. [2020], Bansal et al. [2021], Lai and Tan

[2019b]. The empirical results, however, are mixed at best, because there was either insufficient complementarity in human and AI's domain knowledge or a lack of ability for people to judge the reliability of AI recommendations. This approach tends to focus on high-stakes decisions and are not scalable in the number of decisions because humans are required to make each decision.

Another line of work explores intelligent systems and considers different tasks that AI can perform and the optimal level of automation versus human agency Wang et al. [2021], Mackeprang et al. [2019], Lai and Tan [2019b]. For example, building on a classic model of levels of automation Parasuraman et al. [2000], Mackeprang et al. [2019] proposed a design framework that decomposes the design space of an intelligent system into sub-tasks then allocates human, AI or both to perform each sub-task.

The goal of our work is to have AI partially automate a large volume of decisions rather than assisting individual decisions. Extending existing models of human agency and automation Parasuraman et al. [2000], Mackeprang et al. [2019], we introduce *proactive* human agency, with which human can act and exercise control prior to model deployment, instead of reacting to model outputs. By conducting a controlled experiment, we explore whether this new human-AI collaboration paradigm can achieve complementary performance by outperforming AI and manual approaches. While some prior work also discussed delegation based on predicted outputs (e.g., predicted probability) [Keswani et al., 2021, Chandrasekharan et al., 2019], our work focuses on identifying trustworthy regions in the input space. Furthermore, to the best of our knowledge, our work is the first study with controlled experiments to examine the effect of conditional delegation.

5.3.2 AI explanations for human-AI interaction

Mental model, defined as an understanding of how a system works, is a key concept in human-computer interaction Norman [2013]. Having an appropriate mental model allows people to accurately anticipate a system's behaviors and interact more effectively. People's mental model can be refined by explanations of how the system works. Therefore, explanation and transparency features have long been an interest of HCI research on various technologies Abdul et al. [2018], Herlocker et al. [2000], Lim et al. [2009], Rader et al. [2018].

Recently, AI explanations have gained much attention Lai and Tan [2019b], Liao et al. [2020], Ghai et al. [2020], Dodge et al. [2019], Buçinca et al. [2021], Bansal et al. [2021], Zhang et al. [2020]. The popularity of complex, inscrutable AI models such as deep neural networks make the difficulty of understanding a primary challenge for modern AI technologies. This challenge has given rise to a technical field of explainable AI (XAI), producing an abundance of techniques that aim to make AI more understandable by people. While the landscape of XAI technique is beyond the scope of this paper Guidotti et al. [2018b], Adadi and Berrada [2018], Gilpin et al. [2018], an important distinction relevant to our study is the contrast between *local explanations*, which focus on explaining the rationale for a particular prediction, versus *global explanations*, which aim to give a high-level understanding of how the AI works. We explore the effect of both types of explanation in our study and will discuss the details of the XAI techniques used for our toxicity prediction model in the next section.

HCI studies on XAI have found explanations to improve user understanding of AI systems Cheng et al. [2019], Ghai et al. [2020], Buçinca et al. [2020], and somewhat mixed results on enhancing user trust Cheng et al. [2019], satisfaction Ghai et al. [2020] and willingness to adopt AI systems Tsai et al. [2021]. Moreover, explanations provide additional information that can be utilized to assist the task that people perform. For example, Lai and Tan proposed a spectrum between human agency and full automation for machine learning to assist human decision-making Lai and Tan [2019b], and considered showing explanation as an additional form of machine assistance beyond solely providing prediction labels, and thus increase the level of automated assistance. In interactive machine learning, explanation has been studied as a primary means for people to directly inspect the model limitations, instead of just observing model behaviors, for people to provide feedback Stumpf et al. [2009], Ghai et al. [2020] to improve the model.

For our conditional delegation task, we hypothesize that explanations of AI model predictions, i.e., keywords that the model bases its prediction on, can give hints to people about keyword rules they should consider, and potentially help them judge the effectiveness of a given rule.

5.3.3 Distribution Shift and Experimental Studies on Out-of-distribution Examples

Current AI models rely on identifying patterns in training datasets. In a real-world scenario, it is unlikely that models are used to classify data that is exactly the same as the training dataset. For instance, a moderation team would likely work with a model trained on an existing dataset, then applied to the data on their platform. The difference between the training dataset and the deployment data is called distribution shift, which often results in a performance drop [McCoy et al., 2019, Clark et al., 2019, Jia and Liang, 2017]. For instance, McCoy et al. [2019] find that state-of-the-art models in natural language inference adopt three fallible syntactic heuristics and perform around random chance when tested on examples where these heuristics fail.

Despite substantial interest in distribution shift in the AI community, the effect has been rarely examined in empirical studies of human-AI collaboration, with a few recent exceptions [Liu et al., 2021, Chiang and Yin, 2021]. Liu et al. [2021] demonstrated that there exists a clear difference between in-distribution and out-of-distribution examples when human and AI collaborate to make individual decisions in recidivism prediction and profession prediction. They suggested that complementary performance is more plausible for out-of-distribution examples because of AI's performance drop. Chiang and Yin [2021] examined human reliance on the model in human-AI decision making and found that surprisingly humans rely on AI more out-of-distribution, where the AI performance is worse.

The existence of distribution shift is a strong motivation for some form of conditional delegation so that humans can identify the trustworthy regions. In our setup, however, as we conditionally delegate decisions to AI, strong AI performance in-distribution is likely more critical for the human-AI collaborative performance. We thus hypothesize that it is more challenging to identify the trustworthy regions for out-of-distribution examples because the model behavior is likely more spurious.

5.3.4 Content Moderation

Content moderation has attracted substantial interest from the research community due to its growing importance in online communities [Kiesler et al., 2012]. There is a large body of research studying the effect of moderation on community behavior, including whether one should regulate at all [Chancellor et al., 2016, Chandrasekharan et al., 2017, Srinivasan et al., 2019, Jhaver et al., 2019b, Chang and Danescu-Niculescu-Mizil, 2019, Seering et al., 2017]. In contrast, our work is concerned with the practice of content moderation, i.e., how moderators can efficiently deal with a large number of comments. The scale of content is the most important argument for some form of automation in content moderation [Gillespie, 2018, Gorwa et al., 2020]. Moreover, an active line of research has investigated the "emotional labor" of moderation work by the volunteer moderators [Dosono and Semaan, 2019, Matias, 2016, Roberts, 2014], further highlighting the importance of avoiding burnout for moderators through automation.

One strategy is to use rule-based methods. For instance, Reddit moderators can configure an AutoModerator bot to set rules for reporting or deleting all comments that contain certain words.[2] The key advantage of this method is that it is entirely under the control of moderators. Through interviews with 16 moderators, Jhaver et al. [2019a] found that AutoModerator improves the efficiency of moderation. However, there exists a need for audit tools to monitor the performance of the keyword rules. They also highlight the fact that AutoModerator fundamentally changes the work of moderators and may introduce additional unnecessary work. Chandrasekharan et al. [2019] also found that hard-coded rules are prone to mistakes.

An alternative strategy is to use AI models beyond rule-based approaches. Toxic comment detection or hatespeech detection has attracted a lot of interest from the AI community [Wulczyn et al., 2017, Qian et al., 2019, Wiegand et al., 2019, Nobata et al., 2016]. Notably, the Perspective API is reportedly used by the New York Times, Disqus, and other platforms.[3] However, researchers increasingly recognize the pitfalls of full automation: 1) models are trained with historical data and

[2] `https://www.reddit.com/wiki/automoderator`.

[3] `https://www.perspectiveapi.com/case-studies/`.

can present issues such as gender bias and racial bias in AI models [Sap et al., 2019, Park et al., 2018], potentially exacerbating structural inequalities [Blackwell et al., 2017]; 2) there exist diverse rules and preferences of austerity and value in different communities [Chandrasekharan et al., 2018, Fiesler et al., 2018, Scheuerman et al., 2021, Smith et al., 2020]. Anecdotally, we deployed a version of our model on a subreddit to report comments that are predicted as toxic, and the moderators asked us to shut it down due to high false positive rates (i.e., low precision). Inspired by the diversity of rules, Chandrasekharan et al. [2019] proposed a new system that combines classifiers based on different communities and advocated that this tool be configured as part of moderation workflow.

Our effort represents a new direction in exploring the mixed initiative in content moderation. Conditional delegation combines traditional rule-based approaches and AI models by providing moderators with the ability to decide when to trust or distrust the AI model. Such rules can be created for any model of choice, so it is orthogonal to the research on improving the capability of AI. It can also be used to tailor different requirements of precision and tune the tradeoffs between false positives and false negatives.

5.4 AI Model

A critical component of our study is the model used to assist people in content moderation. In this section, we present details of how we obtain the model used in our study and provide an overview of its properties.

5.4.1 Model Development

Current AI models are driven by the data used to train the model. We choose two datasets to simulate the in-distribution and out-of-distribution scenarios. We then develop an interpretable model that is trained on the in-distribution data and achieves reasonable performance on the out-of-distribution data.

Data. In this work, we use a dataset of Wikipedia comments Wulczyn et al. [2017] (henceforth

WikiAttack), made public by Wikipedia and Google Jigsaw. Notably, Jigsaw powers the Perspective API[4] , a popular free service for toxic comment detection. Therefore, using a model derived from this dataset allows ecological validity to our study as the dataset is used by real-world social media platform and community moderators. We use the original train/test split of Wulczyn et al. [2017], resulting in 70k comments in the training set and 23K comments in the test set. We use the test set to evaluate the ability of participants to create keyword-based rules for conditional delegation for **WikiAttack**.

To simulate the out-of-distribution scenario, we use another dataset of hate speech on Reddit Qian et al. [2019], consisting of 22K comments, on which we apply the same model mentioned above. As a result, the datasets that participants explore to create rules are of comparable size between Wikipedia (in-distribution) and Reddit (out-of-distribution). Throughout the rest of the paper, we will use **in-distribution** and **WikiAttack**, **out-of-distribution** and **Reddit** interchangeably.

Model. We use a rationale-style neural architecture [Lei et al., 2016] as the classifier underpinning our tool, producing both explanations and predictions. Figure 5.2 illustrates our model architecture. It uses one text encoder to identify rationales (i.e., a subset of tokens) from the input, and another text encoder to make predictions based on the rationales.[5] Trained in tandem with a sparsity objective on the rationales, this model attempts to obscure as much of the input as possible while still leaving enough to make an accurate classification. In short, this model achieves competitive accuracy while having the ability to provide explanations directly by showing the rationales on which the prediction is based on.

For the generator and predictor, we use independent, pretrained BERT [Devlin et al., 2019] instances distributed by HuggingFace [Wolf et al., 2019]. We use Pytorch Lightning[6] for fine-tuning. We use Gumbel Softmax Jang et al. [2016] to enforce a binary constraint on the predicted rationale, such that a token is either fully included or fully excluded from the input. As an implementation detail, we find it highly useful to pre-fine-tune the predictor layer on the full (un-masked) input

[4] https://www.perspectiveapi.com/

[5] Technically, the predictor uses the masked input.

[6] `https://www.pytorchlightning.ai/`.

before further training it in tandem with the generator.

Because our task emphasizes precision over accuracy, we experiment with different parameters to trade-off precision and recall. Figure 5.3 shows model performance both in-distribution and out-of-distribution with different parameters. We observe a clear performance drop out-of-distribution (e.g., F1 drops from about ∼0.8 to ∼0.6), which validated our choice of Reddit as an out-of-distribution scenario. In our experiments, we choose the model with recall weight 0.5 (the second bar). Note that participants did not have access to this performance data because our goal is to simulate the scenario where moderators work with a model developed on an existing dataset. It is up to the moderators to figure out how well the model performs and when to trust or distrust the model.

This model can achieve BERT-like accuracy while being able to precisely and parsimoniously identify the rationale responsible for its prediction. Table 5.1 presents example rationales for comments that are predicted toxic, both correctly and incorrectly. We find qualitatively that the model produces sensible rationales in this application. While it identifies some surprising tokens as toxic such as "you", it does succeed in learning that the primary evidence of non-toxicity is a lack of toxic tokens: it retains only 2% of tokens on average for predicted-nontoxic comments, versus 15% for predicted-toxic comments. Note that this explanation method has attracted some criticism for producing rationales that don't necessarily align with human reasoning [Zheng et al., 2021], but it has the advantage of producing rationales that are, by construction, sufficient (in the logical sense) for the model's prediction. Generating high-quality explanations is an active area of research, and our paradigm of conditional delegation can be used for any model of choice.

The rationale produced by this model is a form of **local explanations**, identifying important words in each prediction. Our experiment also includes **global explanations**, which convey an overview of the model behavior across all inputs. We generate these global explanations by identifying the tokens that occur most frequently in the rationales of the model on the in-distribution and out-of-distribution data respectively. We display these top-15 most frequent rationale tokens (Table 5.2) as the "global explanations". We can immediately observe differences between Reddit and WikiAttack: “cunt” and “retard” are not common in rationales in WikiAttack but are among the top five on Reddit.

To provide context for our later findings, it is difficult to produce precise classification out-of-distribution. Figure 5.4 shows that even with a high positive class probability threshold (0.93), the model only climbs to 0.58 precision. Thus, even a very conservative application of this model is still producing 2 false positives for every 3 true positives–impractical for use by real moderators, and something we would like to be able to improve on via conditional delegation.

5.4.2 Performance of Individual Words

The goal of this work is to explore human-created keywords rule for conditional delegation to AI, such as “if a comment contains word X and is predicted toxic, the model will be trusted to report the comment for moderation action”. In comparison, with a manual rule-based approach (e.g., the current AutoModerator system used by content moderators of Reddit), such a rule takes a form like “if a comment contains word X, that comment will be reported”. Our hypothesis is that with the proper choice of rules, humans can produce a system which is more precise than either the manual rule-based approach or the model working alone.

We perform preliminary analysis to characterize the scope of the potential improvement and to contextualize our experimental results. A crucial question in motivating our approach is whether there exist trustworthy regions of the model, i.e., are there certain words that occur systematically in comments where the model achieves high precision.

First, we compute the precision of conditional delegation for all words that show up in at least 100 comments. Figure 5.5 shows the 10 words with the highest precision as conditional delegation rules (i.e., based on model predictions for all comments containing them) on WikiAttack and Reddit respectively. Conditional delegation based on these words leads to greater precision than the model working alone (dashed lines), suggesting that users can improve the precision of the model by identifying these words for conditional delegation. In addition, we compare that with the precision of using the word as a “report all” rule as with manual rule-based approach, by considering all comments containing the word as toxic. We can see generally, for these words with top precision, trusting the model leads to higher precision than “report all”, both in-distribution and out-of-distribution. However, the difference is much smaller for Reddit (out-of-distribution). In particular, “faggot”, “cunt”, “retard”, and “nigger” achieve very high precision on this dataset even if one simply reports all comments that contain any of those words. These results indicate that conditional delegation can outperform both the manual rule-based approach and the model working alone if users are able to make good choices of keywords rules.

Next, we examine the precision of the words that are most frequent in rationales (Table 5.2, to be shown as global explanations). Figure 5.6 shows that on WikiAttack, the majority of global explanations achieve greater precision than the model working alone. However, on Reddit, this is true only for six words (“cunt”, “retard”, “faggot”, “bitch”, “she”, “her”), indicating the challenge of creating good rules for out-of-distribution data.

Figure 5.7 further shows (the number of reported toxic comments - the number of reported non-toxic comments) if that word is chosen as a rule. This measure reflects both the coverage and precision of a keyword rule. We refer to this measure as **reward** because it is used as an incentive in our human subject experiments, to be introduced later. On WikiAttack, most global explanations lead to positive rewards. We also observe the clear advantage of using conditional delegation over “report all”. This advantage becomes smaller on Reddit and disappears for “retard” and “cunt” because these two words have great precision and coverage by simply reporting all comments with them. In fact, rewards on Reddit are dominated by “retard” and “cunt” due to their high coverage. A user could achieve a quite high reward (and outperform the model) simply by reporting all comments with either of these two words.

In summary, there are specific words that delineate trustworthy regions of the AI model, and even certain words (“retard”, “cunt”) where simply reporting all comments containing these words would be more effective in terms of reward than delegating such comments to the model, particularly in the out-of-distribution setting. However, we are able to recognize such words with the benefit of a fully-labeled dataset (i.e., oracle access) — discovering them in a real-world setting could be very challenging. We explore this challenge in a rigorous human subject experiment in the next section.

5.5 Experimental Design

Equipped with the model, one goal of this study is to design interfaces with different support features to enable people to come up with effective keyword-based rules for conditional delegation. We then examine the effect of these support features through a human-subject experiment. In this section, we start by introducing different types of support features that we consider and then explain the study procedure.

5.5.1 Experimental Conditions and Interface Design

In order to enable people to create keyword-based rules for conditional delegation and observe model behaviors with them, the basic function of our tool is to search for a keyword and browse comments that contain it. This allows users to determine whether a keyword would serve as a good rule. For different experimental conditions, our design space mainly involves what information we provide when returning the search results.

Experimental conditions. As discussed in §5.4, in addition to predicting whether a comment is toxic or not, our model can provide local explanations (i.e., which words are used as rationales for the prediction) as well as global explanations (i.e., most frequent words that show up in the rationales). Therefore, we consider the following four conditions:

- **Predicted labels.** Predicted labels are shown along with the searched comments.

- **Predicted labels + local explanations.** In addition to predicted label for each comment, we highlight rationales, i.e., words in the comment the model uses to determine toxicity for comments that are predicted toxic. We refer to this condition as "*local explanations*".

- **Predicted labels + local explanations + global explanations.** Participants have access to all of the features in the previous condition and are also provided a list of words that the model typically uses in determining comment toxicity (Table 5.2). We refer to this condition as "global explanations".

- **Manual condition.** The final condition is designed to simulate the current state of AutoModerator, where moderators come up with "report all" rules. We create a consistent interface where participants have the ability to search comments and browse returned results to assess whether they are indeed toxic, instead of whether the model prediction is precise. Participants do not have access to any model-related information.

In the rest of this paper, we refer to these conditions as **experimental conditions** and WikiAttack vs. Reddit as **distribution types**.

Interface design. We start by introducing the interface for "Predicted label + local explanations + global explanations", which includes all possible components of the other conditions (see Figure 5.8). The widgets in the interface are arranged in two columns, where instructions and comments are displayed on the left, while the search bar and the current set of rules are on the right. The instructions box (Figure 5.8(1)) reminds participants of the task and provides more information about the interface to ensure that they can fully leverage the tool's features. Global explanations are shown below the instructions box (Figure 5.8(3)). When the participant clicks on a rule that is represented by a button, it automatically searches comments with the respective keyword-based rule. In addition to searching particular words, we also allow users to load random comments, which can be used to explore the data (Figure 5.8(4)). Upon a query or loading random comments, the comments are displayed as *cards* below the *load random comments* button. Depending on the condition, a comment could have a predicted label (Figure 5.8(5)) and rationales could be highlighted (Figure 5.8(6)).

On the right side, the first two widgets are the *search bar* (Figure 5.8(7)) and *clear* button (Figure 5.8(8)). The participant enters keyword-based rules and then comments with the respective rule are shown on the left, as described in the previous paragraph. Participants can filter comments by their predicted label (Figure 5.8(9)). By default, both predicted toxic and nontoxic comments are shown. When the participant is satisfied with the rule, they may click on the *add rule* button (Figure 5.8(10)) to add the rule to their list. All of the participant's rules are displayed in the component below the *add rule* button. We also display their total matched comments and predicted toxic matched comments (Figure 5.8(11)). Finally, participants may click on the *finish making rules and go to survey* button to submit their rules and proceed to the exit survey (Figure 5.8(12)).

Figure 5.8 shows interfaces for the other three conditions. "Predicted labels + local explanations" condition (Figure 5.9b) removes the global explanations (Figure 5.8(3)) and is otherwise the same as "Predicted labels + local explanations + global explanations". "Predicted labels" condition (Figure 5.9a) further removes the highlights of rationales when returning search results. "Manual" condition (Figure 5.9c) only has "search" and "load random comments".

We conducted pilot studies with lab mates at the same university and mechanical turkers to gather feedback and ensured that interactions were intuitive and the interface was easy to use. With many widgets in the interface, it could potentially become overwhelming for the user. To overcome that problem, we included a short tutorial walk-through (Figure 5.8(2)) that explains each widget in sufficient details. To minimize confusion, we also included a separate link in the instructions box (Figure 5.8(1)) that enumerates how each widget should be used.

5.5.2 Study Design and Procedure

We employ a between-subject design with the four conditions described above. We aimed to obtain 0.95 power to detect an effect size of 0.1 in precision at the standard 0.01 alpha error probability using F-tests. As such, we recruited 30 participants for each condition, summing up to 120 participants each for in-distribution and out-of-distribution and 240 participants in total. The study is approved by the University IRB.

Task flow. Participants went through four phases during the study: (1) consent and introduction; (2) interface tutorial; 3) the actual task; (3) exit survey. During the introduction phase, participants were shown an accordion of instructions. Due to the nature of toxic comments, they are warned that the task will contain profanity and hate speech and are cautioned to only proceed if they are okay with reading toxic comments.

After the instruction, participants are asked three attention-check questions to allow us to filter out participants who did not pay attention to the instructions. They are given toxicity trial questions with four simple comments to label if they are toxic or nontoxic: (1) *I had a really bad day today.* (2) *I just made a fucking great steak!* (3) *Go fuck yourself.* (4) *I hope you die painfully tomorrow.* These questions allow us to filter participants who finish this task hastily without paying effort and prepare them for toxicity judgment. We remove participants whose accuracy is less than or equal to 50% on these questions. As a result, we filter 6 out of 240 in our analysis.

To familiarize participants with the interface, we include a tutorial walk-through when they first land on the page to instruct them on how to use each feature. We also include a link that featured more detailed instructions and a short demonstration video. Participants are required to submit at least ten rules, and can then exit the task whenever they are satisfied with the set of rules created.

In the exit survey, we collected basic demographic information, their knowledge and familiarity of AI and content moderation, and subjective measures, to be introduced in the later section.

Reward. To motivate quality work, in addition to a base payment, we design a bonus incentive as follows: participants will be awarded \$0.10 for every 100 toxic comment their rules correctly reported, and penalize them \$0.10 for every 100 nontoxic comment their rules mistakenly reported (lower bounded by \$0 and upper bounded by \$2). This bonus thus rewards both precision–how likely comments under the rule (for the manual condition) or conditional delegation with the rule are correctly classified as toxic, and coverage–the quantity of comments covered by the rule. To make the calculation easy to understand for participants, this reward makes a simplified assumption that the cost of wrongly reported non-toxic comments (false positive) equals the benefit of correctly reported toxic comments (true positive).

This reward mechanism is explained to participants, and we include one question in attention check to ensure they understood it. We also explicitly suggest that, to optimize for the reward, their goal should be to come up with keywords that meet the following criteria: (1) that occur in a lot of comments; (2) with which the model makes accurate predictions on the comments, and (3) that are a diverse set so they may cover different kinds of toxic comments.

Participant information. We recruited participants from Amazon Mechanical Turk. We note that while this recruiting choice may limit the generalizability of our results, social media content moderation is often performed by part-time volunteers whose expertise varies. Furthermore, we believe turkers are a sufficiently good sample for us to compare whether conditional delegation improves the content moderation outcomes over the baseline condition, and expert users are likely to further enhance the improvement pattern, if any. We encourage future work to further test the paradigm in realistic social media contexts.

To ensure high quality responses, all participants satisfy the following criteria: (1) performed at least 1000 HITs; (2) approval of 99% performed HITs in previous requesters; (3) reside in the United States; 4) has the adult content qualification since our task shows toxic comments. The experiment follows a between-subject design therefore we do not allow any repeated participants.

There were 115 male, 116 female, 2 non binary, and 1 preferred not to answer. 52 participants are aged 18-29, 114 aged 30-39, 34 aged 40-49, 26 50-59, 7 aged over 59, and 1 I prefer not to answer. Participants rated their knowledge on artificial intelligence (25 had no knowledge, 156 had little knowledge, 49 had some knowledge, 4 had a lot of knowledge), and social media content moderation (36 had no knowledge, 113 had little knowledge, 66 had some knowledge, 19 had a lot of knowledge) on five-point Likert scales. Participants were paid an average wage of $11.80 per hour.

Overall, most turkers are satisfied with our task design and interface. Here are two quotes from their feedback: "*This was super intriguing. I had never participated in an activity like this before. It was hard coming up with bad words since they are not part of my vocabulary. It was interesting to see which words usually coincided with toxic subjects. Overall, very interesting project.*" and "*It was interesting. I see now how difficult moderation can be for some sites.*"

5.5.3 Evaluation Measures

We consider three types of evaluation measures to cover efficacy, efficiency, and subjective perception.

Efficacy. As discussed in §5.2, our main goal is to examine whether humans can improve the precision of the model with a good coverage via conditional delegation. We consider two precision-based measures: *average precision* and *union precision*. For the first three experimental conditions with delegation support features, average precision is formally defined as

$$\frac{1}{|R|}\sum_{r\in R}\frac{|\{x \text{ is toxic } \& \ x \text{ contains } r \ \& \ x \text{ is predicted toxic}\}|}{|\{x \text{ contains } r \ \& \ x \text{ is predicted toxic}\}|},$$

where R is the set of rules that participants choose and x refers to a comment, whereas union precision is formally defined as

$$\frac{|\{x \text{ is toxic } \& \ x \text{ contains any } r \in R \ \& \ x \text{ is predicted toxic}\}|}{|\{x \text{ contains any } r \in R \ \& \ x \text{ is predicted toxic}\}|}.$$

As the manual condition does not have a model, these two definitions become $\frac{1}{|R|}\sum_{r\in R}\frac{|\{x \text{ is toxic } \& \ x \text{ contains } r\}|}{|\{x \text{ contains } r\}|}$ and $\frac{|\{x \text{ is toxic } \& \ x \text{ contains any } r\in R\}|}{|\{x \text{ contains any } r\in R\}|}$.

The difference in the denominators highlights the role of conditional delegation, which only affect the comments that the model predicts as toxic. It follows that the performance with conditional delegation is also determined by the model's base performance, i.e., how well the model can identify toxic comments. Intuitively, average precision reflects the average quality of every single rule a person provides, while union precision measures the performance when using all rules from the person as a set, and can be skewed by the performance of higher-coverage rule in the set. Thus, one's ability to come up with both high-precision and high-coverage rules can lead to better union precision.

Finally, we consider the *reward* participants received, as introduced in §5.5.2, which measures the quantity difference between reported toxic comments (true positive) and reported non-toxic comments (false positive). This measure reflects both precision and coverage. This metric is highly volatile because a small number of keywords can achieve much higher rewards than others, especially out-of-distribution (e.g., “retard” and “cunt” on Reddit as shown in §5.4). We believe that precision is the more reliable measure of efficacy given that our participants tended to only choose about 10 rules.

Engagement and efficiency. We consider number of logged actions a participant took during the experiment task and number of rules they added as measurements for engagement. 13 types of unique actions were logged, including searching a rule, filtering comments by predicted labels (toxic and nontoxic), load random comments, get page comments, etc. We consider the number of actions more indicative of engagement, since participants can search for a rule without adding it. For efficiency, we consider total elapsed time and rules per minute. Elapsed time starts from the moment participants enter the interactive interface until they click on “finish making rules and go to survey”, in minutes. Rules per minute is the number of rules added divided by elapsed time. Since rules are the final product of the task, rules per minute is more indicative of efficiency.

Subjective measures. Finally, we consider the following three categories of subjective perception, all gathered by the exit survey, using a five-point Likert scale (Strongly Disagree to Strongly Agree) for all scale items.

- Subjective workload. We adopt three applicable items from NASA-TLX [Hart, 2006]:
    - **Mental demand.** I felt that the task was mentally demanding.
    - **Feelings of success**. I felt successful accomplishing what I was asked to do.
    - **Negative emotions.** I was stressed, insecure, discouraged, irritated, and annoyed during the task.
- Confidence. There are multiple loci of confidence in this task: in the model, in one’s own ability to create conditional delegation rules, and in the human-AI collaborative outcome. So we consider the following three measure (they were not asked in the manual condition since they do not apply):
    - **Confidence in model.** I trust the model to be able to correctly identify most toxic

comments.

- ∗ **Confidence in created rules** I am confident that my rules significantly improve the model's accuracy in detecting toxic comments.
- ∗ **Confidence in deployment.** I am confident that my moderator team would feel comfortable relying on the AI model combined with the rules I provided.

- Understanding. We are interested in whether global and local explanations could improve people's perceived understanding of the model. We consider both the global understanding of the AI model as a whole and the local understanding on the rationales behind predictions. These questions were skipped in the manual condition.

  - ∗ **Understanding of model.** I felt that I had a good understanding of how the AI works.
  - ∗ **Understanding of prediction.** I felt that I had a good understanding of why the AI identifies a comment to be toxic.

5.6 Experiment

We report results based on the three sets of evaluation measured described above: efficacy, efficiency and engagment, and subjective measures. We refer to the WikiAttack task as in-distribution and Reddit task as out-of-distribution and the terms will be used interchangeably.

5.6.1 Efficacy

Even lay people are able to create rules with higher precision than the model working alone, both in-distribution and out-of-distribution (see Figure 5.10 and Figure 5.11). To determine whether participants are able to create rules that improve model precision, we conduct t-test on the precision of conditional delegation with human-created rules vs. the model working alone. We find that differences are all statistically significant (p <0.001), both on WikiAttack (in-distribution) and Reddit (out-of-distribution), based on average precision and union precision.

In particular, on WikiAttack, the model working alone already outperforms the manual condition, and conditional delegation further improves the precision. These observations demonstrate that humans, in our case turkers who are not experts in content moderation, are able to create rules that improve model precision, suggesting that conditional delegation can be a promising direction to pursue.

Next, we examine the effect of distribution types and experimental conditions on precision. We conduct two-way ANOVA of distribution types and experimental condition in average precision and union precision. We find significant effects in distribution type, experimental condition, and their interaction (p <0.001). The effect of distribution type is the most salient, suggesting a clear difference between in-distribution and out-of-distribution.

Given the significant interaction, we further conduct one-way ANOVA to understand the effect of experimental condition on performance separately for WikiAttack and Reddit, and if significant, conduct post-hoc analysis using Tukey's HSD. For average precision (Figure 5.10), experimental condition has a significant effect in WikiAttack (p <0.001), but not in Reddit (p =0.864). Post-hoc Tukey's HSD shows that the manual condition is significantly worse than all other experimental conditions with delegation support features (p <0.001) on WikiAttack. For union precision (Figure 5.11a and 5.11b) (using rules created by a participant as a set), experimental condition significantly affects performance in both WikiAttack and Reddit (p <0.001). Post-hoc Tukey's HSD shows on WikiAttack, the manual condition is significantly worse than other experimental conditions with delegation support features (p ¡ 0.001). On Reddit, we found the global explanation condition is worse than the manual condition (p =0.004), and only showing prediction labels (p =0.028).

Finally, we examine the effect of distribution types and experimental conditions on reward. Two-way ANOVA finds a statistically significant effect of distribution type and interaction between distribution type and experimental condition (p <0.001). Therefore, we conduct one-way ANOVA to understand the effect of experimental condition on reward separately for WikiAttack and Reddit. On WikiAttack (Figure 5.11c), we find a statistically significant effect of experimental condition

(p =0.01), and post-hoc Tukey's HSD shows that the manual condition is significantly worse than other experimental conditions with delegation support features (p <0.001 for global explanations, p =0.007 for local explanations, and p =0.004 for predicted labels). On Reddit, the experimental condition also has a statistically significant effect (p =0.018). Post-hoc Tukey's HSD shows that only the difference between the manual condition and global explanations is significant (p =0.018).

These results show that, on WikiAttack, where the model performs well, people can easily identify rules with both high precision (average precision) and with high coverage (reflected by union precision and reward), as long as predicted labels are provided, achieving complementary performance. But on Reddit, where the model's base performance is significantly worse, it is more challenging to achieve better human-AI performance over the manual rule-based approach. It follows that in both situations, we do not observe that explanations, either local or global, significantly improve the performance of conditional delegation. However, adding the global explanation feature seems to unexpectedly hurt people's ability in choosing rules with both high coverage and high precision, and lead to slightly lower human-AI performance in union precision and reward.

What rules do people make? To further make sense of their performance, we dive into the content of the rules provided by participants. Table 5.3 lists the top rules in each condition along with the percentage of people who chose that rule. On WikiAttack, participants with delegation support features are more likely to choose "fuck" (above 60%), a high-precision rule to identify toxic content in Wikipedia comments as shown in §5.4, than the manual condition (only 36.7%). In comparison, for Reddit, "fuck" is a less precise rule (i.e., people also use the word in non-toxic comments). The word does not show up in top 10 for the manual condition, but shows up in the other conditions.

This observation suggests that the delegation support features can help users identify good rules when the model performs well, but may mislead people when the model performs poorly. The reason that global explanations slightly hurt the performance in union precision and reward could be that participants were led to choose some high-coverage rules with relatively low precision such as "fuck", which was listed in the global keywords (Figure 5.8).

Figure 5.12 further shows the top words in reward among the rules created by participants. In addition to “fuck” on WikiAttack and “cunt”/“retard” on Reddit, the result highlights the advantage of conditional delegation. Users can achieve high reward by trusting the model beyond swearing words, for instance, “you” on WikiAttack and “her” on Reddit. The reason is that the AI model excels at deciding whether “you” is used for personal attack or simply for referring purposes.

Summary. In short, our results demonstrate that conditional delegation is more effective than the model working alone, and that even laypeople are able to create high-quality conditional delegation rules for content moderation. Compared to the manual rule-based approach currently used for content moderation, advantage of our human-AI collaborative approach via conditional delegation may depend on the base performance of the AI, and may not be sufficient if the AI significantly under-performs, e.g., when used on out-of-distribution comments. Further research is required to understand the necessary conditions for conditional delegation to outperform the manual rule-based approach. Our analysis did not find evidence that model explanations could help people create better rules for conditional delegation. We explore their benefits for other aspects of user experience later.

5.6.2 Efficiency and Engagement

We conduct two-way ANOVA to determine whether distribution type and experimental condition have a significant effect on user engagement (number of actions, number of rules) and efficiency (elapsed time, rules per minute). In all evaluation measures of engagement and efficiency, we only find statistically significant effects of experiment conditions, suggesting that patterns with two distribution types are comparable. Therefore, in this section, we merge the data on WikiAttack and Reddit, and report results from one-way ANOVA on experimental conditions.

Participants working on conditional delegation are more engaged (see Figure 5.13). Figure 5.13a shows that participants with all delegation support features were much more engaged than the manual condition. In particular, predicted labels only condition incurred many more actions than other conditions. One-way ANOVA also finds a statistically significant effect in experimental condition ($p < 0.001$). Post-hoc Tukey's HSD shows the negative difference between the manual condition and other experimental conditions are all statistically significant (p <0.001 for predicted labels and global explanations, p =0.009 for local explanations).

When it comes to number of rules, the outcome of task engagement, the difference is not as salient. Because we require a minimum of 10 rules, every condition leads to a little above 10 rules: the manual condition is just above 10 at 10.6, while global explanations leads to 12.3 rules. That said, one-way ANOVA still finds a significant effect in experimental condition (p =0.021). Post-hoc Tukey's HSD shows that the difference between global explanations and the manual condition is statistically significant (p =0.028).

Explanations improve task efficiency (see Figure 5.14a and 5.14b). Figure 5.14a shows the time spent on the interactive interface in each condition: participants with predicted labels only spent the most time on this task. This result is consistent with the number of actions because it likely requires more time to take more actions. However, the difference is relatively weak: one-way ANOVA shows that the effect of experimental conditions is only borderline significant (p =0.074), and post-hoc Tukey's HSD do not find any statistically different pairs.

Rules per minute is a better measure of efficiency since it reflects how long it takes for people to identify a rule that they are satisfied with. The trend is reversed from the time spent: predicted labels only lead to the lowest number of rules per condition, however, with the help of explanations, humans are able to achieve a greater number of rules per minute. One-way ANOVA confirms a statistically significant effect of experimental condition (p =0.008). Post-hoc Tukey's HSD suggests that the only statistically different pair is predicted labels only and global explanations (p =0.031), suggesting that global explanations helped participants to achieve the highest efficiency to come up with rules.

Global explanations lead to much higher overlap between most frequent words in rationales (Figure 5.14c). To understand this improvement in efficiency, we examine the overlap between human-created rules and the most frequent words in rationales (Table 5.2), which are the words shown in global explanations and also more likely to have appeared in the highlighted words in local explanations. Figure 5.14c shows that global explanations lead to much higher overlap than the other conditions. This observation confirms that global explanations provide direct hints for possible rules, thus improved the efficiency to come up with required number of rules.

Summary. Taken together, our results show that people are more engaged when performing conditional delegation than working on creating manual rules, with more actions and a tendency to spend more time on the task. This tendency comes with a cost of efficiency in creating rules when only showing predicted labels. Showing model explanations, especially global explanations, can significantly improve the efficiency, resulting in comparable efficiency between conditional delegation and the manual rule-based approach. The reason can be attributed to participants leveraging keywords in explanations as hints to create delegation rules. Future research is required to explore means to encourage people to examine the performance of these hinted rules more carefully, to improve both efficiency and efficacy.

5.6.3 Subjective Perception

Finally, we report the subjective perception of participants (subjective workload, confidence, and perceived understanding of AI). All results are based on answers in exit survey, with a five-point Likert scale.

Subjective workload. Overall, participants were neutral about whether the task was mentally demanding (M=3.15, SD=1.08), agreed that they felt relatively successful in accomplishing the task (M=3.95, SD=0.91), and disagreed that they felt negative emotions (M=1.93, SD=1.03). Two-way ANOVA does not show any statistically significant effects of distribution types and experimental conditions. For WikiAttack, we observe a weak trend that local explanations lead to less subjective workload (lower mental demand, more feeling of success, and less negative emotion) while adding global explanation has the opposite effect. These patterns, however, do not hold for Reddit.

Confidence. Overall, participants reported relatively strong confidence in all of our measures: confidence in the model (M=3.63, SD=1.01), confidence in the rules they created (M=3.73, SD=0.98), confidence in the deployment of the system from human-AI collaboration (M=3.61, SD=1.0), leaning towards agreeing with all these statements. We do not find any statistically significant effect of distribution type and experimental condition with two-way ANOVA. There is a non-significant trend that conditions with explanation, especially local explanation, result in better confidence for WikiAttack, but not for Reddit, where the model performs relatively poorly.

Perceived understanding. Overall, participants report a good global understanding of the model (M=3.37, SD=0.97) and local understanding of individual predictions (M=3.56, SD=1.05) on WikiAttack than Reddit, possibly related to the difference in model performance between distribution types. Two-way ANOVA only shows a marginally significant effect of distribution type in global understanding of the model (p =0.063). It is somewhat surprising that model performance leads to this difference in perceived understanding. Interestingly, there is a trend that local explanations lead to better perceived local understanding on predictions, but worse perceived global understanding on the model, but not when global explanations are added.

Summary. Overall, subjective measures show a relatively positive experience across board, but not strong difference between conditions. There is some evidence that when the model performs well (WikiAttack), local explanations provide the best experience: strong performance with relatively high efficiency, less subjective workload and more confidence in the outcomes.

5.7 Conclusion

Through investigating the three research questions introduced in the beginning, our study shows the promise of conditional delegation as a new paradigm for human-AI collaboration. Even with crowdworkers who are not experts of the content moderation task, conditional delegation can achieve better performance than the model working alone. However, whether the human-AI collaboration can outperform the manual rule-based approach varies for in-distribution and out-of-distribution AI. Out-of-distribution AI has significant performance disadvantage that cannot be adequately compensated by conditional delegation. We also found that, in general, providing predicted labels with our keyword search based interface is sufficiently effective in supporting people to create delegation rules. Providing explanations can improve efficiency by hinting on rules to consider, but can also mislead people to use high-frequency but not necessarily high-precision rules. We discuss implications of these results below.

The promise of conditional delegation. Our study is a first step towards understanding and leveraging the promise of conditional delegation. It is an intuitive approach that can be used in a wide variety of domains so that users can proactively decide when to use an AI model and in what ways based on the output of the AI. For instance, judges can specify when to show the risk estimates for recidivism prediction and when to hide the model output. Doctors can identify subsets of patients for which they rely on AI to send alerts. Our study only explored one type of workflow and one type of action. Different applications may require a diverse set of workflows and actions, and have varying tradeoffs between false positives and false negatives.

Moreover, we only begin to define the design space for supporting users in conditional delegation. An essential requirement is to help users make sense of model behaviors under different delegation conditions. Keyword-based rules are a reasonable approach in content moderation given that rule-based methods are already used in AutoModerator. We used a rationale-style model to facilitate this kind of interaction, although we expect post-hoc explanation methods to play similar roles. It is important to empirically validate the effect of the underlying models and explanation methods. We also focused exclusively on conditional delegation based on trustworthy regions in the input space. A promising direction is to investigate the joint effect of delegation based on inputs and outputs [Chandrasekharan et al., 2019, Keswani et al., 2021]. Another limitation of our study lies in that crowdworkers only came up with about 10 rules because of our minimum requirement. Although our results are encouraging with non-expert users, additional experiments are required to validate the potential of conditional delegation with expert users. Notably, future research can explore means to facilitate people to identify a greater number of rules, examine the combined effect of rules, and monitor the performance of rules after model deployment. In long-term deployment, it is especially valuable to investigate how to update the delegation conditions once the model is updated.

Additionally, conditional delegation can potentially alleviate AI bias as we give users the freedom to choose trustworthy regions based on their domain knowledge or notion of fairness. However, the flip side is that this process could introduce human bias, if for example one's notion of fairness is ill conceived. We encourage future work to understand and develop ways to rail-guard the impact of human biases in conditional delegation.

The effect of distribution shift. Our results highlight the importance of considering the effect of distribution shift in designing experiments on human-AI collaboration, to better understand the generalizability of results. We are able to achieve complementary performance in-distribution on WikiAttack, but not out-of-distribution on Reddit. In practice, it is rarely the case that an AI model faces exactly the same distribution in deployment. Therefore, it is critical to understand the outcomes in out-of-distribution contexts to understand the generalizability or applicable scope of a given form of human-AI collaboration.

It is useful to note that although our results are presented as in-distribution vs out-of-distribution, the differences are complicated between WikiAttack and Reddit. First, there exists a clear difference in model performance, so our results can be seen as comparing a high-performance model with a low-performance model. Second, the nature of comments on Wikipedia and Reddit differs substantially. It is possible that crowdworkers are more used to comments on Reddit or that common swearing words such as “retard” and “cunt” happened to work well on the Reddit dataset that we used. This complexity demonstrates that the contrast of in-distribution versus out-of-distribution contexts is not a monolithic dimension, which further adds to the challenge of experimental design to account for the effect of distribution shift.

The priming effect of explanations. While explanations can improve efficiency, global explanations are found to slightly hurt performance when working with out-of-distribution AI, as participants may have chosen the keywords in explanations without carefully examining the model behaviors with them. These observations echo concerns of unintended consequences with the use of explanations in human-AI collaboration [Lai and Tan, 2019b, Bansal et al., 2021, Green and Chen, 2019b].

In other words, for our task of creating keywords rules, keywords-based explanations have a priming effect that leads to biased adoption of presented words. Note that priming, if used appropriately, can shape user behaviors in a positive way. The challenge is that with the technique we used to generate global explanations (i.e., most frequent tokens in rationale), the top tokens do not necessarily correlate with high precision (Figure 5.6). Future work can explore techniques that can exploit some proxy of precision, such as considering the uncertainty or confidence of predictions. Another direction is to utilize de-biasing technique to mitigate the effect of priming, such as explicitly reminding people to attend to wrong predictions with the chosen keywords.

It is worth noting that local explanations seem to have less of a priming effect than global explanations but still improves efficiency. It is possible that the many highlights in search results are too scattered to have a salient effect. Future work can explore other XAI techniques or provide additional support, such as to help users have an overview of the rationales in all search results.

Implications on content moderation. It is impressive that crowdworkers can already create keyword-based rules that achieve greater precision than the model working alone. However, we recognize that our experiment setup is only a first step towards using conditional delegation in content moderation. First, crowdworkers are not representative of moderators, who have way more experience with their platform's data. As moderators are more familiar with the moderation process and more knowledgeable about important words, experts might find the interface more useful than crowdworkers. However, participatory design and future work can develop more serendipitous features. Second, in practice, moderators usually have historical data on which moderation decisions were made. This historical data can be used in the process of creating keyword-based rules. Third, prior work has shown that moderators often update the rules used by AutoModerator [Jhaver et al., 2019a, Chandrasekharan et al., 2019] and our work does not take into account any future updates. Neither do we leverage any existing rules that moderators have created. For future work, we hope to integrate a model that receives feedback from moderators and allow updates to the model to reflect the feedback. The ideal pipeline would require careful development in the model architecture and interface to refrain any unnecessary actions from interfering with moderators' tasks. Last but not least, content moderation involves a wide range of different rules beyond toxicity, and even the policies under the umbrella term of toxicity can vary, so the AI model that we uses represents a narrow component in content moderation. In short, our work uses content moderation as a testbed to illustrate the promise of conditional delegation. Much future work is required to realize the impact of conditional delegation in content moderation.

Limitations. First, our work represents one instantiation of conditional delegation. We emphasize precision and coverage to increase the ability of moderators to deal with a large amount of comments ("true positives") while minimizing unnecessary labor for moderators ("false positives"). This tradeoff between true positives, true negatives, false positives, and false negatives can vary in practice depending on the application and the actions taken according to AI predictions. Second, our participants are not representative of content moderators. It also follows that our evaluations are limited by the number of rules that participants created in about 10 minutes. Our case study shows the promise of conditional delegation, but further study is required in each application domain of interest to develop the best design for human-AI collaboration in identifying delegation conditions. Third, our choice of model, datasets, and explanations affect the experimental outcome. It is important to further dissect the relevant dimensions and investigate the effect of alternative choices.

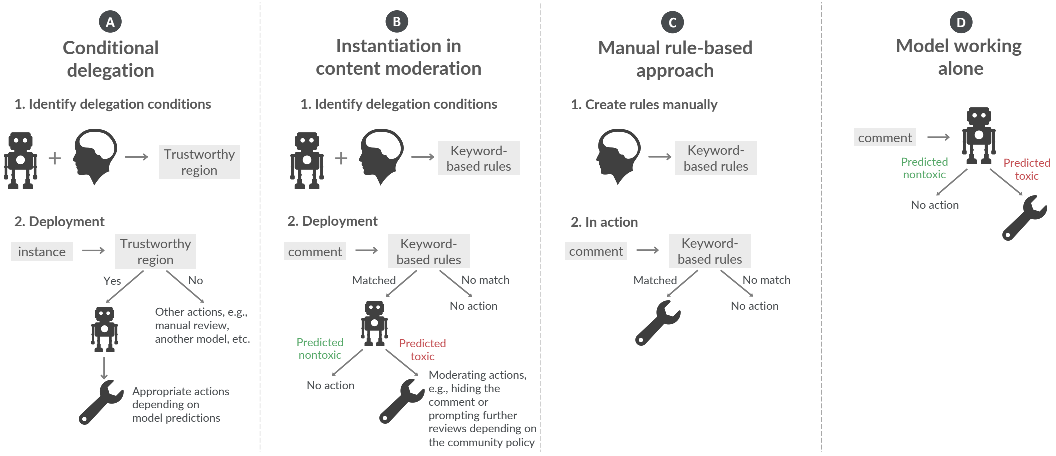


Figure 5.1: Illustration of conditional delegation. Part A shows a general form of conditional delegation. Humans and AI work together to identify trustworthy regions of AI. Then once deployed, the AI model only affects the instances that belong to the trustworthy regions. Part B instantiates conditional delegation in the context of content moderation for this work. The right columns shows the contrast of the current manual rule-based approach for content moderation (Part C) and the model working alone (Part D).

There are four potential approaches for content moderation. They are condition delegation, instantiation in content moderation, manual rule-based approach, and just the model.

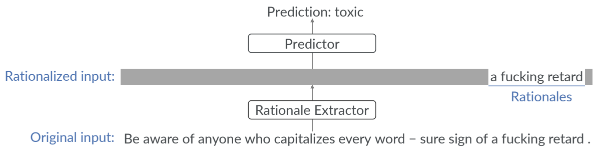
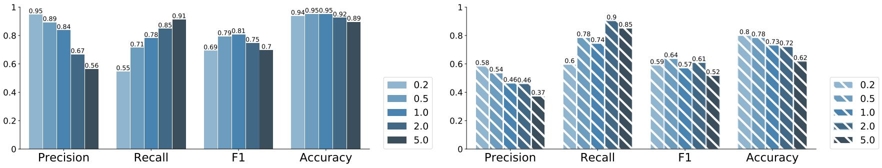


Figure 5.2: Illustration of the model with an example. The rationale extractor first identifies "rationales" in the input, and then the predictor makes the prediction based on the rationales. This model can achieve competitive accuracy while having built-in interpretability because the prediction is made exclusively based on rationales.

The model extracts rationales from the input and makes prediction based on the rationales.



(a) WikiAttack (in-distribution)
(b) Reddit (out-of-distribution)

Figure 5.3: An overview of model performance with different hyper-parameters. The hyperparameter shows the relative weight of recall vs. precision. We choose the model with 0.5 because it achieves competitive performance both in-distribution and out-of-distribution. Throughout this paper, we use solid bars for WikiAttack (in-distribution) and bars with white slashes for Reddit (out-of-distribution).

We experiment with different parameters to trade-off precision and recall. We chose the model with 0.5 as the hyperparameter value as it achieves competive performance in both datasets.

	WikiAttack
True positive	highlightRGB152,193,217highlight== Hello == Your such a fucking freak mate. Get a life. And get a facelift you ugly cunt.
True positive	highlightRGB152,193,217highlight== FUCKYOU == wow,just wow james delenoy. u are a class A faggot. i could pwn u in a coma. i dont give a fuck if u block this account cuz we are a freight train from hell and we will make more and come at u with the ferocity of a mama bear. SUCK OUR GRUNDLES, the entire waltmanist community
False positive	highlightRGB152,193,217highlight::Haha, I fucking pissed myself reading this ') Thanks, i needed that after 3 hours trecking around in warrington with the worst cold ever. I mean it, thanks))
	Reddit
True positive	highlightRGB152,193,217highlightCrawl back to you r/politics safe space. I'm sick of fucking faggot liberals coming into communities where they aren't welcome and trying to ruin them from the inside out. You're on KiA, so you already know how that works. Fuck you
True positive	highlightRGB152,193,217highlightAh yes, coming from the idiot that calls people a retard when he disagrees. Get the fuck out of your internet bubble and look around you irl for a moment.
False positive	highlightRGB152,193,217highlightSomeone in chat baited me into saying "paki". I got banned. I'm fucking Pakistani and that word has never been "offensive" to anyone. I've used it for 14 years myself. I had 1k hours in that game but it's staying uninstalled

Table 5.1: Comments that are predicted toxic from WikiAttack and Reddit with their identified rationales by our model.

WikiAttack	you, fuck, your, suck, die, shit, nigger, faggot, cock, my, bitch, stupid, go, ass, i
Reddit	you, fuck, cunt, retard, shit, your, stupid, faggot, bitch, hate, she, i, guy, her, idiot

Table 5.2: Words that are most frequently used in rationales.

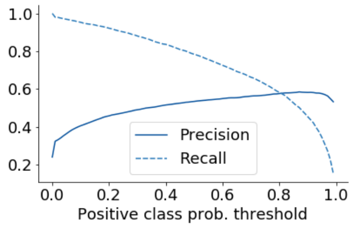
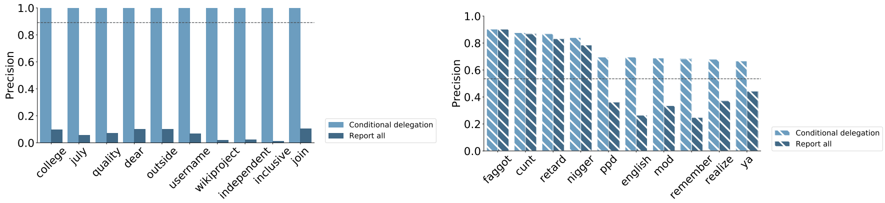


Figure 5.4: Precision-recall plot of model on Reddit dataset (out-of-distribution). Even at high positive class probability thresholds, precision remains low.

The precision-recall plot of model on Reddit dataset shows that even with high positive class probability threshold of 0.93, the model only reaches 0.58 precision.

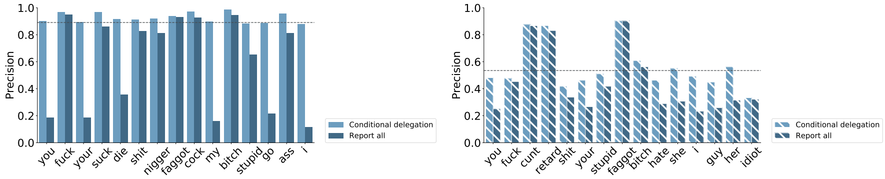


(a) WikiAttack.

(b) Reddit.

Figure 5.5: Words with top precision on WikiAttack (in-distribution) and Reddit (out-of-distribution). "Conditional delegation" shows precision among comments with the word based on model predictions, while "Report all" shows this measure if we consider a comment toxic as long as it contains the word (manual rule-based approach). Dashed lines show model precision on all comments (i.e., the precision of the model working alone). In both cases, all the top 10 words lead to greater precision than the model working alone.

This figure shows the top 10 words with highest precision as conditional delegation rules.

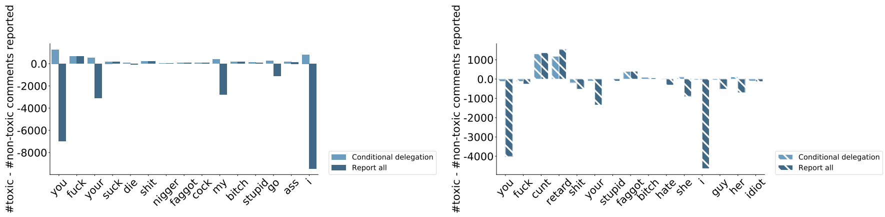


(a) WikiAttack.

(b) Reddit.

Figure 5.6: Precision for words that show up most frequently in rationales on WikiAttack (in-distribution) and Reddit (out-of-distribution) (ordered by frequency). "Conditional delegation" shows precision among comments with the word based on model predictions, while "Report all" shows this measure if we consider a comment toxic as long as it contains the word (the manual rule-based approach). Dashed lines show model precision on all comments (i.e., the precision of the model working alone).

This figure shows that the majority of global explanations on WikiAttack achieve greater precision than the model working alone.

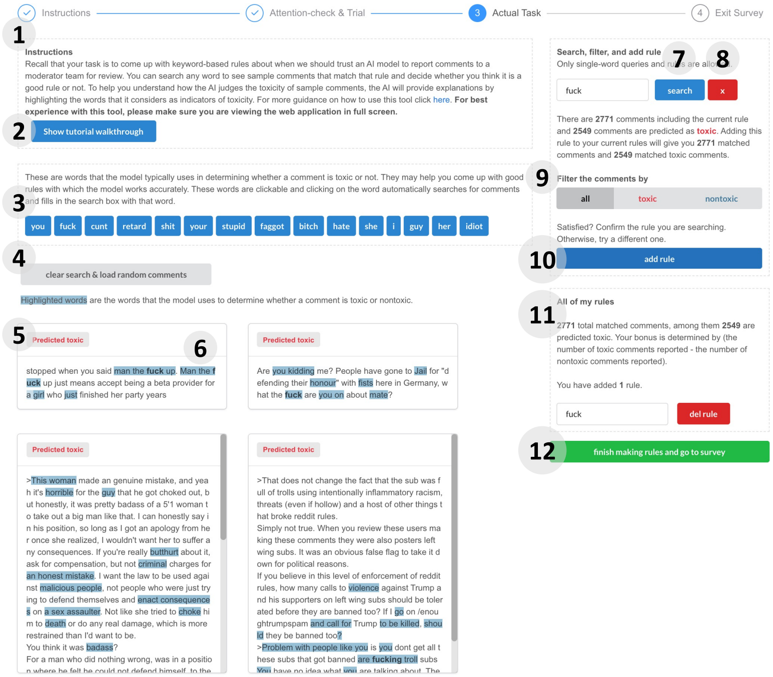


(a) WikiAttack

(b) Reddit

Figure 5.7: Reward (number of reported toxic comments - number of reported non-toxic comments), a measure combining precision and coverage, for words that show up most frequently in rationales on WikiAttack (in-distribution) and Reddit (out-of-distribution) (ordered by frequency). "Conditional delegation" shows reward for comments with the word based on model predictions, while "Report all" shows this measure if we consider a comment toxic as long as it contains a word (the manual rule-based approach).

This figure shows the reward on both datasets. On WikiAttack, most global explanations lead to positive rewards while on Reddit, rewards are dominated by "retard" and "cunt" due to high coverage.



Instructions

Recall that your task is to come up with keyword-based rules about when we should trust an AI model to report comments to a moderator team for review. You can search any word to see sample comments that match that rule and decide whether you think it is a good rule or not. To help you understand how the AI judges the toxicity of sample comments, the AI will provide explanations by highlighting the words that it considers as indicators of toxicity. For more guidance on how to use this tool click here. **For best experience with this tool, please make sure you are viewing the web application in full screen.**

These are words that the model typically uses in determining whether a comment is toxic or not. They may help you come up with good rules with which the model works accurately. These words are clickable and clicking on the word automatically searches for comments and fills in the search box with that word.

Highlighted words are the words that the model uses to determine whether a comment is toxic or nontoxic.

stopped when you said man the **fuck** up. Man the **fuck** up just means accept being a beta provider for a girl who just finished her party years

Are you kidding me? People have gone to Jail for "defending their honour" with fists here in Germany, what the **fuck** are you on about mate?

>This woman made an genuine mistake, and yeah it's horrible for the guy that he got choked out, but honestly, it was pretty badass of a 5'1 woman to take out a big man like that. I can honestly say in his position, so long as I got an apology from her once she realized, I wouldn't want her to suffer any consequences. If you're really butthurt about it, ask for compensation, but not criminal charges for an honest mistake. I want the law to be used against malicious people, not people who were just trying to defend themselves and enact consequences on a sex assaulter. Not like she tried to choke him to death or do any real damage, which is more restrained than I'd want to be.
You think it was badass?
For a man who did nothing wrong, was in a position where he felt he could not defend himself, to the

>That does not change the fact that the sub was full of trolls using intentionally inflammatory racism, threats (even if hollow) and a host of other things that broke reddit rules.
Simply not true. When you review these users making these comments they were also posters left wing subs. It was an obvious false flag to take it down for political reasons.
If you believe in this level of enforcement of reddit rules, how many calls to violence against Trump and his supporters on left wing subs should be tolerated before they are banned too? If I go on /enoughtrumpspam and call for Trump to be killed, should they be banned too?
>Problem with people like you is you dont get all these subs that got banned are **fucking** troll subs
You have no idea what you are talking about. The

Search, filter, and add rule
Only single-word queries and rules are allowed.

fuck

There are **2771** comments including the current rule and **2549** comments are predicted as **toxic**. Adding this rule to your current rules will give you **2771** matched comments and **2549** matched toxic comments.

Filter the comments by

Satisfied? Confirm the rule you are searching. Otherwise, try a different one.

All of my rules

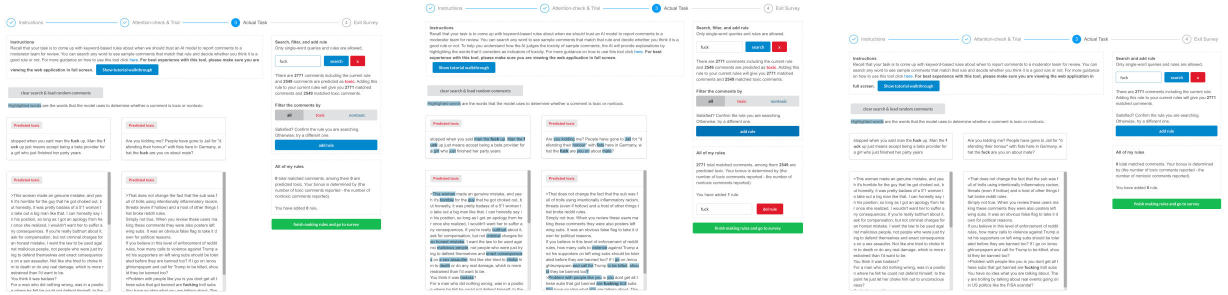
2771 total matched comments, among them **2549** are predicted toxic. Your bonus is determined by (the number of toxic comments reported - the number of nontoxic comments reported).

You have added **1** rule.

fuck

Figure 5.8: Interface for "Predicted label + local explanations + global explanations". We use this interface to go through the design of all delegation support features.

This interface includes all delegation support features.



Instructions
Recall that your task is to come up with keyword-based rules about when we should trust an AI model to report comments to a moderator team for review. You can search any word to see sample comments that match that rule and decide whether you think it is a good rule or not. For more guidance on how to use this tool click here. **For best experience with this tool, please make sure you are viewing the web application in full screen.**

Highlighted words are the words that the model uses to determine whether a comment is toxic or nontoxic.

stopped when you said man the **fuck** up. Man the **f**
uck up just means accept being a beta provider for
a girl who just finished her party years

Are you kidding me? People have gone to Jail for "d
efending their honour" with fists here in Germany, w
hat the **fuck** are you on about mate?

>This woman made an genuine mistake, and yea
h it's horrible for the guy that he got choked out, b
ut honestly, it was pretty badass of a 5'1 woman t
o take out a big man like that. I can honestly say i
n his position, so long as I got an apology from he
r once she realized, I wouldn't want her to suffer a
ny consequences. If you're really butthurt about it,
ask for compensation, but not criminal charges for
an honest mistake. I want the law to be used agai
nst malicious people, not people who were just try
ing to defend themselves and enact consequence
s on a sex assaulter. Not like she tried to choke hi
m to death or do any real damage, which is more r
estrained than I'd want to be.
You think it was badass?
For a man who did nothing wrong, was in a positio
n where he felt he could not defend himself, to the

>That does not change the fact that the sub was f
ull of trolls using intentionally inflammatory racism,
threats (even if hollow) and a host of other things t
hat broke reddit rules.
Simply not true. When you review these users ma
king these comments they were also posters left
wing subs. It was an obvious false flag to take it d
own for political reasons.
If you believe in this level of enforcement of reddit
rules, how many calls to violence against Trump a
nd his supporters on left wing subs should be toler
ated before they are banned too? If I go on /enou
ghtrumpspam and call for Trump to be killed, shou
ld they be banned too?
>Problem with people like you is you dont get all t
hese subs that got banned are **fucking** troll subs
You have no idea what you are talking about. The

Search, filter, and add rule
Only single-word queries and rules are allowed.

There are **2771** comments including the current rule and **2549** comments are predicted as **toxic**. Adding this rule to your current rules will give you **2771** matched comments and **2549** matched toxic comments.

Filter the comments by

Satisfied? Confirm the rule you are searching. Otherwise, try a different one.

All of my rules

0 total matched comments, among them **0** are predicted toxic. Your bonus is determined by (the number of toxic comments reported - the number of nontoxic comments reported).

You have added **0** rule.

(a) Predicted labels condition

Instructions
Recall that your task is to come up with keyword-based rules about when we should trust an AI model to report comments to a moderator team for review. You can search any word to see sample comments that match that rule and decide whether you think it is a good rule or not. To help you understand how the AI judges the toxicity of sample comments, the AI will provide explanations by highlighting the words that it considers as indicators of toxicity. For more guidance on how to use this tool click here. **For best experience with this tool, please make sure you are viewing the web application in full screen.**

Highlighted words are the words that the model uses to determine whether a comment is toxic or nontoxic.

stopped when you said man the **fuck** up. Man the f
uck up just means accept being a beta provider for
a girl who just finished her party years

Are you kidding me? People have gone to Jail for "d
efending their honour" with fists here in Germany, w
hat the **fuck** are you on about mate?

>This woman made an genuine mistake, and yea
h it's horrible for the guy that he got choked out, b
ut honestly, it was pretty badass of a 5'1 woman t
o take out a big man like that. I can honestly say i
n his position, so long as I got an apology from he
r once she realized, I wouldn't want her to suffer a
ny consequences. If you're really butthurt about it,
ask for compensation, but not criminal charges for
an honest mistake. I want the law to be used agai
nst malicious people, not people who were just try
ing to defend themselves and enact consequence
s on a sex assaulter. Not like she tried to choke hi
m to death or do any real damage, which is more
restrained than I'd want to be.
You think it was badass?
For a man who did nothing wrong, was in a positio

>That does not change the fact that the sub was f
ull of trolls using intentionally inflammatory racism,
threats (even if hollow) and a host of other things t
hat broke reddit rules.
Simply not true. When you review these users ma
king these comments they were also posters left
wing subs. It was an obvious false flag to take it d
own for political reasons.
If you believe in this level of enforcement of reddit
rules, how many calls to violence against Trump a
nd his supporters on left wing subs should be toler
ated before they are banned too? If I go on /enou
ghtrumpspam and call for Trump to be killed, shou
ld they be banned too?
>Problem with people like you is you dont get all t
hese subs that got banned are **fucking** troll subs

Search, filter, and add rule
Only single-word queries and rules are allowed.

There are **2771** comments including the current rule and **2549** comments are predicted as **toxic**. Adding this rule to your current rules will give you **2771** matched comments and **2549** matched toxic comments.

Filter the comments by

Satisfied? Confirm the rule you are searching. Otherwise, try a different one.

All of my rules

2771 total matched comments, among them **2549** are predicted toxic. Your bonus is determined by (the number of toxic comments reported - the number of nontoxic comments reported).

You have added **1** rule.

(b) Predicted labels + local explanations condition

Instructions
Recall that your task is to come up with keyword-based rules about when to report comments to a moderator team for review. You can search any word to see sample comments that match that rule and decide whether you think it is a good rule or not. For more guidance on how to use this tool click here. **For best experience with this tool, please make sure you are viewing the web application in full screen.**

Highlighted words are the words that the model uses to determine whether a comment is toxic or nontoxic.

stopped when you said man the **fuck** up. Man the **f**
uck up just means accept being a beta provider for
a girl who just finished her party years

Are you kidding me? People have gone to Jail for "d
efending their honour" with fists here in Germany, w
hat the **fuck** are you on about mate?

>This woman made an genuine mistake, and yea
h it's horrible for the guy that he got choked out, b
ut honestly, it was pretty badass of a 5'1 woman t
o take out a big man like that. I can honestly say i
n his position, so long as I got an apology from he
r once she realized, I wouldn't want her to suffer a
ny consequences. If you're really butthurt about it,
ask for compensation, but not criminal charges for
an honest mistake. I want the law to be used agai
nst malicious people, not people who were just try
ing to defend themselves and enact consequence
s on a sex assaulter. Not like she tried to choke hi
m to death or do any real damage, which is more r
estrained than I'd want to be.
You think it was badass?
For a man who did nothing wrong, was in a positio
n where he felt he could not defend himself, to the
point he just let her choke him out to unconscious
ness?

>That does not change the fact that the sub was f
ull of trolls using intentionally inflammatory racism,
threats (even if hollow) and a host of other things t
hat broke reddit rules.
Simply not true. When you review these users ma
king these comments they were also posters left
wing subs. It was an obvious false flag to take it d
own for political reasons.
If you believe in this level of enforcement of reddit
rules, how many calls to violence against Trump a
nd his supporters on left wing subs should be toler
ated before they are banned too? If I go on /enou
ghtrumpspam and call for Trump to be killed, shou
ld they be banned too?
>Problem with people like you is you dont get all t
hese subs that got banned are **fucking** troll subs
You have no idea what you are talking about. The
y are trolling by talking about real events going on
in US politics like the FISA scandal?

Search and add rule
Only single-word queries and rules are allowed.

There are **2771** comments including the current rule. Adding this rule to your current rules will give you **2771** matched comments.

Satisfied? Confirm the rule you are searching. Otherwise, try a different one.

All of my rules

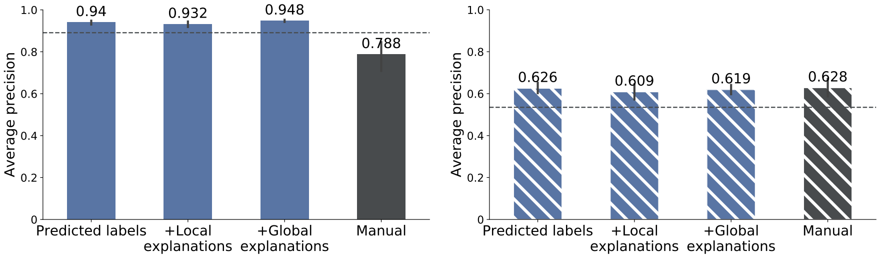
0 total matched comments. Your bonus is determined by (the number of toxic comments reported - the number of nontoxic comments reported).

You have added **0** rule.

(c) Manual condition

Figure 5.9: Interfaces for the other three experimental conditions.

These interfaces include some of the delegation support features.

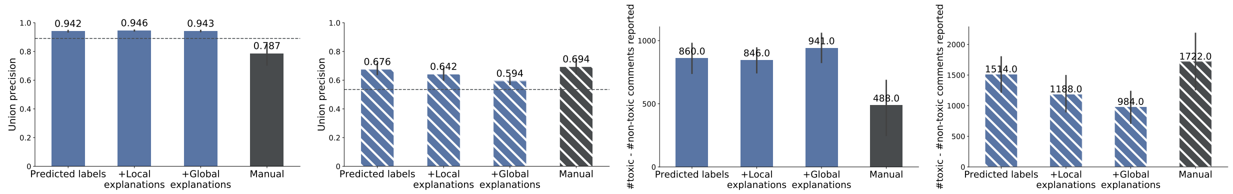


(a) WikiAttack

(b) Reddit

Figure 5.10: Average precision for WikiAttack (in-distribution) and Reddit (out-of-distribution). Error bar shows 95% confidence interval throughout the paper, and the dashed lines show the precision with the model working alone. The first three conditions represent the precision with conditional delegation, while the manual condition reports precision via the manual rule-based approach by reporting all comments that contain any keyword.

This figure shows the average precision for WikiAttack and Reddit. There are four bars in each plot and the first three bars represent conditional delegation conditions while the last bar represents manual condition.



(a) WikiAttack

(b) Reddit

(c) WikiAttack

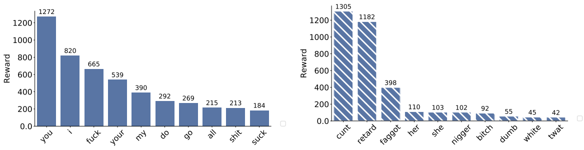
(d) Reddit

Figure 5.11: Union precision and reward for WikiAttack (in-distribution) and Reddit (out-of-distribution). The dashed lines in Figure 5.11a and 5.11b show the precision with the model working alone. Reward is defined as (number of reported toxic comments - number of reported non-toxic comments).

These figure shows the union precision and reward for WikiAttack and Reddit.

	WikiAttack
Predicted labels	bitch (69.0%), asshole (62.1%), fuck (62.1%), cunt (62.1%), nigger (51.7%), dick (48.3%), faggot (44.8%), shit (44.8%), fag (37.9%), motherfucker (31.0%)
+ Local explanations	bitch (71.4%), cunt (71.4%), asshole (67.9%), fuck (60.7%), faggot (53.6%), pussy (42.9%), dick (39.3%), fag (35.7%), cock (35.7%), retard (35.7%)
+ Global explanations	faggot (86.7%), nigger (73.3%), fuck (70.0%), bitch (66.7%), cunt (56.7%), cock (56.7%), ass (50.0%), asshole (46.7%), shit (46.7%), pussy (36.7%)
Manual	cunt (70.0%), nigger (63.3%), faggot (60.0%), fag (56.7%), bitch (53.3%), asshole (46.7%), retard (43.3%), whore (43.3%), fuck (36.7%), pussy (26.7%)
	Reddit
Predicted labels	cunt (86.2%), bitch (72.4%), faggot (62.1%), fuck (58.6%), retard (44.8%), asshole (41.4%), nigger (41.4%), pussy (34.5%), fag (31.0%), whore (31.0%)
+ Local explanations	cunt (72.4%), bitch (65.5%), fuck (55.2%), pussy (55.2%), nigger (48.3%), asshole (41.4%), faggot (41.4%), retard (37.9%), shit (37.9%), dumbass (34.5%)
+ Global explanations	bitch (69.0%), cunt (65.5%), faggot (62.1%), fuck (58.6%), retard (58.6%), nigger (41.4%), pussy (41.4%), shit (37.9%), dick (37.9%), idiot (34.5%)
Manual	nigger (76.7%), cunt (73.3%), faggot (60.0%), bitch (56.7%), retard (43.3%), whore (43.3%), asshole (36.7%), fag (30.0%), spic (30.0%), chink (30.0%)

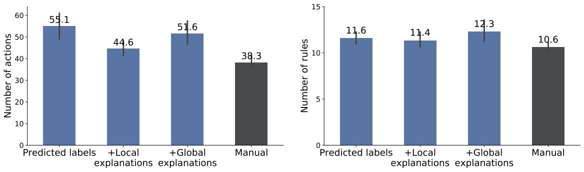
Table 5.3: Most frequent rules chosen by participants.



(a) WikiAttack (b) Reddit

Figure 5.12: Top 10 human-created rules in reward when used for conditional delegation.

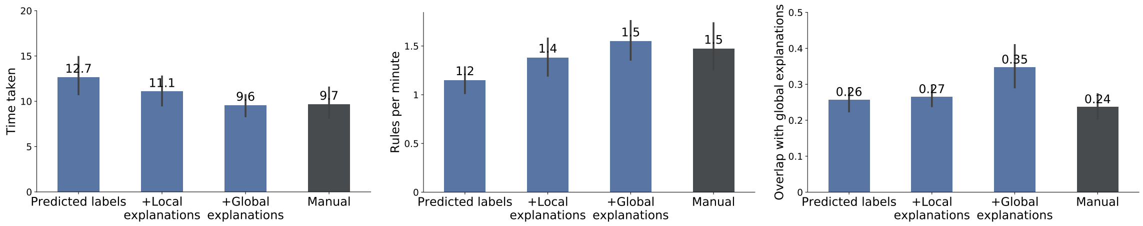
This plot shows the top 10 human-created rules in reward.



(a) Number of actions. (b) Number of rules.

Figure 5.13: Engagement. Conditional delegation with all delegation support features leads to much better engagement and more submitted rules than the manual condition.

The plots show the number of actions taken and number of rules created by different experiment conditions.



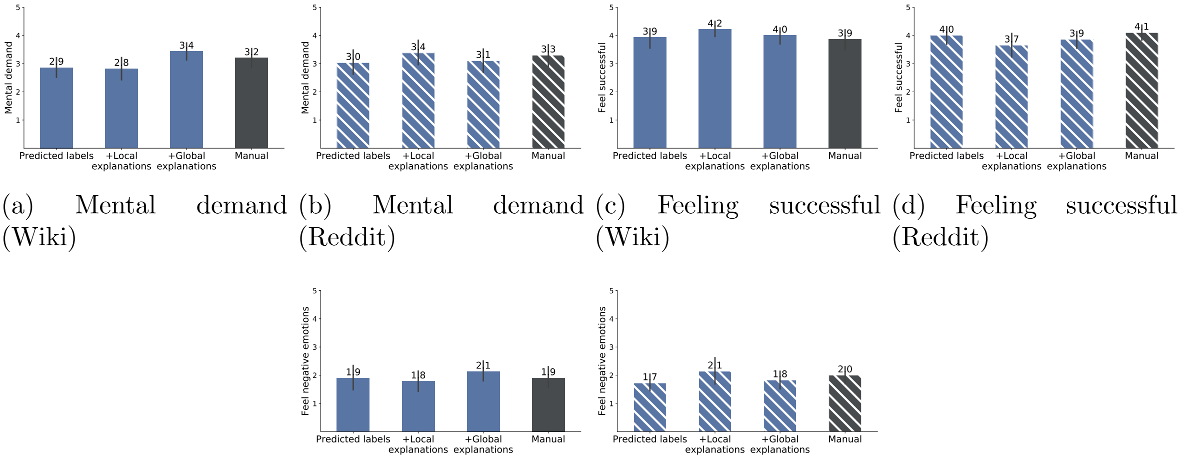
(a) Elapsed time

(b) Rules per minute

(c) Overlap with global explanations.

Figure 5.14: Efficiency. Participants spent more time working on conditional delegation than the manual condition, but the efficiency is improved with explanations, especially global explanations.

The plots show time taken, rules per minute, and overlap with global explanations by different experiment conditions.



(a) Mental demand (Wiki)

(b) Mental demand (Reddit)

(c) Feeling successful (Wiki)

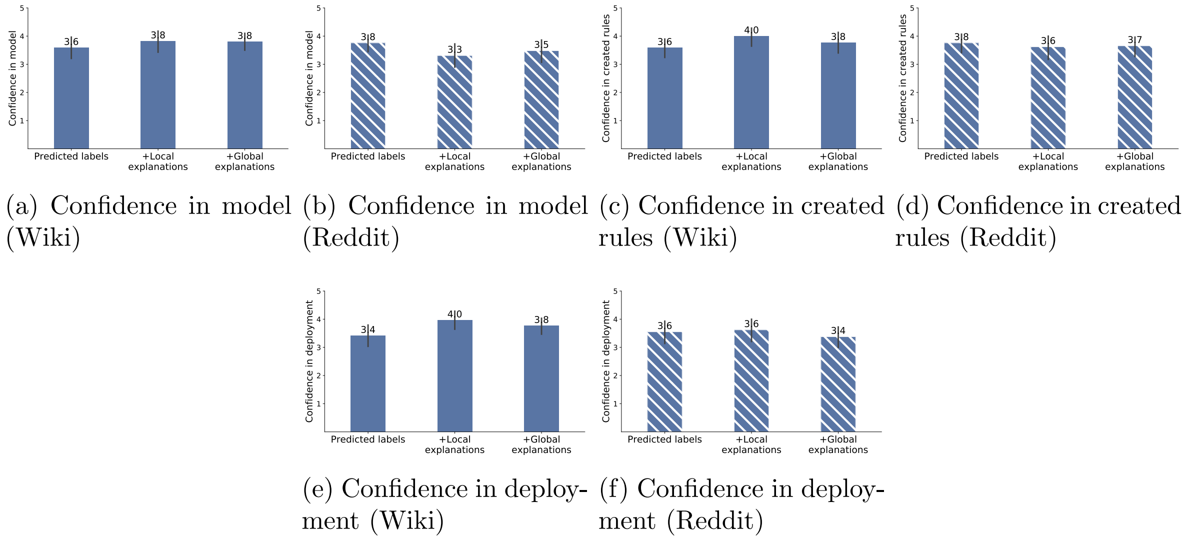
(d) Feeling successful (Reddit)

(e) Negative emotions (Wiki)

(f) Negative emotions (Reddit)

Figure 5.15: Subjective workload. Overall, participants were neutral about whether the task was mentally demanding (M=3.15, SD=1.08), agreed that they felt successful in accomplishing the task (M=3.95, SD=0.91), and disagreed that they felt negative emotions (M=1.93, SD=1.03).

The plots show the different subjective measures' ratings adopted from NASA-TLX by different experiment conditions.

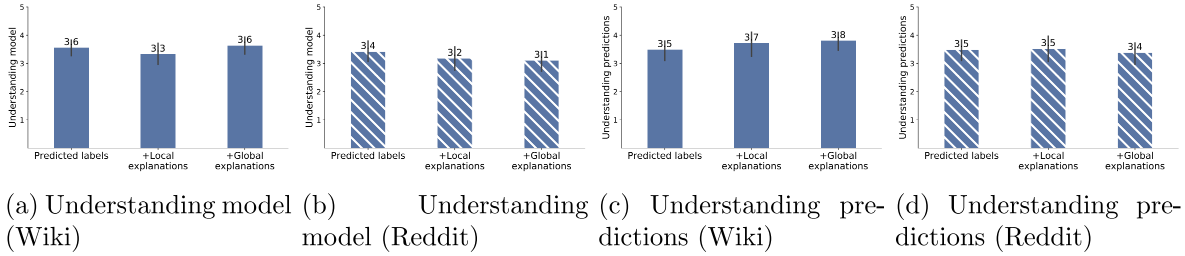


(a) Confidence in model (Wiki) (b) Confidence in model (Reddit) (c) Confidence in created rules (Wiki) (d) Confidence in created rules (Reddit)

(e) Confidence in deployment (Wiki) (f) Confidence in deployment (Reddit)

Figure 5.16: Confidence. Overall, participants show strong confidence in the model, their performance, and moderators' potential adoption.

The plots show confidence in model, confidence in created rules, and confidence in deployment ratings by different experiment conditions.



(a) Understanding model (Wiki) (b) Understanding model (Reddit) (c) Understanding predictions (Wiki) (d) Understanding predictions (Reddit)

Figure 5.17: Understanding. Overall, participants report a better understanding of the model as a whole and individual predictions on WikiAttack than Reddit, although the differences are not statistically significant.

The plots show understanding of model and predictions ratings by different experiment conditions.

Chapter 6

Future work

This thesis demonstrates the potential of how human-AI collaborations can take place in a safe environment without having to worry about losing human agency in the decision making process. However, there are many future directions in improving this process.

The goal of my research is to understand the effects of different aspects affecting human-AI collaborations to create AI-backed interactive systems that assist humans in making better decisions in a wide variety of tasks. Current AI systems consider a one-sided aspect of explanations and fail to understand other elements hindering the process of improving complementary performance. There is room for improvement in creating AI-assisting systems that elucidate weaknesses and biases of AI, allowing humans to make decisions that are not under the influence of AI. This is especially crucial when the systems are deployed in high-stakes domain such as medical and justice.

As mentioned in previous chapters, achieving complementary performance is one of the metrics used to determine smooth human-AI collaborations. One of my future research directions is exploring different factors affecting complementary performance. The current AI setup assumes the most optimistic AI performance. However, in reality, data distribution in the training set and test set differ, resulting in a drop in AI performance. I explored how the difference in data distribution could directly affect complementary performance [Liu et al., 2021]. Besides the widely debated topic of how explanations should be generated and displayed, other factors are under studied, potentially impeding complementary performance in different tasks. It remains an open question on how the research community should set up experiments to emulate real-life scenarios and how systems should be designed to create and bolster meaningful human-AI interactions.

My current work focuses on empowering human decision making on a wide range of tasks spanning from simple to challenging. However, human-AI interactions should also be enhanced in downstream natural language processing tasks, which could potentially help provide meaningful insights and improve current processes. I explored how human-AI collaborations could be helpful in text summarization and found that human-AI collaboration in formal and informal text summarization is helpful and valuable to a certain degree [Lai et al., 2022]. Exploring human-AI interaction in other tasks is one of the first steps in realizing the goal of empowering human decisions and tasks by exploring various tasks and building systems that are made more accessible to experts and non-experts.

Another line of my future work is to make technology more accessible to both AI experts and non-experts. Large language models (LLM) are gaining popularity recently due to their potential and impressive performance on simple tasks. While many companies are making LLMs more accessible to the public, how these models can be fully maximized by both AI experts and non-experts and avoid unintended consequences remain under studied. Wu et al. [2022] took the first step in making LLMs more accessible by proposing the concept of "Chaining" LLM steps. In the grand scheme of building systems that are more accessible, developing reliable systems require extensive testing. How could we extensively test the systems before production deployment? I plan to investigate the concept of a sandbox that allows extensive testing and evaluation of the systems. Sandbox allows a safe environment that could prevent any potential unintended harm or inappropriate responses before deploying into the wild. More importantly, it provides an interactive interface for humans, experts, and non-experts, to examine the effectiveness of the systems.

While technology has been advancing rapidly, there exists a gap between how we can build systems that fully utilize the potential of both technology and humans. We live in interesting times, and it is an exciting time to build systems for people and study humans' usage of these systems. This thesis contributes to a new paradigm of understanding human-AI interactions and building systems that empower the human decision making process.

Bibliography

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI conference on human factors in computing systems, pages 1–18, 2018.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928, 2020.

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

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access, 6:52138–52160, 2018.

Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network effects. In Proceedings of ICWSM, 2013.

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

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. Ai Magazine, 35(4):105–120, 2014.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In Proceedings of the 2019 chi conference on human factors in computing systems, pages 1–13, 2019.

Robert Andrews, Joachim Diederich, and Alan B Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems, 8(6):373–389, 1995.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.

Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. Opencrowd: A human-ai collaborative approach for finding social influencers via open-ended answers aggregation. In Proceedings of The Web Conference 2020, pages 1851–1862, 2020.

W Ross Ashby. An introduction to cybernetics. 1957.

Zahra Ashktorab, Q Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. Human-ai collaboration in a cooperative game setting: Measuring social perception and outcomes. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2):1–20, 2020.

David H Autor, Frank Levy, and Richard J Murnane. The skill content of recent technological change: An empirical exploration. The Quarterly journal of economics, 118(4):1279–1333, 2003.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of ICLR, 2015.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 7, pages 2–11, 2019a.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 2429–2437, 2019b.

Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. arXiv preprint arXiv:2006.14779, 2020.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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, pages 1–16, 2021.

Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C Mitchell. Reinforcement learning for the adaptive scheduling of educational activities. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2020.

Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2020.

James O Berger. Statistical decision theory and Bayesian analysis. Springer Science & Business Media, 2013.

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

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075, 2017.

Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2018.

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW):1–19, 2017.

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

Melissa Boston. Assessing instructional quality in mathematics. The Elementary School Journal, 113(1):76–104, 2012.

Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2019.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. Science, 365(6456): 885–890, 2019.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In Proceedings of the 25th International Conference on Intelligent User Interfaces, pages 454–464, 2020.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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):1–21, 2021.

Patrick Buckley and Elaine Doyle. Gamification and student motivation. Interactive learning environments, 24(6):1162–1175, 2016.

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

Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In Proceedings of the 24th International Conference on Intelligent User Interfaces, pages 258–262. ACM, 2019a.

Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, page 4. ACM, 2019b.

Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. Hello ai: Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):104, 2019c.

Dallas Card and Noah A Smith. The importance of calibration for estimating proportions from annotations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1636–1646, 2018.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. In Advances in Neural Information Processing Systems, pages 5175–5186, 2019.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. arXiv preprint arXiv:1809.01499, 2018.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. Attention-based explanations don't help humans detect misclassifications of online toxicity. In ICWSM, 2020a.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. Feature-based explanations don't help people detect misclassifications of online toxicity. In Proceedings of the International AAAI Conference on Web and Social Media, volume 14, pages 95–106, 2020b.

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

Avner Caspi and Paul Gorsky. Online deception: Prevalence, motivation, and emotion. CyberPsychology & Behavior, 9(1):54–59, 2006.

Noah Castelo, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. Journal of Marketing Research, 56(5):809–825, 2019.

Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In Proceedings of CSCW, CSCW '16, pages 1201–1213, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3592-8.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? arXiv preprint arXiv:1810.12366, 2018.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. In Proceedings of ACM Human Computer Interaction, 1(Computer-Supported Cooperative Work and Social Computing):31:1–31:22, 2017. ISSN 2573-0142.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. In Proceedings of ACM Human Computer Interaction, 2(Computer-Supported Cooperative Work and Social Computing):32:1–32:25, November 2018. ISSN 2573-0142.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A cross-community learning-based system to assist reddit moderators. Proceedings of the ACM on human-computer interaction, 3(CSCW):1–30, 2019.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. Trajectories of blocked community members: Redemption, recidivism and departure. In Proceedings of WWW, pages 184–195, 2019.

Gaowei Chen, Sherice N Clarke, and Lauren B Resnick. Classroom discourse analyzer (cda): A discourse analytic tool for teachers. Technology, Instruction, Cognition & Learning, 10(2), 2015.

Gaowei Chen, Carol KK Chan, Kennedy KH Chan, Sherice N Clarke, and Lauren B Resnick. Efficacy of video-based teacher professional development for increasing classroom discourse and student learning. Journal of the Learning Sciences, pages 1–39, 2020.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of KDD, 2016.

Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, page 559. ACM, 2019.

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

Chun-Wei Chiang and Ming Yin. You'd better stop! understanding human reliance on machine learning models under covariate shift. In 13th ACM Web Science Conference 2021, pages 120–129, 2021.

Dan Claudiu Ciresan, Ueli Meier, Luca Maria Gambardella, and Jurgen Schmidhuber. Convolutional neural network committees for handwritten character classification. In 2011 International Conference on Document Analysis and Recognition, pages 1135–1139. IEEE, 2011.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4069–4082, 2019.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In Proceedings of IUI, 2018a.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In 23rd International Conference on Intelligent User Interfaces, pages 329–340, 2018b.

Cone Communications. Game changer: Cone survey finds 4-out-of-5 consumers reverse purchase decisions based on negative online reviews, 2011.

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

Richard Correnti, Mary Kay Stein, Margaret S Smith, James Scherrer, Margaret McKeown, James Greeno, and Kevin Ashley. Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. Socializing Intelligence Through Academic Talk and Dialogue. AERA, 284, 2015.

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

Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 2382–2393, 2017.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 120–128, 2019.

Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 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, pages 1–12, 2020a.

Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. arXiv:2002.08035 [cs.CY], 2020b.

Ewart J de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. A design methodology for trust cue calibration in cognitive agents. In International conference on virtual, augmented and mixed reality, pages 251–262. Springer, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255, 2009.

Daniel Clement Dennett. The intentional stance. MIT press, 1989.

Bella M DePaulo, G Daniel Lassiter, and Julie L Stone. Attentional determinants of success at detecting deception and truth. Personality and Social Psychology Bulletin, 8(2):273–279, 1982.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL, 2019.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1): 114, 2015.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64 (3):1155–1170, 2018.

Sidney K D'Mello, Andrew M Olney, Nathan Blanchard, Borhan Samei, Xiaoyi Sun, Brooke Ward, and Sean Kelly. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In Proceedings of the 2015 ACM on international conference on multimodal interaction, pages 557–566, 2015.

Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In Proceedings of the 24th International Conference on Intelligent User Interfaces, pages 275–285, 2019.

Patrick J Donnelly, Nathan Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystran, and Sidney K D'Mello. Automatic teacher modeling from live classroom audio. In Proceedings of the 2016 conference on user modeling adaptation and personalization, pages 45–53, 2016.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of ai under the law: The role of explanation. arXiv preprint arXiv:1711.01134, 2017.

Bryan Dosono and Bryan Semaan. Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit. In Proceedings of the 2019 CHI conference on human factors in computing systems, pages 1–13, 2019.

Anca Dragan. Specifying ai objectives as a human-ai collaboration problem. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 329–329, 2019.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1):eaao5580, 2018.

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

John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. ACM Transactions on Interactive Intelligent Systems (TiiS), 8(2):1–37, 2018.

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

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

Pelle Ehn. Scandinavian design-on skill and participation. Usability-Turning technologies into tools. P. Adler and T. Winograd, 1992.

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

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

Douglas C Engelbart. Augmenting human intellect: a conceptual framework (1962). PACKER, Randall and JORDAN, Ken. Multimedia. From Wagner to Virtual Reality. New York: WW Norton & Company, pages 64–90, 2001.

Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating algorithmic process in online behavioral advertising. In Proceedings of the 2018 CHI conference on human factors in computing systems, pages 1–13, 2018.

Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In Proceedings of the 8th international conference on Intelligent user interfaces, pages 39–45, 2003.

Diane Farsetta and Daniel Price. Fake tv news: Widespread and undisclosed. Center for Media and Democracy, 6, 2006.

Shi Feng and Jordan Boyd-Graber. What can ai do for me: Evaluating machine learning interpretations in cooperative play. arXiv preprint arXiv:1810.09648, 2018.

Shi Feng and Jordan Boyd-Graber. What can ai do for me? evaluating machine learning interpretations in cooperative play. In Proceedings of the 24th International Conference on Intelligent User Interfaces, pages 229–239, 2019.

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

Vanessa Wei Feng and Graeme Hirst. Text-level discourse parsing with rich linguistic features. In Proceedings of ACL, 2012.

Vanessa Wei Feng and Graeme Hirst. Detecting deceptive opinions with profile compatibility. In Proceedings of IJCNLP, 2013.

Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. Reddit rules! characterizing an ecosystem of governance. In Proceedings of International AAAI Conference on Web and Social Media(ICWSM), 2018.

Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Fed. Probation, 80:38, 2016.

Carl Benedikt Frey and Michael A Osborne. The future of employment: how susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114:254–280, 2017.

Emma Frid, Ceslo Gomes, and Zeyu Jin. Music creation by example. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.

Sorelle A Friedler, Chitradeep Dutta Roy, Carlos Scheidegger, and Dylan Slack. Assessing the local interpretability of machine learning models. arXiv preprint arXiv:1902.03501, 2019a.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 329–338. ACM, 2019b.

Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4438–4446, 2017.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 219–226. ACM, 2019.

Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. The effects of automatic speech recognition quality on human transcription latency. In Proceedings of the 13th Web for All Conference, pages 1–8, 2016.

Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. Mental models of ai agents in a cooperative game setting. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2020.

Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. arXiv preprint arXiv:2001.09219, 2020.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. Hafez: an interactive poetry generation system. In Proceedings of ACL 2017, System Demonstrations, pages 43–48, 2017.

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

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

Tarleton Gillespie. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press, New Haven, June 2018. ISBN 9780300173130.

Tarleton Gillespie. Content moderation, ai, and the question of scale. Big Data & Society, 7(2): 2053951720943234, 2020.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pages 80–89. IEEE, 2018.

Matthew Gombolay, Xi Jessie Yang, Brad Hayes, Nicole Seo, Zixi Liu, Samir Wadhwania, Tania Yu, Neel Shah, Toni Golen, and Julie Shah. Robotic assistance in coordination of patient care. 2016.

Ian Goodfellow, Y Bengio, and A Courville. Machine learning basics. In Deep learning, volume 1, pages 98–164. MIT press, 2016.

Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In European conference on computer vision, pages 241–257. Springer, 2016.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7 (1):2053951719897945, 2020.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. arXiv preprint arXiv:1410.5401, 2014.

Ben Green. “fair” risk assessments: A precarious approach for criminal justice reform. In 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2018.

Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 90–99. ACM, 2019a.

Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):50, 2019b.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. Science, 363(6425):374–378, 2019.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820, 2018a.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5):1–42, 2018b.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5):93, 2019.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3608–3617, 2018.

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

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

Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 392–402, 2020.

Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In Proceedings of the human factors and ergonomics society annual meeting, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

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

Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. Proceedings of the National Academy of Sciences, 116(6):1844–1850, 2019.

Narayan Hegde, Jason D Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J Cai, Mahul B Amin, Craig H Mermel, et al. Similar image search for histopathology: Smily. NPJ digital medicine, 2(1):1–9, 2019.

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

Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2019.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, page 600. ACM, 2019.

Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics, 3(2):119–131, 2016.

Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. 2017.

Benjamin D Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adali. Rating reliability and bias in news articles: Does ai assistance help everyone? In Proceedings of ICWSM, 2019.

Eric Horvitz. Principles of mixed-initiative user interfaces. In Proceedings of CHI, 1999.

Eric J Horvitz, John S Breese, and Max Henrion. Decision theory in expert systems and artificial intelligence. International journal of approximate reasoning, 2(3):247–302, 1988.

Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, and Adish Singla. Teaching multiple concepts to a forgetful learner. In Advances in Neural Information Processing Systems, pages 4048–4058, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

Ipsos. Socialogue: Five stars? thumbs up? a+ or just average?, 2012.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In Proceedings of NAACL, 2019a.

Sarthak Jain and Byron C Wallace. Attention is not explanation. arXiv preprint arXiv:1902.10186, 2019b.

Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, page 239. ACM, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144, 2016.

Emily Jensen, Meghan Dale, Patrick J Donnelly, Cathlyn Stone, Sean Kelly, Amanda Godley, and Sidney K D'Mello. Toward automated feedback on teacher discourse to enhance teacher learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2020.

Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. ACM Transactions on Computer-Human Interaction (TOCHI), 26(5):1–35, 2019a.

Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–27, 2019b.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, 2017.

Nitin Jindal and Bing Liu. Opinion spam and analysis. In Proceedings of WSDM, 2008.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of ECML, 1998.

AV Kamasheva, ER Valeev, R Kh Yagudin, and KR Maksimova. Usage of gamification theory for increase motivation of employees. Mediterranean Journal of Social Sciences, 6(1 S3):77, 2015.

Kenji Kaneko, Fumio Kanehiro, Shuuji Kajita, Kazuhiko Yokoyama, Kazuhiko Akachi, Toshikazu Kawasaki, Shigehiko Ota, and Takakatsu Isozumi. Design of prototype humanoid robotics platform for hrp. In IEEE/RSJ International Conference on Intelligent Robots and Systems, volume 3, pages 2431–2436. IEEE, 2002.

Sean HK Kang. Spaced repetition promotes efficient and effective learning: Policy implications for instruction. Policy Insights from the Behavioral and Brain Sciences, 3(1):12–19, 2016.

Harmanpreet Kaur, Cliff Lampe, and Walter S Lasecki. Using affordances to improve ai support of social media posting decisions. In Proceedings of the 25th International Conference on Intelligent User Interfaces, pages 556–567, 2020a.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2020b.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. arXiv:1909.12434 [cs, stat], 2019.

Peter GW Keen. Decision support systems; an organizational perspective. Technical report, 1978.

Finn Kensing and Joan Greenbaum. Heritage: Having a say. In Routledge international handbook of participatory design, pages 21–36. Routledge, 2013.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. arXiv preprint arXiv:2102.13004, 2021.

Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. Building successful online communities: Evidence-based social design, pages 125–178, 2012.

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

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

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). arXiv preprint arXiv:1711.11279, 2017.

Gabriela Kiryakova, Nadezhda Angelova, and Lina Yordanova. Gamification in education. Proceedings of 9th International Balkan Education and Science Conference, 2014.

René F Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 2390–2395, 2016.

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

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

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. 2017b.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. The quarterly journal of economics, 133(1):237–293, 2018.

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

Ronald T Kneusel and Michael C Mozer. Improving human-machine cooperative visual search with soft highlighting. ACM Transactions on Applied Perception (TAP), 15(1):3, 2017.

Jim Knight. What can we do about teacher resistance? Phi Delta Kappan, 90(7):508–513, 2009.

Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 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, page 411. ACM, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1885–1894. JMLR. org, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. arxiv, 2020.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In International Conference on Machine Learning, pages 5637–5664. PMLR, 2021.

Karen D Könings, Tina Seidel, and Jeroen JG van Merriënboer. Participatory design of learning environments: integrating perspectives of students, teachers, and designers. Instructional Science, 42(1):1–9, 2014.

Andreas Krause and Daniel Golovin. Submodular function maximization., 2014.

Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 5686–5697. ACM, 2016.

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

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In 2013 IEEE Symposium on Visual Languages and Human Centric Computing, pages 3–10. IEEE, 2013.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In Proceedings of the 20th international conference on intelligent user interfaces, pages 126–137. ACM, 2015.

Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, page 487. ACM, 2019.

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

Richard Ladyshewsky. The impact of peer-coaching on the clinical reasoning of the novice practitioner. Physiotherapy Canada, 56:15–25, 2004.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1902.00006, 2019.

Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In Proceedings of FAT*, 2019a.

Vivian Lai and Chenhao Tan. 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, pages 29–38, 2019b.

Vivian Lai, Zheng Cai, and Chenhao Tan. Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 486–495, 2019.

Vivian Lai, Han Liu, and Chenhao Tan. "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2020a.

Vivian Lai, Han Liu, and Chenhao Tan. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2020b.

Vivian Lai, Alison Smith-Renner, Ke Zhang, Ruijia Cheng, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. An exploration of post-editing effectiveness in text summarization. arXiv preprint arXiv:2206.06383, 2022.

Himabindu Lakkaraju and Osbert Bastani. " how do i fool you?": Manipulating user trust via misleading black box explanations. arXiv preprint arXiv:1911.06473, 2019.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1675–1684, 2016.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154, 2017.

Walter S Lasecki, Christopher D Miller, Iftekhar Naim, Raja Kushalnagar, Adam Sadilek, Daniel Gildea, and Jeffrey P Bigham. Scribe: deep integration of human and machine intelligence to caption speech in real time. Communications of the ACM, 60(9):93–100, 2017.

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. The science of fake news. Science, 359(6380):1094–1096, 2018.

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

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. arXiv preprint arXiv:1606.04155, 2016.

Weiwen Leung, Zheng Zhang, Daviti Jibuti, Jinhao Zhao, Maximilian A Klein, Casey Pierce, Lionel Robert, and Haiyi Zhu. Can user interface design influence hiring bias in the online freelance marketplace? 2020.

Stephan Lewandowsky, Michael Mundy, and Gerard Tan. The dynamics of trust: Comparing humans to automation. Journal of Experimental Psychology: Applied, 6(2):104, 2000.

David Lewis. Causal explanation. 1986.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a general rule for identifying deceptive opinion spam. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1566–1576, 2014.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. In Proceedings of NAACL, 2016.

Li Li. Sex differences in deception detection, 2011.

Gaobo Liang and Lixin Zheng. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Computer methods and programs in biomedicine, 187:104964, 2020.

Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. arXiv preprint arXiv:2001.02478, 2020.

Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 2119–2128, 2009.

Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, Jennifer Skeem, et al. The limits of human predictions of recidivism. Science advances, 6(7):eaaz0652, 2020.

Adam Liptak. Sent to prison by a software program's secret algorithms, 2017.

Peter Lipton. Contrastive explanation. Royal Institute of Philosophy Supplements, 27:247–266, 1990.

Zachary C Lipton. The mythos of model interpretability. arXiv preprint arXiv:1606.03490, 2016.

Zachary C Lipton. The mythos of model interpretability. Queue, 16(3):31–57, 2018.

Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. arXiv preprint arXiv:2101.05303, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151: 90–103, 2019.

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

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

Tania Lombrozo. Explanation and abductive inference. Oxford handbook of thinking and reasoning, pages 260–276, 2012.

Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. Novice-ai music co-creation via ai-steering tools for deep generative models. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2020.

Brian Lubars and Chenhao Tan. Ask not what ai can do, but what ai should do: Towards a framework of task delegability. In Proceedings of NeurIPS, 2019.

Ana Lucic, Hinda Haned, and Maarten de Rijke. Why does my model fail? contrastive local explanations for retail forecasting. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 90–98, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Proceedings of NeurIPS, 2017.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888, 2018a.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering, 2(10):749–760, 2018b.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025, 2015a.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Proceedings of EMNLP, 2015b. URL `https://www.aclweb.org/anthology/D15-1166`.

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

Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. Discovering the sweet spot of human-computer configurations: A case study in information extraction. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–30, 2019.

Dmitry M Malioutov, Kush R Varshney, Amin Emad, and Sanjeeb Dash. Learning interpretable classification rules with boolean compressed sensing. In Transparent Data Mining for Big and Small Data, pages 95–121. Springer, 2017.

Keri Mallari, Kori Inkpen, Paul Johns, Sarah Tan, Divya Ramesh, and Ece Kamar. Do i look like a criminal? examining how race presentation impacts human judgement of recidivism. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2020.

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

J Nathan Matias. The civic labor of online moderators. In Internet Politics and Policy conference. Oxford, United Kingdom, 2016.

Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 922–928. IEEE, 2015.

Maranda McBride and Shona Morgan. Trust calibration for automated decision aids. Institute for Homeland Security Solutions, pages 1–11, 2010.

Jon McCormack, Toby Gifford, Patrick Hutchings, Maria Teresa Llano Rodriguez, Matthew Yee-King, and Mark d'Inverno. In a silent way: Communication between ai and improvising musicians beyond sound. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–11, 2019.

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

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

John M McGuirl and Nadine B Sarter. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human factors, 48(4):656–665, 2006.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. Nature, 577(7788):89–94, 2020.

Albert Mehrabian. Silent messages, volume 8. Wadsworth Belmont, CA, 1971.

Stephanie M Merritt, Deborah Lee, Jennifer L Unnerstall, and Kelli Huber. Are well-calibrated users effective users? associations between calibration of trust and performance on an automation-aided task. Human Factors, 57(1):34–47, 2015.

Sarah Michaels and Catherine O'Connor. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. Socializing intelligence through talk and dialogue, pages 347–362, 2015.

Sarah Michaels, Catherine O'Connor, and Lauren B Resnick. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. Studies in philosophy and education, 27(4):283–297, 2008.

Sarah Michaels, Mary Catherine O'Connor, Megan Williams Hall, and Lauren B Resnick. Accountable talk® sourcebook. Pittsburg, PA: Institute for Learning University of Pittsburgh. Murphy, PK, Wilkinson, IAG, Soter, AO, Hennessey, MN, & Alexander, JF, 2010.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2018.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In Proceedings of the conference on fairness, accountability, and transparency, pages 279–288, 2019.

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. arXiv preprint arXiv:1905.07697, 2019.

Bonnie M Muir. Trust between humans and machines, and the design of decision aids. International journal of man-machine studies, 27(5-6):527–539, 1987.

Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. How data science workers work with data: Discovery, capture, curation, design, creation. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2019.

Charles Munter. Developing visions of high-quality mathematics instruction. Journal for Research in Mathematics Education, 45(5):584–635, 2014.

Charles Munter and Richard Correnti. Examining relations between mathematics teachers' instructional vision and knowledge and change in practice. American Journal of Education, 123 (2):000–000, 2017.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44):22071–22080, 2019.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1802.00682, 2018.

Allen Newell and Herbert Alexander Simon. Human problem solving, volume 104. Prentice-Hall Englewood Cliffs, NJ, 1972.

An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, pages 189–199, 2018.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In Proceedings of NAACL, 2018.

Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In Proceedings of the second Nordic conference on Human-computer interaction, pages 101–110. ACM, 2002.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web, pages 145–153, 2016.

Don Norman. The design of everyday things: Revised and expanded edition. Basic books, 2013.

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

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

Amy Ogan. Reframing classroom sensing: promise and peril. interactions, 26(6):26–32, 2019.

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

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

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

James Overton. Scientific explanation and computation. In ExaCt, pages 41–50, 2011.

Melinda T Owens, Shannon B Seidel, Mike Wong, Travis E Bejines, Susanne Lietz, Joseph R Perez, Shangheng Sit, Zahur-Saleh Subedar, Gigi N Acker, Susan F Akana, et al. Classroom sound can be used to classify teaching practices in college science courses. Proceedings of the National Academy of Sciences, 114(12):3085–3090, 2017.

Catherine O'Connor, Sarah Michaels, and Suzanne Chapin. Scaling down" to explore the role of talk in learning: From district intervention to controlled classroom study. Socializing intelligence through academic talk and dialogue, pages 111–126, 2015.

Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. Human–autonomy teaming: A review and analysis of the empirical literature. Human Factors, page 0018720820960865, 2020.

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

Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30(3):286–297, 2000.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–15, 2019.

Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. What you see is what you get? the impact of representation criteria on human bias in hiring. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 7, pages 125–134, 2019.

William R Penuel, Jeremy Roschelle, and Nicole Shechtman. Designing formative assessment software with teachers: An analysis of the co-design process. Research and practice in technology enhanced learning, 2(01):51–74, 2007a.

William R Penuel, Jeremy Roschelle, and Nicole Shechtman. The whirl co-design process: Participant experiences. Research and Practice in Technology Enhanced Learning, 2(1):51–74, 2007b.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. Verbal and nonverbal clues for real-life deception detection. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2336–2346, 2015.

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

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

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–52, 2021.

Andrew Prahl and Lyn Van Swol. Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting, 36(6):691–702, 2017.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4755–4764, Hong Kong, China, November 2019. Association for Computational Linguistics.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. Dataset shift in machine learning. The MIT Press, 2009.

Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In Proceedings of the 2018 CHI conference on human factors in computing systems, pages 1–13, 2018.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In Proceedings of ACL, 2019.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In Proceedings of the aaai conference on artificial intelligence, volume 33, pages 4780–4789, 2019.

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

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Proceedings of AAAI, 2018a.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018b.

Mark O Riedl. Human-centered artificial intelligence and machine learning. Human Behavior and Emerging Technologies, 1(1):33–36, 2019.

Sarah T Roberts. Behind the screen: The hidden digital labor of commercial content moderation. University of Illinois at Urbana-Champaign, 2014.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2019.

Chris Russell. Efficient search for diverse coherent explanations. In Proceedings of FAT*, 2019.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296, 2017.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The Risk of Racial Bias in Hate Speech Detection. In Proceedings of ACL, 2019.

Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. A framework of severity for harmful content online. arXiv preprint arXiv:2108.04401, 2021.

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

Joseph Seering, Robert Kraut, and Laura Dabbish. Shaping pro and anti-social behavior on Twitch through moderation and example-setting. In Proceedings of CSCW, New York, NY, USA, 2017. ACM.

Leslie E Sekerka and Jason Chao. Peer coaching as a technique to foster professional development in clinical ambulatory settings. Journal of continuing education in the health professions, 23(1): 30–37, 2003.

Sofia Serrano and Noah A. Smith. Is Attention Interpretable? In Proceedings of ACL, 2019.

Samuel Severance, William R Penuel, Tamara Sumner, and Heather Leary. Organizing for teacher agency in curricular co-design. Journal of the Learning Sciences, 25(4):531–564, 2016.

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

Youngsang Shin, Minaxi Gupta, and Steven Myers. Prevalence and mitigation of forum spamming. In 2011 Proceedings IEEE INFOCOM, pages 2309–2317. IEEE, 2011.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Proceedings of ICML, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science, 362 (6419):1140–1144, 2018.

Herbert A Simon. Theories of decision-making in economics and behavioral science. The American economic review, 49(3):253–283, 1959.

Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In Proceedings of ICML, 2014.

Michael E Skinner and Frances C Welch. Peer coaching for better teaching. College Teaching, 44 (4):153–156, 1996.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. How can we fool lime and shap? adversarial attacks on post hoc explanation methods. arXiv preprint arXiv:1911.02508, 2019.

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

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In 23rd International Conference on Intelligent User Interfaces, pages 293–304, 2018.

C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2020.

Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Dan Weld, and Leah Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In Proceedings of CHI, 2020.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of EMNLP, 2013.

Yu Song, Shunwei Lei, Tianyong Hao, Zixin Lan, and Ying Ding. Automatic classification of semantic content of classroom dialogue. Journal of Educational Computing Research, page 0735633120968554, 2020.

Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. In Proceedings of CSCW, 2019.

Marc Steen, Menno Manschot, and Nicole De Koning. Benefits of co-design in service design projects. International Journal of Design, 5(2), 2011.

Hendrik Strobelt, Daniela Oelke, Bum Chul Kwon, Tobias Schreck, and Hanspeter Pfister. Guidelines for effective usage of text highlighting techniques. IEEE transactions on visualization and computer graphics, 22(1):489–498, 2015.

Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE transactions on visualization and computer graphics, 24(1):667–676, 2017.

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. S eq 2s eq-v is: A visual debugging tool for sequence-to-sequence models. IEEE transactions on visualization and computer graphics, 25(1):353–363, 2018.

Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. International journal of human-computer studies, 67(8):639–662, 2009.

Masashi Sugiyama and Motoaki Kawanabe. Machine learning in non-stationary environments: Introduction to covariate shift adaptation. MIT press, 2012.

Supreme Court of the United States. Daubert v. merrell dow pharmaceuticals, inc., 1993. 509 U.S. 579.

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

Abhijit Suresh, Tamara Sumner, Isabella Huang, Jennifer Jacobs, Bill Foland, and Wayne Ward. Using deep learning to automatically detect talk moves in teachers' mathematics lessons. In 2018 IEEE International Conference on Big Data (Big Data), pages 5445–5447. IEEE, 2018.

Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. Automating analysis and feedback to improve mathematics teachers' classroom discourse. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 9721–9728, 2019.

Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Enhancing human learning via spaced repetition optimization. Proceedings of the National Academy of Sciences, 116(10):3988–3993, 2019.

Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. In Proceedings of ACL, 2014.

Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. Investigating human+ machine complementarity: A case study on recidivism. arXiv preprint arXiv:1808.09123, 2018.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 107–118, 2020.

Lisa Torrey and Jude Shavlik. Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pages 242–264. IGI global, 2010.

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

Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–17, 2021.

United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. State court processing statistics, 1990-2009: Felony defendants in large urban counties., 2014.

Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. Contrastive explanations with local foil trees. arXiv preprint arXiv:1806.07470, 2018.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of NeurIPS, 2017.

Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 chi conference on human factors in computing systems, pages 1–14, 2018.

Jesse Vig. Deconstructing BERT: Distilling 6 Patterns from 100 Million Parameters. `https://towardsdatascience.com/\deconstructing-bert-distilling-6-patterns-from\-100-million-parameters-b49113672f77`, 2019. [Online; accessed 27-Apr-2019].

Joke Voogt, Hanna Westbroek, Adam Handelzalts, Amber Walraven, Susan McKenney, Jules Pieters, and Bregje De Vries. Teacher learning in collaborative curriculum design. Teaching and teacher education, 27(8):1235–1244, 2011.

Joke Voogt, Therese Laferriere, Alain Breuleux, Rebecca C Itow, Daniel T Hickey, and Susan McKenney. Collaborative design as a form of professional development. Instructional science, 43 (2):259–282, 2015.

Joke M Voogt, Jules M Pieters, and Adam Handelzalts. Teacher collaboration in curriculum design teams: effects, mechanisms, and conditions. Educational Research and Evaluation, 22(3-4): 121–140, 2016.

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

Aldert Vrij. Detecting lies and deceit: The psychology of lying and implications for professional practice. Wiley, 2000.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. 2017.

Donna L Waddell and Nancy Dunn. Peer coaching: the next step in staff development. The Journal of Continuing Education in Nursing, 36(2):84–89, 2005.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. arXiv preprint arXiv:1807.00199, 2018.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. Allennlp interpret: A framework for explaining predictions of nlp models. arXiv preprint arXiv:1909.09251, 2019.

Margaret Walshaw and Glenda Anthony. The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. Review of educational research, 78(3):516–551, 2008.

Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–24, 2019a.

Dakuo Wang, Q Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini. How much automation does a data scientist want? arXiv preprint arXiv:2101.03970, 2021.

Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, page 601. ACM, 2019b.

Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In 26th International Conference on Intelligent User Interfaces, pages 318–328, 2021.

Andrew Wayne, Michael Garet, Alison Wellington, and Hanley Chiang. Promoting educator effectiveness: The effects of two key strategies. ncee 2018-4009. National Center for Education Evaluation and Regional Assistance, 2018.

Noreen M Webb, Megan L Franke, Marsha Ing, Jacqueline Wong, Cecilia H Fernandez, Nami Shin, and Angela C Turrou. Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. International Journal of Educational Research, 63:79–93, 2014.

Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. A human-grounded evaluation of shap for alert processing. arXiv preprint arXiv:1907.03324, 2019.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. IEEE transactions on visualization and computer graphics, 26(1):56–65, 2019.

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

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In Proceedings of NAACL, 2019.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In Proceedings of EMNLP, 2019a.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. arXiv preprint arXiv:1908.04626, 2019b.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. arXiv preprint arXiv:2005.00582, 2020.

Anna Winterbottom, Hilary L Bekker, Mark Conner, and Andrew Mooney. Does narrative information bias individual's decision making? a systematic review. Social science & medicine, 67 (12):2079–2088, 2008.

Thomas Wolf. huggingface/pytorch-pretrained-bert. `https://github.com/huggingface/pytorch-pretrained-BERT`, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, 2019.

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

Mike Wu, Milan Mosse, Noah Goodman, and Chris Piech. Zero shot learning for code education: Rubric sampling with deep learning inference. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 782–790, 2019a.

Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun. Deep image: Scaling up image recognition. arXiv preprint arXiv:1501.02876, 2015.

Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pages 1180–1192, 2017.

Tongshuang Wu, Daniel S Weld, and Jeffrey Heer. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. ACM Transactions on Computer-Human Interaction (TOCHI), 26(4):1–27, 2019b.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In CHI Conference on Human Factors in Computing Systems, pages 1–22, 2022.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale. In Proceedings of the 26th International Conference on World Wide Web, pages 1391–1399, 2017. ISBN 978-1-4503-4913-0.

Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 842–850, 2015.

Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2020.

Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 4477–4488, 2016.

Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–11, 2019.

Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In Proceedings of the 2020 chi conference on human factors in computing systems, pages 1–13, 2020.

Qiang Ye, Rob Law, Bin Gu, and Wei Chen. 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):634–639, 2011.

Yelp. Yelp dataset 2019. `https://www.yelp.com/dataset`, 2019. Accessed: 2019-01-01.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, page 279. ACM, 2019.

Kyung-Hyan Yoo and Ulrike Gretzel. Comparison of deceptive and truthful travel reviews. Information and communication technologies in tourism 2009, pages 37–47, 2009.

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. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In Proceedings of WWW (Companion), 2018.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. “why should you trust my explanation?” understanding uncertainty in lime explanations. arXiv preprint arXiv:1904.12991, 2019.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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, pages 295–305, 2020.

Ziqiong Zhang, Qiang Ye, Rob Law, and Yijun Li. 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):694–700, 2010.

Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW):1–22, 2017.

Yiming Zheng, Serena Booth, Julie Shah, and Yilun Zhou. The Irrationality of Neural Rationale Models. arXiv:2110.07550 [cs], October 2021. URL `http://arxiv.org/abs/2110.07550`. arXiv: 2110.07550.

Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In Proceedings of AAAI, 2015.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. Proceedings of the IEEE, 2020.