What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play

Shi Feng
University of Maryland
shifeng@cs.umd.edu

Jordan Boyd-Graber
University of Maryland
jbg@cs.umd.edu

Abstract

Machine learning is an important tool for decision making, but its ethical and responsible application requires rigorous vetting of its interpretability and utility: an understudied problem, particularly for natural language processing models. We design a task-specific evaluation for a question answering task and evaluate how well a model interpretation improves human performance in a human-machine cooperative setting. We evaluate interpretation methods in a grounded, realistic setting: playing a trivia game as a team. We also provide design guidance for natural language processing human-in-the-loop settings.

1 Introduction

Machine learning (ML) is integrated into real-world decision making processes as the field makes rapid progress, even surpassing human performance on many tasks, such as image classification [20], playing video games [41], and playing Go [53]. ML could replace humans on these tasks. However, to maximize the social good ML brings, replacing humans completely should not be the goal. In sensitive areas such as medicine and criminal justice, the computational objectives for training ML models cannot fully describe the factors one must consider when making a decision. In some other areas such as natural language processing, the strengths of humans and computers are sometimes complimentary. For example, humans are excellent at reasoning about what we consider "common sense", while tasks such as disambiguating word senses is still difficult for computers [44]; tasks like deceptive review detection is difficult and time consuming for humans, while simple linear ML models achieve high accuracy with little processing time. In these cases, it would be most effective and efficient to have humans and ML models cooperate.

The cooperation is only effective if the two parties can communicate with each other. One direction of this communication is well-studied: ML models can be improved with human feedback with methods such as reinforcement learning [58] and imitation learning [49, 48]. The communication from ML models to humans presents different challenges: the reasoning behind a prediction is hidden in the model's parameters, and humans only see the conclusions without any justification. This challenge is particularly significant for modern neural networks with as many as millions of parameters—they appear as inscrutable blackboxes. Standard output from a neural network classifier is a prediction (e.g., an object class), optionally augmented with the corresponding confidence score (a value between zero and one). This is not informative or intuitive; moreover, due to overfitting, the confidence score is often overly high [18], making the model output even less intuitive for humans. To improve this communication by more expressive feedback to the humans, interpretation methods explain the predictions of a ML model in a human-intelligible way. In a cooperative setting, the role of interpretation is to connect humans and computers and to help humans decide to trust the model prediction or not.

Progress in ML research largely relies on rigorous evaluations. To maximize the effectiveness of this cooperative system, we need to evaluate the interpretation in isolation. However, this is more involved than running models on test data. As accepted as interpretability is as a goal, it remains elusive to measure. The first challenge is the lack of ground truth. As Lipton [38] argues, these no clear agreement on what interpretability means. There is no definitive answer to what interpretation is best at faithful to the model and useful to humans at the same time. Secondly, humans, the eventual consumer of interpretations, are integral to systems [43]. Previous work focused on how humans can use interpretations to help the model do its job better. For example, the interpretability of Local Interpretable Model-Agnostic Explanations [46, LIME] helps humans do feature engineering to improve downstream predictions of a classifier. Alternatively, interpretations are evaluated by how much they help humans debug ML models [47, 12].

Kleinberg *et al.* [28] asks how ML can improve human decision making. Applying this thinking to the evaluation, we measure interpretability by asking what ML can do for humans: a good interpretation is one that *augments* [30] human intelligence. This concept resonates with the seminal work of mixed-initiative user interfaces [23], which emphasizes user interfaces where the human and the computer can drive towards a shared goal and ones that enhance human ability [2].

Specifically, we measure which *form* of interpretation best helps humans on tasks at hand. We focus on three forms of interpretation popular among the interpretable ML community: highlighting important features in the input, showing relevant training examples, and visualizing uncertainty. After introducing common methods for generating interpretation in these forms (Related Work), we detail the evaluation method (Setup). In this work we focus on the comparison *between* forms of interpretation, using one method for each form. However, as we will discuss later, our framework, interface, and experiments can be easily generalized to the comparison *within* each from, between different underlying algorithms.

The testbed for our interpretability evaluation was chosen from the natural language domain—a question answering task, Quizbowl [6]. In addition to being a challenging task for ML, it is also an exciting game that is loved by trivia enthusiasts. Furthermore, it is a task where humans and ML have complementary strengths, so effective cooperation and interpretation have great potential (Interpretation Testbed).

We recruit both Quizbowl enthusiasts and Amazon Mechanical Turkers (novices in comparison) to play Quizbowl on an interactive interface, provide them different combinations of the interpretations and measure how their performance changes. These different user groups reveal imperfections in how we communicate how a computer is answering questions. Experts have enough world and task expertise to confidently overrule when the computer is wrong. However, novices are too trusting. In Discussion, we propose how to can explore new interpretations and visualizations to help humans more confidently interpret ML algorithms.

2 Related Work

2.1 Human-AI Cooperation

While the explosion of ML and explaining computer decisions is recent, explainability stretches back to expert systems [59]. Early work already recognized the potential lack of usability of AI: Suchman [56] criticizes that AI's rigid concepts of "plans and goals" are incompatible with how people behave in the real world. The recent surge of interest in this area was a result of the success of ML models based on neural networks, a.k.a. deep learning [34]. These complicated models have stupendous predictive power, however their fairness, accountability, and transparency remains a concern [61]. Such concern is reflected in the "right to explanation" in GDPR [10]. From a practical standpoint, the inscrutability of these models makes it difficult to integrate into real world decision-making in high risk areas such as urban planning, disease diagnosis, predicting insurance risk, and criminal justice. As an example, Schmidt and Herrmann [51] recognize the importance of interpretability when interacting with autonomous vehicles.

Thus, ML model predictions need explanations. Efforts including the Explainable AI (XAI) initiative [17] led to the conceptualization of a series of human-AI cooperation paradigms, including human-aware AI [7], and human-robot teaming [62]. This also motivated the ML community to develop interpretation methods for deep neural models [4, 54].

Although interpretation methods have rigorous mathematical formulations, some even axiomatically derived [57], it remains unclear how we can evaluate the efficacy of these methods on *real tasks* with *real users*. Lipton [38] argues that there is no clear agreement on what interpretability means: looking at ML models alone, no definitive answer exists as in what would be the best interpretation in both faithfulness to the model and usefulness to humans. In attempt to define interpretability rigorously, Doshi-Velez and Kim [9] provide a taxonomy of evaluation settings, but also highlighted the difficulty of realistic evaluation. To borrow insights from the human side, Miller [40] provides an overview of social science research regarding how people define, generate, select, evaluate, and present explanations.

The HCI community has a rich body of research towards making computers more usable, for example in interaction design [26] and software learnability [16]. Still, interpreting ML models has its unique challenges. Krause *et al.* [32] compared different ML models under one visualization method, partial dependency. Smith *et al.* [55] and Lee *et al.* [35] focused on the interpretation of topic models. In contrast, we compare interpretation of classification models across various forms, making our framework more generalizable to other tasks and interpretation methods.

2.2 Interpretation of Machine Learning Models

Some ML models are inherently interpretable. For example, sparse linear classifiers used by LIME [46], decision trees and association rule lists [33, 36]. However, most state-of-the-art models in vision and language—domains with the widest range of applications—are deep neural models with hundreds of thousands of parameters. To improve their usability, we need interpretation methods. In addition to conveying and visualizing the uncertainty of a prediction, researchers focus on two classes of methods: feature-based and example-based.

Conveying Uncertainty Augmenting the prediction from a neural network classifier with a confidence score (a value between zero and one) conveys the uncertainty of the model. In a cooperative setting, the uncertainty helps humans decide to trust the model or not [3, 50]. To make it more informative, we can also display the confidence scores for the classes other than the top one [39]. Estimating uncertainty for a deep neural model can be challenging: due to overfitting, its confidence can be overly high and requires calibration [18]. Sometimes uncertainty estimate needs to be combined with anomaly detection to be robust [21, 11].

Feature-based Interpretations Model predictions can be explained by highlighting the most salient features in the input, typically visualized by a saliency map. For a linear classifier, the most salient features are the ones with the largest corresponding weights. For non-linear classifiers, the saliency map can be calculated by the gradient of the loss function w.r.t. each input feature [54]. Alternatively, one can locally approximate a non-linear classifier with a simpler linear model, then use the weights to explain the predictions from the non-linear model [46].

Example-based Interpretations We can explain by example, by finding the most influential training examples for the prediction on a test example. The influential examples can be found by nearest neighbor search in the representation space, which is natural to clustering algorithms and their deep variation [45]. Examples can also be found according to other definitions of importance: e.g., influence functions to find examples for neural models [31].

2.3 Evaluation of Interpretation

A fair and accurate assessment of interpretations is crucial for improving the understability of AI and consequently human-AI cooperation. That an interpretation is visually pleasing or confirms existing assumptions is not enough, as it can lead to a false sense of security. Evaluation beyond isolated, natural examples is necessary to ensure the robustness of interpretation, especially for neural networks [13, 27, 22, 1, 25].

Conditioning a more realistic setting, Doshi-Velez and Kim [9] provide an ontology of interpretation evaluations with a human in the loop. Following this framework, Narayanan *et al.* [43] conducted one such evaluation with synthetic tasks and hand-crafted interpretations to study the desirable cognitive properties of interpretations.

Application-grounded is the most realistic setting in the taxonomy of Doshi-Velez and Kim [9], which means evaluating with real users on real tasks. This setting best aligns with what interpretations are intended for—improving human performance on the end task. However, application-grounded evaluation is also challenging because it requires real tasks and real, motivated users. The task needs a large pool of willing human testers, and ideally one that challenges both humans and computers.

3 Interpretation Testbed: Quizbowl

We use Quizbowl as a testbed for evaluating the effectiveness of three forms of interpretation. This section introduces Quizbowl: the task, the model, and how we generate interpretations. We also discuss why Quizbowl is an effective testbed for interpretability.

3.1 Quizbowl and Computer Models

Quizbowl is both a challenging task for ML [6], and a trivia game played by thousands of students around the world each year. Each question consists of multiple clues, presented to the players word-by-word, verbally or in text. The ordering of Quizbowl clues is *pyramidal*—different clues at the beginning, easy clues at the end, and the challenge is to answer with as few clues as possible. For a question with n words, the players have n chances to say *this is all the information I need to answer the question*. The player can do so by buzzing, which interrupts the readout so the player provide an answer. Whoever gets the answer correct first wins that question and receives ten points.[1] But when players buzz and answer incorrectly, they lose five points. Success in Quizbowl requires a player to not only be knowledgeable but also balance between aggressiveness and accuracy [19].

QANTA [24] is a simple, powerful, and interpretable system for Quizbowl. A stripped-down, minimal version of it is provided to participants in the NIPS 2017 Human-Computer Question Answering competition [5]. We use the *guesser* of QANTA, which has a linear decision function built on ElasticSearch [14, ES]. As the name implies, guesser generates guesses for what the answer to a question could be. Despite the simplicity and interpretability, ES-based systems perform very well on Quizbowl, defeating top trivia players.[2]

Our goal, however, is not to defeat humans. We team humans and computers together and use their cooperation to measure the effectiveness of interpretations. In our cooperative setting, instead of having a model to decide when to buzz in, *the human needs to decide when the system has a good guess*.

When answering a Quizbowl question—which takes many steps, the human constantly interacts with the model, which provides many opportunities to evaluate the interpretability of models. Every word provides new evidence

[1] Like previous work, we only consider *toss-up/starter* questions.

[2] `https://youtu.be/bYFqMINXayc`

that can change the underlying interpretation and convince the human that the system has a good answer to offer.

Quizbowl is challenging for humans and computers in different ways [6, 63]. Computers can memorize every poem and book ever written, making it trivial to identify quotes. Computers also can memorize all of the *reflex clues* that point to answers (e.g., if you hear "phosphonium ylide", answer "Wittig") and apply them without any higher reasoning. Humans can chain together evidence ("predecessor of the Queen who pardoned Alan Turing") and solve wordplay ("opera about an enchanted woodwind instrument").

Thus, Quizbowl is representative of tasks where the cooperation between the two has huge potential [60]. This also makes Quizbowl a suitable testbed for interpretation methods designed to better interface humans and computers. Furthermore, the competitiveness of Quizbowl encourages humans to use the help from the computers as much as possible, avoiding a degenerate scenario where the users solve the task on their own. It also attracts a large pool of enthusiastic participants, which is crucial for application-grounded evaluations.

3.2 Interpretation of the QANTA Model

ES, like many other linear models, is interpretable. For a new question, ES mainly uses tf-idf features to find the most relevant training example, which is either a Wikipedia page or a previously seen Quizbowl question, and use the label of that document as the answer.

Our goal is to see which forms of interpretation are most helpful to the users, and using a linear model with natural interpretations makes this easy. Our Quizbowl system supports three forms of interpretation, each corresponding to a class of methods widely studied in recent literature as mentioned in the previous section.

To show the *uncertainty* of the model, we show **guesses and scores** from our guesser. We take the top ten guesses sorted by their corresponding scores. Regular classification models output a probability over all possible answers, however, the relevance score returned by ES is not normalized. We keep them unnormalized to stay true to the model. Despite its simplistic form, these scores provide strong signal about model uncertainty.

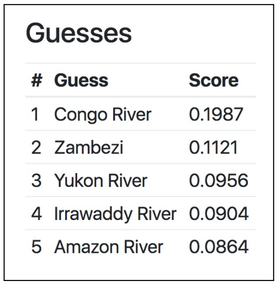
For *example-based interpretation*, we show the **evidence** for the top guess from our guesser. For neural models, extracting salient training examples can be challenging [31, 45], but it is straightforward for our ES-based model. Because the prediction is the label of the most relevant documents, the extracted documents are naturally the most salient training examples, and the overlapped words are the most salient features.

For *feature-based interpretation*, we **highlight** the most salient words in the question, generating a saliency map. In contrast to salient training examples, this interpretation provides a local explanation of what words in the input the model focuses on. For neural models, the calculation of the saliency map usually involves gradients [54, 52, 57]. For our ES-based model, the salient words naturally emerge by comparing the input question against the most salient training example. We use the highlight API[3] to find the most important words in the training example, find their appearances in the input question, and highlight them for feature-based interpretation.

4 Interface Design

We design our Quizbowl interface (Figure 1) to *visualize* the three interpretations described in the previous section. This section introduces the visualizations, placement, and interactivity of the interface.

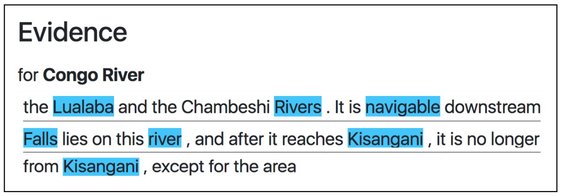
To make Quizbowl players feel at home, we follow the general framework of Protobowl.com, a popular Quizbowl platform that many players actively use for practice. The **Question** area is in the center, and the question is displayed word-by-word in the text box. A **Buzz** button is located close above the question area, and to further reduce the distraction from the question area, players can also buzz in using the space key. After buzzing, the player have eight seconds to enter and select an answer from a drop-down menu.



Guesses

#	Guess	Score
1	Congo River	0.1987
2	Zambezi	0.1121
3	Yukon River	0.0956
4	Irrawaddy River	0.0904
5	Amazon River	0.0864

Guesses show the answers the system is considering along with the associated score. Top ten answers are sorted according to their score (the system prefers higher scores). This helps convey when the model is uncertain (e.g., if all of the guesses have a low score).



Evidence

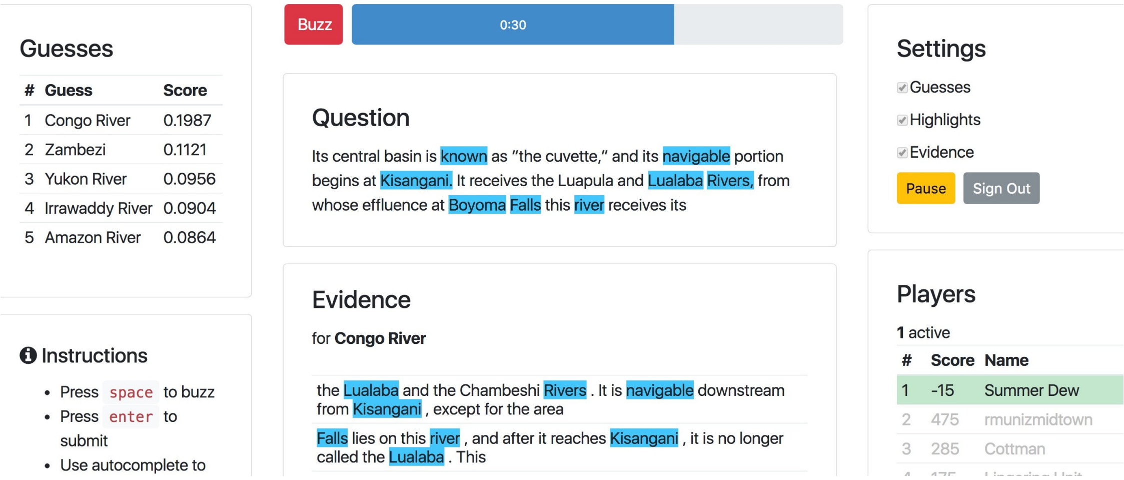
for **Congo River**

the Lualaba and the Chambeshi Rivers . It is navigable downstream

Falls lies on this river , and after it reaches Kisangani , it is no longer

from Kisangani , except for the area

[3] https://www.elastic.co/guide/en/elasticsearch/reference/current/search-request-highlighting.html



Guesses

#	Guess	Score
1	Congo River	0.1987
2	Zambezi	0.1121
3	Yukon River	0.0956
4	Irrawaddy River	0.0904
5	Amazon River	0.0864

Instructions

- Press space to buzz
- Press enter to submit
- Use autocomplete to

Question

Its central basin is known as "the cuvette," and its navigable portion begins at Kisangani. It receives the Luapula and Lualaba Rivers, from whose effluence at Boyoma Falls this river receives its

Evidence

for **Congo River**

the Lualaba and the Chambeshi Rivers . It is navigable downstream from Kisangani , except for the area

Falls lies on this river , and after it reaches Kisangani , it is no longer called the Lualaba . This

Settings

☑ Guesses

☑ Highlights

☑ Evidence

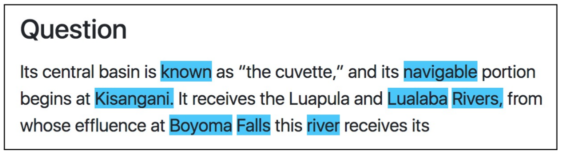
Players

1 active

#	Score	Name
1	-15	Summer Dew
2	475	rmunizmidtown
3	285	Cottman

Figure 1: Screenshot of the interface. Question is displayed in the middle area word-by-word, with question highlights displayed in the same panel. Guesses are listed in the panel on the left. Evidence is in the panel below.

To inform the player of how the model's prediction is supported by training examples, **Evidence** shows the relevant snippets of the most salient training examples for the top guess. It is located below the question area and has the same width to provide a direct comparison against the input question. Each line of the text area shows the snippet of one selected document.



Question

Its central basin is known as "the cuvette," and its navigable portion begins at Kisangani. It receives the Luapula and Lualaba Rivers, from whose effluence at Boyoma Falls this river receives its

We use **Highlight** to visualize the most salient words in both the input question and the evidence snippets. These words are selected for the top guess. As introduced in the previous section, we first highlight important words in the training example snippets using an API of ES, then find their appearances in the input and highlight those too.

These visualizations can be shown in combination. The combination of highlight and evidence has a compounding effect: when both are enabled, players see highlighted words in both the question and the evidence (for example in Figure 1); when highlight is enabled without evidence, players only see highlights in the question.

Our main goal of the layout design is to minimize distraction from the question area while boosting the competitiveness of the player. So we place the question area in the middle and have all visualizations around it. It is difficult to ensure in design that different forms of visualizations are exposed to the users equally, as some forms (e.g., evidence) are inherently less intuitive to visualize. However, all visualizations must be implemented in an interface for a real-world evaluation; we discuss the limitations of our design and future work in Discussion.

5 Setup

This section explains how we use our interface to evaluate the visualizations introduced in the previous section. Human players and the computer guesser play in cooperation, and the effectiveness of each visualization is measured by the change in human performance with and without it. To ensure accuracy and unbiasedness, we control what visualizations each player sees instead of letting them choose.

5.1 Data and Participants

We collected 160 new questions for this evaluation that is not previously seen by the Quizbowl community to avoid bias in players' exposure to questions.

We recruited 40 experts (Quizbowl enthusiasts) by advertising on an online forum, and 40 novices using mechanical Turkers. Experts are free to play as many questions as they want (but each player can only play a question once), and we encourage them to play more by offering monetary prizes for those who finishes the whole question set. We require the novices to each answer at least 20 questions, and require a positive score at the end (according to standard Quizbowl scoring rules) to encourage good faith responses. Online Quizbowl platforms such as `Protobowl.com` are usually anonymous, so we did not collect any information about the participants other than an email address for collecting prizes (optional).

5.2 Human-AI Cooperation on Quizbowl

Unlike previous work where Quizbowl interfaces are used for computers to *compete* with humans [6, 19], our interface aims at human-AI *cooperation*. We let a human player form a team with a computer teammate and put the human is in charge. The visualizations introduced in the previous section are different ways for the computer to communicate with the human.

We have two different settings for the cooperation. In the **novice setting**, we have one player on the interface, with the computer guesser as teammate, but without opponents.

The **expert setting** is more competitive than the novice setting. First, the experts enjoy and are (by definition) experts at the task. To encourage them to play to the best of their ability, we simulate the Quizbowl setting as closely as possible (for novices the simple task is already taxing enough without competition). In a real Quizbowl match, players not just compete against themselves (can I get the question right) but also with each other (can I get the question right before Selene does). Just as Quizbowl's incremental structure helps us evaluate changing visualizations, Quizbowl's structure encourages competition: more difficult clues at the start of the question help determine who knows the most about a subject.

To best recreate that environment, expert players compete with each other. Our site resembles `Protobowl.com`, a popular online Quizbowl platform where players play against each other (but without the computer teammate). The computer provides a consistent guess for all players, but human players might see different visualizations (details in the next subsection). The competitiveness of the expert setting is familiar to Quizbowl enthusiasts, whom we recruit via a Quizbowl community forum.

Our experiment with the experts was possible thanks to Quizbowl's enthusiast community. It was because Quizbowlers love to play this game and to improve their skills by practicing, that they were willing to learn our interface, team up with the computer, and compete under this slightly irregular setting. This provided us new perspectives of how users from a wider range of skill levels use visualizations, compared to many previous work that only had non-expert mechanical Turkers [55, 29, 8].

5.3 Controlling Visualizations

We consider the three visualizations, and in total eight combinations (including no visualization).

To compare within-subjects (players vary greatly based on their innate ability), we vary the visualizations a player sees randomly. We sample the enabled combination with the goal of having, in expectation, a uniform distribution over players, questions, and visualization combinations. For player P at question Q, we sample from an eight-class categorical distribution, with the parameter of each combination C set to $N - \#(C, P)$, where $\#(C, P)$ is the number of times player P has seen the visualization combination C and N is the expected count of each combination (in our case set to the number of questions divided by eight). In the expert setting, visualizations are sampled independently for each player, and players may (and usually do) see different visualizations.

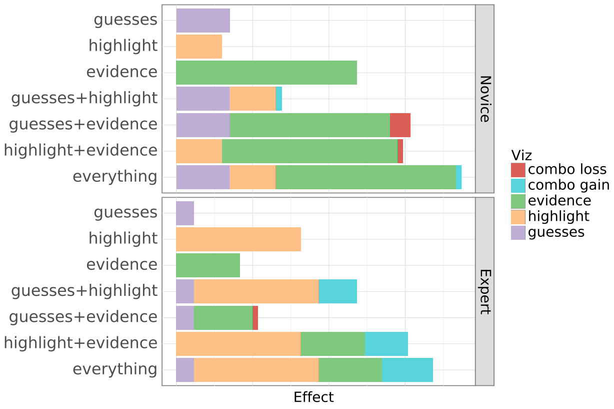


Figure 2: Visualization effects from the regression analysis, for novices (above) and experts (below). Higher value means a visualization helps more in player accuracy. In addition to the individual visualizations, *combo gain* and *combo loss* capture the additional effect of combining of multiple visualizations. *Highlight* and *Evidence* are effective for both novices and experts. Combining leads to more positive effect for experts than novices, potentially because experts can process more information in limited time.

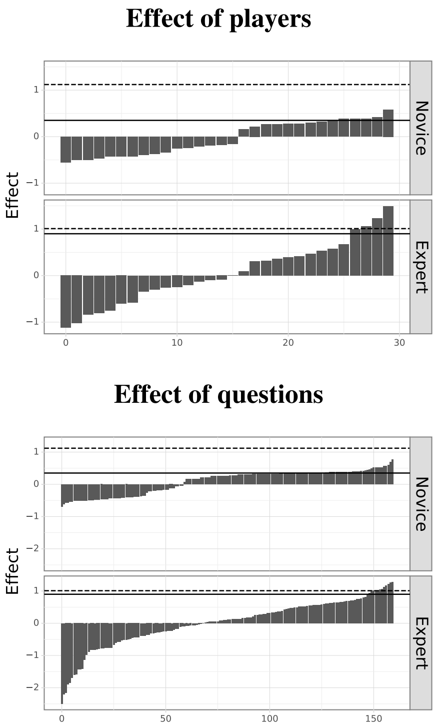
Each of the three visualizations can be turned on or off, so we have in total $2 \times 2 \times 2 = 8$ conditions, including the null condition where all visualizations are hidden.

For all experiments, we only allow each player to answer each question once.

6 Results

With the game play data, our primary goal is determine if the visualizations were helpful or not, and how experts and novices used them differently. We first use a regression analysis to quantitatively determine how much each condition affects the accuracy of the players; then we break down the results to see how the players behave differently under the conditions, including how aggressive they are; we also look at specific cases where some visualization consistently succeeded or failed to convince multiple players of the model prediction.

After filtering players who answered very few questions,


Effect of players

Effect of questions

Figure 3: Effect of players and questions from the regression analysis. Solid horizontal lines show the bias term that captures the baseline accuracy; dashed lines show the effect of combining all visualizations. Experts have a higher bias, thus higher average accuracy; they are also less affected by visualizations.

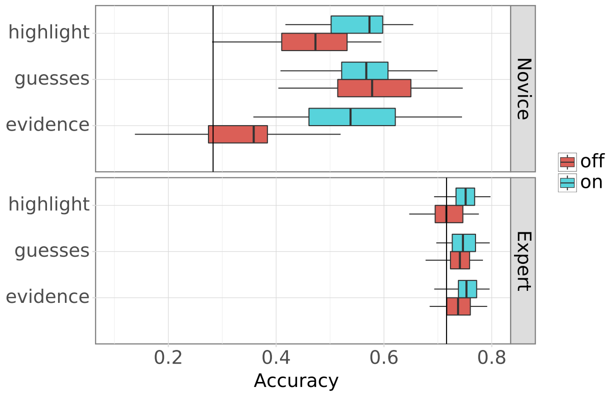


Figure 4: Accuracy of novices (above) and experts (below), with and without each visualization. One visualization being *On* means that the visualization is shown to the players (possibly in combination with others); other visualizations can still be used if one of them is *Off*. Vertical bars show the baseline accuracy without any visualization. Unsurprisingly, experts show higher performance than novices and are more consistent. Among the visualizations, evidence significantly improves novice performance.

we arrive at 30 experts that answered 1983 questions, and 30 novices that answered 600 questions. Turkers usually stopped after answering the required 20 questions, but many experts kept on playing. Among all players, 7 experts answered all 160 questions.

6.1 Regression Analysis

There are many reasons a player might answer a question correctly: their innate skill, an easy question, or the aid of some visualization. To tease apart these factors we follow Narayanan *et al.* [43] and apply a regression analysis.

Each game record is a ⟨player ID, question ID, result ⟩ triplet uniquely identifiable by the two IDs. To convert a triplet into the input to the regression model, we first extract a collection of *features*. We design four main classes of features to capture four main aspect of the game: player's innate skill level, question difficulty, interface condition, and game condition. The first two aspects can be captured by the unique IDs. Interface condition represents the visualizations shown to the player, and we have eight features corresponding to the eight conditions introduced in the previous section. For game condition, the first feature is the relatively position in the question when the player buzzed (to understand how visualizations affect buzzing position as an outcome instead of feature, we use a separate analysis); for the expert setting, we also include extra features to capture the competitiveness: number of active players and the current accuracy of the top active player.

We train a logistic regression to predict the outcome of a player answering some questions: one if the answer is correct and zero otherwise. Specifically, for each game record triplet, we extract the features and feed the input vector to the logistic regression, which then predicts the probability of a positive result; we then compare the prediction against the ground truth, and update the regression with gradient descent.

The weights of the logistic regression encode the importance of the corresponding features: the probability of a positive result increases with features with positive weights, which means these features help the players. Similarly negative weights means the features hurt the player accuracy. To understand which visualizations are most helpful to Quizbowl players, we inspect the sign and magnitude of their corresponding feature weights.

Figure 2 shows the effect of visualizations based on regression weights: a high positive weight means the visualization is useful, zero means it is ineffective, and negative means it is harmful. For a combination of visualizations, the difference between the actual weight and the sum of its parts encodes the *additional* effect of showing them together: a

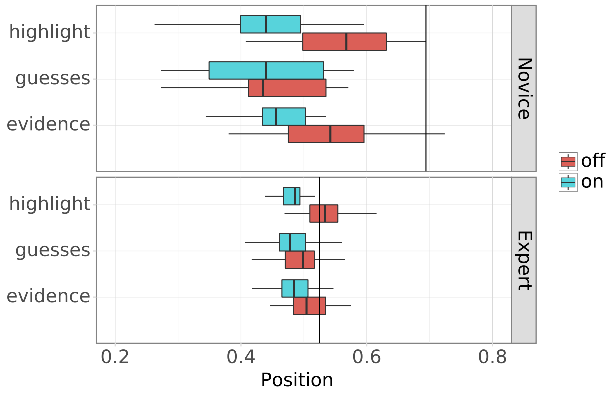


Figure 5: Average buzzing position (divided by the length of question) of novices (above) and experts (below), with and without each visualization: the goal is to buzz as early as possible. Vertical bars show the baseline buzzing position without any visualization. Like Figure 4, experts are better and more consistent. Among the visualizations, *Highlight* is most effective in helping both novices and experts answer faster.

positive difference means showing them in combination has extra benefits.

The visualization that helps novices is not the same as what helps experts. For experts, highlight is the most helpful individual visualization, while for novices, evidence is the most helpful. For experts, the combination of highlight and evidence has a positive additional impact, which is reasonable because this combination adds highlights to the evidence, making the contrast more intuitive. However, the same combination does not show additional benefit for novices.

Figure 3 shows the effect of players and questions, compared to the bias term (baseline accuracy) and the effect of combining all visualizations. Experts have higher baseline accuracy, shown by the higher bias, and are less sensitive to the visualizations.

We hypothesize that this inconsistency is because experts can use evidence more effectively. Question highlighting requires less multitasking than evidence: players have to look away from the question they need to answer to take in the evidence. Quizbowl players likely know when they can glance down to related training data and can also determine whether the training data are helpful.

6.2 How Visualizations Change Player Behavior

The regression analysis provides a quantitative comparison between all visualizations in how they affect the player accuracy. However, accuracy alone does not tell the full story of how they play the game. This section describes how each visualization affects the behavior of the players and how the effect differs for novices and experts. Ideal players should be both aggressive and accurate: seeing very few words and getting the answer right. Visualizations should help them reach this goal.

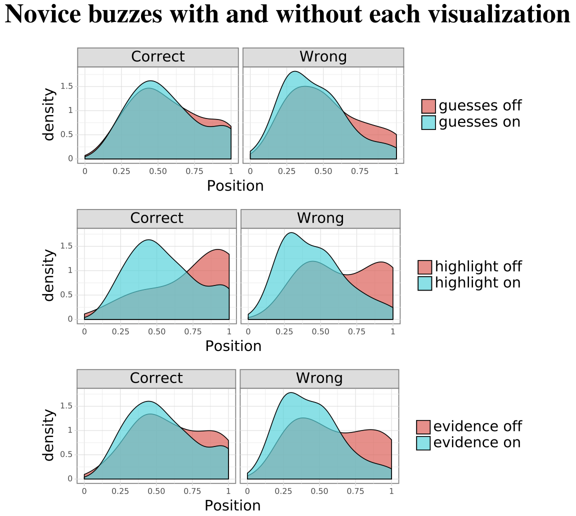


Figure 6: The distribution of buzzes of novices on correct guesses (left) and wrong guesses (left); colors indicate if each visualization is enabled; positions are normalized by question length. With visualizations, novices were significantly more aggressive, but also got more questions correct earlier. *Highlight* is the most effective.

Figure 4 and Figure 5 show the player accuracy and average buzzing position—with and without each visualization—for experts and novices separately. Because expert players are recruited from the Quizbowl community, they have higher accuracy and smaller variance (in both metrics). The effect of each visualization on the accuracy of experts and novices concurs (Figure 2). However, novices generally buzz at about the same point as experts (Figure 5), which suggests the novices are playing too aggressively for their skill level.

We show the difference in aggressiveness by plotting the density of buzzing positions (experts in Figure 7 and novices in Figure 6). In all settings, the density shifts earlier: players are more aggressive with visualizations, especially for novices, which is consistent with Figure 5. The visualizations allow players to answer correctly earlier. Especially for novices with highlights, the distribution of correct buzzing positions shifts significantly earlier in the questions.

Although novices are helped by visualizations, these visualizations are not enough to help them discern useful help from misleading help. Novices are too aggressive at the start of the question with visualizations: they trust the pre-

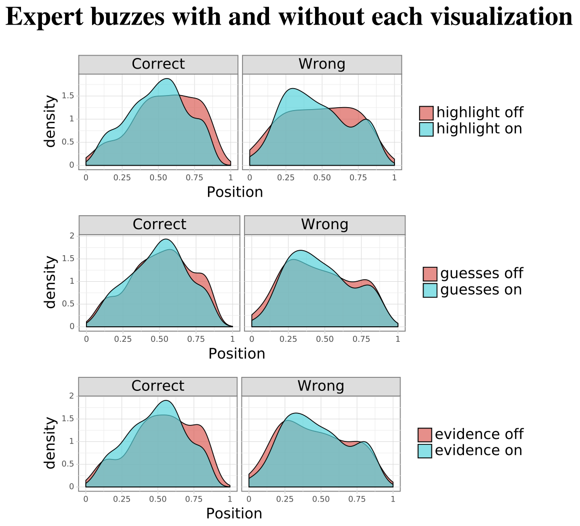


Figure 7: The distribution of buzzes of experts on correct guesses (left) and wrong guesses (left); colors indicate if each visualization is enabled; positions are normalized by question length. Experts were not significantly more aggressive with visualizations, but they did get more answers correct earlier.

dictions of the system too much. While experts mentally tune out bad suggestions, novices are less discerning. Visualizations thus must also convey whether they should be trusted, not just what answer they are suggesting.

6.3 Successes and Failures of Interpretations

We now examine specific cases where interpretations help or hurt players.

Figure 8 shows an example where interpretations enable players to answer correctly. In total of twelve expert players answered the shown question, and eight answered correctly. The earliest an expert answered correctly without the evidence was at 72% of the question, while the three experts with the evidence all answered correctly before 50%. With the evidence and highlight, players can infer from the keywords that the author is “Thoreau”, and that the guess is likely correct. The computer shows a salient training example and is effective in convincing the players that the retrieved evidence is correct.

Question:
(This essay) was composed after its author **refused** to pay a **poll tax** to support the **Mexican-American war**, and its ideology inspired Martin Luther King, Jr. and Mohandas Gandhi.
Evidence:
him to pay six years of delinquent **poll tax**. Thoreau **refused** because of his opposition to the **Mexican-American War** and slavery, and he spent a night in jail because of this refusal.

Figure 8: Interpretations that help players answer a question on “ Civil Disobedience” correctly. With the shown part of the question, three experts answered correctly with the evidence; no expert answered correctly without.

Question:
A **book** by this man was first published with a **preface** by Andreas Osiander titled **Ad Lectorem**.
Evidence:
the **Ad Lectorem preface** to Copernicus’s **book** was not actually by him.

Figure 9: Interpretations that fail to convince players. Three expert players, when presented with the interpretation (some question text and evidence omitted), rejected the computer’s correct guess (Copernicus) and answered differently.

Figure 9 shows a failure to convince, where the combination of highlight and evidence fails to convince the player of the computer’s *correct* guess: three expert players rejected the computer’s prediction and provided different answers, relatively early in the question (before 50%). The information provided by the evidence is that Copernicus has a book with a preface named Ad Lectorem, this piece of evidence strongly supports the computer’s guess “Copernicus”. However, it is expressed differently than the question, with an unrelated but confusing “not” in the middle of the sentence.

7 Discussion

7.1 Experts vs. Novices

In our cooperative setting, experts and novices use model interpretations differently. Novices tend to be too trusting, playing much more aggressively when visualizations are enabled (Figure 5). Experts tend to use interpretations that require less processing (Figure 2). For tasks where computer performance is on par with the best human players, but they have different strengths, interpretations must be intuitive, easy to process, and convey uncertainty. An unstructured saliency mapping over input features might not be effective enough (Figure 9).

7.2 Intrinsic vs. Extrinsic Evaluation

Our approach is an extrinsic evaluation [43]. The task is played by thousands who compete in regularly. Using Quizbowl allows a contextual, motivated evaluation of whether an interpretation is useful. In contrast, intrinsic evaluation relies on the interpretation alone. It is more direct but limited. In tasks where no ground-truth explanation is available, the most tractable and commonly used method is to construct ground-truth using a simpler model as a benchmark for interpretability. For example, weights of linear models are used for evaluating input highlight explanations [37, 42]. This is restricted to tasks where the benchmark model performs similarly to the complex model that requires interpretation, and it does not work in application-grounded setting (Section 3).

Extrinsic evaluations are hard to design, as they are affected by more factors, especially humans' trust. When a user does not trust the model and ignores it, the difference in the performance is not affected by the explanations at all. Narayanan *et al.* [43] uses "alien" tasks to enforce trust, tasks that humans do not have knowledge of. Our approach, in contrast, considers trust as an inherent part of the cooperation: good interpretations should be consistent and intuitive to convince humans to use it.

7.3 Generalizability and Limitations

This paper focuses on comparing forms of interpretations, so we limited the experiment to one method in each form. However, our evaluation framework, the interface, and the experimental setup can be naturally adapted to other comparisons, for example, between different underlying algorithms within each form, or between different models with interpretation methods fixed.

Our method can also be applied to natural language tasks other than Quizbowl, although Quizbowl's characters make it uniquely suitable. To use our interface for some other text classification task, for example sentiment analysis or spam detection, one can convert the task into an incremental version where the input is shown word-by-word. Time limitation or competition can be added to encourage the users to pay attention to visualizations [43]. One task related to Quizbowl has wide real world application: simultaneous interpretation (or simultaneous translation, not to be confused with model interpretation). Interpreters need to trade off between accuracy and delay, much like Quizbowlers need to balance accuracy and aggressiveness. The underlying mechanism of the QANTA buzzer [19] also resembles how simultaneous translation systems handle this trade-off [15].

Limitations First, because we compare visualizations individually and in combinations, their placement were fixed to avoid confusing the players. The fixed placement leads to uneven exposure to the users, so they might pay less attention to some visualizations than others. If we focus on individual visualizations, one way to resolve this issue is to display the interpretation in a single fixed location, for example below the question area. This would lead to a fair display of different visualizations without confusing the users. However, one single location might not suit all visualizations: for example, input highlight should collocate with the input, while evidence is best displayed next to the input for comparison.

Visualizations displayed on our interface change from question to question, and the randomization (Setup) might confuse the users. Before answering questions, each user sees a tutorial that walks through the components of the interface, but this can be improved by a set of warm-up questions to familiarize the users of the interaction, which we will implement in future studies. In addition, we can randomly sort the questions instead of the visualizations, so the users see the same layout for multiple questions, reducing context switches and consequently the cognitive load.

7.4 Future Work

While we focus on broad categories of interpretations to reveal that some visualizations are more effective than others (e.g., highlighting is more useful than guess lists), we can also use this approach to evaluate specific highlighting methods in a task-based setting. This can help reveal how best to choose spans for highlighting, which words are best suited for highlighting, and how to convey uncertainty in highlighting.

While our evaluation has focused on the downstream task, we can expand our analysis to measure how much users look at visualizations and in what contexts (e.g., with an eye tracker). This would reveal situational usefulness of visualization components; if, for example, highlighting were only useful to distinguish when two guesses had similar scores, we could decrease cognitive load by only showing highlights when needed.

A tantalizing extension is to make these modifications automatically, using the reward of task performance to encourage a reinforcement learning algorithm to adjust interface elements to optimize performance: such as changing font sizes, setting buttons for users to explicitly agree or disagree with model predictions, or modifying the highlighting strategy.

8 Conclusion

We focus on the evaluation of interpretation methods in a human-AI cooperative setting. We evaluate model interpretations in the language domain using a competitive question answering task, Quizbowl. Our experiments with both experts and novices reveal how they trust and use interpretations differently, producing a more accurate and realistic evaluation of machine learning interpretability. Our results highlight the importance of taking the skill level of the target user into consideration, and suggests that, combining interpretations more intelligently and adapting to the user, we can further improve the human-AI cooperation.

References

[1] J. Adebayo, B. Kim, I. Goodfellow, J. Gilmer, and M. Hardt. Sanity checks for saliency maps. In *Proceedings of Advances in Neural Information Processing Systems*, 2018.

[2] J. Allen, C. I. Guinn, and E. Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 1999.

[3] S. Antifakos, N. Kern, B. Schiele, and A. Schwaninger. Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the international conference on Human-computer interaction with mobile devices and services*, 2005.

[4] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 2010.

[5] J. Boyd-Graber, S. Feng, and P. Rodriguez. *Human-Computer Question Answering: The Case for Quizbowl*. Springer, 2018.

[6] J. L. Boyd-Graber, B. Satinoff, H. He, and H. D. III. Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of Empirical Methods in Natural Language Processing*, 2012.

[7] T. Chakraborti, S. Kambhampati, M. Scheutz, and Y. Zhang. AI challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775*, 2017.

[8] E. Clark, A. S. Ross, C. Tan, Y. Ji, and N. A. Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In *International Conference on Intelligent User Interfaces*, 2018.

[9] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv: 1702.08608*, 2017.

[10] European Parliament and Council of the European Union. General data protection regulation. 2016.

[11] S. Feng, E. Wallace, A. G. II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of Empirical Methods in Natural Language Processing*, 2018.

[12] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision*, 2017.

[13] A. Ghorbani, A. Abid, and J. Y. Zou. Interpretation of neural networks is fragile. *arXiv preprint arXiv: 1710.10547*, 2017.

[14] C. Gormley and Z. Tong. *Elasticsearch: The Definitive Guide*. O'Reilly Media, Inc., 2015.

[15] A. Grissom II, H. He, J. Boyd-Graber, J. Morgan, and H. Daumé III. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Empirical Methods in Natural Language Processing*, 2014.

[16] T. Grossman, G. Fitzmaurice, and R. Attar. A survey of software learnability: metrics, methodologies and guidelines. In *International Conference on Human Factors in Computing Systems*, 2009.

[17] D. Gunning. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference of Machine Learning*, 2017.

[19] H. He, J. L. Boyd-Graber, K. Kwok, and H. D. III. Opponent modeling in deep reinforcement learning. In *Proceedings of the International Conference of Machine Learning*, 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision*, 2015.

[21] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2017.

[22] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. Evaluating feature importance estimates. In *ICML Workshop on Human Interpretability in Machine Learning*, 2018.

[23] E. Horvitz. Principles of mixed-initiative user interfaces. In *International Conference on Human Factors in Computing Systems*, 1999.

[24] M. Iyyer, J. Boyd-Graber, L. M. B. Claudino, R. Socher, and H. D. III. A neural network for factoid

question answering over paragraphs. In *Proceedings of Empirical Methods in Natural Language Processing*, 2014.

[25] H. Jiang, B. Kim, and M. R. Gupta. To trust or not to trust a classifier. In *Proceedings of Advances in Neural Information Processing Systems*, 2018.

[26] W. Ju and L. Leifer. The design of implicit interactions: Making interactive systems less obnoxious. *Design Issues*, 2008.

[27] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un)reliability of saliency methods. *arXiv preprint arXiv: 1711.00867*, 2017.

[28] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 2017.

[29] R. T. Kneusel and M. C. Mozer. Improving human-machine cooperative visual search with soft highlighting. *ACM Transactions on Applied Perception*, 2017.

[30] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, and J. Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27, sep 2013.

[31] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference of Machine Learning*, 2017.

[32] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *International Conference on Human Factors in Computing Systems*, 2016.

[33] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Knowledge Discovery and Data Mining*, 2016.

[34] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 2015.

[35] T. Y. Lee, A. Smith, K. Seppi, N. Elmqvist, J. Boyd-Graber, and L. Findlater. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 2017.

[36] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 2015.

[37] J. Li, W. Monroe, and D. Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv: 1612.08220*, 2016.

[38] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv: 1606.03490*, 2016.

[39] S. Liu, X. Wang, M. Liu, and J. Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 2017.

[40] T. Miller. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[41] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

[42] W. J. Murdoch, P. J. Liu, and B. Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *Proceedings of the International Conference on Learning Representations*, 2018.

[43] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. 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.

[44] S. Papandrea, A. Raganato, and C. D. Bovi. Supwsd: A flexible toolkit for supervised word sense disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2017.

[45] N. Papernot and P. D. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv: 1803.04765*, 2018.

[46] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining*, 2016.

[47] M. T. Ribeiro, S. Singh, and C. Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the Association for Computational Linguistics*, 2018.

[48] A. S. Ross and F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AIAA*, 2018.

[49] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of Artificial Intelligence and Statistics*, 2011.

[50] E. Rukzio, J. Hamard, C. Noda, and A. De Luca. Visualization of uncertainty in context aware mobile applications. In *Proceedings of the international conference on Human-computer interaction with mobile devices and services*, 2006.

[51] A. Schmidt and T. Herrmann. Intervention user interfaces: a new interaction paradigm for automated systems. *Interactions*, 2017.

[52] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-CAM: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv: 1610.02391*, 2016.

[53] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 2017.

[54] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations*, 2014.

[55] A. Smith, T. Y. Lee, F. Poursabzi-Sangdeh, J. Boyd-Graber, N. Elmqvist, and L. Findlater. Evaluating visual representations for topic understanding and their effects on manually generated labels. *Transactions of the Association for Computational Linguistics*, 2017.

[56] L. A. Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.

[57] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference of Machine Learning*, 2017.

[58] R. S. Sutton and A. G. Barto. *Introduction to reinforcement learning*. 1998.

[59] W. R. Swartout. Xplain: A system for creating and explaining expert consulting programs. Technical report, University of Southern California, 1983.

[60] C. Thompson. *Smarter Than You Think: How Technology is Changing Our Minds for the Better*. The Penguin Group, 2013.

[61] USACM. Statement on algorithmic transparency and accountability. *Public Policy Council*, 2017.

[62] D. W. Vinson, L. Takayama, J. Forlizzi, W. Ju, M. Cakmak, and H. Kuzuoka. Human-robot teaming. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[63] E. Wallace and J. Boyd-Graber. Trick me if you can: Adversarial writing of trivia challenge questions. In *Proceedings of ACL 2018 Student Research Workshop*, 2018.