What can AI do for me?

Evaluating Machine Learning Interpretations in Cooperative Play

Shi Feng
University of Maryland
College Park, Maryland
shifeng@cs.umd.edu

Jordan Boyd-Graber
University of Maryland
College Park, Maryland
jbg@umiacs.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 propose an evaluation of interpretation on a real task with real human users, where the effectiveness of interpretation is measured by how much it improves human performance. We design a grounded, realistic human-computer cooperative setting using a question answering task, Quizbowl. We recruit both trivia experts and novices to play this game with computer as their teammate, who communicates its prediction via three different interpretations. We also provide design guidance for natural language processing human-in-the-loop settings.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**; **Collaborative interaction**.

KEYWORDS

interpretability; natural language processing; question answering

ACM Reference Format:
Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me?: Evaluating Machine Learning Interpretations in Cooperative Play. In *24th International Conference on Intelligent User Interfaces (IUI '19), March 17–20, 2019, Marina del Rey, CA, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3301275.3302265

1 INTRODUCTION

The field of machine learning (ML) is making rapid progress, with models surpassing human performance on many tasks, such as image classification [22], playing video games [45], and playing Go [59]. However, a drop-in replacement for humans—even assuming that it is achievable—is not always the ideal integration of machine learning into real-world decision making. In sensitive areas such as medicine and criminal justice, the computational objectives of ML models cannot yet fully capture the factors one must consider when making a decision, such as fairness and transparency. In some other areas such as natural language processing, the strengths of humans and computers are sometimes complimentary. Humans are excellent at reasoning about what we consider "common sense", while some tasks in this category such as disambiguating word senses are still difficult for computers [48]. Tasks like deceptive review detection is difficult and time consuming for humans while simple linear ML models achieve high accuracy with little processing time [35]. On tasks such as simultaneous interpretation where humans are still far superior than computers, experts can still be assisted on some aspects of the task: interpreters often find certain content such as technical terms, names of people and organizations, and numbers difficult to translate, while computers find that easy. The integration of ML can be more effective and efficient when humans and computers cooperate.

Cooperation is only effective when the two parties communicate well with each other. One direction of this communication, from humans to computers, is well-studied: ML models can be improved with human feedback using reinforcement learning [63] and imitation learning [54, 55]. The other direction of the communication, from ML models to humans, presents different challenges: a standard classification model outputs a prediction (e.g., an object class given an image), but without any justification. Although the prediction can be presented with a confidence score (a value between zero and one), humans struggle to interpret and act on numbers [50, 51]; moreover, due to over-fitting, confidence scores from a neural models can be much higher than the actual prediction uncertainty [20].

To bridge the gap between human and ML models in a cooperative setting, interpretation methods explain the model predictions in a more expressive, human-intelligible way. In a human-centered setting where humans make the final decision, these methods help users decide to trust the model prediction or not. In Section 2 we discuss the existing work of interpreting ML models.

Progress in ML research largely relies on rigorous evaluations, which often relies on standard datasets, for example ImageNet [9] for image classification and Penn Treebank [43] for language modeling. Although interpretability is valued as a laudable goal, it remains elusive to evaluate. We do not have such standard dataset for interpretability—it is not clear what the ground truth should be. As Lipton [41] argues, there is no clear agreement on what interpretability means; there is no definitive answer to what interpretation is most faithful to the model and useful for humans at the same time. Secondly, it is not realistic to evaluate interpretability without humans, the eventual consumer of interpretations [47]. Previous work focuses on how humans can use interpretations to help the model do its job better; for example, interpretations generated by Local Interpretable Model-Agnostic Explanations [52,

IUI '19, March 17–20, 2019, Marina del Rey, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302265

LIME] help humans do feature engineering to improve downstream predictions of a classifier; in other work interpretations are used to help humans debug ML models [13, 53].

Kleinberg *et al.* [30] propose a different perspective and ask how ML can improve human decision making. Applying this thinking, we measure interpretability by asking what ML can do for humans through interpretations: they should *augment* [32] human intelligence. This concept resonates with the seminal work of mixed-initiative user interface [24], which emphasizes user interfaces where the human and the computer can drive towards a shared goal and ones that enhance human ability [2].

Interpretations come in many forms; we focus on three popular options among the interpretable ML community: visualizing uncertainty, highlighting important input features, and retrieving relevant training examples. We measure how they help humans on the tasks at hand and focus on answering the question "how effective can interpretations communicate model predictions to humans". The other question is "how faithful an interpretation is to the model". Section 3 discusses our choice of model to answer the first question; we leave the second question to future work, but discuss in Section 7 how our framework, interface, and experiments can be directly applied.

We choose the testbed for our interpretability evaluation from the natural language domain—a question answering task called Quizbowl [6]. As we discuss in Section 3, in addition to being a challenging task for ML, it is also an exciting game that is loved by human trivia enthusiasts. Furthermore, it is a task where humans and ML have complementary strengths, so effective collaboration with interpretations has great potential.

We recruit both Quizbowl enthusiasts and turkers from Amazon Mechanical Turk (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 the way a computer answers questions. Experts have enough world and task expertise to confidently overrule when the computer is wrong; however, as we will discuss in Section 6, novices are too trusting: they play more aggressively with computer assistance, but are not able to discern useful help from the misleading ones as well as the experts. In Section 7, 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

Explainability is a central problem of applied AI, with research stretching back to the days of expert systems [64]. The recent surge of interest in this area is the result of the success of ML models based on neural networks, a.k.a. deep learning [37]. These complicated models have stupendous predictive power, but at the same time brittle, best demonstrated by the existence of adversarial examples [15], where small perturbation to the input leads to significant change in the model output. 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. The fairness, accountability, and transparency of machine learning remain a concern [66], which is reflected in the "right to explanation" in European Union's new General Data Protection Regulation [11, GDPR].

Thus, ML model predictions need explanations. Efforts including the Explainable AI (XAI) initiative [19] led to the conceptualization of a series of human-AI cooperation paradigms, including human-aware AI [7], and human-robot teaming [67]. As an example, Schmidt and Herrmann [58] recognize the importance of interpretability when interacting with autonomous vehicles. Such need motivated the ML community to develop interpretation methods for deep neural models [4, 60, *inter alia].*

The HCI community has a rich body of research towards making computers more usable, for example in interaction design [28] and software learnability [18]. To borrow insights from the human side, Miller [44] provides an overview of social science research regarding how people define, generate, select, evaluate, and present explanations. Still, interpreting ML models has its unique challenges. Krause *et al.* [34] compare different ML models under one visualization method, partial dependency. Smith *et al.* [61] and Lee *et al.* [38] focus 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

Interpretations can take on several different forms. We focus on interpretation in the form of uncertainty, important input features, and relevant training examples. Some ML models provide canonical interpretations. For models such as decision trees and association rule lists [36, 39], the interpretation is built in the prediction itself. 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. Next we introduce previous work on interpreting both simpler linear models and more complicated neural networks, in each of the three forms.

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, 57]. To make it more informative, we can also display the confidence for the classes other than the top one [42]. Confidence of simple linear models are usually well-calibrated, but estimating uncertainty for a deep neural model is challenging: due to overfitting, they are over-confident and require careful calibration [12, 20].

Highlighting Important Features. Model predictions can be explained by highlighting the most salient features in the input, typically visualized by a heat map. For a linear classifier, the most salient features are the ones with the largest corresponding coefficients; For non-linear classifiers, the relevance of a feature can be calculated by the gradient of the loss function w.r.t. that feature [60]. Alternatively, one can locally approximate a non-linear classifier with a simpler linear model, then use the coefficients to explain the predictions from the non-linear model [52].

Interpretation by Example. We can explain a prediction on a test example by finding the most influential training examples. Various metrics exist for finding important training examples, such as distance in the representation space which is natural to linear models, clustering algorithms and their deep variation [49], and influence functions [33] for non-linear models.

As we discuss in Section 3, although our experiments use a linear classifier, our method can be generalized to evaluating these methods designed for neural models (Section 7).

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. Although interpretation methods have rigorous mathematical formulations, some even axiomatically derived [62], it remains unclear how we can evaluate the efficacy of these methods on *real tasks* with *real users.* Lipton [41] 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.

As it is widely accepted that machine learning models should be evaluated beyond natural examples, e.g., in adversarial settings [15, 26], the evaluation of interpretation should not be limited to being visually pleasing. Indeed, interpretations can be fragile under small input perturbations [14, 29], unfaithful to the model [1, 12, 23], and create a false sense of security [27].

Conditioning a more realistic setting, Doshi-Velez and Kim [10] provide an ontology of various evaluations of interpretation with a human in the loop. Following this framework, Narayanan *et al.* [47] conduct one such evaluation with synthetic tasks and hand-crafted interpretations to study their desirable cognitive properties.

We focus on *application-grounded* evaluation—real tasks with real users. 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 motivated real users. The task needs a large pool of willing human testers, and ideally one that challenges both humans and computers. As we discuss in the next section, Quizbowl is a task that satisfies these conditions.

3 INTERPRETATION TESTBED: QUIZBOWL

This section introduces Quizbowl, our testbed for evaluating the three forms of interpretations. We discuss how the task suits our purposes, which model to use, and how we generate the interpretations.

3.1 Quizbowl and Computer Models

Quizbowl is both a challenging task for machine learning [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*—difficult 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 decide that *this is all the information I need to answer the question.* The player can do so by *buzzing* before the question is fully read, 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 [21].

Quizbowl challenges humans and computers in different ways [6, 68]. Computers can memorize every poem and book ever written, making it trivial to identify quotes. Computers can also 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 human-computer cooperation has huge potential [65]. This also makes Quizbowl a suitable testbed for interpretation methods designed to better interface humans and computers.

Thus, instead of trying to beat humans with computers, we team them 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 that can change the underlying interpretation and convince the human that the system has a good answer to offer. 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. Sesction 5 discusses the cooperation in detail.

As mentioned in Section 1, we focus on the comparison between three forms of interpretation, using one method for each form. But which method to use? Linear models provide canonical interpretations: important features and relevant training examples can be identified based on the coefficients. On the other hand, neural models do not have canonical interpretations: all interpretations are approximations, which by definition are not completely faithful to the model [56].

Luckily in the case of Quizbowl, we have linear models with performance on par or better than neural models. QANTA [25] 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 [16, ES]. As the name implies, guesser generates guesses for what the answer to a question could be. Despite its simplicity, ES-based systems perform very well on Quizbowl, defeating top trivia players.[2]

[1] Like previous work, we only consider *toss-up*/*starter* questions.
[2] https://youtu.be/bYFqMINXayc

3.2 Interpretation of a Question Answering Model

Our goal is to see which forms of interpretation are most helpful to the users, and a linear model with natural interpretations makes this easy. Our ES-based Quizbowl model supports three forms of interpretations, each corresponding to a class of methods widely studied in recent literature as mentioned in the previous section. Given a question never seen in the training set, 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 then uses the label of that document as the answer.

To convey the uncertainty of model predictions, we augment the top ten guesses from our model with their corresponding scores. Unlike regular classification models, ES does not output a probability distribution over all possible answers. Its scores measure the relevance between the question and training examples, but are not normalized. We keep the scores unnormalized to stay true to the model. Despite its simplistic form, these scores provide strong signal about model uncertainty, for example, a large gap between the top two scores usually indicate a confident prediction.

Interpretation by example—getting the *evidence*—is straightforward with our ES-based model. The prediction is the label of the most relevant documents, so the extracted documents are naturally the most salient training examples. We can further identify the most important words in each retrieved training example, using the highlight API[3]. This gives us *evidence highlights*. The player can make a better decision of whether to trust the computer prediction by judging how relevant the evidence is to the question.

To highlight important input features—generating *question highlights*—we build on the previous *evidence highlights*. The most important words in the question naturally emerge when we compare the question against the most salient training example. Specifically, we go through the question and find words that appear highlighted in the evidence. Question highlights inform the player whether the computer is looking at the right keywords in the question.

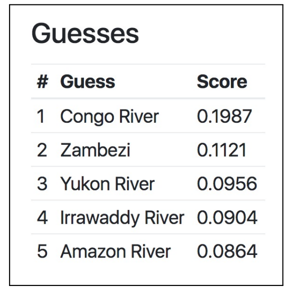
Although generating *question highlights* depends on *evidence highlights*, the former can be displayed without the later. We discuss how we control which interpretation to display in the next two sections.

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.

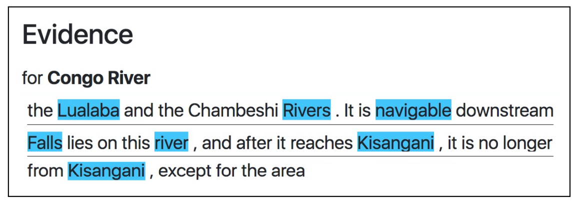
To make Quizbowl players feel at home, we follow the general framework of Protobowl.com, a popular Quizbowl platform that many players actively use for practice. The **Question** area is in the center, and the question is displayed word-by-word in the text box. A **Buzz** button is located close above 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.

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

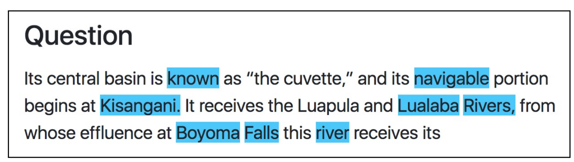


#	Guess	Score
1	Congo River	0.1987
2	Zambezi	0.1121
3	Yukon River	0.0956
4	Irrawaddy River	0.0904
5	Amazon River	0.0864

Guesses show the answers the computer 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).



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.



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.

Multiple interpretations 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 design goal 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 interpretations around it. It is difficult to ensure that different forms of interpretations are exposed to the users equally, as some forms (e.g., evidence) are inherently less intuitive to visualize. However, all interpretations must be implemented in an interface for a real-world evaluation; we discuss the limitations of our design and future work in Section 7.

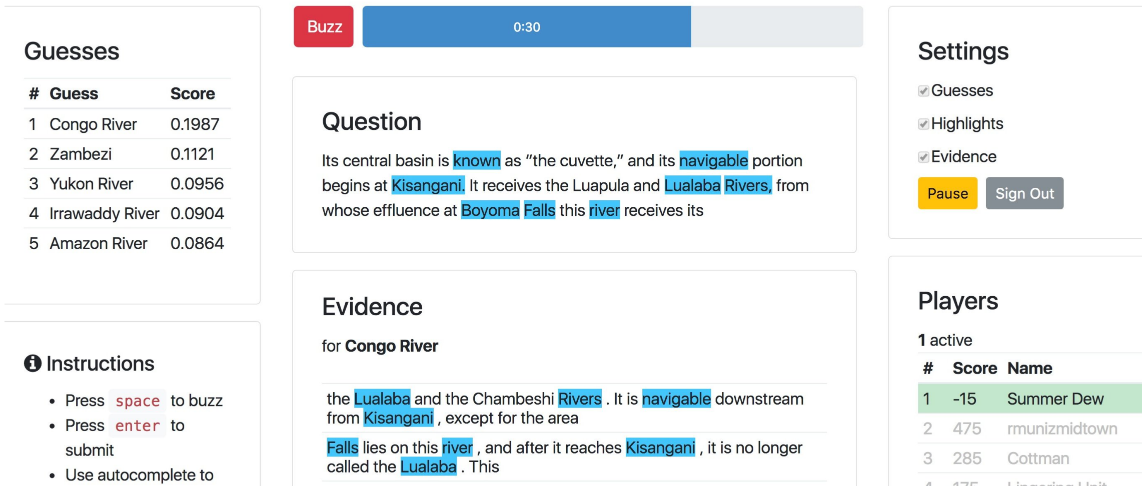


Figure 1: Screenshot of the interface. Question is displayed in the middle area word-by-word, with question highlights displayed in the same panel. Guesses are listed in the panel on the left. Evidence is in the panel below.

5 SETUP

This section explains how human players and the computer guesser play in cooperation. To ensure accuracy and unbiasedness, we control what interpretations each player sees instead of letting them choose.

5.1 Data and Participants

We collect 160 new questions for this evaluation that had not been previously seen by the Quizbowl community to avoid bias in players’ exposure to questions.

We recruit 40 experts (Quizbowl enthusiasts) by advertising on an online forum, and 40 novices using MTurk. 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 finish the whole question set. We require novices to each answer at least twenty 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 do 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, 21], our interface aims at human-AI *cooperation*. We let a human player form a team with a computer teammate and put the human in charge. As the question is displayed word-by-word, the computer periodically updates its guesses and interpretations (every 4 words in our experiments); at any point before the question is fully read, the human can decide to buzz, interrupt the readout, and provide an answer. The interpretations should help the human better decide whether to trust the computer’s prediction or not.

We have two different experimental settings. In the simpler, non-competitive **novice setting**, we have a single turker interact with the interface, with the computer guesser as teammate, but without opponents.

The competitive **expert setting** better resembles real Quizbowl games, and the players in this setting are experts that enjoy the game. 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?). Quizbowl’s pyramidality encourages competition: difficult clues at the start of the question help determine who knows the most about a subject. Our interface resembles Protobowl.com, a popular online Quizbowl platform where players play against each other (but without the computer teammate). The computer generates the same output (both prediction and interpretations), but human players might have access to different interpretations, e.g., David sees evidence while Selene sees question highlights. Next section discusses the setup in detail.

Our experiment in the expert setting 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 interpretations, compared to many previous work that only had non-expert turkers [8, 31, 61].

5.3 Controlling Which Interpretations to Show

Each of the three interpretations can be turned on or off, so we have in total $2 \times 2 \times 2 = 8$ conditions, including the null condition where all interpretations are hidden. To compare within-subjects (players vary greatly based on their innate ability), we vary the interpretations a player sees randomly. We sample the enabled combination

Figure 2: Coefficients of the linear regression showing the effects of interpretations, for novices (above) and experts (below). Higher value means an interpretation improves player accuracy. In addition to the individual interpretations, *combo gain* and *combo loss* capture the additional effect of combining multiple interpretations. *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.

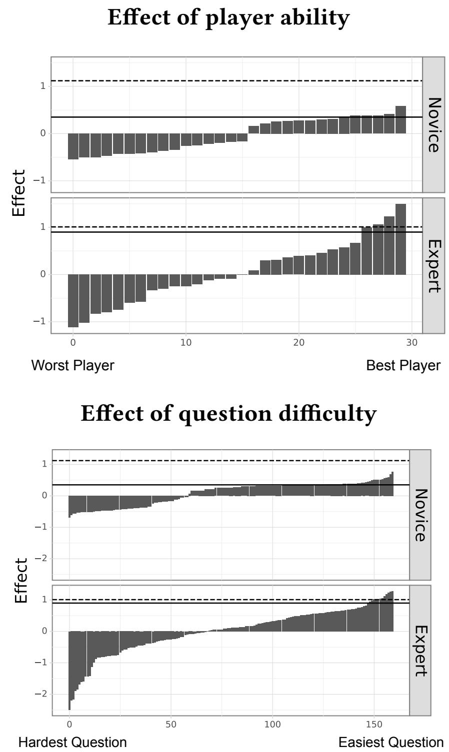


Figure 3: Effect of player ability (above) and question difficulty (below) from the regression analysis. Solid horizontal lines show the bias term that captures the baseline accuracy without any help from the computer; dashed lines show the effect of combining all interpretations. Experts have a higher average accuracy; they are also less affected by interpretations.

with the goal of having, in expectation, a uniform distribution over players, questions, and interpretation 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 interpretation combination C and N is the expected count of each combination (in our case the number of questions divided by eight). In the expert setting, interpretations are sampled independently for each player, and players may (and usually do) see different interpretations. For all experiments, we only allow each player to answer each question once.

6 RESULTS

With data collected from game plays, our primary goal is determine if the interpretations are helpful or not, and how experts and novices used them differently. We first do 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, specifically how aggressive they are; we also look at specific cases where some interpretation consistently succeeded or failed to convince multiple players of the model prediction.

After filtering players who answer very few questions, we arrive at 30 experts that answer 1983 questions, and 30 novices that answer 600 questions. Turkers usually stopped after answering the required twenty questions, but many experts kept on playing. Among all players, seven experts answer all 160 questions.

6.1 Regression Analysis

Whether a player can answer a question correctly is determined by several factors: the player's innate skill, the difficulty of the question, the aid of some interpretation, or the competitive level (in expert setting). To tease apart these factors we follow Narayanan *et al.* [47] and apply a regression analysis.

We describe these factors using the four sets of features listed in Table 1. To capture the player's innate skill and the difficulty of the question, we include the IDs of both in the feature set. Each combination of interpretations has its own features, for example, *guesses*, *evidence*, and *guesses+evidence* are three independent features. For game condition, the first feature is the relative position in the question when the player buzzed (to understand how interpretations 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.

The we use a linear model to predict whether the player can answer the question correctly. Specifically, for each game record, we extract the features and feed the vector as input to the linear model, which then predicts the probability of a positive result; to

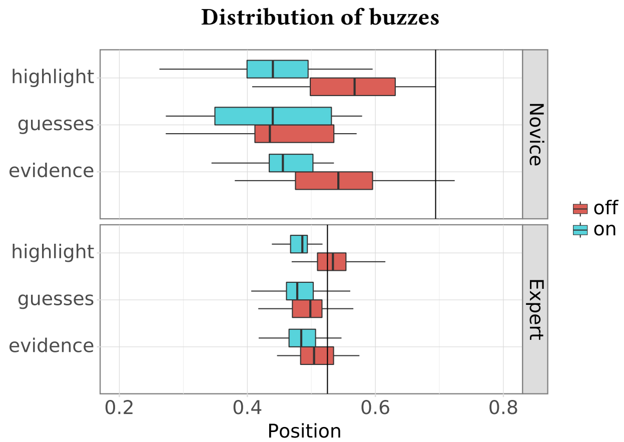


Figure 4: Average buzzing position (relative to question length) of novices (above) and experts (below), with and without each interpretation. The goal is to buzz as early as possible. Vertical bars show the baseline buzzing position without any interpretation. Experts are better and more consistent.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

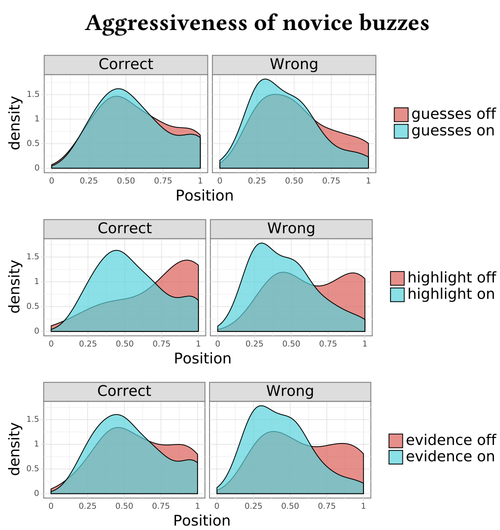


Figure 5: The distribution of buzzes of novices on correct guesses (left) and wrong guesses (left); colors indicate if each interpretation is enabled; positions are normalized by question length. With interpretations, novices are significantly more aggressive, but also get more questions correct earlier. *Highlight* is the most effective.

train the model, we compare the prediction against the ground truth, and update the model with gradient descent. We train this model on the game play data, for experts and novices separately.

The coefficients of the linear model then explains the importance of the corresponding features: the probability of a positive result increases with features with positive coefficients, which means these

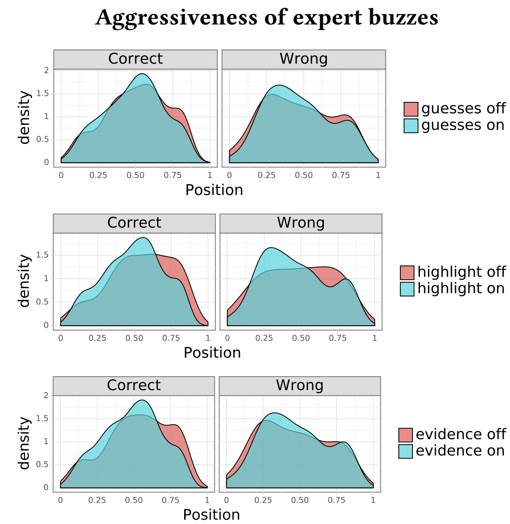


Figure 6: The distribution of buzzes of experts on correct guesses (left) and wrong guesses (left); colors indicate if each interpretation is enabled; positions are normalized by question length. Experts are not significantly more aggressive with interpretations, but they did get more answers correct earlier.

Interpretation (8)	none, guesses, highlight, evidence, guesses + highlight, guesses + evidence, highlight + evidence, guesses + highlight + evidence
Player (30)	player IDs (separate for experts and novices)
Question (160)	question IDs
Others (3)	buzzing position (relative to question length), number of active players (expert only), current accuracy of the top active player (expert only)

Table 1: Our four sets of features used in regression analysis. Numbers in the parentheses indicate the number of features in that set.

features help the players. Similarly negative coefficient means the features hurt the player accuracy. To understand which interpretations are most helpful to Quizbowl players, we inspect the sign and magnitude of their corresponding coefficients.

Figure 2 shows the effect of interpretations based on regression coefficients: a high positive weight means the interpretation is useful, zero means it is ineffective, and negative means it is harmful. It is not guaranteed that the strengths of multiple interpretations are combined when they are displayed at the same time. This is due to confounding factors such as information load—the player might feel distracted when too much information is displayed on the interface and thus perform worse. The additional effects of combining interpretations are "combo gain" and "combo loss" (Figure 2). For example, combining guesses and evidence has a negative effect on novices; the loss is computed by comparing the "guesses+evidence" coefficient with the arithmetic sum of the "guesses" and "evidence" coefficients.

The interpretation that helps novices is not the same as what helps experts. For experts, highlight is the most helpful individual interpretation, while for novices, evidence is the most helpful. For experts, the combination of highlight and evidence achieves extra gain, 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, potentially due to information overload.

We hypothesize that the main difference between experts and novices is that 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.

To understand how much variance players display in their skill and questions in their difficulty, we show their corresponding coefficients (Figure 3). The solid horizontal line shows the baseline accuracy of that player group without any interpretation (the bias term—or the intercept—of the linear model). Experts show a higher baseline accuracy, which is not surprising since they are experts; they also show a larger variance in accuracy within the group, potentially due to the competitive environment; they are also more sensitive to the difference in question difficulty. To compare these factors against the interpretations, we show with the dashed horizontal line the combination of all three interpretations. Experts are less sensitive to the interpretations, potentially due to a higher confidence in their own guesses.

6.2 How Interpretations Change Player Behavior

The regression analysis provides a quantitative comparison between all interpretations 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 interpretation 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 answering correctly. Interpretations should help them reach this goal.

Figure 4 show the average buzzing position of each player group with and without each interpretation. Novices buzz much later than experts when no interpretation is enabled (comparing the solid vertical bars), but buzz at about the same point as experts when interpretations are enabled, despite a lower accuracy (Figure 3). This suggests that the novices are too trusting in the computer teammate, and end up playing too aggressively for their skill level.

We see a similar trend when we plot the density of buzzing positions (experts in Figure 6 and novices in Figure 5). In all settings, the density shifts earlier: players are more aggressive with interpretations, especially for novices, which is consistent with Figure 4. The interpretations allow players to answer correctly earlier.

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 7: Interpretations that help players answer a question on Civil Disobedience correctly. With the shown part of the question, three experts answer correctly with the evidence; no expert answer 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 8: 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.

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 predictions 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 7 shows an example where interpretations enable players to answer correctly. A total of twelve expert players answered the question, and eight answered correctly. The earliest an expert can answer correctly without the evidence was at 72% of the question, while the three experts with the evidence all answer 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.

Figure 8 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

The evaluation we present is grounded in a realistic setting, but also task-specific. This section discusses how our method can be directly applied to other settings, its limitations, and how we can incorporate other components such as an eye tracker to our framework for a more fine-grained assessment of interpretability.

7.1 Forms and Methods of Interpretation

Interpretations take on many forms, and within each form we have multiple methods to generate the interpretation. For example, to highlight salient input features for image classification, we can use variants of input gradient [4, 60]. To optimize the generalizability of our results (despite being task-specific) and demonstrate the flexibility of our method, we focus on a comparison between forms of interpretation. To select one method of each form, we choose a high-performance linear model for its canonical interpretations. Our evaluation framework, including the interface and the regression analysis, can be directly applied to a different comparison—one between multiple methods of the same form. This comparison is particularly useful in the case of neural models, where all existing interpretations are some approximation, and the evaluation of how faithful they are to the model is crucial.

7.2 Intrinsic and Extrinsic Evaluation

Our approach is an extrinsic evaluation [47]. 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 [40, 46]. 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.* [47] 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 Generalizing to Other Tasks

Our method can 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 [47]. 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 [21] also resembles how simultaneous translation systems handle this trade-off [17].

7.4 Limitations

First, because we compare visualizations individually and in combinations, their placement is 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.

Another limitation of our study is that, when a player's performance improves with some interpretation, we cannot tell how much of that improvement comes from the player using that interpretation. We cannot derive causality from correlation. The key missing factor is how much attention the player gave the interpretation, and how much the decision is based on that. The attention the player gave each interpretation could be measured using an eye tracker, and we leave this to future work.

7.5 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 focuses 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 propose and demonstrate an evaluation of interpretation methods in a human-AI cooperative setting. We focus on the natural language domain and use a question answering task derived from a popular trivia game, 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.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful and constructive comments. Additionally, we would like to thank Alison Smith, Leah Findlater, Hernisa Kacorri, Alina Striner, and Andreas Mathisen for their valueble input. This work was supported NSF Grant IIS-1822494 and subcontract to Raytheon BBN Technologies by DARPA award HR0011-15-C-0113. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

REFERENCES

[1] Julius Adebayo, Been Kim, Ian Goodfellow, Justin Gilmer, and Moritz Hardt. 2018. Sanity Checks for Saliency Maps. In *Proceedings of Advances in Neural Information Processing Systems*.
[2] JE Allen, Curry I Guinn, and E Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* (1999).
[3] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. 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*.
[4] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to Explain Individual Classification Decisions. *Journal of Machine Learning Research* (2010).
[5] Jordan Boyd-Graber, Shi Feng, and Pedro Rodriguez. 2018. *Human-Computer Question Answering: The Case for Quizbowl*. Springer.
[6] Jordan L. Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. 2012. Besting the Quiz Master: Crowdsourcing Incremental Classification Games. In *Proceedings of Empirical Methods in Natural Language Processing*.
[7] Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, and Yu Zhang. 2017. AI challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775* (2017).
[8] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *International Conference on Intelligent User Interfaces*.
[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*.
[10] Finale Doshi-Velez and Been Kim. 2018. Towards A Rigorous Science of Interpretable Machine Learning. *Springer Series on Challenges in Machine Learning* (2018).
[11] European Parliament and Council of the European Union. 2016. General data protection regulation. (2016).
[12] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of Empirical Methods in Natural Language Processing*.
[13] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision*.
[14] Amirata Ghorbani, Abubakar Abid, and James Y. Zou. 2018. Interpretation of Neural Networks is Fragile. *Association for the Advancement of Artificial Intelligence*.
[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *Proceedings of the International Conference on Learning Representations*.
[16] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide*. O'Reilly Media, Inc.
[17] Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation. In *Empirical Methods in Natural Language Processing*. docs/2014_emnlp_simtrans.pdf
[18] Tovi Grossman, George Fitzmaurice, and Ramtin Attar. 2009. A survey of software learnability: metrics, methodologies and guidelines. In *International Conference on Human Factors in Computing Systems*.
[19] David Gunning. 2017. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).
[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the International Conference of Machine Learning*.
[21] He He, Jordan L. Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent Modeling in Deep Reinforcement Learning. In *Proceedings of the International Conference of Machine Learning*.
[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision*.
[23] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2018. Evaluating Feature Importance Estimates. In *ICML Workshop on Human Interpretability in Machine Learning*.
[24] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *International Conference on Human Factors in Computing Systems*.
[25] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. 2014. A Neural Network for Factoid Question Answering over Paragraphs. In *Proceedings of Empirical Methods in Natural Language Processing*.
[26] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of Empirical Methods in Natural Language Processing*.
[27] Heinrich Jiang, Been Kim, and Maya R. Gupta. 2018. To Trust Or Not To Trust A Classifier. In *Proceedings of Advances in Neural Information Processing Systems*.
[28] Wendy Ju and Larry Leifer. 2008. The design of implicit interactions: Making interactive systems less obnoxious. *Design Issues* (2008).
[29] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2017. The (Un)reliability of saliency methods. *arXiv preprint arXiv: 1711.00867* (2017).
[30] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The quarterly journal of economics* (2017).
[31] Ronald T Kneusel and Michael C Mozer. 2017. Improving Human-Machine Cooperative Visual Search With Soft Highlighting. *ACM Transactions on Applied Perception* (2017).
[32] Kenneth R. Koedinger, Emma Brunskill, Ryan S.J.d. Baker, Elizabeth A. McLaughlin, and John Stamper. 2013. New Potentials for Data-Driven Intelligent Tutoring System Development and Optimization. *AI Magazine* 34, 3 (sep 2013), 27. https://doi.org/10.1609/aimag.v34i3.2484
[33] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the International Conference of Machine Learning*.
[34] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *International Conference on Human Factors in Computing Systems*.
[35] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of ACM FAT**.
[36] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Knowledge Discovery and Data Mining*.
[37] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* (2015).
[38] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* (2017).
[39] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* (2015).

[40] Jiwei Li, Will Monroe, and Daniel Jurafsky. 2016. Understanding Neural Networks through Representation Erasure. *arXiv preprint arXiv: 1612.08220* (2016).

[41] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *arXiv preprint arXiv: 1606.03490* (2016).

[42] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. 2017. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics* (2017).

[43] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics* (1993).

[44] Tim Miller. 2017. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269* (2017).

[45] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* (2015).

[46] W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. In *Proceedings of the International Conference on Learning Representations*.

[47] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint arXiv: 1802.00682* (2018).

[48] Simone Papandrea, Alessandro Raganato, and Claudio Delli Bovi. 2017. SUPWSD: A Flexible Toolkit for Supervised Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

[49] Nicolas Papernot and Patrick D. McDaniel. 2018. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *arXiv preprint arXiv: 1803.04765* (2018).

[50] Ellen Peters, Daniel Västfjäll, Paul Slovic, CK Mertz, Ketti Mazzocco, and Stephan Dickert. 2006. Numeracy and decision making. *Psychological science* (2006).

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

[52] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Knowledge Discovery and Data Mining*.

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP Models. In *Proceedings of the Association for Computational Linguistics*.

[54] Andrew Slavin Ross and Finale Doshi-Velez. 2018. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients.

[55] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of Artificial Intelligence and Statistics*.

[56] Cynthia Rudin. 2018. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv preprint arXiv:1811.10154* (2018).

[57] Enrico Rukzio, John Hamard, Chie Noda, and Alexander De Luca. 2006. Visualization of uncertainty in context aware mobile applications. In *Proceedings of the international conference on Human-computer interaction with mobile devices and services*.

[58] Albrecht Schmidt and Thomas Herrmann. 2017. Intervention user interfaces: a new interaction paradigm for automated systems. *Interactions* (2017).

[59] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of Go without human knowledge. *Nature* (2017).

[60] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Proceedings of the International Conference on Learning Representations*.

[61] Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, and Leah Findlater. 2017. Evaluating visual representations for topic understanding and their effects on manually generated labels. *Transactions of the Association for Computational Linguistics* (2017).

[62] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the International Conference of Machine Learning*.

[63] Richard S Sutton and Andrew G Barto. 1998. *Introduction to reinforcement learning*.

[64] William R Swartout. 1983. *Xplain: A system for creating and explaining expert consulting programs.* Technical Report. University of Southern California.

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

[66] USACM. 2017. Statement on algorithmic transparency and accountability. *Public Policy Council* (2017).

[67] David W Vinson, Leila Takayama, Jodi Forlizzi, Wendy Ju, Maya Cakmak, and Hideaki Kuzuoka. 2018. Human-Robot Teaming. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*.

[68] Eric Wallace and Jordan Boyd-Graber. 2018. Trick Me If You Can: Adversarial Writing of Trivia Challenge Questions. In *Proceedings of ACL 2018 Student Research Workshop*.