Shi Feng and **Jordan Boyd-Graber**. **Learning to Explain Selectively: A Case Study on Question Answering**. *Empirical Methods in Natural Language Processing*, 2022, 9 pages.


```
    @inproceedings{Feng:Boyd-Graber-2022,
Author = {Shi Feng and Jordan Boyd-Graber},
Url = {http://umiacs.umd.edu/~jbg//docs/2022_emnlp_augment.pdf},
Booktitle = {Empirical Methods in Natural Language Processing},
Location = {Abu Dhabi},
Year = {2022},
Title = {Learning to Explain Selectively: A Case Study on Question Answering},
}
```


Links:

- Research Teaser [https://youtu.be/27BsqhJajWs]
- Code and Data [https://bit.ly/selective_explanation]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Learning to Explain Selectively: A Case Study on Question Answering

Anonymous EMNLP submission

Abstract

Explanations promise to bridge the gap between human and AI, yet AI-augmented human decision making proves difficult: explanations are helpful in some cases but harmful in others (Bansal et al., 2021; Lai et al., 2021). The effect of explanation depends on many factors, such as human expertise (Feng and Boyd-Graber, 2019), human agency (Lai and Tan, 2019), and explanation format (Gonzalez et al., 2020; Smith-Renner et al., 2020a). Using a uniform setup—always showing the same type of explanation in all cases—is suboptimal, but it's also hard to rely on heuristics to adapt the setup for each scenario. We propose learning to explain selectively using human feedback to directly optimize human accuracy. We formulate selective explanation as a contextual bandit problem, train a model to learn users' needs and preferences online, and use the model to choose the best combination of explanations to provide in each scenario. We experiment on question answering following the evaluation protocol of Feng and Boyd-Graber (2019) and show that selective explanations further improve human accuracy for both experts and amateurs.

1 Introduction

Recent advances in machine learning (ML) (Silver et al., 2017; Brown et al.; Jumper et al., 2021; Ramesh et al., 2021) sparked new life in **intelligence augmentation**—the vision that computers are not mere number-crunching tools, but also interactive systems that can augment humans at problem solving and decision making (Engelbart, 1962). The hope is to combine the complimentary strengths of machine and human, and to fully harness the capabilities of these models with human intuitions and oversight (Dafoe et al., 2020; Amodei et al., 2016). But this agenda is obstructed by the many counterintuitive traits of neural networks (NNs) (Szegedy et al., 2014; Goodfellow et al., 2015; Zhang et al., 2017) and our lack of theoretical understanding (Belkin et al., 2019): these models are not interpretable to humans by default and it is difficult to foresee when they will fail. This lack of interpretability also amplifies the risk of model bias (Angwin et al., 2016; Bolukbasi et al., 2016; Caliskan et al., 2017), making it difficult to use NN-powered AIs in real-world decision making.

To bridge the gap between human and machine, various methods attempt to explain model predictions in human-interpretable terms, e.g., by providing more context to the model's uncertainty estimation (Gal et al., 2016; Bhatt et al., 2021), by highlighting the most important part of the input (Ribeiro et al., 2016; Lundberg and Lee, 2017; Ebrahimi et al., 2017), and by retrieving the most relevant training examples (Renkl, 2014; Koh and Liang, 2017). Grounded in psychology (Lombrozo, 2006, 2007; Kulesza et al., 2012), these explanations promise to augment human decision making.But when tested in application-grounded evaluations—with real problems and real humans (Doshi-Velez and Kim, 2018), it proves difficult for any single explanation method to achieve consistent improvement in disparate context (Bansal et al., 2021; Buçinca et al., 2020).

A major contributor to this problem is the breadth of context that the explanation method is applied to. Internally, the explanation method is faced with shifts in the input distribution which the model can react badly to (Goodfellow et al., 2015; Liu et al., 2021); externally, it needs to deal with human users with diverse levels of expertise (Feng and Boyd-Graber, 2019), engagement (Sidner et al., 2005), and general trust in AI (Dietvorst et al., 2015). Our current use of explanations demands an one-size-fits-all solution, but existing methods cannot provide that as they are largely oblivious to the above mentioned variables.

Selective explanations Each person is unique, and the right explanation will also vary from one deci-

sion to another, so we propose to show explanations selectively to maximize their utility as decision support. Concretely, we assume a given set of explanation methods, but instead of showing all of them for every decision that the human user makes, we use a *selector policy* to choose a subset of the explanations to display. We can think of the selector as controlling an on/off switch for each explanation method. The selector is allowed, for example, to show three types of explanations for one example but withhold all of them for the next one.

Online optimization In order for the policy to accurately estimate the utility of explanations in each context, its training data must offer a reasonable coverage over the joint distribution of all types of explanations, human users, and examples, which means that the dataset will have to include cases where the human user receives suboptimal decision support, e.g., with excessive explanations causing information overload (Doshi-Velez and Kim, 2018). We focus on the online setting which represents real-world scenarios where the opportunity cost of giving suboptimal support cannot be ignored. In this setting, a good policy must balance the trade-off between exploring new combinations of explanations and sticking to explanations with good observed performance; we model this trade-off by formulating the selective explanation problem as a multi-armed bandit (Robbins, 1952).

We evaluate selective explanations on Quizbowl using the same platform as Feng and Boyd-Graber (2019). To mimic real-world decision making as well as possible, we recruited twenty trivia enthusiasts and ran a multi-player, real-time Quizbowl tournament. We compare our method head-to-head against baselines such as showing all explanations for all examples. Selective explanations out-perform all other strategies, including the best subset of explanations identified by Feng and Boyd-Graber (2019). We also evaluate our method with mechanical turkers—amateurs whose performance without assistance is far worse than the AI. Explanations significantly boost their performance, but only selective explanations can help them reach performance comparable with the AI.

2 Selective Explanations as Decision Support

Explanations have many uses in human-AI cooperation; this paper focuses on using explanations as decision support—to improve the quality of human decisions under machine assistance. Not all problems benefit from machine assistance (Doshi-Velez and Kim, 2018)—in this section, we identify three criteria for decision support testbeds. We then introduce our setup based on Quizbowl (Rodriguez et al., 2019), a competitive trivia game.

2.1 Criteria for Decision Support Testbeds

It is not uncommon to use low-stake and synthetic tasks to evaluate machine assistance, but it's important to find tasks where results can generalize. Building on existing work (e.g. Lee and See, 2004; Lim et al., 2009; Yin et al., 2019), we identify the three criteria for suitable tasks.

Clear objectives The task must have well-defined metrics, and the standard for good decisions must be clear to all participants. With unreliable metrics, a well-optimized decision support will still fail to improve decision quality (Amodei et al., 2016).

Diversity of context A reliable testbed should be diverse in terms of both participants (e.g., their skill levels) and test examples (e.g., their difficulty level). As discussed in Section 1, the lack of diversity contributes to the inconsistent results.

Incentives to be engaged The participants must be incentivized to pay attention to model outputs in order to establish proper reliance (Lim et al., 2009). As a corollary, the model should demonstrate complementary strengths and provide information that participants cannot extract by themselves. In terms of the setup, engagement can also be improved by imposing time limits (Doshi-Velez and Kim, 2018) and introducing competition (Bitrián et al., 2021).

We choose Quizbowl (Rodriguez et al., 2019)—a task that roughly satisfies all three criteria—as our testbed. Compared to previous work that uses Quizbowl to evaluate explanations (Feng and Boyd-Graber, 2019), we make several changes to the setup for evaluating online selective explanations. In the following, we first introduce the most basic setting with only human players and build up our system one component at a time.

2.2 Human-only Quizbowl

We start with the most basic (and traditional) setting: Quizbowl with only human players. Quizbowl is a trivia game popular in the English-speaking world where players compete to answer questions from all areas of academic knowledge, including history, literature, science, sports, and

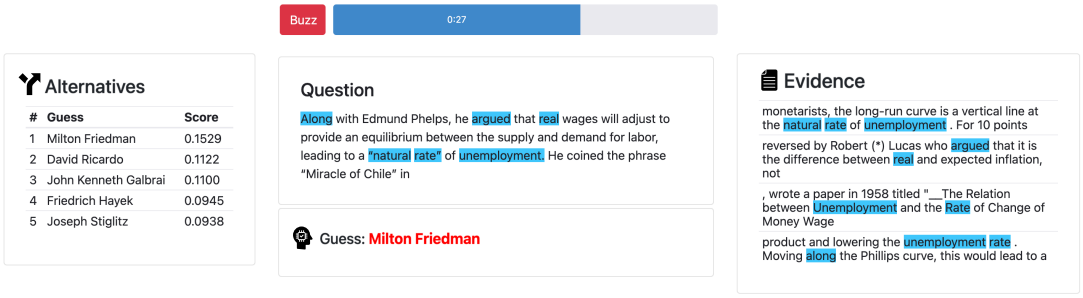

Figure 1: Our Quizbowl web interface when all four explanations are displayed. In the middle we show the question word-by-word; below, we show the current best model guess, which is colored red when the Autopilot is confident, otherwise gray; on the left we show Alternatives, including confidence scores; on the right we show snippets of relevant training examples as Evidence; finally we show Highlights for the question and the evidence, respectively.

more.[1] Each Quizbowl question consists of four to five clues. The clues are organized by their difficulty in each question: starting with the clue that's most difficult and obscure, and finishes with the one that's easiest and most telling. The clues are presented to all players *word-by-word* in real-time, verbally or in text (e.g. web interfaces). And players compete to answer as early as possible.

To signal that they know the answer, players interrupt the question with a *buzz*, which takes its name from the sound the device makes. Whoever buzzes needs to answer: ten points for a correct answer, and five-point penality for a wrong one. A player only gets one chance at each question.

To win Quizbowl, you need to answer quickly *and* correctly. This game requires not only trivia knowledge but also an accurate assessment of confidence and risk (He et al., 2016). We formally discuss the evaluation metric in Section 3.1.

2.3 Human + AI + Explanations

In our Quizbowl games, human players augmented with AI decision support compete against each other. In each human-AI team, the human player is still in charge of making decisions of when to buzz and what to answer, but now with the help of a machine learning *guesser* which predict an answer given a question (we provide details about the guesser in Section 3). In addition to showing the guesser's current best guess, we show four types of explanations: Alternatives (Lai and Tan, 2019), salient word Highlights (Ribeiro et al., 2016), relevant training examples as Evidence (Wallace et al., 2018), and a new explanation that we call Autopilot. As the name suggests, Autopilot assumes the role of the human player and make suggestions on *whether* to buzz or to wait (details in Section 2.5). We build our interface (Figure 1) by extending the interface of Feng and Boyd-Graber (2019). We discuss these changes in detail next and in Section 3.

2.4 Human + AI + *Selective* Explanations

With selective explanations, the decision support is customized for each player and each question. For each new question, we use a selector policy (or *selector* for short) to control the on/off switch for each explanation. We refer to a combination of explanations as a *configuration*; for example one configuration could be showing Highlights and Evidence but hiding Alternatives. A configuration is selected at the beginning of each question and kept constant throughout the question, but the content of each explanation is still updated dynamically. For example, Highlights will always available when its turned on for a question, but the exact words being highlighted can change as more clues are revealed.

We make two important changes to the setting of Feng and Boyd-Graber (2019) to accurately estimate the effect of selectivity.

- **The guesser prediction is always available.** We make this design choice in order to better isolate the effect of the explanations.
- **Separate highlights for the question and the evidence.** Highlights can be applied to

[1] While these games often have collaboration on questions, we consider only individual players on tossup (US) or starter (UK/INDIA) questions. Likewise, throughout this paper we assume each human-AI team has a single human player. The extension to multiple humans is non-trivial and is thus left for future work.

#	Evidence	Highlights	
		Question	Evidence
1			
2	✓		
3	✓	✓	
4	✓	✓	✓
5		✓	

Table 1: Each configuration is a set of visualizations shown to users, and our policy leanrs which configuration helps users the most. Most visualizations can be turned on or off indepedently, but some only makes sense in the presence of others, e.g. we cannot highlight the evidence if we do not show evidence at all. This table summarizes the available configurations for two visualizations: `Autopilot` and `Highlights` which are dependent on each other. Combined with the other two explanations (`Alternatives`and `Autopilot`) which can be turned on or off indepedently, we have in total twenty possible configurations.

#	Description
1	Confidence of current top guesses.
2	Previous confidence of current top guesses.
3	Change in confidence of top guesses.
4	Gap in confidence between top guesses.
5	If top guesses maintained their rank.
6	If top guesses appear in previous step.
7	User's accuracy.
8	User's average relative buzzing position.
9	User's average EW score.
10	Gap in EW compared to optimal buzzer.
11	Portion of words highlighted in question.
12	Portion of words highlighted in evidence.
13	Longest highlighted substring in question.
14	Longest highlighted substring in evidence.

Table 2: The user model uses the above features in addition to BERT representations of the questions. The three categories capture information about the guesser's current prediction, the user, and the explanations. These features let the selector predict which explanations will be most useful for a human-AI team.

both the question and the evidence. In Feng and Boyd-Graber (2019), the two are treated as one explanation. However, their experiments confirm that highlighting the question alone is already effective. In this paper we separate the two and the policy can control them individually. Table 1 lists the available configurations for `Highlights` and `Evidence`.

2.5 A New Explanation: Autopilot

While most of our explanations were used in previous work, we introduce a more assertive explanation we call the `Autopilot`. At each time step during the question, `Autopilot` gives the human player one bit of information: should you buzz or not. The suggestion is based on the binary prediction of whether the guesser's current top answer is correct or wrong, just as how human players assesses their own confidence.

An autonomous AI could use `Autopilot` to decide when to buzz. But in a human-AI team, it's just a suggestion, and the decision is still left to the human. If the human blindly follows the suggestion, the human-AI team reduces to an autonomous AI trying to win by itself, hence the name.

Both `Autopilot` and the selector are trying to maximize the chance of winning. Whereas `Autopilot` is optimizing for the AI only, the selector optimizes for the team. And this is no coincidence—we design `Autopilot` to test if selective explanation goes beyond implicit calibration: the hope is for it to outperform both human–`Autopilot` team and a fully-autonomous AI using `Autopilot` to decide when to buzz.

We use a simple, threshold-based model for `Autopilot` similar to Yamada et al. (2018): it looks at the normalize confidence scores of the top five guesses, and recommends buzzing if the gap between the top two is larger than 0.05 (a threshold tuned on the dev set from Rodriguez et al. (2019)). Despite its simplicity, this model is very efficient at chooshing the right time to buzz (Yamada et al., 2018; Rodriguez et al., 2019).

2.6 Training the Explanation Selector

Our goal is to build effective human-AI teams whose cooperation requires the selector to select which explanations to show to the human. This section describe the machine learning model—learned from users' preferences in behavioral data—of the user which lets the selector pick user-specific explanations to show the user. Finally, to model the exploration-exploitation trade-off, we use multi-armed bandits to learn the selector policy and maximize the accumulated EW score.

2.6.1 User Model

Given a human player, a question, and one of the available explanation configurations, the user

model predicts the the EW score received from this question. To model aspects of the human player as well as properties of each specific question, the user model uses both manually crafted features and BERT representations. Table 2 shows the full list of features. The user model can also be veiwed as a value function in reinforcement learning.

2.6.2 Optimizing Accumulated EW Score

Our goal is to empower humans to complete the task at hand as accurately and as efficiently as possible. Given a new question, the selector should choose the best configuration based on its model of the user; however, to learn this model, the selector needs to test how well each of configuration works for each type of questions. This presents an exploration-exploitation trade-off, which we model with multi-armed bandits (Robbins, 1952). We optimize the accumulated reward—the accumulated EW score of the team. In the experiments, we compare several bandit algorithms.

3 Experiments

We run two experiments with real human participants: a single-player experiment with amateurs, and a multi-player real-time Quizbowl tournament with experts. This section first introduces the metric for evaluating Quizbowl competency, then provides details about the human players, the AI player, the explanation methods, and the baselines. We show that selective explanation provides personalized decision support and leads to the best augmented human performance.

3.1 Evaluating accuracy and efficiency using one metric without an opponent

Winning in Quizbowl requires you to answer correctly before your opponent. In real Quizbowl games with two or more players, a high score is a proof that a player is both accurate and efficient—in the sense that they require little information to get the answer right. In a perfect assessment of Quizbowl player, we would control for factors such as question topic and have a head-to-head competition between every pair of players. In an ideal evaluation of decision support, we need to control for confounders such as player skill, and have a head-to-head comparison between every possible pair of differently-augmented players, e.g., strong player with no support vs. weak player with selective explanations, and vice versa. However, this is infeasible due to the number of confounders.

We would like a single metric to evaluate both accuracy and efficiency without running head-to-head competitione. Accuracy is trivial to evaluate by itself, but efficiency is not as simple as counting the number of words that the player had seen when they answered a question correctly because not all words have the same value: answering earlier by one word is much more difficult at the beginnning of the question than at the end. The reward for answering earlier should be proportional to the increase in the chance of beating an opponent.

The expected wins (EW) metric implements this idea. Concretely, it assigns a weight to each correct answer depending on the percentage of the question revealed. The higher the percentage, the lower the assigned weight. For example, answer answering correctly halfway through the question counts as 0.3 points in EW, while a correct answer at the end only counts as 0.05 points. We use weights provided by Rodriguez et al. (2019) which are estimated using maximium likelihood from previous game data (Boyd-Graber et al., 2012).

3.2 Setup: Mechanical Turkers as Amateurs

We recruit twenty amateur players on Amazon Mechanical Turk. Each amateur player answers a set of sixty Quizbowl questions, and the questions are randomly permuted for each player. Each player is randomly assigned to either the experimental group with selective explanations or a control group with a baseline policy; more on these conditions later.

Before the user answers questions, we familiarize the user with the interface. During that period, the user can explore the interface without restriction (e.g., they can turn explanations on and off), and we switch to the assigned setting after the user clicks a button to indicate that they are ready.

3.3 Setup: Quizbowl Enthusiasts as Experts

We recruit twenty expert Quizbowl players from online forums. For these experts, we use a newly commissioned set of 144 questions no participant has seen before. The questions are divided into six rounds with twenty-four questions each.

Unlike the amateur experiment, the experts play a real multi-player Quizbowl game. To make sure that our game is fair and competitive, we divide players into three rooms. The initial assignment uses players' self-reported skill level. We subsequently adjust the assignment at the end of each round by promoting the top 20% players in each room and relegating the bottom 20%.

Condition	Description
`None-fixed`	Display no explanation.
`Everything-fixed`	Display all explanations.
`Random-dynamic`	Choose a new random configuration for each question.
`Selective-dynamic`	Selector chooses the configuration for each question.
`Autopilot-fixed`	Display `Autopilot` suggestions only.
`AI-only`	`Autopilot` replaces human player.

Table 3: Conditions in the randomized controlled trial. Under `fixed` conditions, one configuration is used for all questions; under `dynamic` conditions, the enabled configurations could change from one question to another. In all conditions the human player has access to the guesser's prediction. In the baseline `AI-only` condition, no human player is involved.

3.4 Setup: AI Guesser and Explanations

The human player is assisted by a machine learning guesser. Given a question, the guesser produces a multinomial distribution over the set of possible answers (Boyd-Graber et al., 2012); we update this prediction after every four question words. We use the BERT-based guesser from Rodriguez et al. (2019), and refer readers to that paper for model details and standard evaluation results. Next we discuss how we generate explanations for the guesser.

- `Alternatives`: We show the guesser's current top five predictions along with their confidence scores.
- `Evidence`: We retreive four training examples that are most similar to the current question. To measure similarity we use cosine distance between question representations by the guesser (Wallace et al., 2018).
- `Highlights` on **question**: We use HotFlip (Ebrahimi et al., 2017) and show tokens with a nomalized attribution score higher than 0.15.
- `Highlights` for **evidence**: We search for the highlighted question tokens in the retrieved training examples, and highlight them.
- `Autopilot`: We colorize the guesser's prediction based on the `Autopilot`'s current decision: red if buzzing, gray if not. When `Autopilot` is disabled, the color is always black.

Hyperparameters of an explanation (e.g. number of highlighted tokens) affect its effectiveness. Here we choose a fix set of hyperparameters based on feedback from internal trial runs. However, the choice of hyperparamters can also be considered as part of the explanation configuration. Then, we can use the selector with an expanded action space to, for example, also choose the number of tokens to highlight. We discuss this more in Section 5.2.

3.5 Setup: Selector policy

As the user plays, we train their personalized selector policy using LinUCB (Auer, 2002). The parameters of the user model are not updated during bandit training; new information gathered about the user is incorporated into the user model via features (Table 2).

3.6 Setup: Conditions and Baselines

Table 3 lists the conditions of our randomized controlled trial. The experimental condition is selective explanations. The control conditions include baseline policies such as using a fixed explanation configuration for all questions. To control the number of conditions, we omit conditions with fixed configurations, e.g. `Alternatives`+`Evidence`-`fixed`. Instead, we include `Everything-fixed`, which Feng and Boyd-Graber (2019) show to be most effective at improving user accuracy.

The guesser's accuracy is on par with the experts. So if the amateur players are *willing* and *able* to *blindly* and precisely follow the `Autopilot`, they could achieve good scores. But we consider this as a degenerate solution to human-AI cooperation.

To account for this issue, we include two special settings. In `Autopilot-fixed`, we display `Autopilot` suggestions as the only explanation for the human player. In `AI-only`, we *replace* the human player with `Autopilot` to make decisions. Using these two settings, we can quantify to what degree the human player follows `Autopilot`.

In our forum post for expert recruitment, we promise an "interface to augment human players explanations of AI predictions". To stay true to

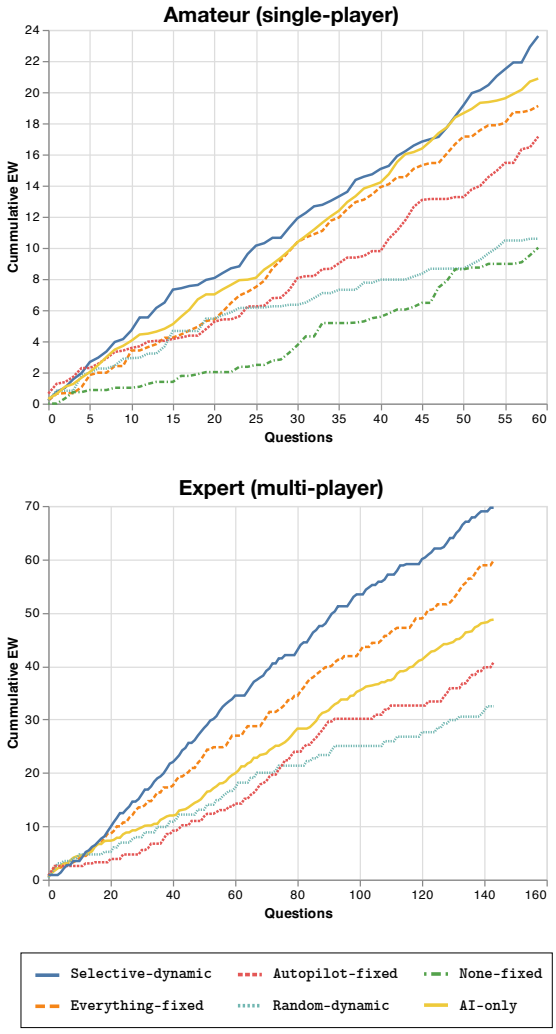

Figure 2: Mean cumulative EW score under each condition by amateurs (top) and experts (bottom). The selective condition performs the best amongst all human-AI cooperative settings.

this promise and ensure a good experience for the experts (who participates in the game of Quizbowl for fun), we omit the baseline `None-fixed` condition in our expert experiments. This omission should not affect our results since the baseline is already compared to other conditions in Feng and Boyd-Graber (2019).

3.7 Evaluation: Does mediation improve performance measure by EW?

We use the mean cumulative EW score over the course of the game (144 questions for experts and 60 for amateurs) for our quantitative comparison. If the human-AI team with a tailored selector can improve their EW score, this suggests explanations are helping the users more than other conditions.

Figure 2 shows how the mean EW score under each condition increases as the players answer more questions. Among all human-AI coopeartive settings, the `Selective-dynamic` condition performs the best. Especially for experts, selective explanation by the selector is better than both showing all explanations and `AI-only`. Importantly, as our model acquires more data for each the each user with more questions (and as the user acclimates to their teammate), the gap between `Selective-dynamic` and `Everything-fixed` grows.

Without explanations, amateurs are much wose than `AI-only`. With selective explanations, amateurs are comparable to `AI-only` and only slightly better than showing all explanations.

Under the `Autopilot-fixed` condition, if players blindly follow the AI's suggestion—buzz when the `Autopilot` says so and provide the AI prediction as the answer—they should match the `AI-only` baseline. However, both experts and amateurs lose to the `AI-only` under this condition. This indicates that the other conditions evince a synergy: humans are not simply blindly following the AI suggestions more closely. Rather, the diverse and selective explanations allow the players to better decide when to follow and when to use their own knowledge.

3.8 Analysis: What does selector choose to show?

We are interested in what the selector learns as most effective and what it chooses to show to players. Figure 3 visualizes the evolving distribution of configurations selected by the bandit selector and that by the random selector.

First, the selector did not learn to show all explanations for all questions—it learned to be selective. And by comparing to the random selector, we see that the selector formed a clear preference among explanations. In fact, at the end of the game, the selector—learning purely from interaction—recovers the ranking of individual explanations reported by Feng and Boyd-Graber (2019): highlight > evidence > alternatives. Interestingly, the selector did not coverge to this ranking until the players finished about 60 questions: initially the list of alternatives was the preferred explanations, possibly because it is easier for the players to interpret than the others. Eventually as the players get more used to the other explanations and the selector continues to learn about the players, it converges.

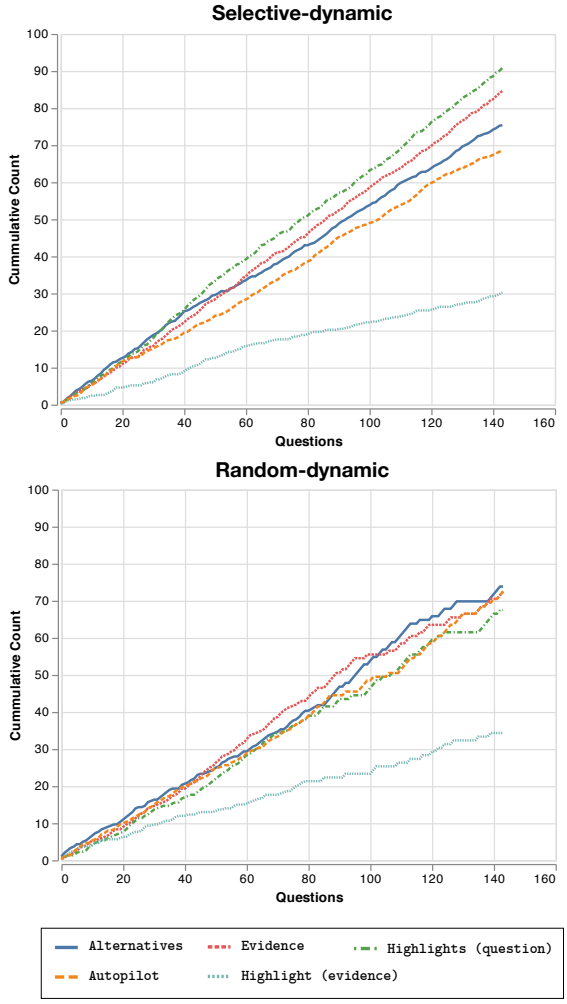

Figure 3: Mean cumulative count of explanations being shown to **experts**. Here we compare the explanations selected by the selector (left) and by random (right). Based on the frequency, we see that the selector learned a ranking of explanations consistent with the effectiveness reported in Feng and Boyd-Graber (2019): question highlights is most effectiv, then evidence, then alternatives.

4 Discussion and Related Work

In this section, we discuss related work and possible extensions of selective explanations.

4.1 Who should drive?

Clearly defining the shared obligations of the team is crucial to the success of the team. By design, we keep ultimate control of decision making with the human. However, this may not be optimal; a distracted, overloaded, or hesitating human might be better served by an AI "taking the wheel" if it is certain. The most relevant work to ours is Gao et al. (2021), which similarly uses bandit feedback to optimize team performance. Whereas our policy chooses from the set of explanation configurations, their policy makes a binary decision: whether to delegate a decision to the human or leave it to the AI. Our `Autopilot` explanation can be seen as "soft" delegation. Future work should compare selective explanation with more methods for delegation and deferral (Madras et al., 2018; Lubars and Tan, 2019; Kamath et al., 2020; Lai et al., 2022).

4.2 Alignment, and learning to optimize human objectives

Typically, ML algorithms optimize automatic metrics: how well can a machine replace or emulate a human. However, this is inconsistent with how humans and machines interact in the real world; often models need to be personalized to users (Zhou and Brunskill, 2016). The research area that deals with the general problem of optimizing human's objectives is alignment (Amodei et al., 2016). Specifically for human-AI teams, an unsettled question is how to optimize for that partnership; while we optimize for short-term accuracy, a reasonable alternative would be to optimize for longer-term learning Bragg and Brunskill (2020). An interesting direction would be to take a real-world task and directly optimize the underlying model (not just the selector) to create tailored explanations, as Lage et al. (2018) did for synthetic tasks.

5 Conclusion: Explanations Tailored for Users

Users benefit from collaborating with AI, and this collaboration can be improved by explaining the AI well. Moreover, the this benefit is not universal, some users need or thrive with different explanations. However, finding the right combination is not easy; while our bandit approach can find useful explanations, it requires both the user to become acclimated to human-AI teaming and the bandit to explore the space of explanations. As human-AI collaborations become more common, we must continue to search for better signals and methods to help the teaming minimize stress and acclimation but maximize fun and productivity.

Limitations

5.1 Limited Modeling of Factors in Human-AI Cooperation

As we discussed in Section 1, a major contributor to the inconsistency of human-AI experimental results is the large number of factors that can influence the cooperative effectiveness. One of those factors that's relatively easy to model is the human's skill level. In theory, selective explanation should be able to model that: if we optimize selective explanation jointly for experts and amateurs, the selector should be able to learn and choose different explanations for the two different groups of players. Unfortunately we couldn't have done that experiment because Quizbowl is too challenging for mechanical turkers without any assistance, and when they compete head-to-head the game is made more difficult by the element of competition.

There are other factors of human-AI cooperation that has been identified by previous work but we couldn't model: the level of human agency (Lai and Tan, 2019; Bansal et al., 2021) the model's predictive accuracy (Bansal et al., 2020), the user's mental model of machine learning (Bansal et al., 2019), and the amount of interactivity (Smith-Renner et al., 2020a,b). Even within limited interactions, there is significant variation about the optimal modality of explanations (Gonzalez et al., 2020). Other factors, such as the distribution of test examples and model architecture, affect the quality of output from various post-hoc explanation methods (Ghorbani et al., 2019; Jones et al., 2020).

Another major limitation of our evaluation is that we only experimented with one question answering problem, Quizbowl. Our method is generally applicable to decision making problems. But finding another suitable task and adapting our infrastructure, experiment design, incentive strutures is highly non-trivial. We are actively looking for other problems to experiment on and hope to conduct more extensive experiments in the future.

5.2 Selector's Action Space is Limited

We present this work as another step towards learned explanations that are more aligned with human values (Amodei et al., 2016). Our method seeks to maximize a human objective, not heuristic proxies of that (Doshi-Velez and Kim, 2018), and not the objective of the solo machine. In this work we focus on a simplified setting with a limited design and action space, but our experimental setting closely mimics how a human-AI team would operate in a real-world task; in particular, our testbed, Quizbowl, bears merits that are essential for a task to have in order to benefit from this idea.

We focus on this restricted selector to keep the sample complexity for multi-armed bandit under control. In principle the selector could be more fine-grained if we allow it to dynamically change the configuration as the clues in the question are revealed. Despite challenges with regards to sample complexity, we believe that this expansion of action space is a logical next step.

Ethics Statement

The general ethical concerns of explainable artificial intelligence (XAI) apply to this work, and we refer readers to Miller (2019) and Gunning et al. (2019) for a more detail account of those concerns.

A special concern with this work is what counts as explanations. This paper studies exclusively post-hoc explanations that do not have theoretical guarantees. These ad-hoc explanations might appear reasonable—and they do, in some sense, since they improve human performance in our experiments, but there is no telling whether the information conveyed by the explanations is reliable. In other words, it is equally justifiable to interpret these so-called explanations as persuasions or even deceptions—in the sense that the model and the explanation method are collectively trying to convince the human to agree with them. To hedge against this concern, we do not make any claims about the nature of these explanations in this paper. Instead, we study the empirical properties of them, and whether they can be useful.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv: 1606.06565*.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: Risk assessments in criminal sentencing.

Peter Auer. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2020. Is the most accurate AI the best teammate? optimizing ai for team-

work. In *Association for the Advancement of Artificial Intelligence*.

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

Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *International Conference on Human Factors in Computing Systems*.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413.

Paula Bitrián, Isabel Buil, and Sara Catalán. 2021. Enhancing user engagement: The role of gamification in mobile apps. *Journal of Business Research*, 132:170–185.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of Advances in Neural Information Processing Systems*.

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*.

Jonathan Bragg and Emma Brunskill. 2020. Fake it till you make it: Learning-compatible performance support. In *Proceedings of Uncertainty in Artificial Intelligence*.

TB Brown, B Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. Language models are few-shot learners. arxiv 2020. In *Proceedings of Advances in Neural Information Processing Systems*.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *International Conference on Intelligent User Interfaces*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*.

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

Finale Doshi-Velez and Been Kim. 2018. Towards a rigorous science of interpretable machine learning. *Springer Series on Challenges in Machine Learning*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the Association for Computational Linguistics*.

Douglas C Engelbart. 1962. Augmenting human intellect: A conceptual framework. *Menlo Park, CA*, page 21.

Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me: Evaluating machine learning interpretations in cooperative play. In *International Conference on Intelligent User Interfaces*.

Yarin Gal, Yutian Chen, Roger Frigola, S. Gu, Alex Kendall, Yingzhen Li, Rowan McAllister, Carl Rasmussen, Ilya Sutskever, Gabriel Synnaeve, Nilesh Tripuraneni, Richard Turner, Oriol Vinyals, Adrian Weller, Mark van der Wilk, and Yan Wu. 2016. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Oxford.

Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI collaboration with bandit feedback. In *International Joint Conference on Artificial Intelligence*.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. 2019. Interpretation of neural networks is fragile. In *Association for the Advancement of Artificial Intelligence*.

Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer. 2020. Human evaluation of spoken vs. visual explanations for open-domain qa. *arXiv preprint arXiv:2012.15075*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120.

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*.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. 2020. Selective classification can magnify disparities across groups. *arXiv preprint arXiv:2010.14134*.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the Association for Computational Linguistics*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference of Machine Learning*.

Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *International Conference on Human Factors in Computing Systems*.

Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2018. Human-in-the-loop interpretability prior. *arXiv preprint arXiv:1805.11571*.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *International Conference on Human Factors in Computing Systems*.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*.

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**.

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

Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *International Conference on Human Factors in Computing Systems*.

Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–45.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*.

Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology*.

Brian Lubars and Chenhao Tan. 2019. Ask not what ai can do, but what ai should do: Towards a framework of task delegability. In *Proceedings of Advances in Neural Information Processing Systems*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of Advances in Neural Information Processing Systems*.

David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. In *Proceedings of Advances in Neural Information Processing Systems*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference of Machine Learning*.

Alexander Renkl. 2014. Toward an instructionally oriented theory of example-based learning. *Cognitive science*, 38(1):1–37.

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*.

Herbert Robbins. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*.

Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*.

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*, 550(7676):354–359.

Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020a. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *International Conference on Human Factors in Computing Systems*.

Alison Smith-Renner, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2020b. Digging into user control: perceptions of adherence and instability in transparent models. In *International Conference on Intelligent User Interfaces*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*.

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Ikuya Yamada, Ryuji Tamaki, Hiroyuki Shindo, and Yoshiyasu Takefuji. 2018. Studio ousia's quiz bowl question answering system. *arXiv preprint arXiv: 1803.08652*.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *International Conference on Human Factors in Computing Systems*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *Proceedings of the International Conference on Learning Representations*.

Li Zhou and Emma Brunskill. 2016. Latent contextual bandits and their application to personalized recommendations for new users.