

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

Links:

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

Downloaded from http://umiacs.umd.edu/~jbg/docs/2022_emnlp_augment.pdf

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

001 Explanations promise to bridge the gap be- 042
002 tween human and AI, yet AI-augmented hu- 043
003 man decision making proves difficult: expla- 044
004 nations are helpful in some cases but harm- 045
005 ful in others (Bansal et al., 2021; Lai et al., 046
006 2021). The effect of explanation depends on 047
007 many factors, such as human expertise (Feng 048
008 and Boyd-Graber, 2019), human agency (Lai 049
009 and Tan, 2019), and explanation format (Gon- 050
010 zalez et al., 2020; Smith-Renner et al., 2020a). 051
011 Using a uniform setup—always showing the 052
012 same type of explanation in all cases—is sub- 053
013 optimal, but it’s also hard to rely on heuris- 054
014 tics to adapt the setup for each scenario. We 055
015 propose learning to explain selectively using 056
016 human feedback to directly optimize human 057
017 accuracy. We formulate selective explanation 058
018 as a contextual bandit problem, train a model 059
019 to learn users’ needs and preferences online, 060
020 and use the model to choose the best combi- 061
021 nation of explanations to provide in each sce- 062
022 nario. We experiment on question answering 063
023 following the evaluation protocol of Feng and 064
024 Boyd-Graber (2019) and show that selective 065
025 explanations further improve human accuracy 066
026 for both experts and amateurs.

027 1 Introduction

028 Recent advances in machine learning (ML) (Sil- 069
029 ver et al., 2017; Brown et al.; Jumper et al., 2021; 070
030 Ramesh et al., 2021) sparked new life in **intelli- 071
031 gence augmentation**—the vision that computers 072
032 are not mere number-crunching tools, but also 073
033 interactive systems that can augment humans at 074
034 problem solving and decision making (Engelbart, 075
035 1962). The hope is to combine the complimen- 076
036 tary strengths of machine and human, and to fully 077
037 harness the capabilities of these models with hu- 078
038 man intuitions and oversight (Dafoe et al., 2020; 079
039 Amodei et al., 2016). But this agenda is obstructed 080
040 by the many counterintuitive traits of neural net- 081
041 works (NNS) (Szegedy et al., 2014; Goodfellow

et al., 2015; Zhang et al., 2017) and our lack of the- 042
oretical understanding (Belkin et al., 2019): these 043
models are not interpretable to humans by default 044
and it is difficult to foresee when they will fail. 045
This lack of interpretability also amplifies the risk 046
of model bias (Angwin et al., 2016; Bolukbasi et al., 047
2016; Caliskan et al., 2017), making it difficult to 048
use NN-powered AIs in real-world decision making. 049

To bridge the gap between human and machine, 050
various methods attempt to explain model predic- 051
tions in human-interpretable terms, e.g., by pro- 052
viding more context to the model’s uncertainty 053
estimation (Gal et al., 2016; Bhatt et al., 2021), 054
by highlighting the most important part of the 055
input (Ribeiro et al., 2016; Lundberg and Lee, 056
2017; Ebrahimi et al., 2017), and by retrieving 057
the most relevant training examples (Renkl, 2014; 058
Koh and Liang, 2017). Grounded in psychol- 059
ogy (Lombrozo, 2006, 2007; Kulesza et al., 2012), 060
these explanations promise to augment human 061
decision making. But when tested in application- 062
grounded evaluations—with real problems and real 063
humans (Doshi-Velez and Kim, 2018), it proves 064
difficult for any single explanation method to 065
achieve consistent improvement in disparate con- 066
text (Bansal et al., 2021; Buçinca et al., 2020). 067

A major contributor to this problem is the 068
breadth of context that the explanation method is 069
applied to. Internally, the explanation method is 070
faced with shifts in the input distribution which the 071
model can react badly to (Goodfellow et al., 2015; 072
Liu et al., 2021); externally, it needs to deal with 073
human users with diverse levels of expertise (Feng 074
and Boyd-Graber, 2019), engagement (Sidner et al., 075
2005), and general trust in AI (Dietvorst et al., 076
2015). Our current use of explanations demands 077
an one-size-fits-all solution, but existing methods 078
cannot provide that as they are largely oblivious to 079
the above mentioned variables. 080

Selective explanations Each person is unique, and 081
the right explanation will also vary from one deci- 082

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

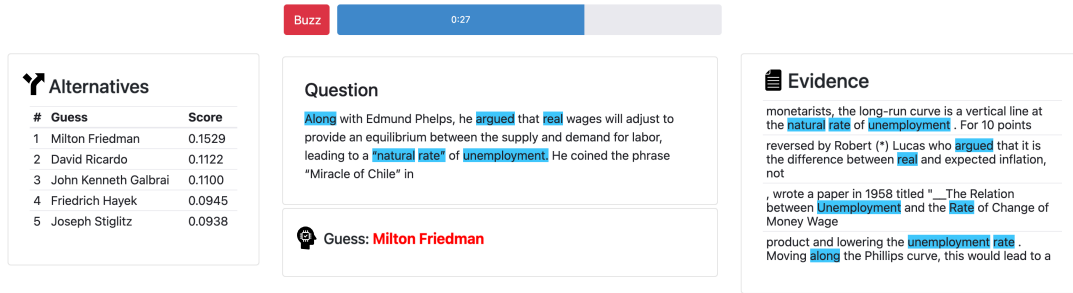


Figure 1: Our Quizbowl web interface when all four explanations are displayed. In the middle we show the question word-by-word; below, we show the current best model guess, which is colored red when the Autopilot is confident, otherwise gray; on the left we show Alternatives, including confidence scores; 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 penalty 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:

¹While 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.

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

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

Table 1: Each configuration is a set of visualizations shown to users, and our policy learns which configuration helps users the most. Most visualizations can be turned on or off independently, 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 independently, we have in total twenty possible configurations.

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 calibra-

#	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.

tion: 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 normalized 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 choosing 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 competition. 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 beginning 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 maximum 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 retrieve 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 Hot-Flip (Ebrahimi et al., 2017) and show tokens with a normalized 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. . 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

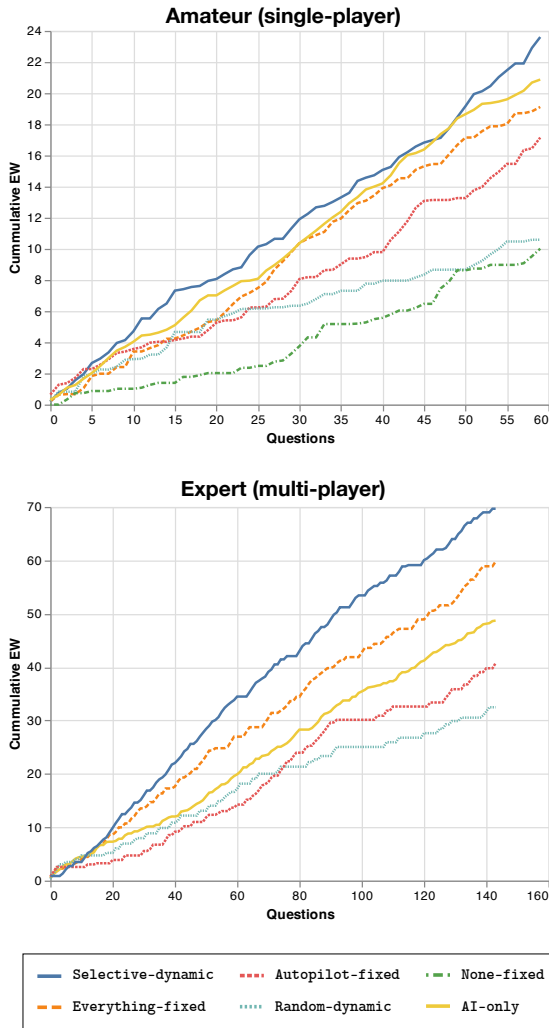


Figure 2: Mean cumulative EW score under each condition by amateurs (top) and experts (bottom). The selective condition performs the best amongst all human-AI cooperative settings.

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 un-

der each condition increases as the players answer more questions. Among all human-AI cooperative 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 coverage 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.

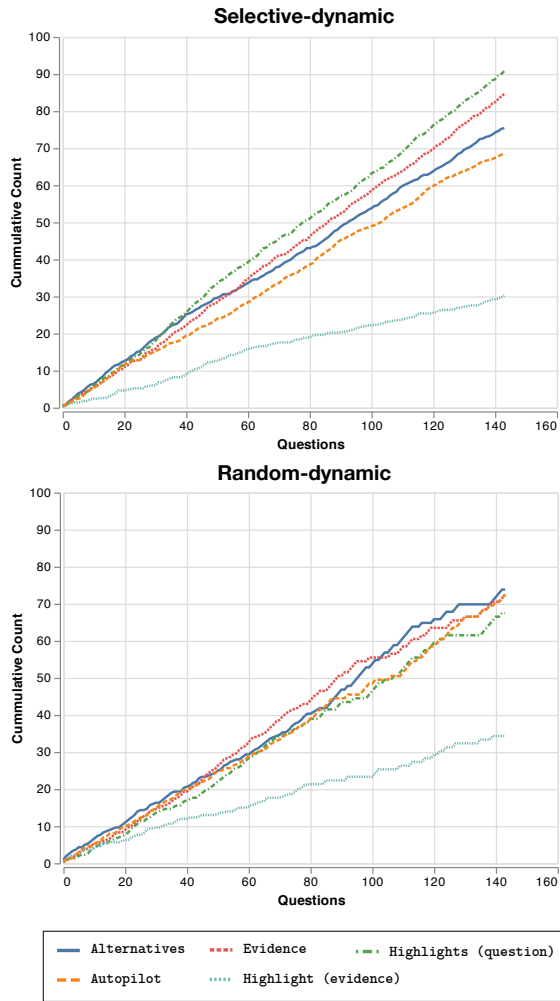


Figure 3: Mean cumulative count of explanations being shown to **experts**. Here we compare the explanations selected by the selector (left) and by random (right). Based on the frequency, we see that the selector learned a ranking of explanations consistent with the effectiveness reported in [Feng and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

594 Limitations

595 5.1 Limited Modeling of Factors in 596 Human-AI Cooperation

597 As we discussed in Section 1, a major contribu- 644
598 tor to the inconsistency of human-AI experimen- 645
599 tal results is the large number of factors that can influ- 646
600 ence the cooperative effectiveness. One of those 647
601 factors that’s relatively easy to model is the hu- 648
602 man’s skill level. In theory, selective explanation 649
603 should be able to model that: if we optimize selec- 650
604 tive explanation jointly for experts and amateurs, 651
605 the selector should be able to learn and choose dif- 652
606 ferent explanations for the two different groups of 653
607 players. Unfortunately we couldn’t have done that 654
608 experiment because Quizbowl is too challenging 655
609 for mechanical turkers without any assistance, and 656
610 when they compete head-to-head the game is made 657
611 more difficult by the element of competition.

612 There are other factors of human-AI cooperation 644
613 that has been identified by previous work but we 645
614 couldn’t model: the level of human agency (Lai 646
615 and Tan, 2019; Bansal et al., 2021) the model’s 647
616 predictive accuracy (Bansal et al., 2020), the user’s 648
617 mental model of machine learning (Bansal et al., 649
618 2019), and the amount of interactivity (Smith- 650
619 Renner et al., 2020a,b). Even within limited in- 651
620 teractions, there is significant variation about the 652
621 optimal modality of explanations (Gonzalez et al., 653
622 2020). Other factors, such as the distribution of 654
623 test examples and model architecture, affect the 655
624 quality of output from various post-hoc explana- 656
625 tion methods (Ghorbani et al., 2019; Jones et al., 657
626 2020).

627 Another major limitation of our evaluation is 644
628 that we only experimented with one question an- 645
629 swering problem, Quizbowl. Our method is gener- 646
630 ally applicable to decision making problems. But 647
631 finding another suitable task and adapting our in- 648
632 frastructure, experiment design, incentive strutures 649
633 is highly non-trivial. We are actively looking for 650
634 other problems to experiment on and hope to con- 651
635 duct more extensive experiments in the future.

636 5.2 Selector’s Action Space is Limited

637 We present this work as another step towards 644
638 learned explanations that are more aligned with 645
639 human values (Amodei et al., 2016). Our method 646
640 seeks to maximize a human objective, not heuristic 647
641 proxies of that (Doshi-Velez and Kim, 2018), and 648
642 not the objective of the solo machine. In this work 649
643 we focus on a simplified setting with a limited de-

644 sign and action space, but our experimental setting 645
646 closely mimics how a human-AI team would oper- 647
648 ate in a real-world task; in particular, our testbed, 649
649 Quizbowl, bears merits that are essential for a task 650
650 to have in order to benefit from this idea. 651

652 We focus on this restricted selector to keep the 653
653 sample complexity for multi-armed bandit under 654
654 control. In principle the selector could be more 655
655 fine-grained if we allow it to dynamically change 656
656 the configuration as the clues in the question are 657
657 revealed. Despite challenges with regards to sam- 658
658 ple complexity, we believe that this expansion of 659
659 action space is a logical next step. 660

661 Ethics Statement

662 The general ethical concerns of explainable artifi- 644
663 cial intelligence (XAI) apply to this work, and we 645
664 refer readers to Miller (2019) and Gunning et al. 646
665 (2019) for a more detail account of those concerns. 647

668 A special concern with this work is what counts 644
669 as explanations. This paper studies exclusively 645
670 post-hoc explanations that do not have theoretical 646
671 guarantees. These ad-hoc explanations might ap- 647
672 pear reasonable—and they do, in some sense, since 648
673 they improve human performance in our experi- 649
674 ments, but there is no telling whether the informa- 650
675 tion conveyed by the explanations is reliable. In 651
676 other words, it is equally justifiable to interpret 652
677 these so-called explanations as persuasions or even 653
678 deceptions—in the sense that the model and the 654
679 explanation method are collectively trying to con- 655
680 vince the human to agree with them. To hedge 656
681 against this concern, we do not make any claims 657
682 about the nature of these explanations in this paper. 658
683 Instead, we study the empirical properties of them, 659
684 and whether they can be useful. 660

685 References

- 686 Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. 644
687 Christiano, John Schulman, and Dan Mané. 2016. 645
688 Concrete problems in AI safety. *arXiv preprint* 646
689 *arXiv: 1606.06565*. 647
- 690 Julia Angwin, Jeff Larson, Surya Mattu, and Lauren 648
691 Kirchner. 2016. Machine bias: Risk assessments in 649
692 criminal sentencing. 650
- 693 Peter Auer. 2002. Using confidence bounds for 644
694 exploitation-exploration trade-offs. *Journal of Ma-* 645
695 *chine Learning Research*, 3(Nov):397–422. 646
- 696 Gagan Bansal, Besmira Nushi, Ece Kamar, Eric 647
697 Horvitz, and Daniel S Weld. 2020. Is the most accu- 648
698 rate AI the best teammate? optimizing ai for team- 649
699 650

693	work. In <i>Association for the Advancement of Artificial Intelligence</i> .	
694		
695	Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S	
696	Lasecki, Daniel S Weld, and Eric Horvitz. 2019.	
697	Beyond accuracy: The role of mental models in	
698	human-ai team performance. In <i>Proceedings of</i>	
699	<i>the AAAI Conference on Human Computation and</i>	
700	<i>Crowdsourcing</i> .	
701	Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond	
702	Fok, Besmira Nushi, Ece Kamar, Marco Tulio	
703	Ribeiro, and Daniel S Weld. 2021. Does the whole	
704	exceed its parts? the effect of ai explanations on	
705	complementary team performance. In <i>International</i>	
706	<i>Conference on Human Factors in Computing Sys-</i>	
707	<i>tems</i> .	
708	Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik	
709	Mandal. 2019. Reconciling modern machine-	
710	learning practice and the classical bias-variance	
711	trade-off. <i>Proceedings of the National Academy of</i>	
712	<i>Sciences</i> .	
713	Umang Bhatt, Javier Antorán, Yunfeng Zhang,	
714	Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato,	
715	Gabrielle Melançon, Ranganath Krishnan, Jason	
716	Stanley, Omesh Tickoo, et al. 2021. Uncertainty as	
717	a form of transparency: Measuring, communicating,	
718	and using uncertainty. In <i>Proceedings of the 2021</i>	
719	<i>AAAI/ACM Conference on AI, Ethics, and Society</i> ,	
720	pages 401–413.	
721	Paula Bitrián, Isabel Buil, and Sara Catalán. 2021. En-	
722	hancing user engagement: The role of gamifica-	
723	tion in mobile apps. <i>Journal of Business Research</i> ,	
724	132:170–185.	
725	Tolga Bolukbasi, Kai-Wei Chang, James Y Zou,	
726	Venkatesh Saligrama, and Adam T Kalai. 2016.	
727	Man is to computer programmer as woman is to	
728	homemaker? debiasing word embeddings. In <i>Pro-</i>	
729	<i>ceedings of Advances in Neural Information Pro-</i>	
730	<i>cessing Systems</i> .	
731	Jordan L. Boyd-Graber, Brianna Satinoff, He He, and	
732	Hal Daumé III. 2012. Besting the quiz master:	
733	Crowdsourcing incremental classification games. In	
734	<i>Proceedings of Empirical Methods in Natural Lan-</i>	
735	<i>guage Processing</i> .	
736	Jonathan Bragg and Emma Brunskill. 2020. Fake it	
737	till you make it: Learning-compatible performance	
738	support. In <i>Proceedings of Uncertainty in Artificial</i>	
739	<i>Intelligence</i> .	
740	TB Brown, B Mann, N Ryder, M Subbiah, J Kaplan,	
741	P Dhariwal, A Neelakantan, P Shyam, G Sastry,	
742	A Askell, et al. Language models are few-shot learn-	
743	ers. arxiv 2020. In <i>Proceedings of Advances in Neu-</i>	
744	<i>ral Information Processing Systems</i> .	
745	Zana Bućinca, Phoebe Lin, Krzysztof Z Gajos, and	
746	Elena L Glassman. 2020. Proxy tasks and subjec-	
747	tive measures can be misleading in evaluating ex-	
748	plainable ai systems. In <i>International Conference</i>	
749	<i>on Intelligent User Interfaces</i> .	
	Aylin Caliskan, Joanna J Bryson, and Arvind	750
	Narayanan. 2017. Semantics derived automatically	751
	from language corpora contain human-like biases.	752
	<i>Science</i> , 356(6334):183–186.	753
	Allan Dafoe, Edward Hughes, Yoram Bachrach, Tan-	754
	tum Collins, Kevin R McKee, Joel Z Leibo, Kate	755
	Larson, and Thore Graepel. 2020. Open problems	756
	in cooperative ai. <i>arXiv preprint arXiv:2012.08630</i> .	757
	Berkeley J Dietvorst, Joseph P Simmons, and Cade	758
	Massey. 2015. Algorithm aversion: people er-	759
	roneously avoid algorithms after seeing them err.	760
	<i>Journal of Experimental Psychology: General</i> ,	761
	144(1):114.	762
	Finale Doshi-Velez and Been Kim. 2018. Towards a	763
	rigorous science of interpretable machine learning.	764
	<i>Springer Series on Challenges in Machine Learning</i> .	765
	Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing	766
	Dou. 2017. HotFlip: White-box adversarial exam-	767
	ples for text classification. In <i>Proceedings of the As-</i>	768
	<i>sociation for Computational Linguistics</i> .	769
	Douglas C Engelbart. 1962. Augmenting human in-	770
	tellect: A conceptual framework. <i>Menlo Park, CA</i> ,	771
	page 21.	772
	Shi Feng and Jordan Boyd-Graber. 2019. What can AI	773
	do for me: Evaluating machine learning interpreta-	774
	tions in cooperative play. In <i>International Confer-</i>	775
	<i>ence on Intelligent User Interfaces</i> .	776
	Yarin Gal, Yutian Chen, Roger Frigola, S. Gu, Alex	777
	Kendall, Yingzhen Li, Rowan McAllister, Carl Ras-	778
	mussen, Ilya Sutskever, Gabriel Synnaeve, Nilesch	779
	Tripuraneni, Richard Turner, Oriol Vinyals, Adrian	780
	Weller, Mark van der Wilk, and Yan Wu. 2016. <i>Un-</i>	781
	<i>certainty in Deep Learning</i> . Ph.D. thesis, University	782
	of Oxford.	783
	Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-	784
	Arteaga, Ligong Han, Min Kyung Lee, and Matthew	785
	Lease. 2021. Human-AI collaboration with bandit	786
	feedback. In <i>International Joint Conference on Arti-</i>	787
	<i>ficial Intelligence</i> .	788
	Amirata Ghorbani, Abubakar Abid, and James Y. Zou.	789
	2019. Interpretation of neural networks is fragile. In	790
	<i>Association for the Advancement of Artificial Intelli-</i>	791
	<i>gence</i> .	792
	Ana Valeria Gonzalez, Gagan Bansal, Angela Fan,	793
	Robin Jia, Yashar Mehdad, and Srinivasan Iyer.	794
	2020. Human evaluation of spoken vs. visual ex-	795
	planations for open-domain qa. <i>arXiv preprint</i>	796
	<i>arXiv:2012.15075</i> .	797
	Ian J. Goodfellow, Jonathon Shlens, and Christian	798
	Szegedy. 2015. Explaining and harnessing adversar-	799
	ial examples. In <i>Proceedings of the International</i>	800
	<i>Conference on Learning Representations</i> .	801

802	David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—explainable artificial intelligence. <i>Science robotics</i> , 4(37):eaay7120.	854
803		855
804		856
805		857
		858
806	He He, Jordan L. Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In <i>Proceedings of the International Conference of Machine Learning</i> .	859
807		860
808		861
809		862
810	Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. 2020. Selective classification can magnify disparities across groups. <i>arXiv preprint arXiv:2010.14134</i> .	863
811		
812		
813		
814	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. <i>Nature</i> , 596(7873):583–589.	
815		
816		
817		
818		
819		
820	Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In <i>Proceedings of the Association for Computational Linguistics</i> .	864
821		865
822		
823		
824	Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In <i>Proceedings of the International Conference of Machine Learning</i> .	866
825		867
826		
827		
828	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 <i>International Conference on Human Factors in Computing Systems</i> .	868
829		869
830		870
831		871
832		
833	Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2018. Human-in-the-loop interpretability prior. <i>arXiv preprint arXiv:1805.11571</i> .	872
834		873
835		874
836		875
837	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 <i>International Conference on Human Factors in Computing Systems</i> .	876
838		877
839		878
840		879
841		
842		
843	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. <i>arXiv preprint arXiv:2112.11471</i> .	880
844		881
845		882
846		
847	Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In <i>Proceedings of ACM FAT*</i> .	883
848		884
849		885
850		886
851	John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. <i>Human factors</i> , 46(1):50–80.	887
852		888
853		889
		890
		891
		892
		893
		894
		895
		896
		897
		898
		899
		900
		901
		902
		903
		904
		905

906 David Silver, Julian Schrittwieser, Karen Simonyan,
907 Ioannis Antonoglou, Aja Huang, Arthur Guez,
908 Thomas Hubert, Lucas Baker, Matthew Lai, Adrian
909 Bolton, et al. 2017. Mastering the game of go with-
910 out human knowledge. *nature*, 550(7676):354–359.

911 Alison Smith-Renner, Ron Fan, Melissa Birchfield,
912 Tongshuang Wu, Jordan Boyd-Graber, Daniel S
913 Weld, and Leah Findlater. 2020a. No explainability
914 without accountability: An empirical study of expla-
915 nations and feedback in interactive ml. In *Interna-
916 tional Conference on Human Factors in Computing
917 Systems*.

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

923 Christian Szegedy, Wojciech Zaremba, Ilya Sutskever,
924 Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and
925 Rob Fergus. 2014. Intriguing properties of neural
926 networks. In *Proceedings of the International Con-
927 ference on Learning Representations*.

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

932 Ikuya Yamada, Ryuji Tamaki, Hiroyuki Shindo, and
933 Yoshiyasu Takefuji. 2018. Studio ousia’s quiz bowl
934 question answering system. *arXiv preprint arXiv:
935 1803.08652*.

936 Ming Yin, Jennifer Wortman Vaughan, and Hanna Wal-
937 lach. 2019. Understanding the effect of accuracy on
938 trust in machine learning models. In *International
939 Conference on Human Factors in Computing Sys-
940 tems*.

941 Chiyuan Zhang, Samy Bengio, Moritz Hardt, Ben-
942 jamin Recht, and Oriol Vinyals. 2017. Understand-
943 ing deep learning requires rethinking generalization.
944 In *Proceedings of the International Conference on
945 Learning Representations*.

946 Li Zhou and Emma Brunskill. 2016. Latent contextual
947 bandits and their application to personalized recom-
948 mendations for new users.