

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 complementary 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

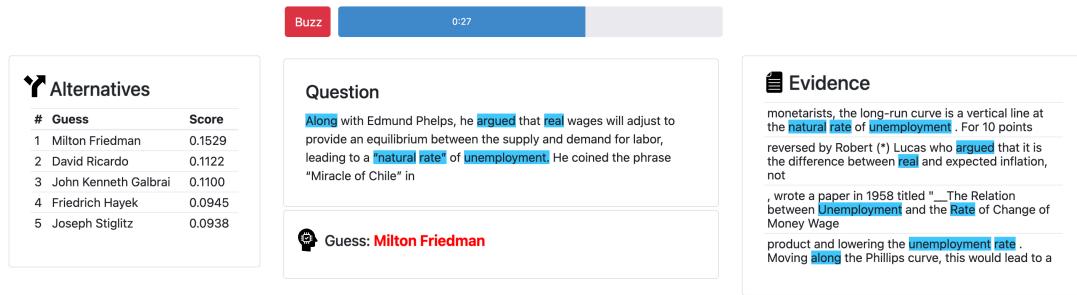


Figure 1: Our Quizbowl web interface when all four explanations are displayed. In the middle we show the question word-by-word; below, we show the current best model guess, which is colored red when the Autopilot is confident, otherwise gray; on the left we show Alternatives, including confidence scores;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

246 both the question and the evidence. In [Feng](#)
 247 and [Boyd-Graber \(2019\)](#), the two are treated
 248 as one explanation. However, their experi-
 249 ments confirm that highlighting the question
 250 alone is already effective. In this paper we sep-
 251 arate the two and the policy can control them
 252 individually. Table 1 lists the available config-
 253 urations for Highlights and Evidence.

2.5 A New Explanation: Autopilot

255 While most of our explanations were used in pre-
 256 vious work, we introduce a more assertive expla-
 257 nation we call the Autopilot. At each time step
 258 during the question, Autopilot gives the human
 259 player one bit of information: should you buzz or
 260 not. The suggestion is based on the binary predi-
 261 cation of whether the guesser’s current top answer
 262 is correct or wrong, just as how human players
 263 assesses their own confidence.

264 An autonomous AI could use Autopilot to
 265 decide when to buzz. But in a human-AI team, it’s
 266 just a suggestion, and the decision is still left to the
 267 human. If the human blindly follows the sugges-
 268 tion, the human-AI team reduces to an autonomous
 269 AI trying to win by itself, hence the name.

270 Both Autopilot and the selector are trying
 271 to maximize the chance of winning. Whereas
 272 Autopilot is optimizing for the AI only, the
 273 selector optimizes for the team. And this is no
 274 coincidence—we design Autopilot to test if se-
 275 lective 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.

276 tion: the hope is for it to outperform both human-
 277 Autopilot team and a fully-autonomous AI us-
 278 ing Autopilot to decide when to buzz.

279 We use a simple, threshold-based model for
 280 Autopilot similar to [Yamada et al. \(2018\)](#): it
 281 looks at the normalized confidence scores of the top
 282 five guesses, and recommends buzzing if the gap
 283 between the top two is larger than 0.05 (a threshold
 284 tuned on the dev set from [Rodriguez et al. \(2019\)](#)).
 285 Despite its simplicity, this model is very efficient
 286 at choosing the right time to buzz ([Yamada et al., 2018](#);
 287 [Rodriguez et al., 2019](#)).

2.6 Training the Explanation Selector

288 Our goal is to build effective human-AI teams
 289 whose cooperation requires the selector to select
 290 which explanations to show to the human. This sec-
 291 tion describes the machine learning model—learned
 292 from users’ preferences in behavioral data—of the
 293 user which lets the selector pick user-specific ex-
 294 planations to show the user. Finally, to model the
 295 exploration-exploitation trade-off, we use multi-
 296 armed bandits to learn the selector policy and max-
 297 imize the accumulated EW score.

2.6.1 User Model

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

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

309 2.6.2 Optimizing Accumulated EW Score

310 Our goal is to empower humans to complete the
311 task at hand as accurately and as efficiently as pos-
312 sible. Given a new question, the selector should
313 choose the best configuration based on its model
314 of the user; however, to learn this model, the se-
315 lector needs to test how well each of configuration
316 works for each type of questions. This presents an
317 exploration-exploitation trade-off, which we model
318 with multi-armed bandits (Robbins, 1952). We op-
319 timize the accumulated reward—the accumulated
320 EW score of the team. In the experiments, we
321 compare several bandit algorithms.

322 3 Experiments

323 We run two experiments with real human partici-
324 pants: a single-player experiment with amateurs,
325 and a multi-player real-time Quizbowl tournament
326 with experts. This section first introduces the met-
327 ric for evaluating Quizbowl competency, then pro-
328 vides details about the human players, the AI player,
329 the explanation methods, and the baselines. We
330 show that selective explanation provides person-
331 alized decision support and leads to the best aug-
332 mented human performance.

333 3.1 Evaluating accuracy and efficiency using 334 one metric without an opponent

335 Winning in Quizbowl requires you to answer cor-
336 rectly before your opponent. In real Quizbowl
337 games with two or more players, a high score is a
338 proof that a player is both accurate and efficient—
339 in the sense that they require little information to
340 get the answer right. In a perfect assessment of
341 Quizbowl player, we would control for factors such
342 as question topic and have a head-to-head com-
343 petition between every pair of players. In an ideal
344 evaluation of decision support, we need to control
345 for confounders such as player skill, and have a
346 head-to-head comparison between every possible
347 pair of differently-augmented players, e.g., strong
348 player with no support vs. weak player with selec-
349 tive explanations, and vice versa. However, this is
350 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 in-
crease 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 ques-
tion revealed. The higher the percentage, the lower
the assigned weight. For example, answer answer-
ing 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 esti-
mated using maximum likelihood from previous
game data (Boyd-Graber et al., 2012).

373 3.2 Setup: Mechanical Turkers as Amateurs

We recruit twenty amateur players on Amazon Me-
chanical 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.

387 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 subse-
quently 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 Guessers 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 hyperparameters 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. `Y+D-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

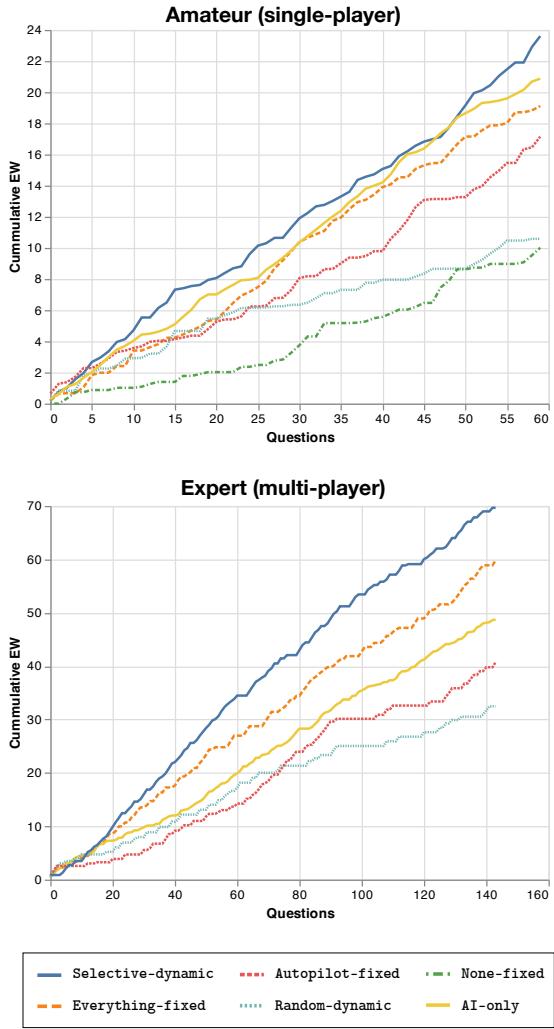


Figure 2: Mean cumulative EW score under each condition by amateurs (top) and experts (bottom). The selective condition performs the best amongst all human-AI cooperative settings.

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 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 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 worse 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 converge 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.

473
474
475
476
477
478
479

480
481

482
483
484
485
486
487
488

489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515

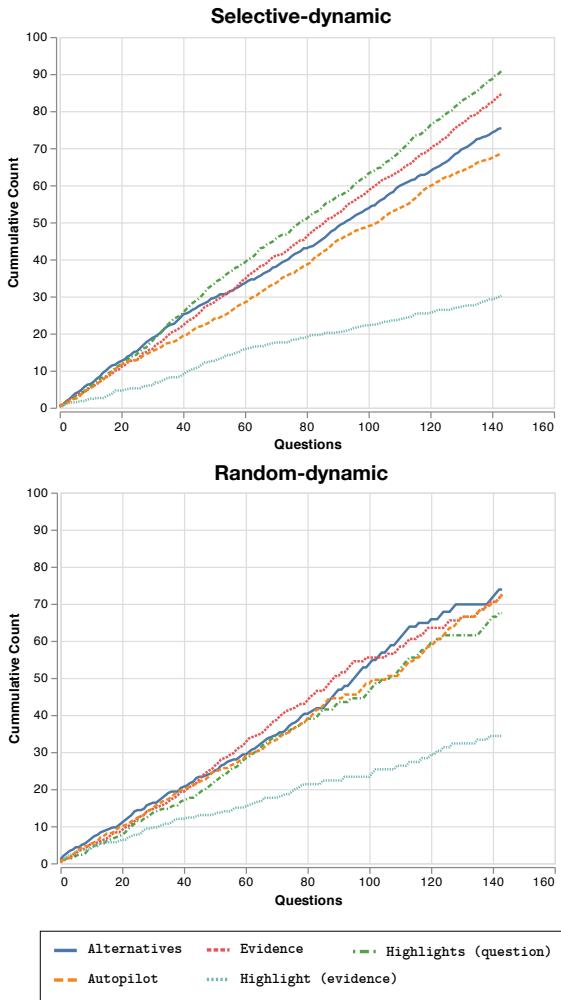


Figure 3: Mean cumulative count of explanations being shown to **experts**. Here we compare the explanations selected by the selector (left) and by random (right). Based on the frequency, we see that the selector learned a ranking of explanations consistent with the effectiveness reported in [Feng and 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.

539

540
541

542

543
544
545
546
547
548
549
550

551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578

594 Limitations

595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 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 structures is highly non-trivial. We are actively looking for other problems to experiment on and hope to conduct more extensive experiments in the future.

636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 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 de-

sign 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.

657 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.

679 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-

693	work. In <i>Association for the Advancement of Artificial Intelligence</i> .	750
694		751
695		752
696		753
697	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 <i>Proceedings of the AAAI Conference on Human Computation and Crowdsourcing</i> .	754
698		755
699		756
700		757
701	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 <i>International Conference on Human Factors in Computing Systems</i> .	758
702		759
703		760
704		761
705		762
706		763
707		764
708	Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias-variance trade-off. <i>Proceedings of the National Academy of Sciences</i> .	765
709		
710		
711		
712		
713	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 <i>Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society</i> , pages 401–413.	766
714		767
715		768
716		769
717		
718		
719		
720		
721	Paula Bitrián, Isabel Buil, and Sara Catalán. 2021. Enhancing user engagement: The role of gamification in mobile apps. <i>Journal of Business Research</i> , 132:170–185.	770
722		771
723		772
724		
725	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 <i>Proceedings of Advances in Neural Information Processing Systems</i> .	773
726		774
727		775
728		776
729		
730		
731	Jordan L. Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. 2012. Besting the quiz master: Crowdsourcing incremental classification games. In <i>Proceedings of Empirical Methods in Natural Language Processing</i> .	777
732		778
733		779
734		780
735		781
736	Jonathan Bragg and Emma Brunskill. 2020. Fake it till you make it: Learning-compatible performance support. In <i>Proceedings of Uncertainty in Artificial Intelligence</i> .	782
737		783
738		
739		
740	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 <i>Proceedings of Advances in Neural Information Processing Systems</i> .	784
741		785
742		786
743		787
744		788
745	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 <i>International Conference on Intelligent User Interfaces</i> .	789
746		790
747		791
748		792
749		
693	Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. <i>Science</i> , 356(6334):183–186.	750
694		751
695		752
696		753
697		
698		
699		
700		
701		
702		
703		
704		
705		
706		
707		
708		
709		
710		
711		
712		
713		
714		
715		
716		
717		
718		
719		
720		
721		
722		
723		
724		
725		
726		
727		
728		
729		
730		
731		
732		
733		
734		
735		
736		
737		
738		
739		
740		
741		
742		
743		
744		
745		
746		
747		
748		
749		

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
806		858
807	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
808		860
809		861
810		862
811	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
812		864
813		865
814		866
815	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.	867
816		868
817		869
818		870
819		871
820		872
821	Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In <i>Proceedings of the Association for Computational Linguistics</i> .	873
822		874
823		875
824		876
825	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> .	877
826		878
827		879
828		880
829	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> .	881
830		882
831		883
832		884
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> .	885
834		886
835		887
836		888
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> .	889
838		890
839		891
840		892
841		893
842		894
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> .	895
844		896
845		897
846		898
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> .	899
848		900
849		901
850		902
851	John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. <i>Human factors</i> , 46(1):50–80.	903
852		904
853		905
854	Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In <i>International Conference on Human Factors in Computing Systems</i> .	906
855		907
856		908
857		909
858		910
859	Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. <i>Proceedings of the ACM on Human-Computer Interaction</i> , 5(CSCW2):1–45.	911
860		912
861		913
862		914
863		915
864	Tania Lombrozo. 2006. The structure and function of explanations. <i>Trends in cognitive sciences</i> .	916
865		917
866	Tania Lombrozo. 2007. Simplicity and probability in causal explanation. <i>Cognitive psychology</i> .	918
867		919
868	Brian Lubars and Chenhao Tan. 2019. Ask not what ai can do, but what ai should do: Towards a framework of task delegability. In <i>Proceedings of Advances in Neural Information Processing Systems</i> .	920
869		921
870		922
871		923
872	Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In <i>Proceedings of Advances in Neural Information Processing Systems</i> .	924
873		925
874		926
875		927
876	David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. In <i>Proceedings of Advances in Neural Information Processing Systems</i> .	928
877		929
878		930
879		931
880	Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. <i>Artificial Intelligence</i> .	932
881		933
882		934
883	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 <i>Proceedings of the International Conference of Machine Learning</i> .	935
884		936
885		937
886		938
887		939
888	Alexander Renkl. 2014. Toward an instructionally oriented theory of example-based learning. <i>Cognitive science</i> , 38(1):1–37.	940
889		941
890		942
891	Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “why should i trust you?”: Explaining the predictions of any classifier. In <i>Knowledge Discovery and Data Mining</i> .	943
892		944
893		945
894		946
895	Herbert Robbins. 1952. Some aspects of the sequential design of experiments. <i>Bulletin of the American Mathematical Society</i> .	947
896		948
897		949
898	Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. <i>arXiv preprint arXiv:1904.04792</i> .	950
899		951
900		952
901		953
902	Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. <i>Artificial Intelligence</i> .	954
903		955
904		956
905		957

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 *Inter-
916 ternational 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 *Inter-
922 ternational 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 Chiyan Zhang, Samy Bengio, Moritz Hardt, Ben-
942 jamin Recht, and Oriol Vinyals. 2017. Under-
943 standing 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.