From "No Clear Winner" to an Effective XAI Process: An Empirical Journey

Jonathan Dodge | Andrew Anderson | Roli Khanna | Jed Irvine | Rupika Dikkala | Kin-Ho Lam | Delyar Tabatabai | Anita Ruangrotsakun | Zeyad Shureih | Minsuk Kahng | Alan Fern | Margaret Burnett

School of Electrical Engineering and Computer Science, Oregon State University, Oregon, USA

Correspondence
Margaret Burnett, Oregon State University.
Email: burnett@oregonstate.edu

Summary

"In what circumstances would you want this AI to make decisions on your behalf?" We have been investigating how to enable a user of an AI-powered system to answer questions like this through a series of empirical studies, a group of which we summarize here. We began the series by (1) comparing four explanation configurations of saliency explanations and/or reward explanations. From this study we learned that, although some configurations had significant strengths, no one configuration was a clear "winner." This result led us to hypothesize that one reason for the low success rates Explainable AI (XAI) research has in enabling users to create a coherent mental model is that the AI itself does not have a coherent model. This hypothesis led us to (2) build a model-based agent, to compare explaining it with explaining a model-free agent. Our results were encouraging, but we then realized that participants' cognitive energy was being sapped by having to create not only a mental model, but also a process by which to create that mental model. This realization led us to (3) create such a process (which we term *After-Action Review for AI* or "AAR/AI") for them, integrate it into the explanation environment, and compare participants' success with AAR/AI scaffolding vs. without it. Our AAR/AI studies' results showed that AAR/AI participants were more effective assessing the AI than non-AAR/AI participants, with significantly better precision and significantly better recall at finding the AI's reasoning flaws.

KEYWORDS:
Explainable AI, After-Action Review for AI (AAR/AI), empirical studies, human-computer interaction

1 | INTRODUCTION

In recent years, *Explainable Artificial Intelligence (XAI)* has begun to focus on the actionability of explanations. Each explanation has its own strengths and weaknesses, but for some users' tasks an explanation's particular strengths might not be valuable or its weaknesses might be critical.

For our research team, the kind of actionability we seek is explanations enabling a user to *assess* a particular AI's suitability for performing tasks on behalf of that user. Since we have been working in the context of Real-Time Strategy (RTS) games, the

outcome we seek is for the user to be able to answer this question: "In what circumstances would you want the agent to play this game on your behalf?" This paper summarizes multiple empirical investigations of how we might achieve this outcome, and what they together revealed.

2 | STUDY 1: EXPLAINING THE 4-TOWERS RTS GAME

How should reinforcement learning (RL) agents explain themselves to domain experts attempting to assess an AI's suitability, if these domain experts are not trained in AI? To investigate this question, we conducted a 124-participant experiment and measured their mental models by: 1) their ability to predict the AI's actions and 2) a written "rulebook" of how they think the AI made its decisions. We did this to compare participants' mental models of an RL agent in the context of a simple Real-Time Strategy (RTS) game. Our goal was to compare participants' mental models of the AI agent under four configurations of explanations: (Control) no explanations; (Saliency) saliency maps, which explain the AI's focus of attention; (Reward) reward-decomposition bars, which explain the AI's predictions of future types of rewards; and (Everything) both Saliency and Rewards. These explanations are shown in Figure 1.

To compare these explanation configurations, we designed a between-subject controlled lab study to measure differences in how people would respond to different combinations of explanations in a Real-Time Strategy (RTS) game. In order to control the space of possible actions, we built our own game, which we refer to as the 4-Towers game. In the 4-Towers game, the agent controlled a kite-shaped tank, placed in the middle of four quadrants (Figure 1, top left). The agent had to decide which of the four quadrants it would attack. Its goal was to maximize its score.
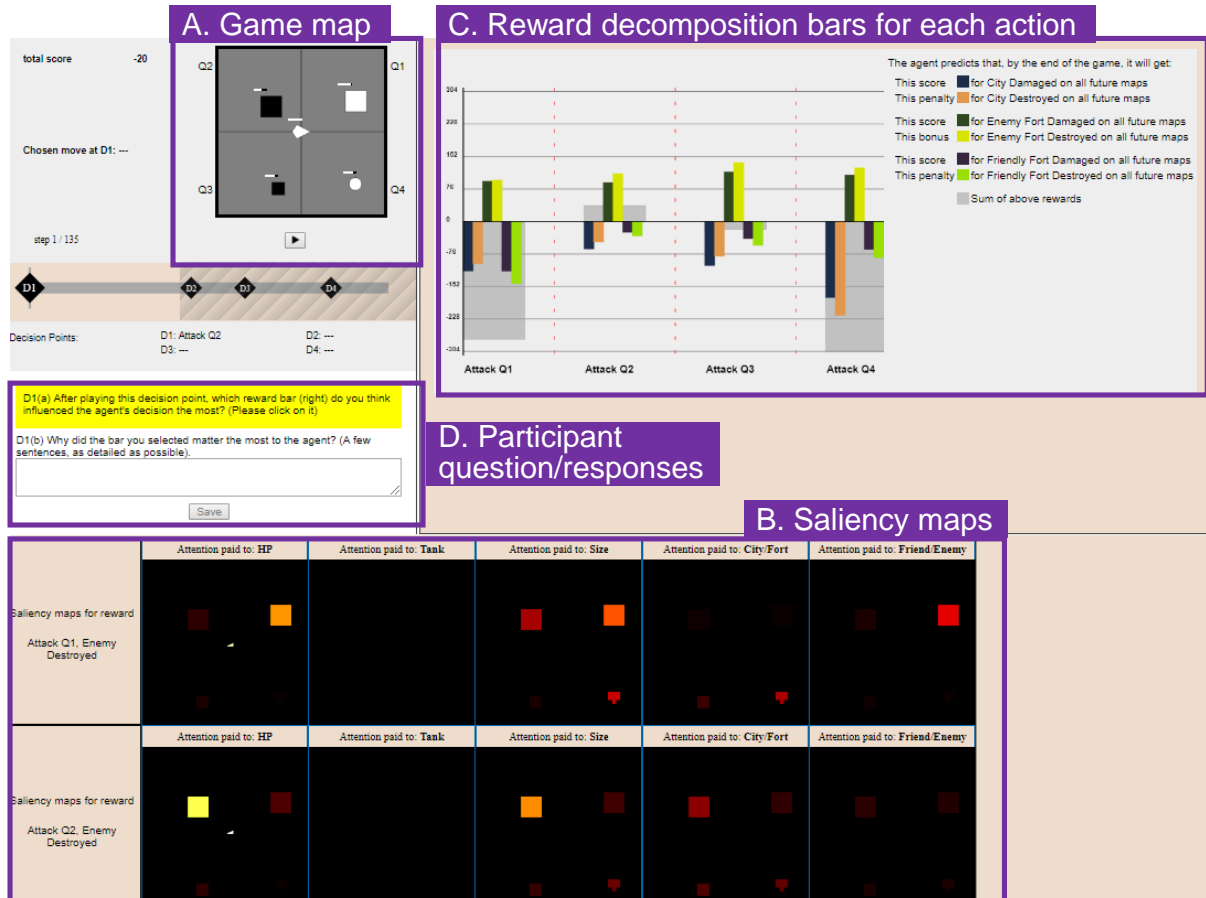

FIGURE 1 The 4-Towers interface used for participants in the "Everything" treatment [1].

FIGURE 2 The 4-Towers participants' final mental model scores for the four treatments.
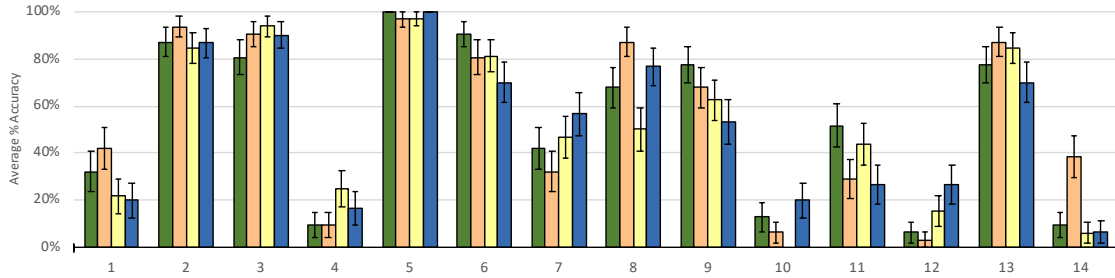

FIGURE 3 Percentage of 4-Towers participants who successfully predicted the AI's next move at each decision point (DP), prior to receiving any explanation [1]. Bar colors denote treatment (from left to right): Control, Saliency, Rewards, and Everything, for each of the DPs (14 DPs presented here). Participants' results varied markedly for the different situations these DPs captured, and there is no evidence that any of the treatments' participants got better over time. All error bars ($SE = \sigma / \sqrt{n}$) are under 10%.

Our 124 participants took a hands-on tutorial to learn how the game worked and how to navigate the system. They then worked their way through 14 decisions the agent made. At each decision point, participants attempted to predict the agent's next decision, and then saw the decision and the explanation and commented on them. At the end of the two-hour session, participants attempted to describe how the agent works. Full details of the study set-up and game can be found in Anderson et al. [1]

We measured participants' efficacy with the decisions by both analyzing their written descriptions using a rubric to arrive at mental model scores, and by their success rates at predicting the AI's next decision. Our results showed that a combined explanation that included saliency and reward bars was needed to achieve a statistically significant difference in participants' mental model scores over the no-explanation treatment (Figure 2). For participants' abilities to predict the agent's next decisions, the reward explanation's presence was more helpful than a saliency map's presence, but participants in *every* treatment had highly inconsistent results (Figure 3).

3 | MOVING TO MODEL-BASED EXPLANATIONS

The 4-Towers results suggest that there was no single "best" explanation able to properly accommodate the participants' diversity and the decision situations' diversity. We speculated that a possible reason might be an inherent difficulty of a human creating a coherent mental model of an AI that does not *itself* have a coherent model.

To investigate whether an AI with a model of itself would improve our ability to explain the AI to humans, we needed a model-based agent and an environment in which the agent should operate. A *model-based* reinforcement learning agent learns an explicit *model* of the environment dynamics (in the form of a search tree for our agent), and it plans its actions via a look-ahead tree search. In contrast, a model-free agent does not explicitly learn a model.

Thus, we created a new environment based on StarCraft 2; specifically, a Tug-of-War environment and a model-based reinforcement learning (RL) agent that works in this environment. We then iteratively prototyped and evaluated several variants of visual explanations and user workflow supports. This section summarizes the game and agents we created [2,3], which we then used in the remainder of the studies described in this paper.
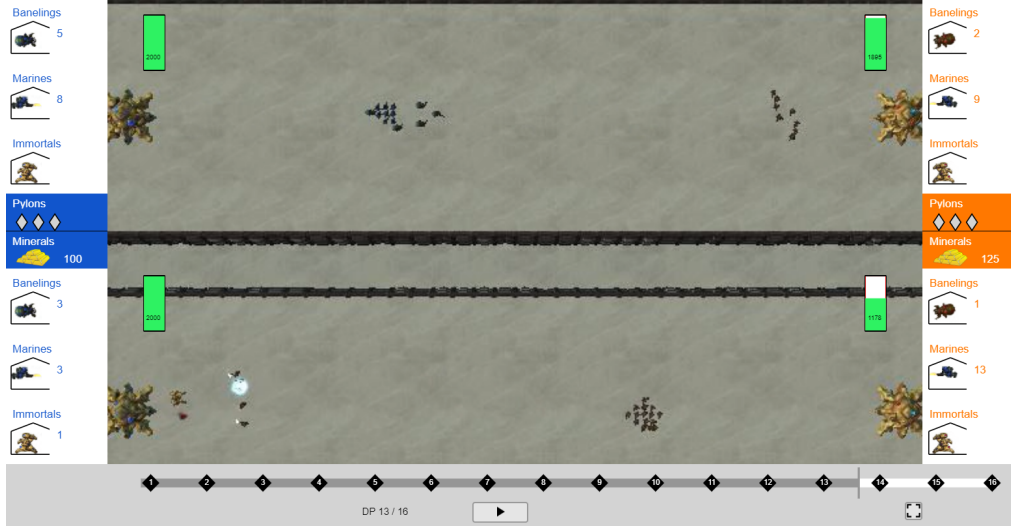
FIGURE 4 The game replay interface for users to watch the game in action.
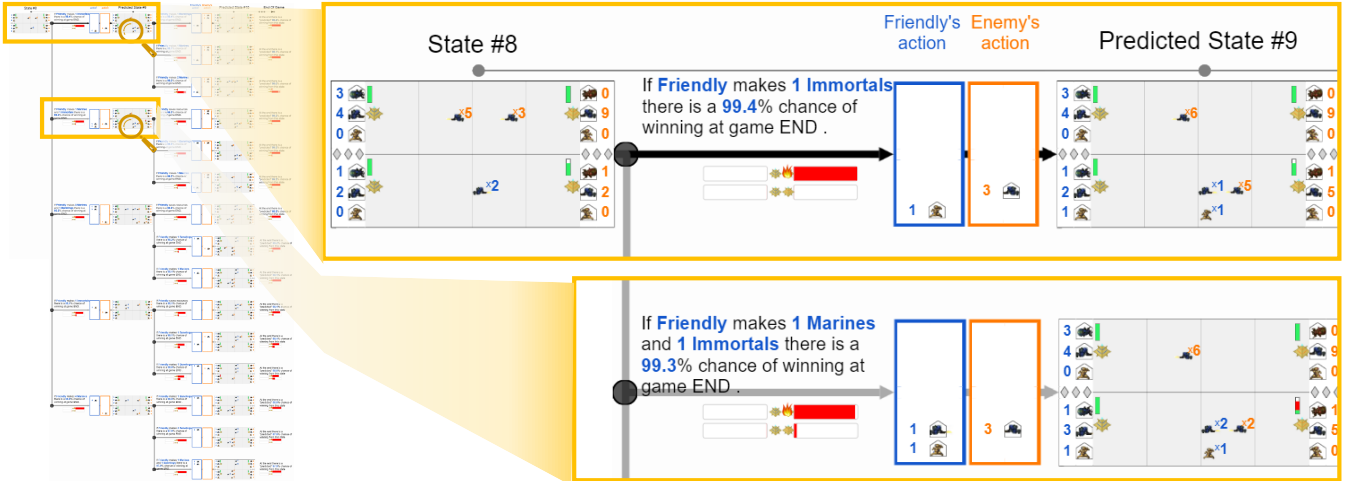

FIGURE 5 Explanation interface to visualize the reasoning process of the reinforcement learning agent. For the state shown (State #8), it shows a number of possible actions (e.g., purchasing an Immortal) and the predicted next states for these actions as a tree structure. States are "thumbnail" images of game states as per Figure 4.

Game Environment: Tug-of-War. The environment we created is a two-player adversarial environment requiring real-time strategic decision making (Figure 4). We created the game using a StarCraft 2 game editor. The game involves a mixture of managing an economy while making offensive/defensive military decisions. Each player can take an action (e.g., purchases resources) at the end of each 30-second wave. In our environment, the players are AI agents.

Model-Based RL Agent Architecture. We used a model-based architecture that resembles the one used by AlphaZero for Go, Chess, and Shogi[4]. In particular, the agent uses the following three learned components to conduct a game-tree search at each decision point to select an action:

- *Learned Model.* This model enables predicting the next state given the current history and actions of the two players. It makes predictions every 30 seconds.
- *Action Ranker.* A score is calculated for each possible action in a game state, which allows actions to be ranked. Our agent uses this ranking whenever it decides to prune actions from consideration during the tree search.
- *Learning Value Function.* The value function estimates the value of a state, which is used to estimate leaf values in our tree search.

Explanations and Explanation Interface. The rationale for each decision made by our model-based agent is completely captured by the search tree constructed for that decision. In this sense, the full search tree constitutes a sound and complete explanation[5] of the agent's decision making process. However, it is challenging to visualize the model's huge tree structure that will enable a human user to understand any one decision the agent made. After multiple iterations of design and testing with users, we arrived at a visualization showing the look-ahead tree structure in Figure 5.

From a user's perspective, the initial interaction begins in the game navigation interface, which allows the user to navigate a game replay (Figure 4). At any decision point, the user can choose to enter the explanation interface to see the multiple possible actions the agent considered, and why it chose the one it chose (Figure 5). (Note: We iterated through multiple versions for our studies; the figure shows the most recent version.) Specifically, the figure shows the game state at the user-selected decision point as the root of the tree (top left). Its child nodes are a subset of possible actions the agent could take from that state. Users can navigate this tree to review the AI's reasoning process.

4 | STUDY 2: MODEL-FREE VS. MODEL-BASED EXPLANATIONS

Using the new StarCraft-based prototype, we designed a qualitative study to compare how 22 human participants would react to model-based vs. model-free explanations of a reinforcement learning agent. Participants were university students with no expertise in AI or machine learning, but with at least some experience with RTS games. Participants in the two treatments (11 in model-free vs. 11 in model-based) used the same underlying explanation interface, except that the model-free explanation's lack of a model of the future meant it did not have a search tree, which amounted to displaying a tree of depth 1. Participants were tasked with making sense of the agents (assisted by the explanations we provided them) in order to assess the agents. We summarize the study's results here; for details, see Dodge et al.[2].

For purposes of comparing model-free with model-based explanations, we defined the concept of an *explanation-informed statement* to be a participant's statement that is both (1) informed by an explanation and (2) about the AI's reasoning process at some decision point. Such statements are in contrast to statements that are only about the AI's behavior, or about how the human thinks the AI should behave, or pure speculation (not drawing upon the explanations) about the AI's reasoning. We categorized these explanation-informed statements into the three types shown in Table 1.

As Table 1 shows, the 11 participants in the model-based treatment generated more explanation-informed statements than the 11 participants in the model-free treatment did. This was also true of each type of explanation-informed statement: model-based participants made about 3 times as many explanation-informed observations, about twice as many explanation-informed inferences, and about twice as many explanation-informed bug reports, as model-free participants did. Finally, the differences in individual participants who contributed one or more explanation-informed statements were consistent with the above statement-count differences. Specifically, twice as many model-based unique participants (10/11) made explanation-informed statements as model-free participants (5/11) did.

Type	Description	Example from our study	MF	MB
Explanation-Informed Observation	Participant strictly interprets the explanation.	"The friendly AI bought 1 baneling predicting that the opponent marines would increase."	7	23
Explanation-Informed Inference	Participant forms or adjusts their mental model explanation, judge the explanation.	"The AI is thinking ahead of how to win the game in the shortest number of rounds."	9	16
Explanation-Informed Bug Report	Participant identifies flaw/bug in agent's reasoning from explanation OR finds explanation confusing.	"Only 2 immortals were created. Prediction of marines was wrong."	4	9
Total			20	48

TABLE 1 Three classes of explanation-informed statements we used for coding participant responses, and the number of statements for the MF (model-free) vs. MB (model-based) treatment[2].

Together, these results show consistent evidence of the model-based participants basing more understanding on the *explanations* than the model-free participants did. We view these results, along with the qualitative results detailed in Dodge et al.[2], as encouraging evidence of the efficacy of model-based explanations, so we continued with the model-based context for the investigations we describe next.

5 | MULTIPLE STUDIES: AFTER-ACTION REVIEW FOR AI

As we considered the diverse ways by which participants did or did not engage with the explanations in the above study, we began to realize that they not only were having to build mental models of the AI agents, but also to build a process/workflow by which to do it. This realization led us to hypothesize that, if we provided a process to scaffold the explanations, participants might use more effective processes to make use of the explanations.

Toward that end, we devised a new workflow to support users' ability to gain actionable insights from XAI explanations. We term the process AAR/AI[2,6]—After-Action Review for AI. AAR/AI's inspiration, the manual After-Action Review process, was devised by the U.S. Army in the mid-70's[7], and has been a success in various branches of the military. After-Action Reviews have also been adapted for other domains including medical treatments[8], transportation services[9], and fire-fighting[10].

AAR/AI is a *process* for *application domain experts* to use in assessing whether and under which circumstances to rely upon an AI agent. We envision AAR/AI to be suitable for sequential domains, guiding the human through a series of steps to evaluate an AI agent's actions using explanations. AAR/AI is a flexible framework that can be tailored to the setting and granularity desired. Our first realization of it was as a 7-step process, which are conducted with a human Facilitator and one or more human Assessors as follows: (1) The Facilitator defines the domain. (2) The Facilitator explains the agent's objective. Steps 3-6 constitute an "inner loop" for each decision to be assessed: (3) The Facilitator reviews what was supposed to happen at this decision. (4) An Assessor identifies what actually happened at this decision. (5) An Assessor describes why it happened. (6) An Assessor formalizes learning from this decision. (7) Finally, an Assessor formalizes learning holistically from every decision they analyzed.

5.1 | The AAR/AI Study Series

We began our investigation of AAR/AI with a series of formative studies[2,6], in which humans were trying to assess whether and when they would trust the AI agent to play on their behalves. We then ran a summative study.

For the formative studies, our overall research question was:

RQ. What are the strengths and weaknesses of guiding human assessment using AAR/AI?

FIGURE 6 AAR/AI questionniare. Participants are asked to *describe* "what happened" by answering to a series of questions, after they filled out the AAR/AI questions on *predicting* "what was supposed to happen" and watch the game.

The studies gathered qualitative data, and were set in early variants of our model-based explanations in the Tug-of-War game described in Section 3. We began with a low-fidelity prototype, because our goal was to inform design decisions about the system we would need to build. The format of the earliest study was a think-aloud study involving 11 participants, with one participant per study session. We collected participants' verbalizations and written responses to the AAR/AI questions throughout their sessions, as well as their detailed behavior data. (Descriptions of the methodologies we used for some of these studies can be found in Dikkala et al.[11] and Dodge et al.[2])

In the earliest study, we "implemented" the explanation system with only predrawn explanation images. (Figure not shown here; however, Figure 5 shows the final version of the explanation after it evolved further.) The AAR/AI process scaffolding was a set of index cards on which participants answered the AAR/AI questions. We asked the AAR/AI questions at every 3rd decision point of a 22 decision point Tug-of-War game (Section 3). For example, for AAR/AI Step 3, we asked at the beginning of a decision point: "What do you think should happen in the next three rounds?" After participants wrote down their answers, they watched the next three rounds, at which point we asked them to "briefly explain what happened in the past three rounds" (Figure 6). Later studies in the formative series used prototypes with higher fidelity, eventually with the AAR/AI questions embedded in the computer prototype.

We then ran a quantitative study to compare the effectiveness of participants using the AAR/AI process with explanations against that using the same explanations but without AAR/AI. The participants were domain-knowledgeable users, and the task was to localize the AI's naturally-occurring reasoning "bugs" (reasoning flaws). Due to COVID-19, we conducted sessions over teleconference (Zoom) and a browser-based custom combination of the platform (game and explanation system described in Section 3, including AAR/AI features for the AAR/AI treatment). The participants were experienced with RTS games but had no AI or machine learning (ML) background. For this study, our overall research question was:

RQ. Given an AI agents and explanations of its actions, can domain-knowledgeable users find an AI agent's reasoning flaws more effectively with or without the AAR/AI process?

5.2 | Results

These studies produced three key findings.

Result 1: Encouraging a diversity of perspectives

One finding from these studies was that AAR/AI encouraged participants to think about the AI agent from multiple perspectives. When categorized according to the Lim-Dey intelligibility types[12], participants' remarks showed evidence of thinking about the agent via *all* of the Lim-Dey intelligibility types (Table 2). Further, the particular intelligibility type(s) considered were different for the different AAR/AI questions, suggesting that each question helped move participants to a new way of thinking about the AI's behavior. For example, responses to the "what do you think should happen..." AAR/AI question (top row) were

	What	What Could	How To	Judgement	Why Did	Why Didn't	Input	Model	Outputs	sum
"What do you think should happen in the next 3 rounds?" (Before watching them)	2	71	16	1	0	0	24	6	2	122
"Could you briefly explain about what actually happened in these past three rounds?" (After watching them)	13	6	2	6	18	2	53	12	74	186
"Why do you think the rounds happened the way they did?"	2	6	3	1	32	2	24	31	30	131
"Why do you think the Friendly AI did what it did?" (After seeing the explanation)	2	8	8	0	55	1	60	27	36	197
"What changes would you make in the decisions made by the Friendly AI to improve it?"	3	8	56	2	2	0	38	3	2	114
sum	22	99	85	10	107	5	199	79	144	750

TABLE 2 Lim-Dey categorization of participant responses to each AAR/AI question[2].

Bloom's level & description [16]	Example observations from our AAR/AI participants
1. Remembering: Have students acquired the ability to correctly recall information?	(Many examples of very basic knowledge)
2. Understanding: Can students understand the information they have learned to recall?	(Many examples of very basic knowledge)
3. Applying: Can students apply their newly learned knowledge?	"I ...like it how <the explanation diagram> is, because like I could try to draw my own conclusions from it."
4. Analyzing: Can students see patterns and make inferences about a problem?	"So we have almost same amount of health on top and bottom. So, for them to defeat us ... they have to focus on either one of these. So I guess they will focus more on bottom, because they have to save them..."
5. Evaluating: Do (can) students take a stand or decision, and justify it?	(watches the replay) "This is gonna be sad. ... Yep. It's all downhill from here. (after watching the replay) Uh, the friendly AI lost, uh, due to their misinvestment in the top row, um, and only increasing their baneling count, which only works at melee range which is ineffective..."
6. Creating: Can students create a new point of view?	"...I feel that the friendly's will invest in marines, especially more in the top row, since it is more damage..."

TABLE 3 Bloom's taxonomy applied to the AAR/AI participants. The first two levels are very basic, so we provide examples at only the higher levels. All participants reached Level 5, and all except one participant reached Level 6 [2].

mainly of the What-Could intelligibility type; responses to the "Why" questions landed fairly evenly across the Why-Did, Input, Model, and Output intelligibility types; and responses to the "What changes" AAR/AI question landed mainly in the How-To intelligibility type. Other research has shown each intelligibility type has its own advantages and disadvantages (e.g. [13,14]), so we see the diversity of perspectives that AAR/AI seemed to elicit as a particular strength of AAR/AI.

Result 2: The role of "active" learning

Traditional after-action reviews rest on active engagement by the assessors, and the same is needed with AAR/AI users. A key element of our strategy to promote users' active engagement in AAR/AI was to ask the user to predict the AI agent's action *before* the action—and before any explanation of that action. Our reason for this strategy was to engage the user in second-guessing the agent and thinking about what would be the right action, rather than to passively go along with whatever the agent did. We did not isolate this element of the interface to see its impact alone, but the data suggest that the AAR/AI interface did engage the participants.

One way this came out was in the self-explanations our participants verbalized. Creating a mental model is one kind of (human) learning, and one of the pillars of learning effectively is self-explaining [15]. One example of this phenomenon was the following participant's verbalization: *"I think the aim of the AI is to increase the number of minerals, and then go to the last one that is immortals, so that they can make a great damage to the nexus."*

Another way of thinking about participants' mental models, self-explanations, and learning is through Bloom's Taxonomy [16]. This taxonomy is a well-known set of levels of (human) learning mastery in education. Applying it to our participants' learning, understanding, self-explaining of the AI can be seen in Table 3. All participants achieved Level 5 at least once during the study, and all except one achieved level 6.

Result 3: High recall in finding "bugs"

The AAR/AI approach greatly helped domain-knowledgeable users find reasoning flaws made by AI agents. In our quantitative study on discovering these flaws, participants in the AAR/AI treatment reported a significantly larger number of problem reports than participants in the non-AAR/AI treatment (13.2 vs. 6.3, $p < .0001$). To investigate the validity of their problem reports, we used two metrics, recall and precision, commonly used in machine learning— recall measures how many of the ten bugs were correctly detected by the participants, and precision measures the proportion of problem reports that were actually bugs. Participants in the AAR/AI treatment achieved significantly higher recall *and* precision. Figure 7 illustrates these results.

FIGURE 7 Participants' recall (left) and precision (right). AAR/AI participants performed significantly better than than Non-AAR/AI participants with both measures.

6 │ CONCLUDING REMARKS

Our empirical journey ultimately led us to the realization that consuming AI's explanations is a learning activity for humans. Thus, rather than viewing explanations as objects to be delivered to humans, we have come to see XAI as a particular type of education—educating human users on how to accomplish some particular task involving understanding *this* AI agent.

This education-oriented view of XAI brings new insights from the results of the studies reported here. For example, it raises the question of which *learning goals* a set of explanations targets. In education, learning goals determine how to measure students' attainment of those goals. For XAI, an explanation system's learning goals can help empiricists choose *appropriate measures* of explanations' efficacy. For example, if the learning goal is to enable humans to interpret an AI, then an interpretation-oriented measure (e.g., Figure 2) may be more appropriate than a prediction-oriented measure (e.g., Figure 3).

Another insight derives from ways educators adjust their pedagogy for the particular content being taught, the particular audience being targeted, and for multiple forms of delivery needed to accommodate diverse humans' learning styles (e.g., reading, hearing, doing). This insight suggests that no particular explanation is likely to be "best" for all situations a particular AI encounters, or for all users trying to interpret that AI (consistent with Figure 3's results). The same insight also suggests that bringing elements of education's active learning techniques may greatly enhance humans' ability to acquire adequate mental models of how the AI works, as in our results on keeping the users active (Section 5's "Result 2") and in recent work on interactive visualization for deep learning education [17,18,19].

Perhaps most important, the education perspective highlights the importance of scaffolding explanations with a *process*, such as AAR/AI, by which humans can work through explanations. A corollary to this insight further suggests the importance of such processes in serving a diversity of learning styles and perspectives. AAR/AI's ability to do so, as illustrated in Table 2, may have been key to the high Bloom-taxonomy levels achieved by the AAR/AI participants and the significantly higher effectiveness of AAR/AI compared to non-AAR/AI (Figure 7). Together, these insights from an education perspective suggest that AI explanations' potential may be limited in XAI approaches that treat explanations as producible/consumable objects. A more promising future for XAI may lie in creating explanations as education experiences for diverse individuals.

ACKNOWLEDGMENTS

References

1. Anderson A, Dodge J, Sadarangani A, et al. Mental Models of Mere Mortals with Explanations of Reinforcement Learning. *ACM Transactions on Interactive Intelligent Systems* 2020; 10(2).

2. Dodge J, Khanna R, Irvine J, et al. After-Action Review for AI (AAR/AI). *ACM Transactions on Interactive Intelligent Systems* 2021 (to appear).

3. Lam KH, Lin Z, Irvine J, et al. Identifying Reasoning Flaws in Planning-Based RL Using Tree Explanations. In: IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI). ; 2021.

4. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016; 529(7587): 484.

5. Kulesza T, Burnett M, Wong W, Stumpf S. Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th ACM International Conference on Intelligent User Interfaces (IUI). ACM. ; 2015: 126–137.

6. Mai T, Khanna R, Dodge J, et al. Keeping it "organized and logical" after-action review for AI (AAR/AI). In: Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI). ACM. ; 2020: 465–476.

7. Morrison JE, Meliza LL. Foundations of the After Action Review Process. tech. rep., Institute for Defense Analyses; : 1999.

8. Sawyer TL, Deering S. Adaptation of the US Army's After-Action Review for Simulation Debriefing in Healthcare. *Simulation in Healthcare* 2013; 8(6): 388–397. doi: 10.1097/SIH.0b013e31829ac85c

9. Lloyd Baird SD. Learning from action: Imbedding more learning into the performance fast enough to make a difference. *Organizational Dynamics* 1999; 27: 19-32. doi: https://doi.org/10.1016/S0090-2616(99)90027-X

10. Ishak A, Williams E. Slides in the Tray: How Fire Crews Enable Members to Borrow Experiences. *Small Group Research* 2017; 48(3): 336–364. doi: 10.1177/1046496417697148

11. Dikkala R, Khanna R, Matthews C, et al. Doing Remote Controlled Studies with Humans: Tales from the COVID Trenches. In: ACM/IEEE 14th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE). ; 2021.

12. Lim B, Dey A, Avrahami D. *Why* and *Why Not* Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In: Proceedings of the 2009 SIGCHI Conference on Human Factors in Computing Systems. ACM. ; 2009: 2119–2128.

13. Lim BY. *Improving understanding and trust with intelligibility in context-aware applications*. PhD thesis. Carnegie Mellon University, ; 2012.

14. Cotter K, Cho J, Rader E. Explaining the News Feed Algorithm: An Analysis of the "News Feed FYI" Blog. In: ACM CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM. ; 2017: 1553–1560.

15. Chi MT, Bassok M, Lewis MW, Reimann P, Glaser R. Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science* 1989; 13(2): 145–182.

16. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. *Taxonomy of Educational Objectives*. London, England: Longmans, Green and Co LTD . 1956.

17. Kahng M, Thorat N, Chau DH, Viégas FB, Wattenberg M. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics* 2019; 25(1): 310–320.

18. Kahng M, Chau DH. How Does Visualization Help People Learn Deep Learning? Evaluating GAN Lab with Observational Study and Log Analysis. In: Proceedings of the 2020 IEEE Visualization Conference (VIS). IEEE. ; 2020: 266–270.

19. Wang ZJ, Turko R, Shaikh O, et al. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 2021; 27(2): 1396–1406.