After-Action Review for AI (AAR/AI)

JONATHAN DODGE, ROLI KHANNA, JED IRVINE, KIN-HO LAM, THERESA MAI, ZHENGXIAN LIN, NICHOLAS KIDDLE, EVAN NEWMAN, ANDREW ANDERSON, SAI RAJA, CALEB MATTHEWS, CHRISTOPHER PERDRIAU, MARGARET BURNETT, and ALAN FERN, Oregon State University, USA

Explainable AI (XAI) is growing in importance as AI pervades modern society, but few have studied how XAI can directly support people trying to *assess* an AI agent. Without a rigorous process, people may approach assessment in ad hoc ways—leading to the possibility of wide variations in assessment of the same agent due only to variations in their processes. AAR, or After-Action Review, is a method some military organizations use to assess human agents, and it has been validated in many domains. Drawing upon this strategy, we derived an AAR for AI, to organize ways people assess reinforcement learning (RL) agents in a sequential decision-making environment. We then investigated what AAR/AI brought to human assessors in two qualitative studies[1]. The first investigated AAR/AI to gather formative information, and the second built upon the results, and also varied the type of explanation (model-free vs. model-based) used in the AAR/AI process. Among the results were: (1) participants reporting that AAR/AI helped to *organize their thoughts* and *think logically* about the agent; (2) AAR/AI encouraged participants to reason about the agent from a *wide range of perspectives*; and (3) participants were able to leverage AAR/AI with the model-based explanations to *falsify* the agent's predictions.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Explainable AI, After-Action Review

ACM Reference Format:
Jonathan Dodge, Roli Khanna, Jed Irvine, Kin-Ho Lam, Theresa Mai, Zhengxian Lin, Nicholas Kiddle, Evan Newman, Andrew Anderson, Sai Raja, Caleb Matthews, Christopher Perdriau, Margaret Burnett, and Alan Fern. 2021. After-Action Review for AI (AAR/AI). *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 1 (January 2021), 33 pages. https://doi.org/10.1145/3453173

(accepted, to appear)

1 INTRODUCTION

By design, AI systems perform decision-making on behalf of a human user. This means that in safety-critical applications such as self-driving cars, vendors may take on additional liability when things go wrong. Failures may have such grave consequences that they are likely to wind up in court [11]. Was the accident caused by the driver not reacting in time, or a defective AI? [48]. How can AI stakeholders best determine that an AI system is safe and regulation compliant?

[1]This paper is a revised and expanded version of [43].

Authors' address: Jonathan Dodge, dodgej@oregonstate.edu; Roli Khanna, khannaro@oregonstate.edu; Jed Irvine, irvine@oregonstate.edu; Kin-Ho Lam, lamki@oregonstate.edu; Theresa Mai, maithe@oregonstate.edu; Zhengxian Lin, linzhe@oregonstate.edu; Nicholas Kiddle, kiddlen@oregonstate.edu; Evan Newman, newmanev@oregonstate.edu; Andrew Anderson, anderan2@oregonstate.edu; Sai Raja, rajasa@oregonstate.edu; Caleb Matthews, mattheca@oregonstate.edu; Christopher Perdriau, perdriac@oregonstate.edu; Margaret Burnett, burnett@oregonstate.edu; Alan Fern, Oregon State University, Corvallis, OR, USA, afern@oregonstate.edu.

2160-6455/2021/1-ART1 $15.00
https://doi.org/10.1145/3453173

Given that intelligent agents interact with the world in ways analogous to those of human agents, could established techniques for evaluating the quality of human performance be applied to an AI? In this paper, we investigate this approach by adapting a technique called After-Action Review (AAR) for use with AI. AAR was devised by the U.S. Army in the mid-70's [46], and has been a success in various branches of the military. It has also been adapted for other domains including medical treatments [61], transportation services [42], and fire-fighting [30]. Apparently, most of these adaptations have proven successful, as a recent meta-analysis of 61 studies reporting effect size for use of AAR found a moderate practical effect overall [32].

We term our adaptation AAR/AI (pronounced "arf-eye", short for "AAR for AI"). 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.

To investigate AAR/AI in the hands of human users, we set it in the context of a real-time strategy (RTS) game as the sequential system. We created a custom game in StarCraft II (Section 4.1). Then, we created a reinforcement learning (RL) agent that yielded high-quality actions in the domain (Section 4.2). For this agent, we also devised two types of explanations of the agent's actions (Section 3.4)—one type was a Model-Free explanation and the other was a Model-Based explanation—so as to observe AAR/AI with two types of agents.

Model-Free and Model-Based agents work differently, yielding different possibilities for explanation. Model-Free agents simply compute a value for each considered action and then select the maximum. In contrast, Model-Based agents expand a search tree as they perform action selection. The root of the tree reflects the current state of the system. The agent considers many actions by predicting the state transition each action will cause. The process is then repeated using each predicted state as a starting state. Model-Based agents offer a richer space for explanation because the action and state information available in the tree is human-interpretable.

To improve experimental control, we needed Model-Free and Model-Based explanations which select the same actions. To accomplish this, we used the same Model-Based agent for both, so that they encoded the same policy, then heavily pruned the Model-Based explanation tree to form the Model-Free one, only exposing the information that an Model-Free system would have.

AlphaZero [64] is a classic example of a Model-Based system. It uses MCTS to expand a game search tree—the agent uses its model of the game rules to recursively predict subsequent states as part of the decision process (given the current state and potential agent actions as input). Model-Free agents too can be applied in domains which are sequential, but are more common in domains which are not, such as image classification (e.g. VGG-19, illustrated by [25]'s Figure 3).

To investigate AAR/AI in the context of these explanations, we conducted two qualitative studies. Study One employed a one-on-one in-lab think-aloud design with paper prototypes, and focused primarily on the process. Using our Study One results, we implemented an interactive prototype, and ran Study Two. Study Two allowed us to both triangulate with our preliminary results, and to consider AAR/AI with two types of explanations: a Model-Free explanation and a Model-Based explanation.

With our two studies, we investigated the following research questions:

RQ1 (Study One) To what extent are participants able to make sense of and learn from our explanations while using AAR/AI for assessment?

RQ2 (Studies One and Two) Which actions should be included in search tree explanations? How do these design choices affect user interaction patterns?

RQ3 (Studies One and Two) How did the aspects of theory present in our explanations affect participants' ability to make explanation-informed statements?

RQ4 (Study Two) How did differences in how participants engaged with our explanations affect their cognitive load?

2 BACKGROUND & RELATED WORK

There are many papers describing the challenges of evaluating AI systems' quality (e.g. [9, 22]), including specific attacks (e.g. [18]). Rising to meet these challenges, approaches like DeepTest [71] attempt to utilize concepts from software engineering to improve testing of deep neural networks. In particular, they seek to measure and improve "neuron coverage" (proposed by Pei et al. [52], similar to code coverage). To accomplish this, they apply a series of transformations to the input, a form of data augmentation conceptually similar to fuzzing. However, these approaches are *system*-oriented in terms of exposing problems, not *human*-oriented by giving an assessor the tools to determine appropriate use for the AI.

2.1 People Analyzing AI

Human-oriented evaluation of AI is an active area of research, though much of it is at a different granularity than we needed. For example, Lim et al. researched how their participants sought information in context-aware systems powered by decision trees. The result of their research was a code set of several "intelligibility types" describing the information. They discovered that their participants demanded *Why* and *Why Not* information, especially when the system behaved unexpectedly [40]. Using Lim's code set, Penney et al. studied how experienced RTS players looked for information when understanding and evaluating an "AI," but they found that participants preferred *What* information over *Why* information and that the large action space of StarCraft II led to high navigation costs, which meant missing important game events [53]. Dodge et al. analyzed how shoutcasters (human expert explainers, like sports commentators) assessed competitive StarCraft II players. They showed the ways that shoutcasters present information that they thought their human audiences needed [16]. Kim et al. gathered 20 experienced StarCraft II players to play against competition bots and rank them based on performance criteria. They noted how human evaluations of the AI bots differ from the evaluations used for AI competitions and that the human player's ability plays a huge role in their evaluations of the AI's overall performance and human-likeness [34]. These studies found how people evaluate an AI, but they did not present a *structured process* for assessment.

There are several models which consider system assessment in a human-oriented way; however, these works do not provide an assessment process for AI, but rather on whether humans will *adopt* systems or not. One such framework is Technology Acceptance Modeling (TAM) [14]. TAM can predict how well a system will be accepted by a user group and explain differences between individuals or subgroups. More recently, the UTAUT (Unified Theory of Acceptance and Use of Technology) model was proposed as an acceptance evaluation model [26]. Carrying on this spirit, recently researchers have produced a spate of publications based on need-finding or perception interviews meant to identify barriers to adoption (e.g. [5, 8, 28, 80]). While these techniques can assist in assessment, they do not offer a concrete *process* for human assessors to enact.

More recently, a few researchers (e.g. [1, 76]) have made first forays into guidelines that can be used to assess explanations without the user-in-context[2] required by adoption models. Yang et al., identified two main challenges that may explain the rarity of this kind of fundamental AI usability research, *"uncertainty surrounding AI's capabilities… [and]... AI's output complexity."* [81]. Similarly

[2]Hoffman et al. [27] describe two kinds of AI evaluation, based on "Satisfaction" (roughly speaking, 'does it help a user complete a task?'), and "Goodness" (roughly speaking, 'does a panel of experts think it is good?').

to adoption models, guidelines can assist in explanation assessment, but do not offer a concrete process to follow.

2.2 People Explaining AI

The primary purpose of explanations is their ability to improve the mental models of the AI systems' users. Mental models are *"internal representations that people build based on their experiences in the real world"* that assist users in predicting system behavior [47].

Devising explanations that actually lead to better mental models is an active area of research. One such example is Kulesza et al.'s proposed principles for explaining (in a "white box" way) machine learning (ML) based systems, wherein the system made its predictions more transparent to the user [37], which in turn improved the quality of their participants' mental models. Another study by Anderson et al. [2, 3] provided insights into the variability of changes in the mental models of participants with different explanation strategies of an AI agent. One promising explanation strategy is to manage users' expectations. For example, Kocielnik et al. found that interventions, like adding a gauge, helped participants estimate the system's accuracy [36].

Another direct consequence of altering the mental models of users is the improvement in their ability to control the system. According to a study by Kulesza et al. [38], participants with the most improved mental models were able to customize the system's recommendations best. Roy et al. found that participants preferred high controllability, even in low accuracy settings [58]. Wang et al. set up an accuracy-control tradeoff explicitly in their auto ML system, allowing users to search longer for higher accuracy, or adjust the search constraints for higher control [77]. Another kind of tradeoff, posed by Smith-Renner et al., reports that the systems *adhering* to the user input more often can increase *instability* with respect to other changes that occur when the model updates to incorporate that input [66], a problem also reported by Stumpf et al. [68]. Still, the preference for controllable systems seems to hold even when the controls *do not work* [73].

Explanations in the domain of AI agents in RTS games have been gaining traction in recent years. In a study by Metoyer et al. [44], they present a format where experienced gamers played while providing explanations to non-RTS players, finding that one key to the explanation process was the manner in which expert players communicated while demonstrating how to play. The study by Kim et al. [35] had experienced gamers play against AI bots in order to assess the bot's skill levels and overall performance. However, despite the research mentioned above, there is a dearth of literature concerning what humans really *need* in order to understand and assess such systems [50].

2.3 After-Action Review

To structure our assessment method, we turned to processes that have been used for humans to assess *other humans*, including Post-Control, Post-Project Appraisal and After-Action Review (AAR) [62]. Our criteria for the process to use as our basis included: (1) have a structured and logical flow, (2) be well established, and (3) be suitable for evaluation *during* a task, not just useful at the end of a task. We selected the AAR method as the one that best fulfilled these criteria.

AAR is a debriefing method created by the United States Army, and it has been used by military and civilian organizations for decades [60], to encourage objectivity [42]. The purpose is to understand what happened in a situation and provide feedback, so people can meet or exceed their performance standards by going through the structured series of steps shown in Table 1.

The AAR process was primarily used as a method to provide performance feedback after soldier training sessions. Before starting an evaluation session, the leader (a designated individual across all sessions) performs groundwork to collect and aggregate data from the training session for further analysis. The leader enters that session with a pre-planned mechanism to collect data. The

evaluation session begins by reiterating the objectives of the analyzed exercise. From there, the leader asks a series of open-ended and leading questions about what happened during the training session, making sure to encourage a diverse range of perspectives. These responses are then filtered into a recapitulation that the group collectively agrees on, and the discussion is shifted to any shortcomings in performance. This is followed by brainstorming solutions to avoid or improve responses to problematic outcomes. The evaluation concludes by delineating an action plan to adhere to for future training [72].

AAR showed effectiveness for combat training centers [60], and the military still uses it, with a recent investigation of current methodologies for simulation-based training [23]. Outside military applications, AAR has been used in other domains, from medical treatment [55, 61], emergency preparedness [13], and emergency response [30, 39]. The closest research to ours discusses how AAR will be different for manned-unmanned team compositions, but focused on the technologies needed to support the AAR process, not the process itself [7].

3 THE AAR/AI PROCESS

Our After-Action Review for AI (AAR/AI) is an assessment method for a human assessor to judge an AI. We base the steps of our method from Sawyer et al's DEBRIEF adaptation from the Army's AAR [61]. In their adaptation, they Define rules, Explain objectives, Benchmark performance, Review what was supposed to happen, Identify what happened, Examine why, and Formalize learning. Table 2 outlines our AAR/AI adaptation.

The original AAR method is a facilitated, team-based approach, but our AAR/AI method is for an individual that is reviewing, learning the AI's behavior, and assessing its suitability [62]. The outcomes are different for the approaches: AAR aims for transfer of knowledge within a team, and AAR/AI aims for individual acquisition of knowledge and assessment of an AI. These two primary differences between AAR and AAR/AI are what generated the specific ways AAR/AI (Table 2's columns 2 and 3) carries out the original method's steps (Table 2's column 1).

3.1 AAR/AI: Defining Rules & Objectives

A facilitator starts each session with a tutorial on the user interface, domain, explanations, and the objectives of the assessment (Steps 1-2, Table 2). This contextualizes the discussion in terms of what the assessor is supposed to do and the objectives of agent that they are assessing. After that, the facilitator begins the AAR/AI "inner loop" (discussed next), and after every loop is done the assessor completes a questionnaire.

US Army AAR Process
Introduction and rules.
Review of training objectives.
Commander's mission and intent (what was supposed to happen).
Opposing force commander's mission and intent (when appropriate).
Relevant doctrine and tactics, techniques, and procedures (TTPs).
Summary of recent events (what happened).
Discussion of key issues (why it happened and how to improve).
Discussion of optional issues.
Discussion of force protection issues (discussed throughout).
Closing comments (summary).

Table 1. Steps of the US Army AAR process [72].

3.2 AAR/AI's Inner-Loop: What, Why, How

During each iteration of the inner loop, the facilitator asks the assessor, "*What was supposed to happen?*", "*What happened?*", "*Why did it happen?*", and "*How can it be improved?*" (Steps 3-6, Table 2). The assessor also summarizes what happened in the past three rounds and writes down anything they observed that was good, bad, or interesting on an index card. At Step 5, we provided the assessor with the AI's explanation for the most recent round, and asked them to explain why the AI did the things it did, according to the process in Table 2. Following this, to formalize learning about this particular decision, the facilitator asks the assessor the questions listed in Table 2 step 6, (e.g. whether they would allow the AI to make these decisions on their behalf). Thus ends the inner loop, which would repeat until the end of that analysis session.

3.3 AAR/AI's Artifacts

Part of AAR/AI involves creating materials to help keep everyone on task during the assessment. The US Army AAR uses cards in order to log observations [72], though the information collected is

	AAR Steps	AAR/AI ?s Answered	AAR/AI Empirical Context
	1. Define the rules	How are we going to do this evaluation? What are the details regarding the situation?	We established the rules of evaluation and the domain (see Supplemental Materials).
	2. Explain the agent's objectives	What is the AI's objective or objectives for this situation?	We explained the AI's objectives for the situation (see Supplemental Materials).
AAR/AI Inner Loop	3. Review what was supposed to happen	What did the evaluator intend to happen?	We asked, *"What do you think should happen in the next round(s)?"*.
	4. Identify what happened	What actually happened?	The participant watched the required number of rounds. Then, we asked, *"Could you briefly explain what actually happened in these past rounds?"*.
	5. Examine why it happened	Why did things happen the way they did?	We asked, *"Why do you think the rounds happened the way they did?"*. Next, the participant summarized anything good, bad, or interesting on an index card. Last, we provided the participant the agent's explanation.
	6. Formalize learning (end inner loop)	Would the evaluator allow the AI to make these decisions on their behalf? What changes would they make in the decisions made by the AI to improve it?	We asked three questions: *"Would you allow the AI to make these decisions on your behalf?"*, *"What changes would you make in the decisions made by the AI to improve it?"*, *"Would you allow the Friendly AI to make this category of decisions on your behalf?"*.
	7. Formalize learning	What went well, what did not go well, and what could be done differently next time?	The participant completed a post-task questionnaire (see Supplemental Materials).

Table 2. How AAR/AI (right two columns) adapts the original After-Action Review debrief steps (left column). The "Empirical Context" column explains how we realized it in Study One. (Study Two's realization of AAR/AI was almost the same, except that we shortened Step 6 to just the "What Changes" question.) Note that steps 3-6 form an "inner loop" that we repeated every three decisions. The parts outside the inner loop are documented in our Supplemental Materials (tutorials, questionnaires, etc), so we describe them only briefly.

largely focused on personnel and their positioning. Since the AI performs within the RTS domain, we turned to how professional shoutcasters analyze AI, like AlphaStar [67]. They used formatted text for actions that they found "good," "bad," or "interesting," which we replicated in the AAR/AI's index cards. This prevents assessors, regardless of the AI's use, from relying on memorizing when a decision is good or not. By using such written artifacts, the AAR/AI process has the benefit of gaining retrospective feedback on the process itself or the explanations used in it. Further, artifacts like these can assist in comparing the assessment results from multiple individuals or be released with the system as a means to document the kind of validation conducted and the results from it, akin to Model Cards [45].

3.4 AAR/AI: Explanation Component

AAR/AI evaluators, like the AAR equivalent, require information on what happened, so our process requires an embedded Explanations Component, since the evaluators not only must they know *what* happened, but the agent must be able to explain *why* it performed an action. In both studies, we used a model-based agent to enable a model-based explanation.

A model-based agent (and its explanation) offers the benefit of explicitly representing its emerging model of the world, such as the future states the agent is trying to reach or avoid—in essence, an underlying rationale for its decisions[3]. For example, consider Model-Based agents that expand a search tree as they select actions, allowing them to fill in an explanation template [15] like the one shown in Figure 1b. On the other hand, Model-Free agents can only fill in a more limited explanation template, illustrated in Figure 1a. In particular, Model-Free explanations do not expand a search tree—instead more of a search stump—by attaching to each action only a single number for its value.

We therefore prototyped a Model-Based explanation for Study One, capturing a portion of the agent's search tree. That explanation is shown in Appendix B, and here we focus on the revised version of our Model-Based explanation, used in Study Two, shown in Figure 2. We described the search tree to participants as, *"...a diagram of decisions, where the Friendly AI decides what actions or decisions it must take to complete a round in the game."*

The explanation lays out the agent's "explanatory theory" [65] of how the game could play out in different situations. In essence, the theory's "constructs" of that theory are: game states, roles (e.g. friends or enemies), actions available to various roles, and (estimated) values of different states and actions.

In AAR/AI then, the evaluator's central mission is to evaluate one aspect of the AI agent's theory: its falsifiability [54]. To carry out this mission, the evaluator answers the AAR/AI questions (e.g., What just happened? Why? ...) by gathering information from a combination of game behavior and the explanation's diagram of actions and states among which the agent is deliberating (Figure 1).

4 METHODOLOGY SHARED BY STUDY ONE AND STUDY TWO

To inform our design of AAR/AI, we ran two in-lab studies: Study One, a one-on-one think-aloud qualitative study and Study Two, a two-treatment qualitative study run in small groups. The main goal of Study One was to formatively investigate participants' sensemaking attempts when doing AI assessment and how AAR/AI came together with those attempts. Additionally, since the AAR/AI process embeds an explanation, we designed both studies to include investigating the explanation strategy in the context of the process.

[3]To compensate for model-free agents' lack of underlying rationale, one body of research attempts to generate approximations of an underlying rationale, e.g., [17].

Study Two moved beyond the sensemaking goal of Study One to gain insights into how the explanations might help humans with *failure detection* or *fault localization*. To illustrate, Study One featured *only* the Model-Based explanation strategy in paper prototype form, whereas Study Two's treatments included software implementations of both the Model-Based explanation and the (simulated) Model-Free (detailed later in Section 4.2). Notionally, since Model-Free agents do not expand a search tree with any depth and can be thought of as a search stump, Model-Free explanations are limited to fewer interactions than Model-Based (as illustrated in Figure 1).

Both studies used the same domain and agent implementation, which we describe next.

4.1 The Domain

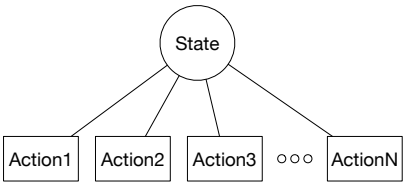
StarCraft II is a popular Real-Time Strategy (RTS) game that offers hooks for AI development ([74, 75]) and a flexible engine for map creation[4]. Using this engine, we built a custom game called Tug of War, shown in Figure 3. The objective of the game is to destroy either of the opponent's Nexus in the top lane or bottom lane. If no Nexus is destroyed after 40 rounds (or decision points, which we denote as DPs), the player whose Nexus has the lowest health will lose.

At every DP of the game:

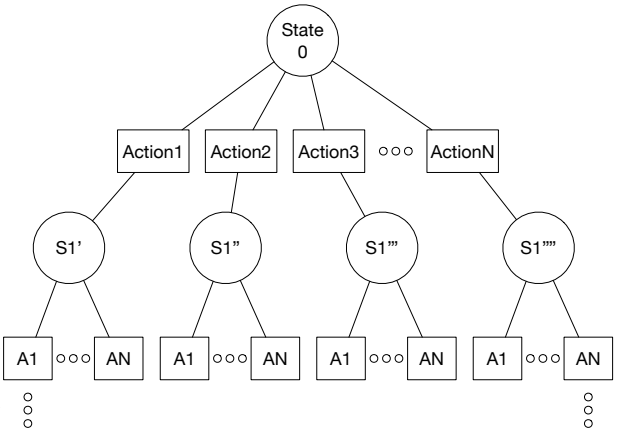
- Each player receives income (100 minerals, +75 per pylon)
- The player chooses to build any combination of unit production facilities (i.e. barracks) to be added for the next DP, subject to the following constraints:

(1) Total cost cannot exceed current mineral count
(2) Players are only allowed to build in *one* lane at a time

[4]Many map creation resources are available at places such as [70].



(a) Template for a Model-Free explanation, notionally showing its "search stump," because it has no depth—just values associated with taking an action in a state. Here, since we are showing the template, we have simplified away details like state/action representation and presentation (addressed in Appendix B).



(b) Template for a Model-Based explanation, notionally showing the full search tree. In general, these search trees do not reach terminal nodes, and when that is the case, they must use a heuristic function or Model-Free system to evaluate the quality of that state.

Fig. 1. A comparison of explanation templates which can be filled out by Model-Free and Model-Based explanations, respectively. Note that if one imagines starting with just a root node and the best action, and iteratively revealing the tree via interaction, Model-Free explanations would only need to "widen" the tree, while Model-Based explanations also support "deepening". We will return to this in Section 8.1.

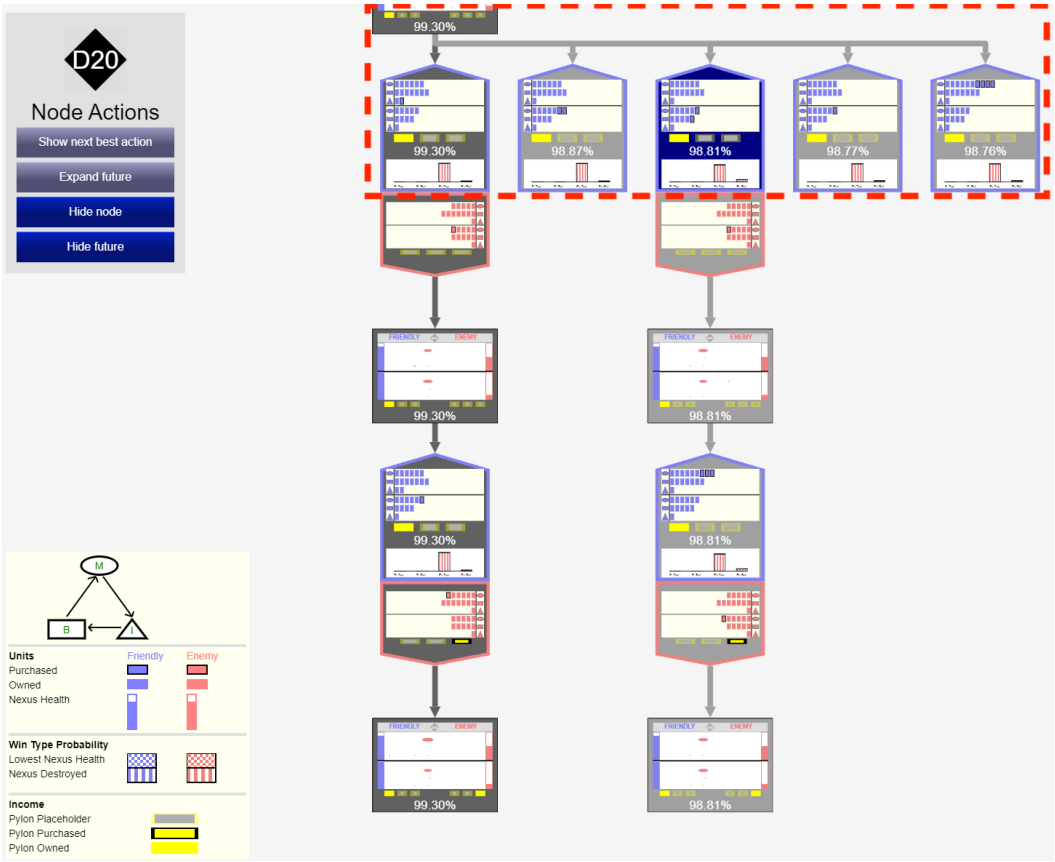


Fig. 2. Interactive Model-Based explanation for DP 20 in Study Two, as observed by participant S1MB20. The Model-Based explanations, shown above, starts at the top with the current state node. Next are the top five predicted friendly agent actions considered, each followed by the enemy agent action that is predicted to be most effective. Next down are the predicted states that are consequent to those actions. The cycle is then repeated again. We refer to a fully-rendered prediction trajectory as a "future". The principal variation, the best predicted future, is at left, with actions decreasing in estimated value to the right. Participants could choose to adjust visibility of individual nodes or future trajectories of particular actions by selecting nodes and clicking buttons on the Node Actions menu. The legend at bottom left reminded participants of the meaning of each rendered state and action detail, as well as showing the rock-paper-scissors relationship among unit types. We simulated Model-Free explanations (dashed red box) by featuring only the root node and the friendly agent action nodes directly below, essentially truncating the depth of the tree. Even though participants in both treatments were able to increase or adjust the *width* of the tree, the Model-Free explanations were essentially different in that they provided no information beneath the friendly agent action node. Since Model-Based explanations included the tree at greater depth, they allowed participants to expand the width of the tree at any internal node—as opposed to just the top level actions.

(3) Players do not know the opponent's action until both actions are finalized

- Players spawn units equal to the total number of unit production facilities currently held (i.e., 5 barracks $\implies$ 5 marines)

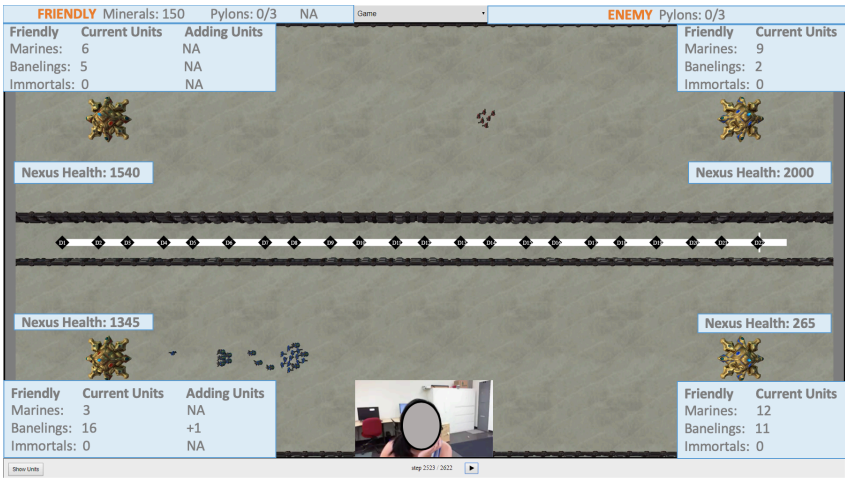


Fig. 3. Game screen at decision point 22 during Study One. Note the text boxes offering state information (current units, nexus health, etc) as well as action information (adding units). The evaluation interface primarily adds a time slider (shown in the middle of the screen with a diamond for each DP) and the blue overlays to increase visibility of fonts presenting information available in the in-game interface.

At each DP, both players choose which lane to build in and the number of unit-producing buildings to spend resources on for each of 3 unit types which share a rock-paper-scissors relationship. **Marines** (50 minerals) are low health units that attack in small quick shots. They are effective against immortals. **Banelings** (75 minerals) are medium health units that attack by exploding on contact. Banelings are effective against marines. Lastly, **Immortals** (200 minerals) are high health units that attack in large slow shots. Immortals can inflict significant damage on a Nexus. Players may also choose to build a pylon to increase their income. The maximum number of pylons they can build is 3, and the cost of a pylon increases each time one is purchased. Note that an action in this context is essentially an integer vector representing the intended purchase of unit-producing buildings and/or pylons, meaning the branching factor is combinatorial with respect to minerals possessed.

Once units spawn, the players can no longer control them; they will move toward the enemy Nexus and attack any enemies along the way. Also, units *always* spawn at the same location each wave.

4.2 The Agent Implementation

We have pointed out that the agent used for both studies is Model-Based, meaning it has access to a transition function that maps a state-action tuple to the successive state. Applying the transition function allows the agent to expand a search tree, and perform minimax search on it[5] on it. The system uses three learned components (all represented by neural networks): the transition model, the heuristic evaluation performed at leaf nodes, and the action ranking at the top level.

The heuristic evaluation function estimates the value, or quality, of non-terminal leaf nodes in the search tree. This function is necessary to address the depth of the full game tree, since the search will rarely be able to expand the tree until all leaf nodes are terminals. The action ranking function provides a fast estimate of the value associated with taking each action in a state. This function is

[5]For more information on game tree search, see Russell and Norvig, Chapter 5 [59].

necessary to address the large action-branching factor by only performing the more expensive tree expansion under some number of top-ranked actions to improve estimates (similar to AlphaGo and AlphaZero [63, 64]). A big difference, however, is that our system uses a learned transition model, due to the stochastic and complex nature of the transitions between states; whereas Silver et al.'s used a perfect move-transition model (e.g., Chess's deterministic rules).

We actually used the same Model-Based agent in both treatments of Study One and Study Two, simulating the Model-Free treatment's agent for the Model-Free treatment. We simulated the Model-Free explanation simply by withholding the Model-Based agent's learned "model of the world," which amounted to less completeness, excluding information past the value associated with the actions (illustrated in Figures 1 and 2). This design enabled scientific control, with the Model-Free and Model-Based agent choosing from (and selecting) the same actions given the same state. This level of control would have been extremely difficult to ensure via independent training processes.

5 METHODOLOGY SPECIFIC TO STUDY ONE

For Study One, we recruited 11 students at Oregon State University who had not taken classes in AI/ML or participated in our previous studies. Since our game is based on StarCraft II, we recruited those familiar with real-time strategy games, to ensure that participants could understand the game sufficiently to assess the AI.

A researcher served as facilitator with one participant (assessor) during the AAR/AI process, starting with a tutorial on the interface, domain, and task (covering AAR/AI Steps 1/2). Since each session was limited to 2 hours, we wanted to ensure that each participant reached the end of the replay with time for our post-task questionnaire. Thus, we decided to have them analyze every third decision point of the 22 available, including the last one (e.g. 3,6,...,21,22). This allowed up to 5-7 minutes for each iteration of the AAR/AI inner loop—though it was rarely necessary to enforce limits during the studies. We chose to sample these decisions because timing our pilot participants revealed that we would not have time to cover all of them, and we wanted the participants to see the full evolution of a game.

At each iteration, the researcher asked the assessor a structured series of open-ended questions to elicit their thoughts as they performed their assessment of the AI's actions (Steps 3-6). Additionally, the participant wrote on index cards (Section 3.3) to help them formalize thoughts and offer the option to refer back to previous notes.

Upon completion of the task (Step 7), we asked: *"Did the process of the questions I asked you help you understand and assess the AI better?"*, *"Do you think the AI's diagrams have enough detail?"*, *"Would you prefer the width of the diagram to be narrower or wider? Or do you like the way it is?"*, *"What kind of actions would you have liked to see on the diagram?"*, and *"In the main task, did you find these cards useful?"*. Finally, we compensated participants $20.

Each session spent ~30 minutes for the briefing/tutorial (pre-task), ~50 minutes on the inner-loop (the main-task), and ~25 minutes on the post-task questionnaire. This timing was consistent with Sawyer et al.'s recommendations (25/50/25%, respectively) [61].

5.1 Analysis Methods

To answer **RQ1**, we drew from a code set that Dodge et al. used in their StarCraft II study, which had been adapted from Lim et al.'s work [16, 41]. Dodge et al. also added a "judgment" code, which the AAR/AI needed because of the nature of assessment. Individually, the two researchers coded 20% of the data corpus, achieving an inter-rater reliability (IRR) of 76.4%, computed via Jaccard Index [31]. Given this level of reliability, they then split up the remaining coding.

To answer **RQ2** and **RQ3**, two researchers applied content analysis [29] to the coded statements from the post-task questions about helpful or problematic elements of the process or explanations, resulting in the code set in Appendix A, Table 9. The two researchers coded 21% of the data corpus separately, achieving inter-rater reliability (IRR) of 82.4% (Jaccard). Given this level of reliability, they then split up the remaining coding.

We enumerate these code sets in the context of the relevant results sections.

6 RESULTS: STUDY ONE

We begin with participants' sensemaking attempts when they were using AAR/AI and our explanations, deferring Study One results that intertwine with Study Two results to Section 8.

6.1 Using AAR/AI to Learn

The goal of our project was to enable participants to understand how the AI agent is "thinking" well enough to evaluate how suitable the agent is for different situations that arise—which involves people building mental models of the AI agent. In this subsection, we consider what the AAR/AI process brought to our Study One participants' mental-model building.

In a post-study questionnaire, we directly asked Study One participants what was helpful about AAR/AI and what was not. In their responses, many of the Study One participants commented on how AAR/AI's "structuredness" helped their understanding by keeping their thinking organized, structured, and/or logical. (Only a single Study One participant said it was not helpful, but this was because they believed that with their experience in RTS games, they already understood the AI's behavior without the need of any assistance.) For example:

S1MB8: *"Uh, yes, I would say <AAR/AI was helpful>. It definitely directed me towards what I should be paying attention to."*
S1MB18: *"I could think what it should improve on and why the previous round happened the way it did. So, when those questions were broken down... Really helped in following the game."*
S1MB14: *"...it categorized the flow of logic that we should've had in analyzing the prediction and what actually happened, so it kept it more organized, and therefore, more logical."*
S1MB17: *"I know it was too much information ... it helped me understand it better. ...it just helps me ... to understand it better, and makes it more logical."*

To understand the level of our Study One participants' mastery of understanding the agent, we applied Bloom's Taxonomy [6], which is a framework used by educators to categorize the different levels of learning. The taxonomy has six levels [4], ranging from basic understanding of a concept (level 1), through a fairly advanced understanding (level 6). Each level requires learners to engage with a higher level of abstraction than the last. The application of Bloom's taxonomy to our context is detailed in Table 3.

As Table 3 shows, subsets of Study One participants showed mastery of every Bloom's level. In fact, each of these participants achieved Bloom's Level 5 at least once during the study. Further, all except one of them achieved Bloom's Level 6 at some point.

Bloom's Level 5 is of particular interest to our project: it is the level of understanding that allows evaluation. Evaluation is precisely the level of understanding needed for assessing an AI.

In considering how the participants who reached Bloom's Level 5 managed to do so, we turned to the Lim-Dey intelligibility types, which we used as a codeset for our qualitative coding (Table 4). As the results show, each of AAR/AI steps guided participants' thinking (according to their self-reports) toward different Lim/Dey perspectives [40]. The wording of the AAR/AI questions compared with the Lim/Dey type names may explain some of this result.

For example, the AAR/AI question in the top row of Table 4, "What ...should happen," guided most participants to focus on "What Could Happen"—an almost syntactic match between the

Level: [6]'s Description	**How it applies to understanding the AI**	**Examples from our participants**
1. **Remembering**: Have students acquired the ability to correctly recall information?	Participants *recall* domain information, such as game rule(s), what an agent can do with particular game units, etc. (Supported by AAR/AI's questions about the game.)	+S1MB20: *"It'd probably buy another baneling... to counter the marines..."*
2. **Understanding**: Can students understand information they have learned to recall?	Participants *understand* the domain information provided. (Supported by AAR/AI's "What" and first "Why" question.)	+S1MB8: *"...you <the AI> don't necessarily know which lane they're coming through... it's not much of an informed decision until the first round happens."*
3. **Applying**: Can students apply their newly learned knowledge?	Participants *apply* the explanation of the AI to the game. (Supported by second "Why" question.)	+S1MB2: *"I ...like it how <the explanation diagram> is, because like I could try to draw my own conclusions from it rather than just like 'oh this is just what happened'."*
4. **Analyzing**: Can students see patterns and make inferences about a problem?	Participants *analyze* the AI's problems in the game, and reason about solving them. (Supported by the prediction task and the "What changes would you make" question.)	+S1MB2: *"So the bottom one did pretty well like overpowering the enemy AI and even attacking nexus, lowering its health while the top one, the enemy AI did a better job sending more marines and the friendly AI sent banelings which got overpowered by the marines."* +S1MB19: *"So we have almost same health on top and bottom. So, to defeat us, they have to focus on either one. So I guess they will focus bottom, because they have to save them at the time. I guess we have to use minerals to buy immortal here, so that we can save ourselves and at the same time, kill the enemy."*
5. **Evaluating**: Can students take a stand or decision, and justify it?	Participants *evaluate* the AI agent, and judge if they would allow the agent to make decisions on their behalf in this or similar situations. (Supported by the "Would you allow..." question series.)	+S1MB5: *"Producing these banelings <in both> lanes allowed nexus damage bottom lane, and then having the one or two marines do consistent damage on the nexus really took down the nexus health, so that was actually a really good decision."* +S1MB20: *"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, and only increasing their baneling count, which only works at melee range which is ineffective to marines if there's already a baneling wall in front of them."*
6. **Creating**: Can students create a new point of view?	Participants *create* new points of view by generalizing upon, abstracting above, or recommending differences in the AI's behaviors.	+S1MB14: *"Well, the enemies will invest in banelings, and I feel that the friendly's will invest in marines, especially more in the top row, since it is more damage..."* +S1MB21: *"I would consistently save a small quantity of minerals each round, rather than trying to save them all in a single round."*

Table 3. Bloom's taxonomy levels Study One participants achieved in learning the agent's behavior.

AAR/AI question and that Lim/Dey type. The "Why...did" AAR/AI question (fourth row) also featured a strong syntactic match with the Lim/Dey "Why did" type. While not a near-syntax match, the AAR/AI question on the last row, "What changes would you make...to improve it," is still semantically a reasonable match to the "How To" Lim/Dey type.

The AAR/AI question on the second row, "what ... actually happened," is more subtle. This question guided many participants to focus on Output types of information. In the context of a computer system, this still seems a fairly direct semantic match between the question and Lim/Dey type. However, this question also guided over one-fourth of the responses toward the Input type, which has neither a syntactic nor semantic match to the Lim/Dey type. It could be an example of these participants working through a cause/effect connection.

Other research has shown each intelligibility type has its own advantages and disadvantages (e.g. [12, 41]), so we see the diversity of perspectives that AAR/AI seemed to elicit as a particular strength of AAR/AI.

6.2 Participants' Views of Model-Based Explanations

Participants' mental-model building with AAR/AI relied upon the presence of an explanation. In Study One, the model-based tree diagrams provided participants with a global view of the agent's decision process, supplementing the local-only "right now" view provided by the game state. As two participants put it:

> S1MB2: *"I kinda of like it how it <explanation diagram> is, because like I could try to draw my own conclusions from it rather than just like 'oh this is just what happened'."*
>
> S1MB14: *"<In the game state>... difficult to grasp the whole situation, so having the graph gave me a chance to get my footing on overall trends and options."*

This way of using the explanation was a theme which was echoed in a post-task response from another participant:

	What	What Could	How To	Judgment	Why Did	Why Didn't	Inputs	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 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 4. Lim Dey coding of participant responses, sliced by question asked during the AAR/AI.

S1MB17: *"The diagrams used to make it easier also helped to understand the predictions. To look at one thing from many angles and make appropriate predictions."*

However, a pitfall some participants fell into was extrapolating too much information from the tree diagrams. Several participants seemed *certain* about the agent's long-term plan, which was troubling because the explanation did not make such a plan explicit, if the agent even had one.

S1MB21: *"At this point, I feel certain that the friendly's trying to destroy the bottom nexus of the enemy."*

S1MB10: *"I think it's because it was a whole game plan from the beginning. ... like from the beginning of the bottom lane, the friendly AI started increasing the troop numbers."*

However, the explanation could not possibly have shown a many-step game plan, because the agent was only looking head two states.

Another participant also expressed difficulty in seeing long term strategies, but for a different reason—granularity mismatches between moves, tactics, and strategies:

S1MB20: *"There are subtasks and decisions that go into making a strategy and not being able to see this had me make less informed assumptions about the future decisions."*

In Study Two, we built interactive software, in part to alleviate the problem of too much or the wrong information at the wrong time.

7 METHODOLOGY SPECIFIC TO STUDY TWO

Study Two used a similar protocol as Study One, but in lab sessions with up to 5 participants at a time and without the think-aloud protocol, to allow more participants than are viable with think-aloud studies. Also, Study Two utilized interactive software that we built, using the results we had just learned from Study One. Also, the fact that Study Two's prototype was implemented enabled participants to perform actions like expanding the tree.

We recruited 22 participants for Study Two at Oregon State University using the same criteria as before, and randomly assigned them to our two treatments, Model-Free and Model-Based. Each participant made predictions, viewed the replay, then viewed the associated explanation for seven decision points (DPs 6, 7, 11, 17, 20, 26, and 36), selected due to their having sizeable impacts on the game outcome, which is the friendly AI winning the game at DP 37. We gave participants four minutes to fill out the prediction sheet, two minutes to understand the explanation for each DP, and an additional four to complete the questionnaire with questions from Table 2's Step 5. To ensure that they did not advance to the explanation before we were ready, we had participants type a short unlock code into the interface after the researcher provided it verbally.

7.1 Analysis Methods

We analyzed **RQ1, RQ2, and RQ3** using the same codesets described in Section 5.1. For **RQ4**, two researchers looked for evidence of participants having been somehow informed by the explanations, in the written responses to our AAR/AI questions (What happened, what was Good/Bad/Interesting about it, Why did it happen, and What changes would you make). Specifically, we removed responses without clear evidence that participants had been informed by the *explanation*, as opposed to the game state or a participant's domain knowledge. Each part of good/bad/interesting was a separate response, so the 22 participants answered 6 questions each for a total of 132 responses, of which 50 passed the filter to be considered "Explanation-Informed Statements". Using content analysis [29], we then derived the code set shown in Table 6. A different two researchers coded 44% of the data corpus independently, achieving IRR of 81.2% (Jaccard). (Usually, researchers use a smaller subset of the data for agreement, but we expanded beyond the more typical 20% in order to include more instances of rare codes.)

8 RESULTS: BOTH STUDIES

Since Study Two was intended to complement and triangulate with Study One, we present Study Two's results in combination with the pertinent results from Study One[6].

8.1 Which Information to Show?

To answer the AAR/AI questions, participants needed information from the explanations, but which information and how much of it to show is a question XAI researchers have been wrestling with for years (e.g., [16, 37, 38, 40, 41, 49, 76]). One participant simply wanted to see everything—corresponding to an explanation with maximum completeness:

S1MB5: *"All the possible actions and all possible outcomes."*

Unfortunately, with the agent considering combinatorial action spaces, showing the full search tree all at once (statically at least, it might be possible to navigate via dynamic mechanisms) would have been too large for humans to process. Thus, we needed to choose a smaller set of noteworthy actions to show—but which ones and how many?

Recall that the explanations showed only four actions (Figure 5). Some participants thought there should be more and/or different actions. For example:

S1MB5: *"... since there are only four options ... if it was a possibility for more options 'cause there was definitely more possibilities."*

However, these four options were only "top" as per the agent's estimations, which may not have been the right four:

S1MB5: *"I would think the AI would have the best four, which it didn't have the best four."*

One participant proposed also showing the *worst* possible choice:

S1MB20: *"I'd like to see ... what the friendly AI thinks is the ... choice that would give them the least chance of winning as well as their greatest chance of winning..."*

Study Two participants seemed to need information about another class of action as well—actions that spend *all* resources—since not explaining this class led them to believe the AI did not consider these actions carefully enough:

S2MF46: *"Why didn't AI use all remaining resources?"*

S2MF38: *"It's unreasonable to not purchase buildings when you've got no reason to save and invest in pylons."*

S2MB30: *"There is no reason that I can think of for it to have not spent minerals."*

Despite the fact that this class of actions was in these participants' world view, the AI does not include this human-created abstraction in its world view. That said, the agent *does* consider each available action, so the "complete" search tree contains at least some information about the kind of actions the participants describe—even if they were pruned away. The importance of this class of actions to the participants suggests that participants need this information, but answering this question might require finding more than just *one* action from the class, but instead *many* of them to reason about as a set. This suggests that participants might benefit from query systems built to select *all* instances of a class of action interesting to their world view.

As to how many actions to show, seven of the participants indicated that they liked the tree—but one wanted a smaller one, and three wanted a larger tree.

S1MB8: *"I liked the way it is. It's easy to read."*

S1MB21: *"I do not have any problem with narrow diagram..."*

S1MB11: *"I would just have more options available..."*

[6]Keeping context explicit is the reason we prefix each participant ID with the appropriate study number; e.g., S1MB5 denotes "Study One, Model-Based participant 5".

The previous paragraphs discussed participants' self-reported responses. Now we turn to what Study Two participants *actually did* when provided with an interface, enabled by watching the screen capture videos from 21 of the participants[7].

Even given the interactive explanation, 10 of the participants (5 in each treatment) did not interact with the explanation beyond panning and zooming—in effect treating it as a static diagram too large for the screen. The behaviors of the remaining participants are aggregated in Table 5. Thus, the experiences of the 10 pan/zoom-only Study Two participants with the interactive explanations prototype were similar to what Study One's participants experienced with their paper-prototyped explanations. This suggests the importance of the system's initial/default presentation of explanations—for about half of our participants, our choice of initial presentation was the only one they ever looked at.

Of the Model-Free participants who did tree manipulations, most were to widen the tree more, which was one of the few interactions available. Tree widening usually occurred in one or two short bursts of 3–8 node additions to the tree. However, they also dragged more nodes around than the Model-Based participants did, presumably for the purpose of comparing actions. For example, at one point, S2MF40 performed a series of drag operations to visually group similar action nodes (characterized by a top lane action making 3–5 marines and 1–2 banelings).

Model-Based participants also manipulated tree width, but they also took advantage of the Model-Based capability of going deeper into the tree, to peer into the AI's predictions of the future. Expanding depth adds 5 nodes, but expanding width adds just 1 node, so the amount of additional information Model-Based participants added per "deepen" manipulation was 5 times as many as with a "widen" manipulation, so the participants who used "deepen" processed a great deal more information than those who did not.

Of these "deepen"s, the most popular among the participants was the one that expanded the second-best action, then the third-best, and so on (17 second-best, 14 third-best, 8 fourth-best, 3 fifth-best, 8 beyond fifth-best). (Recall from Figure 2 that the actions were ordered best to 5th-best.) One pattern shared by four participants (S2MB20, S2MB21, S2MB31, and S2MB36) was to expand the top k futures, for some k, then visually scan it up and down. This behavior "filled the screen" with information, suggesting that our explanation's default presentation did not adequately fill up its rectangular viewing region with nodes (it started out as roughly a ⊢ shape). Had we done

[7]One participant's data (S2MB8) was not included in this analysis, due to corruption of their video file.

MF-PID	Widen	Drag
S2MF1	39	
S2MF32	15	
S2MF40	6	7
S2MF42	54	8
S2MF43	41	2
Totals	155	17

MB-PID	Widen	Deepen	Drag
S2MB20	14	12	
S2MB21		10	2
S2MB31	46	20	
S2MB35	1	2	
S2MB36	7	6	
Totals	68	50	2

Table 5. Interaction totals from participants who interacted with the explanation by: "widening" the tree by adding an action node (at any location), "dragging" a node in the tree by shifting its position, presumably to better enable comparison, or "deepening" the tree by expanding the future associated with a top-level action (refer to Figure 2). (Participant S2MB8's data was damaged, and thus excluded from this analysis.) The following 10 participants did not interact with the explanation beyond pan and zoom operations: S2MF38, S2MF41, S2MF44, S2MF45, S2MF46, S2MB23, S2MB26, S2MB28, S2MB30, S2MB39.

so in this system, the "static diagram" participants might have passively consumed more, and the "screen fillers" would not have had to manually fill the screen.

8.2 Explanations as Theory

One way to think about how participants worked with the explanations to answer AAR/AI questions, is to view the explanations as the agent's "theories". In the explanation trees, upon reaching leaf nodes, the agent used a neural network to evaluate the quality of states. These estimates were, in essence, *axioms* and the minimax search that proceeds atop those values are akin to *theorems*. Thus, if the axioms hold true, then the theorems were true.

In Study One, we saw that not all participants were willing to "grant the axioms." Some were:

S1MB14: *"I mean because, those are the ones with greater scores. So I guess that is why it chose those decisions."*

Others did not grant them and found themselves not understanding or possibly disbelieving parts of the diagram.

S1MB10: *"I think diagram needs improvement, because those are not that clear at some times. ...It does have enough details, but the decisions were, not made... according to the diagram."*

Two participants identified the issue well: that the win probabilities have no clear provenance.

S1MB8: *"... If there's any easy way to say why it came up with these numbers... there were several steps that I just didn't know why it was taking that action..."*

We found that RTS experience seemed to be a potential driver for rejecting the heuristic evaluation function, with S1MB5 and S1MB20 being particularly critical of the agent's decisions:

S1MB5: *"Wow, rewards went down... A baneling is better than a marine by rewards points, but there's clearly a better answer."*

Those with less RTS experience seemed less critical of the agent's explanation, but they still compared the agent's actions to the tree:

S1MB14: *"Information didn't always line up with what occurred. Therefore, it gives a false belief on what/how the AI is doing."*

When we conducted Study Two, through use of the Model-Free and Model-Based explanations, we offered two very different presentations of the agent's theory. In particular, the Model-Free explanations are mostly leaf nodes, meaning they are almost entirely *axiomatic*. Despite this, some of Study Two's participants did find the Model-Free explanations helpful:

S2MF43: *"Decision tree helped understand logic of AI better."*

Further, they were able to use Model-Free explanations to compare different actions, for example:

S2MF41: *"It was very helpful to be able to see multiple potential game paths side by side."*

However, the Model-Free participants did not have access to the information that would allow them to "disprove" deeply nested theorems by following them all the way down the tree. Recall from Figure 2 that Study Two's Model-Free explanations provide the current state at the root node, and then the top k actions and their values beneath that. In contrast, Model-Based explanations allow explorations all the way down the tree, eventually running into an axiomatic value, where we see the same curiosity about provenance that we saw in Study One:

S2MB30: *" Where does the % come from?"*

Thus, Model-Free explanations lacked some information that the Model-Based participants appeared to value highly:

S2MB21: *"Ability to see additional buildings for the next round gave insight on future AI actions. Explanation elements were easy to read and understand."*

S2MF41: *" I'm not 100% sure the information given in the explanations necessarily completely reflected the AI's decisions."*

One way of considering the value Model-Free vs. Model-Based participants obtained from the explanations is to consider their Explanation-Informed Statements. These are defined in Table 6 and, as the table indicates, Model-Free participants made fewer Explanation-Informed Statements of every type than Model-Based participants did (20 vs. 48). Further, Model-Free participants not only provided fewer of them than Model-Based participants did, but also did not even attempt Explanation-Informed bug reports until near the end of the task, as Figure 4 illustrates.

The bug reports were also different. Below are all four reports from Model-Free participants:

S2MF46: *"Good choice, but in bottom nexus is much lower. Why not commit to destroying it?"*
S2MF46: *"Why didn't AI use all remaining resources <at round 36>?"*
S2MF37: *"no round 36 purchase? Why?"*
S2MF38: *"It's unreasonable to not purchase buildings when you've got no reason to save and invest in pylons <at round 36 of 40>. I guess there is a bias introduced on how many buildings it can buy at a time."*

In essence, the Model-Free bug reports above are simply disagreements with high-level strategic choices the agent makes, as opposed to *falsifications* of the logic contained in individual nodes or transitions. A few Model-Based participants also gave those kinds of bug reports, such as:

S2MB30: *"There is no reason that I can think of for it to have not spent minerals."*

However, in addition to these strategy disagreements, Model-Based participants considered the correctness of individual predictions that go into the overall action selection:

S2MB26: *"Only 2 immortals were created. Prediction of marines was wrong."*
S2MB28: *"I think friendly AI is not able to assess that bottom lane is better. It is doing very well in bottom lane. But end result predicted is wrong."*

Or, stated just as logically but less passionately:

S2MB36: *"The friendly AI decided to fortify the bottom lane assuming an attack. The attack actually came from the top, where the enemy now has the advantage."*
S2MB28: *"The enemy AI outsmarted friendly AI. It sent marines along with banelings. Friendly AI thought enemy AI will send marines so it bought baneling producing building."*
S2MB36: *"It's interesting that the AI keeps assuming an attack on the bottom and not defense on the top."*
S2MB36: *"It still assumes it will win by destroying the top nexus."*
S2MB28: *"Friendly AI still predicts it will win by destroying top enemy nexus."*

Code: Description	**Example**	**MF**	**MB**
Explanation-Informed Observation: Participant strictly interprets the explanation.	S2MB39: *"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.	S2MB23: *"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.	S2MB26: *"Only 2 immortals were created. Prediction of marines was wrong."*	4	9
	Totals	20	48

Table 6. Explanation-Informed Statements code set, as applied to Study Two's 22 participant responses to our decision questionnaire (What happened, what was Good/Bad/Interesting about it, Why did it happen, and what Changes would you make), with examples and counts from each treatment.

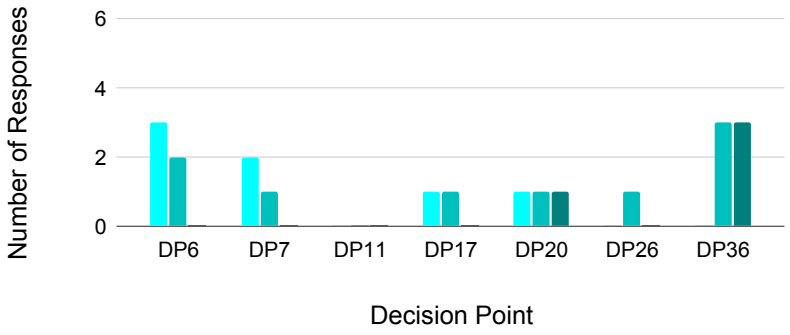
These differences in the Model-Free vs. Model-Based participants' Explanation-Informed Statements illustrate a key strength of Model-Based explanations. They enabled Model-Based participants to "disprove" aspects of the agent's decisions by seeing inconsistencies and logic errors in the path propagating the *axiomatic* values computed at the leaves to the *theorems* about action selection. This process brings explicit falsification [54] capabilities to the system's users.

We observed participants engaging in falsification in both studies. In particular, Model-Based explanations make part of the search tree explicit and include concrete predictions about the future, including states. These concrete predictions allowed participants to falsify those predictions:

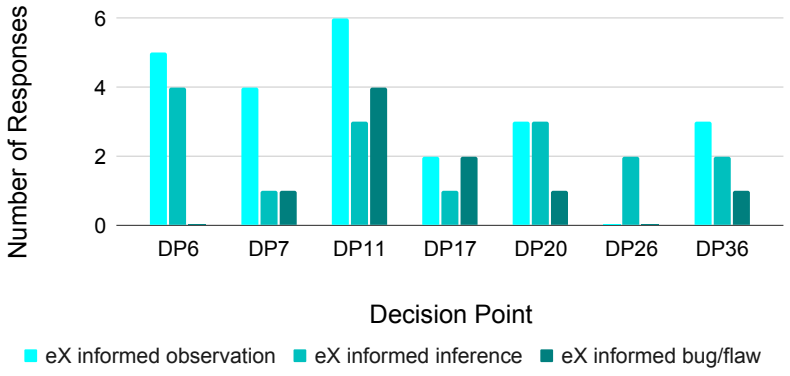
S1MB14: *"So the friendly had ... two banelings, so one baneling and some marines. Yes, that seems right. ... it predicted that the enemy would buy two more marines, and it ended up being so. Yep, it was right ... it was predicted that they would buy a baneling, and they did ... so far, it's going as predicted."*

S2MB26: *" Only 2 immortals were created. Prediction of marines <from the previous state> was wrong...<later>...Prediction was correct."*

We explicitly crafted parts of the process to allow the human to reflect on *their* past thoughts, but this participant focused on the accuracy of the *agent's* predictions about the future. Notably, this



(a) Explanation-Informed Statements from Study Two participants using **Model-Free** explanations.



(b) Explanation-Informed Statements from Study Two participants using **Model-Based** explanations.

Fig. 4. Explanation-Informed Statements from Study Two participants interacting with **Model-Free (top)** and **Model-Based (bottom)** explanations. Statements are broken down into 3 categories. DPs (each bar cluster) are time ordered and aligned.

type of assessment was made possible by the Model-Based agent, and our explanations revealed relevant information to be able compare different time slices.

Thus far, we have focused on viewing explanations as theory in terms of their composition and falsification of elements. However, there are other criteria that can be used to evaluate theories [65]. In Table 7, we consider how to apply these criteria to evaluate *this* agent's Model-Based explanation, this *style* of Model-Based explanations, and in some ways, even *all* Model-Based explanations.

	"The degree to which..." [65]	**Applicable to...**	**Evidence to date for or against**
Testability	...empirical refutation is *possible*: constructs and <predictions> are understandable, internally consistent, free of ambiguity	...this *explanation* of the agent's model of the world.	*Empirical*: The agent's explanations were found to be understandable by several participants, as described in Section 6.1. The diagrams were clear and explicit in their information from most, but not all, participants' reports.
Falsifiability /Empirical Support	...is supported by empirical studies that confirm its validity	...this *explanation* of the agent's model of the world.	*Empirical*: Our explanations explicitly represented the agent's predictions about likely future states and their values, which participants could falsify.
		...this *style* of Model-Based explanation.	*Empirical*: AAR/AI evaluators (one instance: our participants).
Explanatory Power	...accounts for and predicts all known observations within its scope	...this *explanation* of the agent's model of the world.	*Empirical*: One measure is whether the agent's theory and explanation correctly predicted everything. In our study, the agent did not achieve this. *Criteria-based*: Whether its constructs are sufficient to express every possible action and state, i.e. completeness. In this study, the constructs have full explanatory power—but our explanation limited the number, so the actual explanation was not complete.
		...this *style* of Model-Based explanation.	
		...*all* Model-Based explanations.	
Parsimony	...<has> a minimum of concepts and propositions	...this *explanation* of the agent's model of the world.	*Criteria-based*: This explanation had 4 constructs/concepts that do not overlap, so cannot be reduced further.
Generality	...breadth of scope... and independent of specific settings	...this *explanation* of the agent's model of the world.	*Criteria-based*: This explanation's scope is limited to explaining this particular domain.
		...this *style* of Model-Based explanation.	*Criteria-based*: The style of explanation is not restricted to games, and should be usable for any sequential setting of Model-Based AI.
		...*all* Model-Based explanations.	Model-Based explanations are restricted to Model-Based agents.
Utility	...supports the relevant areas	...this *explanation* of the agent's model of the world.	*Empirical*: Most, but not all, participants reported the agent's explanations to be useful to understanding its actions.
		...this *style* of Model-Based explanation.	*Empirical*: AAR/AI evaluators (one instance: our participants).

Table 7. Applying Sjøberg et al.'s Evaluation Criteria for Theories [65] to the agent's Model-Based explanation

8.3 Participants' Cognitive Load and Performance

How did the differences in how participants engaged with the different explanations play out in participants' views of the challenge, effort and frustration levels of the entire AAR/AI process they experienced? To provide insights into this question, we turn to the NASA Task-Load indeX (TLX) responses from 15 of the Study Two participants (not all participants provided this data) at the end of the session. The TLX is a validated post-task survey to measure cognitive load [24]. As shown in Table 8, participants rated Physical Demand very low. There was also no difference in Temporal Demand (in either the medians or the distribution of data points) and little difference in Performance. Thus, we ignore those three and shift focus to the remaining factors.

The remaining three factors reflect participants' perceptions of cognitive load. Table 8 suggests that the participants who saw the Model-Free treatment tended to feel more Mental Demand[8] than the Model-Based participants in their AAR/AI-based evaluations. Consistent with this result, Model-Free participants also reported higher Effort[9] than the Model-Based participants. However, the Model-Free participants reported *less* Frustration[10]. This observation was unexpected for us, since Model-Free explanations contain a smaller amount of information.

These three results conceptually relate to Sweller's influential cognitive load theory [69]. Cognitive load theory includes three concepts: *intrinsic load*, i.e., the cognitive work that is inherent in the task for everyone; *germane load*, i.e., helpful additional cognitive work that may be necessary for that individual (e.g., inferring helpful new abstractions, such as by comparing a past item with a current item to abstract above the current situation); and *extraneous load*, i.e., extra, *un*helpful cognitive work that hampers the individual in performing the task (e.g., having to continually look up the meaning of different UI widgets) [51, 69].

Using these concepts, Mental Demand ("task-inherent" load) approximates intrinsic load, and Effort ("your" load) approximates the sum[11] of intrinsic + germane + extraneous load [51]. Our results suggest that some participants decided that Mental Demand matched Effort (i.e., I had to do it, so it must have been what the task needed). Frustration is an interesting side-effect relating to Demand and Effort—our data suggested that it reflected participants' reaction to excessive load, especially extraneous load.

9 DISCUSSION

9.1 Future AAR/AI Adaptations

AAR/AI is highly adaptable, and this provides leeway to iteratively improve it. Two areas for improvement that we observed were that participants thought they could remember what happened in the past, and that participants found questions/artifacts repetitive and burdensome at times. For example:

[8]TLX question: *"How mentally demanding was the **task**?"* (emphasis added)
[9]TLX question: *"How hard did you have to **work** to accomplish your level of performance?"* (emphasis added)
[10]TLX question: *"How insecure, discouraged, irritated, stressed, and annoyed were you?"*
[11]Orru et al. discussed a version of the NASA-TLX modified to equate the Effort question specifically with extraneous load [21, 51]. However, without that modification, NASA-TLX's Effort question is not confined to extraneous load

	Mental	Physical	Temporal	Performance	Effort	Frustration
Model-Free	14.5	1	13	13.5	15.5	3
Model-Based	11	3	13	14	13	9

Table 8. Median results of the NASA TLX. Our discussion focuses on the responses with the greatest differences between the two treatments (highlighted): Mental Demand, Effort, and Frustration.

S1MB20: *"... I am fairly confident in my ability to remember what occurred."*
S1MB5: *"Some of this stuff kind of repeats..."*

An alternative might be to instead enable people to decide where to pause, in an approach similar to the empirical mechanism used by Penney et al. [53]. In that study, their participants watched a replay until they came to a decision that seemed important, at which point they could pause, consider our questions, and write down their thoughts. In essence, blending this device with our inner loop would give more control to the evaluators as to how often and exactly where the evaluation questions need to be answered.

As a meta-analysis of AAR by Keiser and Arthur [32] observed, it *"was initially operationalized with high administrative and content structure..."* with the goal being that *"higher administrative structure is expected to free up cognitive resources that would otherwise be spent on how to conduct the AAR."* Further, the authors go on to describe situations with less structure, and offer a flowchart (see Figure 10 from [32]) to help select the appropriate flavor of AAR for a variety of use cases.

Our short series of studies left many open questions about AAR/AI's efficacy in different possible usages. Among them, to what extent is it: ...*rigorous* enough to support examining catastrophic failures that will necessarily consume hours of time from investigators? ...*efficient* enough for real time analysis, akin to sports commentary? By investigating open questions like these, researchers will be able to discover shortcomings and devise adaptations to improve fitness for different usages—and possibly illuminate other evaluation processes in so doing.

9.2 Prediction as Explanation

Trend 1: People used explanations as prediction tools. Reed et al. suggested that explaining a solution to a problem helps people solve similar problems [56]. Our strategy followed a similar approach, where participants predicted the agent's action (i.e., the problem), saw the action (i.e., the solution), and then provided an explanation to the action (i.e., explanation of the solution). Some participants even began using the explanations as the basis for their prediction:

S1MB8: *"Understanding the diagram gave some insight into how the AI thought, which made predicting its next move easier."*

Participants engaging with the Model-Based explanation reported attitudes consistent with a series of studies Kelleher and Hnin observed, *"suggest that learners who attempt to understand the steps of a problem solution may have higher germane load but improved ability to apply these elements in novel situations."* [33].

Trend 2: The process of having participants predict the actions first, and then showing them the actions, was powerful. Another trend we observed was that predicting the AI agent's decisions *prior* to observing the AI agent's actual actions turned out to be part of our *explanation strategy*. One of the pillars of learning effectively is self-explaining [10]. Those researchers describe how students who learn with understanding the material and forming self-explanations on their own achieve better outcomes than those relying heavily on examples to learn and struggling to generate explanations on their own. Positioning the prediction task before the observation task effectively caused participants to create self-explanations for the AI agent's actions. Participants used the process and the explanation, to generate their own explanation for predicting the agent's actions:

S1MB10: *"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."*

Participants who answered AAR/AI questions perform a "rationale generation" [17] task, which appears to offer some benefits as an AI evaluation strategy.

Renkl et al. found that acquisition of transferable knowledge can be supported by eliciting self-explanations [57]. Learners with low levels of prior topic knowledge profit from such an elicitation procedure. We observed this effect in our study, as participants with little experience in

RTS comfortably navigated through the process of assessing the AI's actions—even forming their own explanations.

9.3 Encouraging Metacognition

Researchers in the field of education have long pointed to the benefits of metacognition, in which learners evaluate the success of their own learning/understanding processes [19]. Metacognitive activity is well-established as an important influence on learning and understanding [78].

Participants in Study One and Study Two, with both the Model-Free and the Model-Based explanations, showed several instances of metacognition that seemed to come from the integration of AAR/AI, the explanations they saw, and the "active user." For example:

S1MB5: *"It made me think of it like how the AI is thinking. Is it thinking long term? Is it thinking short term? Thinking about the two different lanes each time?... what the best decision would be or what I would make as the decision, so you asking that question made me think 'was my own decision better?'."*

S1MB8: *"...it was good to kind of evaluate myself where I was at when thinking about what decisions the AI was doing, so I can better evaluate the next stage."*

S2MF41: *" Being able to compare the AI's choices in the explanation graphs made it helpful in seeing what may have been a stronger choice (AI vs yourself)."*

S2MB35: *"Friendly AI bought 1 baneling building in bottom lane. I'm unable to notice all possible changes at a decision point."*

One form of metacognition is self-explanation, and our approach encouraged some participants to generate their own explanations:

S1MB10: *"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."*

S2MB8: *"Plans to distract Enemy AI in bottom lane."*

S2MF42: *" AI doesn't appear to consider killing bottom lane to be an avenue to victory."*

Finally, while our process promoted thinking about the future, the artifacts also supported participants' ability to reflect on the *past*:

S1MB19: *"These cards? It's good to write good points and bad points for every three rounds, so that we can go back and see what mistakes we did from the bad."*

10 THREATS TO VALIDITY

Any study has threats to validity, which can skew results towards particular conclusions [79].

One such threat was the participants' amount of domain expertise. Evaluators of an AI system need domain knowledge to evaluate the AI's performance in the domain, and some of the participants may not have had enough RTS experience. As an example, 46% of Study One's participants had at least 10 hours of RTS gaming experience. It is possible that these participants' experience levels may have impacted their ability to evaluate an AI in that domain. Also, it was not clear how to interpret large decreases in the number of clarifications a participant requested early vs. late in the process. It could have meant that the participants understood the explanations over time, or alternatively that they simply gave up. The question wording could also have influenced participants' responses. Many were written and uniformly worded in a balanced set of positive, negative, and neutral wording, but the verbal post-task interview wording was informal, so more subject to individual variation.

The reliability of qualitative coding rests upon inter-rater reliability (IRR) measures. We used Jaccard [31], and 80% is considered good agreement, but for one code set we achieved only 76%. Another hindrance to the generalizability of our findings is the circumscribed design and small size of our study—preventing comparative statistics from yielding meaningful contrast between Study Two's treatments. Similarly, the sensemaking task we have chosen focuses on the *depth* and

breadth of participants' constructed mental models—but says little to nothing about their *accuracy* or *usefulness*.

Also, qualitative studies are intended to reveal phenomena on approaches that have not been investigated before, and are not suitable for generalization. That said, our study can still inform model-based explanations for domains where the branching factor is small (or can be made small via pruning, as we have done).

11 CONCLUSION

In this paper, we have presented AAR/AI (After-Action Review for AI), a new assessment method to bring accountability to both AI agents and to the humans who must assess them. To inform the design of AAR/AI, we present results from two qualitative lab studies to learn what people need when assessing an AI agent, as well as pros/cons of both the AAR/AI process and the explanations embedded in the process. Among the phenomena we found were:

- *"Organized," "Logical," and..."Repetitive":* Some participants remarked that AAR/AI process helped them think logically and stay organized. Some appreciated its support for reflection on past thoughts. Notably, the process helped participants generate rationale for events with long time lags. However, some bemoaned the repetitiveness of the AAR/AI questions.
- *Explanation complexity:* Our search tree explanations for a model-based agent were approximately the right complexity for some of the participants to understand. They reported being able to *"draw their own conclusions"* from them, and appeared to be using them to align the agent's prediction with the actual future. Other participants did not fully understand the diagram. This mix of attitudes toward the same explanation corroborates other research reporting that explanations are not "one size fits all" (e.g. [2]), and suggests allowing people to access different actions and/or explanation types on demand.
- *Model-Free or Model-Based:* In Study Two we had both Model-Free and Model-Based explanations. Study Two participants who used the Model-Free explanations expressed less than half as many explanation-informed statements as the Model-Based participants did. More critically, the Model-Free participants' bug reports were merely participants' disagreements with the agent's strategy, whereas some Model-Based participants were able to point explicitly to logic errors in the explanations.
- *Diversity of perspectives:* As we observed and participants reported, AAR/AI's questions encouraged participants to consider their observations from multiple, different perspectives, which research suggests may produce problem-solving benefits [20].
- *How many and which:* To answer some of the AAR/AI questions, participants needed to compare items in the explanation from a very large set of options, the sheer quantity of which made them hard to co-locate. We provided the AI's most promising options, but some participants wanted to see options the AI considered *bad*, as well as actions that spend all resources. Accommodating different people's comparison needs to answer the AAR/AI questions is an unresolved issue—so methods to support scalable comparisons of items in large datasets (e.g. [49]) is an active area of Info Viz research.
- *From whence:* Some participants needed to know the *provenance* of axiomatic values (value estimations at the leaf nodes). That said, if people are to be held accountable for relying on an AI agent, then the ability to "audit" its decision making by allowing the ability to trace provenance may be a requirement.

Overall, AAR/AI's ability to organize participants' work with our agent's explanations assisted the participants in the assessment process. Our results are particularly promising when combining AAR/AI with Model-Based explanations. Still, developing useful explanations and rigorously

measuring their quality remains quite difficult and, as our participants pointed out, there is much work still to be done. Ultimately, we hope that AAR/AI's framework around explanations can help people like S1MB14 see *"the flow of logic that we **should've** had"* when assessing AI systems that impact us daily.

ACKNOWLEDGEMENTS

This work was supported by DARPA #N66001-17-2-4030. Any opinions, findings, conclusions, or recommendations expressed are the authors' and do not necessarily reflect the views of the DARPA, Army Research Office, or US government.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett. 2019. Explaining Reinforcement Learning to Mere Mortals: An Empirical Study. In *International Joint Conference on Artificial Intelligence*. IJCAI, Macau, China.

[3] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental Models of Mere Mortals with Explanations of Reinforcement Learning. *ACM Trans. Interact. Intell. Syst.* 10, 2, Article 15 (May 2020), 37 pages. https://doi.org/10.1145/3366485

[4] Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives*. Pearson, New York, NY, USA.

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[6] Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives*. Longmans, Green and Co LTD, London, England.

[7] Ralph Brewer, Anthony Walker, E. Ray Pursel, Eduardo Cerame, Anthony Baker, and Kristin Schaefer. 2019. Assessment of Manned-Unmanned Team Performance: Comprehensive After-Action Review Technology Development. In *2019 International Conference on Human Factors in Robots and Unmanned Systems (AHFE '19)*. Springer Nature Switzerland AG, Cham, CHE, 119–130.

[8] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. https://doi.org/10.1145/3359206

[9] Nicholas Carlini and David Wagner. 2016. Towards Evaluating the Robustness of Neural Networks. arXiv:cs.CR/1608.04644

[10] Michelene T.H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science* 13, 2 (4 1989), 145–182. https://doi.org/10.1207/s15516709cog1302_1

[11] CNN. 2016. Who's responsible when an autonomous car crashes? http://money.cnn.com/2016/07/07/technology/tesla-liability-risk/index.html

[12] Kelley Cotter, Janghee Cho, and Emilee Rader. 2017. 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, 1553–1560.

[13] Robert Davies, Elly Vaughan, Graham Fraser, Robert Cook, Massimo Ciotti, and Jonathan E. Suk. 2019. Enhancing Reporting of After Action Reviews of Public Health Emergencies to Strengthen Preparedness: A Literature Review and Methodology Appraisal. *Disaster Medicine and Public Health Preparedness* 13, 3 (june 2019), 618–625. https://doi.org/10.1017/dmp.2018.82

[14] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13 (1989), 319–340. https://doi.org/doi:10.2307/249008.

[15] Jonathan Dodge and M. Burnett. 2020. Position: We Can Measure XAI Explanations Better with Templates. In *ExSS-ATEC@IUI*.

[16] Jonathan Dodge, Sean Penney, Claudia Hilderbrand, Andrew Anderson, and Margaret Burnett. 2018. How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games. In *2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 562, 12 pages.

[17] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA, 263–274. https://doi.org/10.1145/3301275.3302316

[18] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Xiaodong Song. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 1625–1634.

[19] Donna-Lynn Forrest-Pressley and GE MacKinnon. 1985. *Metacognition, Cognition, and Human Performance: Theoretical Perspectives*. Vol. 1. Academic Pr.

[20] Hershey H Friedman, Linda W Friedman, and Chaya Leverton. 2016. Increase diversity to boost creativity and enhance problem solving. *Psychosociological Issues in Human Resource Management* 4, 2 (2016), 7.

[21] Peter Gerjets, Katharina Scheiter, and Richard Catrambone. 2004. Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures. *Instructional Science* 32, 1-2 (2004), 33–58.

[22] Ian Goodfellow and Nicolas Papernot. 2017. The challenge of verification and testing of machine learning. http://www.cleverhans.io/security/privacy/ml/2017/06/14/verification.html

[23] Samer Hanoun and Saeid Nahavandi. 2018. Current and Future Methodologies of After Action Review in Simulation-based Training. In *2018 Annual IEEE International Systems Conference (SysCon) (SysCon '18)*. IEEE, New York, NY, USA, 1–6.

[24] S. G. Hart and L. E. Staveland. 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Adv. Psychol.* 52 (1988), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[25] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[26] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2010. Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model. *International Journal of Social Robotics* 2, 4 (01 Dec 2010), 361–375. https://doi.org/10.1007/s12369-010-0068-5

[27] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *CoRR* abs/1812.04608 (2018). arXiv:1812.04608 http://arxiv.org/abs/1812.04608

[28] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[29] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[30] Andrew Ishak and Elizabeth Williams. 2017. Slides in the Tray: How Fire Crews Enable Members to Borrow Experiences. *Small Group Research* 48, 3 (March 2017), 336–364. https://doi.org/10.1177/1046496417697148

[31] Paul Jaccard. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44 (1908), 223–270.

[32] Nathanael Keiser and Winfred Arthur, Jr. 2020. A meta-analysis of the effectiveness of the after-action review (or debrief) and factors that influence its effectiveness. *Journal of Applied Psychology* (08 2020). https://doi.org/10.1037/apl0000821

[33] Caitlin Kelleher and Wint Hnin. 2019. Predicting Cognitive Load in Future Code Puzzles. In *2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 257, 12 pages.

[34] M. Kim, K. Kim, S. Kim, and A. K. Dey. 2018. Performance Evaluation Gaps in a Real-Time Strategy Game Between Human and Artificial Intelligence Players. *IEEE Access* 6 (2018), 13575–13586.

[35] Man-Je Kim, Kyung-Joong Kim, SeungJun Kim, and Anind K Dey. 2016. Evaluation of StarCraft Artificial Intelligence Competition Bots by Experienced Human Players. In *ACM CHI Conference Extended Abstracts*. ACM, 1915–1921.

[36] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[37] T. Kulesza, M. Burnett, W. Wong, and S. Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *ACM International Conference on Intelligent User Interfaces*. ACM, 126–137.

[38] 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 *ACM Conference on Human Factors in Computing Systems*. ACM, 1–10.

[39] Adam Lareau and Brice Long. 2018. The Art of the After-Action Review. *Fire Engineering* 171, 5 (May 2018), 61–64. http://search.proquest.com/docview/2157468757/

[40] Brian Lim, Anind Dey, and Daniel Avrahami. 2009. *Why* and *Why Not* Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *2009 SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2119–2128.

[41] Brian Y Lim. 2012. *Improving understanding and trust with intelligibility in context-aware applications*. Ph.D. Dissertation. Carnegie Mellon University.

[42] Sandra Deacon Lloyd Baird, Phil Holland. 1999. Learning from action: Imbedding more learning into the performance fast enough to make a difference. 27 (1999), 19–32. https://doi.org/10.1016/S0090-2616(99)90027-X

[43] Theresa Mai, Roli Khanna, Jonathan Dodge, Jed Irvine, Kin-Ho Lam, Zhengxian Lin, Nicholas Kiddle, Evan Newman, Sai Raja, Caleb Matthews, Christopher Perdriau, Margaret Burnett, and Alan Fern. 2020. Keeping It "Organized and Logical": After-Action Review for AI (AAR/AI). In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 465–476. https://doi.org/10.1145/3377325.3377525

[44] Ronald Metoyer, Simone Stumpf, Christoph Neumann, Jonathan Dodge, Jill Cao, and Aaron Schnabel. 2010. Explaining how to play real-time strategy games. *Knowledge-Based Systems* 23, 4 (2010), 295–301.

[45] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[46] John E. Morrison and Larry L. Meliza. 1999. *Foundations of the After Action Review Process*. Technical Report. Institute for Defense Analyses. https://apps.dtic.mil/docs/citations/ADA368651

[47] Donald A Norman. 1983. Some observations on mental models. *Mental Models* 7, 112 (1983), 7–14.

[48] N.Y. Times. 2017. Tesla's Self-Driving System Cleared in Deadly Crash. https://www.nytimes.com/2017/01/19/business/tesla-model-s-autopilot-fatal-crash.html

[49] Oluwakemi Ola and Kamran Sedig. 2016. Beyond simple charts: Design of visualizations for big health data. *Online journal of public health informatics* 8 (28 12 2016). Issue 3. https://doi.org/10.5210/ojphi.v8i3.7100

[50] S. Ontañón, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill, and M. Preuss. 2013. A Survey of Real-Time Strategy Game AI Research and Competition in StarCraft. *IEEE Transactions on Computational Intelligence and AI in Games* 5, 4 (Dec 2013), 293–311. https://doi.org/10.1109/TCIAIG.2013.2286295

[51] Giuliano Orru and Luca Longo. 2018. The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and Germane loads: a review. In *International Symposium on Human Mental Workload: Models and Applications*. Springer, 23–48.

[52] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. DeepXplore. *Proceedings of the 26th Symposium on Operating Systems Principles - SOSP '17* (2017). https://doi.org/10.1145/3132747.3132785

[53] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. 2018. Toward Foraging for Understanding of StarCraft Agents: An Empirical Study. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 225–237. https://doi.org/10.1145/3172944.3172946

[54] Karl R Popper. 1963. Science as falsification. *Conjectures and refutations* 1 (1963), 33–39.

[55] John Quarles, Samsun Lampotang, Ira Fischler, Paul Fishwick, and Benjamin Lok. 2013. Experiences in mixed reality-based collocated after action review. *Virtual Reality* 17, 3 (Sept. 2013), 239–252. https://doi.org/10.1007/s10055-013-0229-6

[56] Stephen Reed, Alexandra Dempster, and Michael Ettinger. 1985. Usefulness of Analogous Solutions for Solving Algebra Word Problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11, 1 (Jan. 1985), 106–125. https://doi.org/10.1037/0278-7393.11.1.106

[57] Alexander Renkl, Robin Stark, Hans Gruber, and Heinz Mandl. 1998. Learning from Worked-Out Examples: The Effects of Example Variability and Elicited Self-Explanations. *Contemporary Educational Psychology* 23, 1 (Jan. 1998), 90–108. https://doi.org/10.1006/ceps.1997.0959

[58] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation Accuracy Is Good, but High Controllability May Be Better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290605.3300750

[59] Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[60] Margaret Salter and Gerald Klein. 2007. *After Action Reviews: Current Observations and Recommendations*. Technical Report. U.S. Army Research Institute for the Behavioral and Social Sciences.

[61] Taylor Lee Sawyer and Shad Deering. 2013. Adaptation of the US Army's After-Action Review for Simulation Debriefing in Healthcare. *Simulation in Healthcare* 8, 6 (Dec. 2013), 388–397. https://doi.org/10.1097/SIH.0b013e31829ac85c

[62] Martin Schindler and Martin J Eppler. 2003. Harvesting project knowledge: a review of project learning methods and success factors. *International Journal of Project Management* 21, 3 (2003), 219 – 228. https://doi.org/10.1016/S0263-

7863(02)00096-0

[63] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484.

[64] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144. https://doi.org/10.1126/science.aar6404 arXiv:https://science.sciencemag.org/content/362/6419/1140.full.pdf

[65] Dag IK Sjøberg, Tore Dybå, Bente CD Anda, and Jo E Hannay. 2008. Building theories in software engineering. In *Guide to advanced empirical software engineering*. Springer, 312–336.

[66] Alison Smith-Renner, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2020. Digging into User Control: Perceptions of Adherence and Instability in Transparent Models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 519–530. https://doi.org/10.1145/3377325.3377491

[67] Dan "Artosis" Stemkoski. 2019. AlphaStar - Analysis by Artosis. https://www.youtube.com/watch?v=_YWmU-E2WFc.

[68] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662.

[69] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4, 4 (1994), 295–312.

[70] The StarCraft II Community. 2019. Tutorials - Sc2MapsterWiki. https://sc2mapster.gamepedia.com/Tutorials.

[71] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. 2017. DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars. arXiv:cs.SE/1708.08559

[72] U.S. Army. 1993. *Training Circular 25-20: A Leader's Guide to After-Action Reviews*. Technical Report. Department of the Army, Washington D.C., USA.

[73] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173590

[74] Oriol Vinyals. 2017. DeepMind and Blizzard open StarCraft II as an AI research environment. https://deepmind.com/blog/deepmind-and-blizzard-open-starcraft-ii-ai-research-environment/

[75] Oriol Vinyals, David Silver, et al. 2019. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii.

[76] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI*, Vol. 19.

[77] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J. Smith, Kalyan Veeramachaneni, and Huamin Qu. 2019. ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300911

[78] Franz Emanuel Weinert and Rainer H Kluwe. 1987. Metacognition, motivation, and understanding. (1987).

[79] Claes Wohlin, Per Runeson, Martin Höst, Magnus Ohlsson, Björn Regnell, and Anders Wesslén. 2000. *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, Norwell, MA, USA.

[80] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174230

[81] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

A HELPFUL/PROBLEMATIC CODE SET FOR EXPLANATIONS

Code: Description	Example	#
Explanation Overall Quality: Participant found explanation useless or helpful in a vague sense, or in determining reasons for actions in the decision process (clarity, or lack thereof).	S1MB2: *"I think it's pretty easy to understand, like, after looking at for a little while."*	8
Diagram Color Coding: Participant comments on the manner in which an explanation object is colored.	S1MB17: *"The color coding is okay, it's pretty distinctive. I don't know if the background is gray or...even the marines are gray... it was confusing because if it was different color..."*	4
Changing Diagram Data Contents: Participant talks about changing data in the diagram (such as changing the node definitions, changing the key, etc). This is NOT about showing an action/state node that is not present.	S1MB18: *"How much minerals it has, something like that. I would like that to be represented on the diagram."*	7
Diagram Node Contents: Participant wants the diagram to contain more/fewer nodes, (e.g. interactively expand a node, request a specific action be examined, or have a "wider/narrower" tree) OR thinks it contains the right amount.	S1MB11: *"I would just have more options available, you know. ... So sometimes, there are missing... missing options which should be taken."*	16
Diagram Glyph Presentation: Participant comments on the glyphs for the action or state nodes, referring to the way the state information is presented in the glyph	S1MB10: *"As the number of units goes on increasing, the line goes on increasing. And that is why it's short. That's clear, but vertical lines are... if it would have been 1, it would have been great. Just 1 line."*	6

Table 9. Helpful/Problematic code set for the *explanations*. Frequencies are from Study One's post task three questions centered on the explanation and its contents. (*"What was helpful about the information given to you?"*, *"What was problematic about the information given to you?"*, and *"Under what circumstances is the agent likely to make bad decisions?"*)

B MORE ABOUT THE EXPLANATION AND ITS DESIGN EVOLUTION

In Figure 5, the root node (region 1) shows the current game state and its estimated value. One layer down (region 2) shows the 4 best actions available to the friendly AI in the current game state–and their values, as estimated by the agent based on the tree expansion. The third level of the tree (region 3) shows actions available to the opponent—again, the 4 best actions and their values as estimated by the agent. The fourth level of the tree (region 4) shows the *predicted* state that the agent thinks will ensue based on the current state, taken together with the simultaneous actions from itself and the opponent. From that level, the agent performs another round of search in the same way, resulting in an agent that looks ahead 2 rounds. Each node is shown with the state or action that node depicts, alongside the estimated value of that state/action, shown in more detail in Figures 6a and 7a. If that value is part of the principal variation (colloquially, the most likely trajectory given "optimal" play from both sides), its value is shown in green instead of blue.

Figure 5 depicts the explanation used in Study One. For Study Two, we used the explanations shown in Figure 2. These Study Two explanations were implemented in an interactive prototype, hence offering interactions not possible in Study One's paper-prototyped explanations. Also, drawing from our observations of Study One participants, in Study Two's explanations we changed the glyphs used to represent states and actions, including outcome bars, which are shown in more

detail in Figures 6b and 7b. Based on Study One's results, we made the default tree more complex by increasing the branching factor at the root from 4 to 5, but eliminated the branching between friendly and enemy actions, instead including only the option estimated to be the best.

We improved the explanation in other ways between the studies. For example showing an estimation of the resources available to both the friendly AI and its opponent, as requested by a Study One participant:

> S1MB20: *"I would enjoy to see ... the AI's, calculation of their minerals. ...further extrapolation of getting this many more minerals allows you to buy these units. ...Because in RTS games you think about the enemy's resources as well and how to manage those as well as your own."*

For Study Two, we incorporated much of this Study One feedback into our explanation design, but Study Two participants were not entirely satisfied. Some wanted information that still went beyond that available in the explanations:

> S2MB30: *"It would have been helpful to know how many immortals are effective against a baneling and number of marines, effective against immortals etc. Instead of which ones are effective against each other."*
>
> S2MB28: *"There was no info on what enemy AI is thinking. Also both lanes play at same time so hard to focus on both."*

These quotes suggest that finding, processing, and sorting out high-level information intermingled with low-level information was cognitively burdensome. Adding to this cognitive burden, some Study Two participants pointed to the cognitive work of comprehending certain glyphs and layout:

> S2MB35: *"I didn't explore all parts of the explanations. Couldn't relate shapes with marines, banelings, or immortal buildings."*
>
> S2MF38: *"Object shapes and names were pretty hard to remember, should have simplified to basic code shapes used in the explanation. Damage powers weren't displayed."*
>
> S2MB26: *"The position of the boxes are wide apart so it takes time to visually go from one box to another. Keeping track of both lanes wide apart is difficult."*

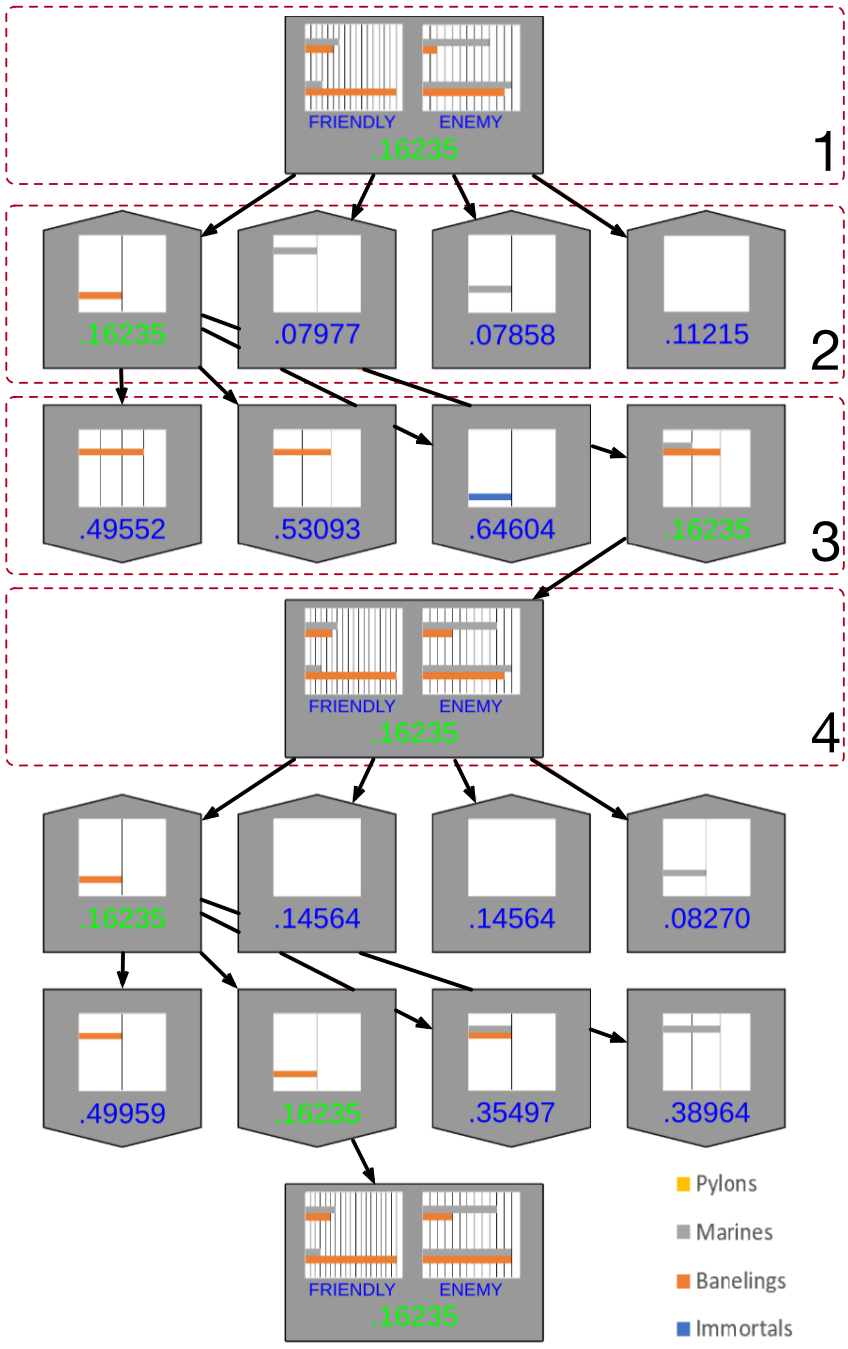
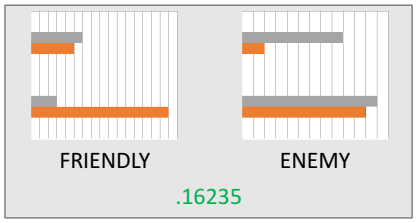
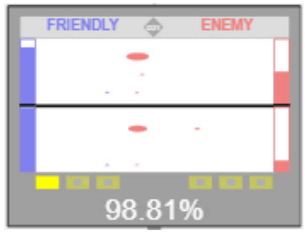


Fig. 5. Search tree explanation for decision point 22 in Study One, presented to participants as a paper prototype. Dashed red boxes show: **(1)** game state at decision point 22, **(2)** top 4 most rewarding actions, as estimated by the AI, **(3)** top 4 most rewarding actions *for the enemy* in response to its "best" action, as estimated by the AI, and **(4)** predicted game state at decision point 23. Our agent searches to depth 2, so the explanation includes another turn of search from the *predicted* state (box 4). Note that all states below the root (box 1) are predicted by the agent. Green highlighted numbers indicate parts of the principal variation.

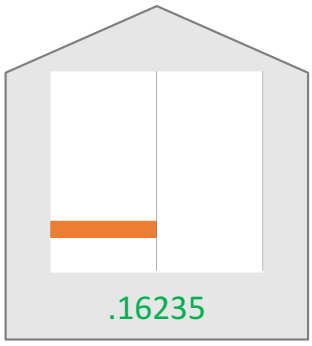


(a) Study One example of State node presentation.

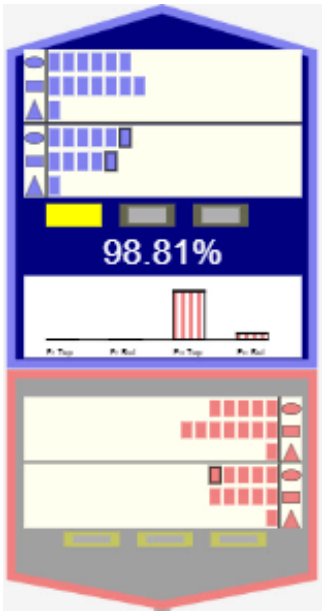


(b) Study Two example of State node presentation.

Fig. 6. In Study One **(left)**, we represented the state with a bar showing a number of unit production facilities for each lane and type. Here, the Friendly AI has 6 marines (gray bar) and 5 banelings (orange bar) in the top lane—with 3 marines and 16 banelings bottom. In Study Two **(right)**, we improved the state representation by including nexus health information via the bars at the edges, as well as pylon count with the yellow/grey rectangles along the bottom. Also the state node, instead of showing troop production facilities, now shows troops that are on the map. This is presented by dividing each lane evenly into four parts, each containing a single shape (oval, square, or triangle) for each type of troop, whose size reflects the number of troops in that part.



(a) Study One example of Action node presentation.



(b) Study Two example of Action node presentation.

Fig. 7. In Study One **(left)**, we used a design similar to states, with bars split by lane and by unit. Each node gives the agent's estimate of the win probability associated with that action (number at the bottom.) In Study Two **(right)**, we improved the action node representation by including both the friendly (top, blue outline) and enemy actions (bottom, red outline) and which lane they are in, with total troop production facilities shown in each lane, and newly acquired production facilities bordered in black. The stacked bar chart illustrates the AI's expectations for likely game outcomes. Each bar shows a nexus's probability of causing a player to lose, with the bar's texture indicating *why* that nexus causes a loss (being destroyed, having lowest health at game end).