Keeping It "Organized and Logical": After-Action Review for AI (AAR/AI)

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
[maithe,khannaro,dodgej,irvine,lamki,linzhe,kiddlen,newmanev,rajasa,mattheca,perdriac,burnett,afern]@oregonstate.edu
Oregon State University

ABSTRACT

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. The results of our qualitative study revealed several strengths and weaknesses of the AAR/AI process and the explanations embedded within it.

CCS CONCEPTS

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

KEYWORDS

Explainable AI, After-Action Review

ACM Reference Format:
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 *25th International Conference on Intelligent User Interfaces (IUI '20), March 17–20, 2020, Cagliari, Italy.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3377325.3377525

1 INTRODUCTION

Consider people tasked with assessing AI systems—specifically those responsible for asserting that the technology is safe and regulation-compliant. An example of such a technology is a self-driving car, where the importance of evaluating its safety is paramount, especially since failures have such grave consequence that they are likely to wind up in court [7]. Assessing accidents caused by self driving cars increasingly tread into legal grey areas. Who is held liable? The driver for failing to react in time, or the company for delivering defective code? [35].

When considering the question of *how* to do assessment, we note that an intelligent agent interacts with the world in ways analogous to those of a human agent. Thus, perhaps we could adapt established techniques for evaluating the quality of *human agents* for use on AI. The technique we specifically refer to is the After-Action Review (AAR), devised by the U.S. Army in the mid-70's [33]. The AAR was a success in various branches of military, and has also been adapted for other domains including medical treatments [47], transportation services [31], and fire-fighting [21].

We term our adaptation AAR/AI ("AAR for AI"). AAR/AI is a *process* for *domain experts* to use in assessing whether and under what circumstances to rely upon an AI agent. We envision AAR/AI to be suitable for sequential domains, such as real-time strategy (RTS) games. It contains a series of steps the human takes to evaluate an AI agent and the explanations it provides of its behaviors.

To investigate AAR/AI, 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 an explanation for the model-based agent to show its search tree (Section 3.3). To evaluate the AAI/AR process in the context of this domain, explanation, and agent, we conducted a qualitative study designed to investigate these RQs:

- **RQ1** When using AAR/AI for assessment, what do people need to make good assessment decisions?
- **RQ2** What are the strengths and weaknesses of guiding human assessment in this way?
- **RQ3** What are strengths and weaknesses of search tree explanations, as we have designed them?

2 BACKGROUND & RELATED WORK

There are many papers describing the challenges of evaluating AI systems' quality (e.g. [5, 15]), including specific attacks (e.g. [12]). Rising to meet these challenges, approaches like DeepTest [54] 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. [38], 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. Other approaches have transported software engineering concepts, such as test selection [16, 20] and formal verification [41]. 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.

IUI '20, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7118-6/20/03…$15.00
https://doi.org/10.1145/3377325.3377525

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 [29]. 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 [39]. 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 [10]. 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 [24]. 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) [9]. 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 [18]. These approaches could be used to examine the quality of AI systems, but they do not offer a concrete *process* for human assessors to enact.

2.2 People Explaining AI

Our process has an explanation explicitly embedded within it, so we briefly survey explanation strategies for AI. The primary purpose of explanations is in 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 to predict system behavior [34].

Explanations are also a powerful tool for shaping the attitudes and skills of users. One such example is Kulesza et al.'s proposed principles for explaining (in a "white box" way) machine learning based systems, wherein the system made its predictions more transparent to the user [26], which in turn improved the quality of their participants' mental models. Another study by Anderson et al. [1] provided insights on the varying changes in the mental models of participants with different explanation strategies of an AI agent.

Another direct consequence of altering the mental models of users is the improvement in their ability to command the system. According to a study by Kulesza et al. [27], participants with the most improved mental models were able to customize the system's recommendations the best, accommodating for the explanations that the researchers provided.

Explanations in the domain of AI agents in RTS games have been gaining traction over the years. In a study by Metoyer et al. [32], they present a format where experienced players played while providing explanations to non-RTS players. The strategy that expert players used while demonstrating how to play the game was found to be key to the explanation process. The study by Kim et al. [25] had experienced players play against AI bots in order to assess the bot's skill levels and overall performance. However, despite the existing research mentioned above, there is a dearth in literature concerning what humans really *need* in order to understand and assess such systems [37].

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 [55].

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) [48]. 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 [46], to encourage objectivity [31]. The purpose is to understand what happened in a situation and give feedback, so people can meet or exceed their performance standards by going through a 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 session for further analysis. The leader enters the session with a pre-planned mechanism to collect data and begins the session 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 scrutinizing any shortcomings in performance. This is followed by brainstorming solutions to avoid or improve responses to problematic outcomes. The session concludes by delineating an action plan to adhere to for future training [55].

	AAR Debrief Steps	AAR/AI Questions 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 three rounds?"*.
AAR/AI Inner Loop	4. Identify what happened	What actually happened?	The participant watched three rounds. Then, we asked, *"Could you briefly explain what actually happened in these past three rounds?"*.
AAR/AI Inner Loop	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 for that decision (Figure 1), and requested they, *"Think aloud about why the Friendly AI did the things it did."*.
AAR/AI Inner Loop	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 steps (left column). The "Empirical Context" column explains how we realized it in our empirical study. 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 here.

AAR showed effectiveness for combat training centers [46], and the military still uses it, with a recent investigation of current methodologies for simulation-based training [17]. Outside military applications, AAR has been used in other domains, from medical treatment [42, 47], emergency preparedness [8], and response [21, 28]. The closest research to ours discusses how AAR will be different for manned-unmanned teams, but focused on the technologies needed to support the AAR process, not the process itself [4].

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 [47]. 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 reviewing, learning the AI's behavior, and assessing its suitability [48]. 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 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.

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?*", "*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 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: Explanation Component

AAR/AI evaluators, like the AAR equivalent, require information on what happened, so our process requires an 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 our evaluation study, we used a model-based agent, so we prototyped a model-based explanation.

A model-based agent (and its explanation) offers the benefit of explicitly representing the future states the agent is trying to reach or avoid. Our model-based explanation captures the agent's search tree, shown in Figure 1. We described the search tree to participants

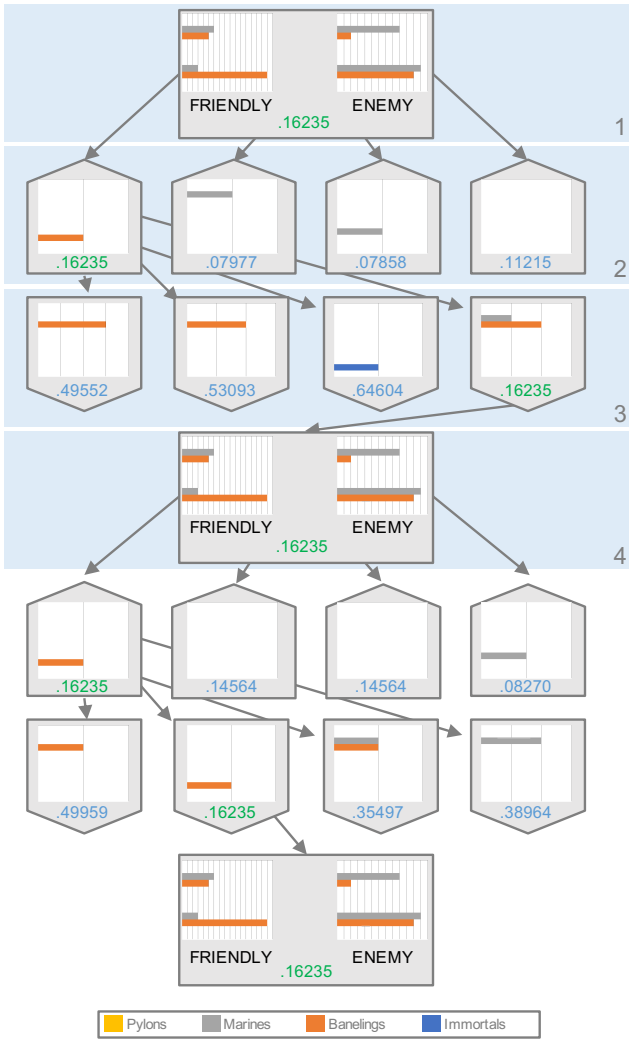

Figure 1: Search tree explanation for decision point 22. Blue background boxes show: (1) game state at decision point 22, (2) top 4 most rewarding actions, as estimated by the AI, (3) top 4 most rewarding actions *for the enemy* in response to its "best" action, as estimated by the AI, and (4) predicted 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.

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" [51] 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 Figure 1, the root node (region 1) shows the current game state and its estimated value. One layer down (region 2) shows the 4 best

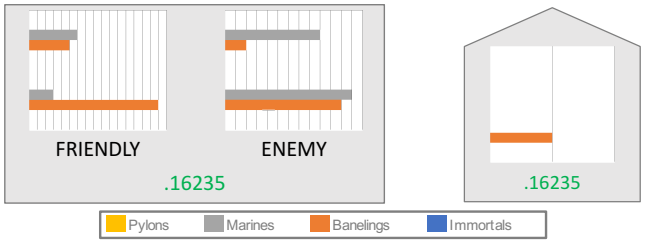

Figure 2: *Left:* An example of State node presentation. Each bar shows a number of unit production facilities for each lane and type. Here, the Friendly AI has 6 marines and 5 banelings in the top lane—with 3 marines and 16 banelings bottom. *Right:* An example of Action node presentation. Similar to the state, bars are split by lane and by unit. Each node is given with the agent's estimate of the win probability associated with that action (number at the bottom.)

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 with more detail in Figure 2. 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.

In AAR/AI then, the evaluators' mission is to evaluate one aspect of the AI agent's theory: its falsifiability [40]. We explain our strategy for doing so next.

3.4 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 [55], though the information collected is 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 [52]. 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 process or explanations. Further, artifacts like these can assist in comparing the assessment results from multiple different individuals.

4 EMPIRICAL STUDY: METHODOLOGY

To inform our design of AAR/AI, we ran an in-lab think-aloud qualitative study. One goal was to investigate what participants needed when doing AI assessment, alongside strengths and weaknesses

of our process. Additionally, since the AAR/AI process embeds an explanation, our other goal was to obtain feedback about the model-based explanation strategy we described in Section 3.3.

We recruited 11 students at Oregon State University who had not taken classes in AI or ML. 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 facilitated for the participant (assessor) during the AAR/AI process, starting with a tutorial about the interface, domain, and task (Steps 1/2). Since each session was limited to 2 hours, we wanted to ensure that each participant reached the end of the replay and had time for our post-task questionnaire. Thus, we decided to have them analyze every third decision point out 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 study.

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 3.4) to help them formalize thoughts and offer the option to refer back to previous ones.

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) [47].

4.1 The Domain

StarCraft II is a popular Real-Time Strategy (RTS) game that offers hooks for AI development ([56, 57]) and a flexible engine for map creation[1]. The game used for this study is a tug-of-war like customized game based on StarCraft II, shown[2] in Figure 3. The objective of the game was to destroy either of the opponent's Nexus in the top lane or bottom lane . If no Nexus is destroyed after 40 rounds, the player whose Nexus has the lowest health will lose.

At every round 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) which will exist for the next round, 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
  (3) Players do knot 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)

[1]Many map creation resources are available at places such as [53].
[2]Materials to replicate this state are freely available in our Supplementary Materials.

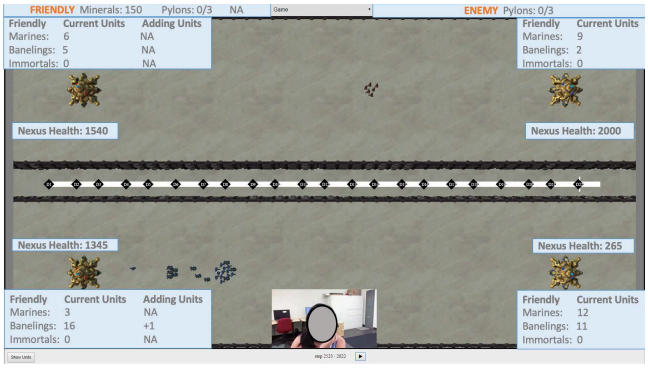

Figure 3: Game screen at decision point 22. Note the text boxes offering state information (current units, nexus health, etc) as well as action information (adding units).

Each round, 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, who 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 per round. 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, meaning the branching factor is combinatorial with respect to minerals possessed.

Once a unit spawns, the players can no longer control it; 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

The agent is model-based, so 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 move tree, and perform minimax search[3] 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 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 [49, 50]). A big difference, however, is that our system uses a learned transition model, due to the stochastic and complex

[3]For more information on game tree search, see Russell and Norvig, Chapter 5 [45].

Code: Description	Example	#
Explanation Overall Quality: Participant found the explanation useless or helpful in a general sense (very vague), or in determining reasons for actions in the decision process (clarity, or lack thereof).	P2: *"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.	P17: *"The color coding is okay. Um, it's pretty distinctive. Um, I don't know if the background is gray or- and 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.	P18: *"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. wishes to interactively expand a node, request a specific action be examined, or have a "wider/narrower" tree) OR thinks it contains the right amount.	P11: *"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	P10: *"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 3: Helpful/Problematic code set for the *explanations*. Frequencies are from three post task questions centered on the explanation and its contents.

nature of the transitions between states, whereas Silver et al.'s used a perfect move-transition model (e.g., Chess's deterministic rules).

4.3 Analysis Methods

To answer **RQ2** and **RQ3**, two researchers applied content analysis [19] 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 Table 3. The two researchers coded 21% of the data corpus separately, achieving inter-rater reliability (IRR) of 82.4%, computed via Jaccard Index [22]. Given this level of reliability, they then split up the remaining coding.

To answer **RQ1**, we drew from a code set that Dodge et al. used in their StarCraft II study, who had adapted from Lim et al.'s work [10, 30]. Dodge et al. also added in 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%. Given this level of reliability, they then split up the remaining coding.

5 RESULTS

Our explanation strategy consists of three components: the AAR/AI process itself, the specific explanation content and presentation, and a "keep the user active" tactic to facilitate their learning of the agent's behaviors. Accordingly, this section has three parts: 1) how the AAR/AI process affected participants' understanding, 2) how the Explanation (tree diagram) content and presentation affected participants' understanding, and 3) how the integration of all three elements of our strategy affected participants' understanding.

5.1 Results: The AAR/AI Process

The goal of our project's explanation strategy is to enable participants to understand how the AI agent is "thinking" well enough to evaluate how suitable the AI agent is for different situations that arise. In essence, our explanation strategy aims to help people build mental models. In this subsection, we consider what the AAR/AI process itself brought to our participants' mental-model building.

Many of the participants commented on how AAR/AI's "structuredness" helped their understanding by keeping their thinking organized, structured, and/or logical. (Only 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:

> P8: *"Uh, yes, I would say <AAR/AI was helpful>. It definitely directed me towards what I should be paying attention to."*

> P18: *"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."*

> P14: *"...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."*

> P17: *"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 participants' mastery of understanding the agent, we applied Bloom's Taxonomy [3], which is a framework used by educators to categorize the different levels of learning. The taxonomy has six levels [2], 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 4.

As Table 4 shows, subsets of participants showed mastery of every Bloom's level. In fact, all participants achieved Bloom's Level 5 at least once during the study. Further, all except one of the participants 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 a form of problem-solving—working out whether the AI agent is "capable enough" for a particular situation—and problem-solving greatly benefits from diversity of thought [14]. Although most research into diversity of thought is in the context of team problem-solving, at an abstract level it amounts to bringing diverse perspectives to a problem (e.g. [14]).

Level: [3]'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.)	+P20: *"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.)	+P8: *"...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.)	+P2: *"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.)	+P2: *"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."* +P19: *"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.)	+P5: *"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."* +P20: *"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.	+P14: *"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..."* +P21: *"I would consistently save a small quantity of minerals each round, rather than trying to save them all in a single round."*

Table 4: Bloom's taxonomy levels participants achieved in learning the agent's behavior.

To consider whether the AAR/AI process was able to elicit diverse perspectives from our participants, we turned to the Lim-Dey intellibiligy types, which we used as a codeset for our qualitative coding (Table 5). As the results show, each of AAR/AI steps guided participants' thinking (according to their self-reports) toward different Lim/Dey perspectives [29]. For example, the first question guided most participants to focus on "What Could Happen," the second on "Input" and "Output" types of information, and the last on "How To" information. Since other research has shown each intelligibility types has its own advantages and disadvantages, we see the diversity of perspectives that AAR/AI seemed to elicit as a particular strength of AAR/AI [2].

5.2 Results: Explanation Content and Presentation

The tree diagrams provided participants with a more 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 aptly:

> P2: *"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'."*

> P14: *"<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 in a post-task response from another participant:

> P17: *"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.

> P21: *"At this point, I feel certain that the friendly's trying to destroy the bottom nexus of the enemy. "*
> P10: *"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:

> P20: *"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."*

Part of P20's complaint above also was a desire for more information, and this issue arose in multiple ways. One participant wanted the explanation to show an estimation of the resources available to both the friendly AI and its opponent:

P20: *"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 is the enemy's resources as well and how to manage those as well as your own."*

5.2.1 How Much More/Less/Different to Show? Addressing the previously described requests for more or different information is not straightforward. With the agent considering combinatorial action spaces, showing the full search tree all at once 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?

To situate the "which" question, the explanations participants saw showed only four actions (recall Figure 1). Some participants thought there should be more and/or different ones. For example:

P5: *"... 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:

P5: *"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:

P20: *"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..."*

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.

P8: *"I liked the way it is. It's easy to read."*

P21: *"I do not have any problem with narrow diagram... "*

P11: *"I would just have more options available... "*

Finally, one wanted everything—which is of course an infeasibly large amount of information to present statically, but might be possible to at least navigate via dynamic mechanisms:

P5: *"All the possible actions and all possible outcomes."*

5.2.2 Explanations as Axioms and Theorems. In the explanation trees, leaf nodes used a neural network to evaluate the quality of states. These estimates were, in essence, axiomatic and the minimax search that proceeds atop those values are akin to theorems. Thus, if the axioms hold true, then the theorems were true. Some participants were open to "grant the axioms."

P14: *"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.

P10: *"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."*

In fact, one participant identified the issue quite well: that the win probabilities have no clear provenance.

P8: *"... 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 P5 and P20 being particularly critical of the agent's decisions:

P5: *"Wow, rewards went down... A baneling is better than a marine by rewards points, but there's clearly a better answer."*

	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 5: Lim Dey coding of participant responses, sliced by question asked during the AAR/AI.

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

P14: *"Information didn't always line up with what occurred. Therefore, it gives a false belief on what/how the AI is doing."*

5.3 Results: Combined Explanation Strategy

Some results seemed directly tied to the integration of all three aspects of our explanation strategy: the AAR/AI process to provide structure, the tree diagrams to provide content, and the tactic of keeping the user active along the way to encourage engagement.

5.3.1 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 [13]. Metacognitive activity is well-established as an important influence on learning and understanding [58].

Our participants showed several instances of metacognition that seemed to come from the integration of AAR/AI, the tree explanation, and the "active user". For example:

P5: *"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."*

P8: *"...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."*

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

P10: *"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"*

.

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

P19: *"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."*

	"The degree to which..." [51]	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 5. 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 6: Applying Sjøberg et al.'s Evaluation Criteria for Theories [51] to the agent's model-based explanation

5.3.2 Falsifying the Agent's Predictions. One of the strengths of the model-based explanations was that it made part of the search tree explicit and that the agent made concrete predictions about the future. However, we observed that this allowed participants to falsify [40] those predictions:

> P14: *"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."*

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

6 DISCUSSION

6.1 Future AAR/AI implementations

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:

> P20: *"... I am fairly confident in my ability to remember what occurred."*
>
> P5: *"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. [39]. In that study, 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.

6.2 Prediction as Explanation

6.2.1 Trend 1: People used explanations as prediction tools. Reed et al. suggested that explaining a solution to a problem helps people to solve similar problems [43]. 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:

> P8: *"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."* [23].

6.2.2 Trend 2: The process of predicting the actions, and then showing the actions, was powerful. Another trend we observed is 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 [6]. "Good" students learn with understanding the material and forming self-explanations on their own, while "poor" students rely heavily on examples to learn and struggle 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:

> P10: *"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" [11] 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 [44]. 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.

6.3 The agent's explanations as theory

Recall from Section 3.3 that the agent's explanations are its explanatory theory of the game. Since it is a theory, we draw upon criteria that can be used to evaluate theories [51]. In Table 6, 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.

7 THREATS TO VALIDITY

Any study has threats to validity, which can skew results towards particular conclusions [59].

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. 46% of 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 [22], and 80% is considered good agreement, but for one code set we achieved only 76%. Other hindrances to the generalizability of our findings include the small size of our study and circumscribed design.

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

8 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 a qualitative in-lab study 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 people" (e.g. [1]), and suggests allowing people to access different actions and/or explanation types on demand.
- *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 [14].
- *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 four most promising options, but some participants wanted to see options the AI considered *bad* as well. 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. [36]) 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.

While AAR/AI was useful in guiding participants to think logically, adding explanations assisted participants in the overall assessment process. Notably, developing useful explanations and rigorously measuring their quality remains quite difficult. We hope that, by appealing to educational frameworks (e.g. Bloom's Taxonomy), we can help people like P14 see *"the flow of logic that we **should've** had"*, a benefit we hope our process will be able to extend to others tasked with assessing AI systems that impact us daily.

9 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] 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.

[2] 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.

[3] 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.

[4] 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 *(AHFE '19)*. Springer Nature Switzerland AG, Cham, CHE, 119–130.

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

[6] 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

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

[8] 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

[9] 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.

[10] 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 *(CHI '18)*. ACM, New York, NY, USA, Article 562, 12 pages.

[11] 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 *(IUI '19)*. ACM, New York, NY, USA, 263–274. https://doi.org/10.1145/3301275.3302316

[12] 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.

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

[14] 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.

[15] 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

[16] A. Groce, T. Kulesza, C. Zhang, S. Shamasunder, M. Burnett, W. Wong, S. Stumpf, S. Das, A. Shinsel, F. Bice, and K. McIntosh. 2014. You Are the Only Possible Oracle: Effective Test Selection for End Users of Interactive Machine Learning Systems. *IEEE Transactions on Software Engineering* 40, 03 (mar 2014), 307–323. https://doi.org/10.1109/TSE.2013.59

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

[18] 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

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

[20] Sandy H. Huang, Kush Bhatia, Pieter Abbeel, and Anca D. Dragan. 2018. Establishing Appropriate Trust via Critical States. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), 3929–3936.

[21] 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

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

[23] Caitlin Kelleher and Wint Hnin. 2019. Predicting Cognitive Load in Future Code Puzzles *(CHI '19)*. ACM, New York, NY, USA, Article 257, 12 pages.

[24] Man-Je Kim, Kyung-Joong Kim, SeungJun Kim, and Anind Dey. 2016. Evaluation of StarCraft Artificial Intelligence Competition Bots by Experienced Human Players *(CHI EA '16)*. ACM, New York, NY, USA, 1915–1921.

[25] 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.

[26] 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.

[27] 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.

[28] 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/

[29] Brian Lim, Anind Dey, and Daniel Avrahami. 2009. *Why* and *Why Not* Explanations Improve the Intelligibility of Context-Aware Intelligent Systems *(CHI '09)*. ACM, New York, NY, USA, 2119–2128.

[30] Brian Y Lim. 2012. *Improving understanding and trust with intelligibility in context-aware applications*. Ph.D. Dissertation. figshare.

[31] 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

[32] 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.

[33] 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

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

[35] 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

[36] 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

[37] 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

[38] 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

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

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

[41] Luca Pulina and Armando Tacchella. 2010. An Abstraction-refinement Approach to Verification of Artificial Neural Networks *(CAV'10)*. Springer-Verlag, Berlin, Heidelberg, 243–257. https://doi.org/10.1007/978-3-642-14295-6_24

[42] 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

[43] 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

[44] 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

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

[46] 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.

[47] 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

[48] 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

[49] 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.

[50] 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

[51] 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.

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

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

[54] 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

[55] 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.

[56] 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/

[57] 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.

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

[59] 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.