Wait, But Why?: Assessing Behavior Explanation Strategies for Real-Time Strategy Games

Justus Robertson
Kennesaw State University
Marietta, Georgia, USA
jrobe380@kennesaw.edu

Athanasios V. Kokkinakis
University of York
York, UK
athanasios.kokkinakis@york.ac.uk

Jonathan Hook
University of York
York, UK
jonathan.hook@york.ac.uk

Ben Kirman
University of York
York, UK
ben.kirman@york.ac.uk

Florian Block
University of York
York, UK
florian.block@york.ac.uk

Marian F. Ursu
University of York
York, UK
marian.ursu@york.ac.uk

Sagarika Patra
University of York
York, UK
sagarika.patra@york.ac.uk

Simon Demediuk
University of York
York, UK
simon.demediuk@york.ac.uk

Anders Drachen
University of York
York, UK
anders.drachen@york.ac.uk

Oluseyi Olarewaju
University of York
York, UK
oluseyi.olarewaju@york.ac.uk

ABSTRACT

Work in AI-based explanation systems has uncovered an interesting contradiction: people prefer and learn best from *why* explanations but expert esports commentators primarily answer *what* questions when explaining complex behavior in real-time strategy games. Three possible explanations for this contradiction are: 1.) broadcast audiences are well-informed and do not need *why* explanations; 2.) consuming *why* explanations in real-time is too cognitively demanding for audiences; or 3.) producing live *why* explanations is too difficult for commentators. We answer this open question by investigating the effects of explanation types and presentation modalities on audience recall and cognitive load in the context of an esports broadcast. We recruit 111 *Dota 2* players and split them into three groups: the first group views a *Dota 2* broadcast, the second group has the addition of an interactive map that provides *what* explanations, and the final group receives the interactive map with detailed *why* explanations. We find that participants who receive short interactive text prompts that provide *what* explanations outperform the no explanation group on a multiple-choice recall task. We also find that participants who receive detailed *why* explanations submit reports of cognitive load that are higher than the no explanation group. Our evidence supports the conclusion that informed audiences benefit from explanations but do not have the cognitive resources to process *why* answers in real-time. It also supports the conclusion that stacked explanation interventions across different modalities, like audio, interactivity, and text, can aid real-time comprehension when attention resources are limited. Together, our results indicate that interactive multimedia interfaces can be leveraged to quickly guide attention and provide low-cost explanations to improve intelligibility when time is too scarce for cognitively demanding *why* explanations.

IUI '21, April 14–17, 2021, College Station, TX, USA

ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450699

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Explainable Artificial Intelligence, Esports, Data-Driven Storytelling

ACM Reference Format:
Justus Robertson, Athanasios V. Kokkinakis, Jonathan Hook, Ben Kirman, Florian Block, Marian F. Ursu, Sagarika Patra, Simon Demediuk, Anders Drachen, and Oluseyi Olarewaju. 2021. Wait, But Why?: Assessing Behavior Explanation Strategies for Real-Time Strategy Games. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3397481.3450699

1 INTRODUCTION

Explainable AI (XAI) is the ability for artificial intelligence systems to explain the hidden reasoning behind behavior to external observers. This ability is especially important for building trust and understanding in the context of human collaboration with black-box machine learning models. Machine learning has made many

recent breakthroughs in mastering complex *sequential decision making systems*, like Atari games [26], turn-based board games [35], and real-time strategy games [27, 38]. As similar sophisticated machine learning models are applied to decision making tasks in real world contexts it will become increasingly important to explain agent behavior to human operators and collaborators. One class of approaches to XAI are *model-agnostic* explanation agents that ignore internal model details in order to generalize their explanations [31]. One of these approaches externally monitors model behavior and explains its decisions with natural language. This approach is very similar to how expert humans explain behavior in sequential decision systems, like a sports commentator who interprets and explains game events to a broadcast audience.

Using this analogy, Dodge et al. studied how expert human esports commentators, called *shoutcasters*, forage for information, synthesize observations, and present explanations to their audience during real-time strategy game broadcasts in order to curate explanation strategies for use in automated XAI systems [12]. Prior work suggests many different types of questions can be answered to make complex systems and behavior intelligible, but *why* questions are the best. In a series of studies, Lim and Dey showed that not only do users primarily request *why* answers [22] but providing *why*-type explanations to participants learning the internal rules of a machine learning model leads to better output predictions, understanding, and trust of the system compared to participants who receive no, *what*, or *how* explanations [23]. Based on Lim and Dey's work, Dodge et al. expected to find that esports shoutcasters primarily present their audiences with implicit answers to *why*-type questions when explaining game events. On the contrary, Dodge et al. found that shoutcasters overwhelmingly answer *what*-type questions the most often and *why*-type questions the least often during live matches. Lim and Dey found that 19% of their participants requested a *why* explanation but only 3% of Dodge et al.'s shoutcaster utterances were *why* answers.

If *why* explanations are desired by audiences and lead to better intelligibility, why do shoutcasters utilize them so infrequently? Dodge et al. offer three possibilities: 1.) well-informed audiences are capable of predicting player actions and therefore do not need game events to be explained; 2.) audiences wants *why* answers but time limits them from consuming complex explanations during live games; or 3.) audiences are capable of consuming *why* explanations in real-time but these explanations are too difficult for shoutcasters to produce live. We test these three possibilities in an online web browser environment with a simple companion application for the game *Dota 2* [37]. *Dota 2* is a multiplayer online battle arena (MOBA) where two teams of five players fight to gather resources and accomplish objectives on a fixed game map. Second screen companions are applications that provide additional commentary, statistics, and analysis alongside a traditional esports broadcast. These companions are an emerging type of application, powered by machine learning models, and have received positive receptions from fans in real-world tournament environments [20]. In our study, we provide a simple companion application in a web browser alongside a *Dota 2* broadcast video to provide interactive text explanation interventions to participants as they watch.

Figure 1 shows our testing environment with a displayed a text explanation intervention. The broadcast video is situated on the left side of the browser window. On the right side of the window, participants are shown an image of the in-game world map. A marker icon appears during certain game events (e.g. battles, player deaths, uncommon actions) on the interactive world map. When the marker is clicked, a text explanation of its game event is shown in a window superimposed over the game map. The text remains onscreen until the participant closes the display window. The point of commentary and explanation is to help audiences comprehend what is happening in a game and why. Formative work in reading comprehension shows there is a link between domain knowledge, game event comprehension, and recall [9]. We use recall to test how well a participant directs their attention and comprehends game events with the aid of different explanation interventions. We separate participants into three groups: 1.) the first group watches a 10 minute excerpt of a *Dota 2* esports broadcast with no map, 2.) the second group watches the same broadcast with an interactive map that provides *what* explanations, and 3.) the final group watches the broadcast with a map that provides *why* explanations synthesized from expert *Dota 2* player feedback. All groups hear broadcast audio with shoutcaster commentary. After the 10 minute session, we ask participants to rate their mental effort and answer multiple choice questions about game events.

We expect one of three possible outcomes for this experiment, one for each of the three participant groups outperforming the others on the recall task along with expected results from the cognitive demand reports:

Possibility 1 If well-informed audiences do not benefit from additional explanations, there are two possible outcomes: **Possibility 1.1** If additional explanations have no effect on well-informed audiences, there will be no differences between the recall task outcomes. **Possibility 1.2** If additional explanations distract well-informed audiences with unnecessary information, there will be higher performance in the group without interactive explanations. We would expect higher reports of cognitive load in both interactive groups.

Possibility 2 If well-informed audiences benefit from explanations but *why* answers are too cognitively demanding to process during a live match compared to other explanation types, we expect the group with interactive prompts and *what* text explanations to outperform the other two groups on the recall task. In this case, we also expect higher reports of cognitive load from those who receive *why* explanations.

Possibility 3 The final possibility is that well-informed audiences are under-served by shoutcasters, who have limited ability to produce *why* explanations. If the audience benefits from real-time *why* explanations, we expect participants who receive *why* interventions to outperform the other two groups on the recall task. This may or may not be accompanied by higher reports of cognitive load.

We recruited 111 *Dota 2* players from the crowdsourcing website Prolific to participate in our study. We find that participants given interactive prompts and *what* explanations outperform the no explanation group on the recall task. We also find that participants given *why* explanations report higher cognitive load than the no explanation group. Both of these results are predicted by **Possibility 2** and support the conclusion that well-informed audiences benefit

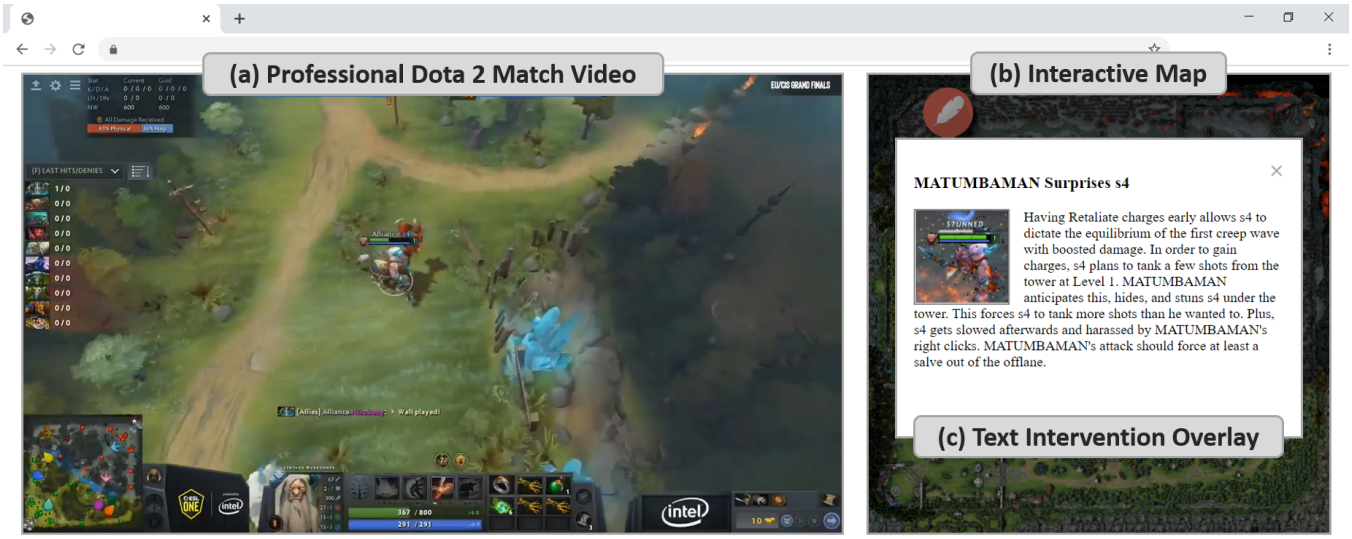


Figure 1: Our online interactive testing environment running in the Google Chrome browser with its three components labeled: (a) a video recording of a professional *Dota 2* esports match, (b) a map with interactive icons that appear according to game events, and (c) a map overlay that provides text interventions to the viewer when icons are clicked. The text intervention shown is a *why* explanation synthesized from expert *Dota 2* player feedback.

from explanations but *why* answers are too cognitively demanding to process when time is scarce. This possibility is anticipated by Dodge et al. who discuss at length how shoutcasters may use *what* and *what-could-happen* answers to approximate *why* explanations during live broadcasts. Our results suggest multimedia interfaces can be leveraged to provide low-cost interventions that direct attention and summarize events when time-scarcity makes *why* answers prohibitively expensive. These interventions, when paired with traditional broadcast media, result in better recall for well-informed audiences than the broadcast alone.

The rest of this paper provides a detailed account of related work, presents the experiment design, provides the study results, and discusses the broader impacts and future directions of our work.

2 BACKGROUND AND RELATED WORK

In this section we present a detailed account of recent machine learning advances in *sequential decision making systems*, the particular structure and challenges of *Dota 2*, an explanation of *intelligibility types*, and how a branch of *explainable artificial intelligence* explains black-box machine behaviors through natural language.

Machine learning has recently made tremendous advances in many areas, including image recognition [21, 36], language modeling [6, 11, 24], and novel artifact generation [16, 39, 42]. Some of the most impressive advancements have been in the context of *sequential decision systems*, where the outcome of past choices influence future states and decisions. Deep neural models have learned to play at human or better levels in classic Atari games [26] without hard-coded domain knowledge or supervised training. *AlphaZero* achieved a similar feat on the classic games Chess, Shogi, and Go [35]. Many experts believed it would be years before any artificial agent performed at a human level in Go, so these results helped capture the public's imagination and led to a successful documentary film, *AlphaGo* [19].

Recent work in sequential decision systems has shifted focus to digital strategy games. In 2019, *AlphaStar* [38] defeated an expert human player in the real-time strategy game *StarCraft II* [5]. Later that year, *OpenAI Five* [27] became the first AI system to defeat reigning world champions at an esports game by beating Team OG in *Dota 2* [37]. *Dota 2* is a multiplayer online battle arena (MOBA) game where two teams of five players battle over resources and objectives in a fixed arena. The game is a complex, difficult challenge due to its long time horizons, partially-observable game states, high-dimensional action space, large pool of playable characters, and the need for team coordination. The game has a large competitive user base and is actively played by full time professionals in regular events and tournaments with prizes worth millions of dollars [30]. *Dota 2* has also been used as a testbed for second screen data-driven storytelling applications [3, 20], similar to how our study environment delivers interactive text explanation interventions on a companion map. We chose to use *Dota 2* as our test game because of its applicability to high-profile machine learning systems, complexity, strong user base, professional scene, and use in second screen applications.

Our study is based on *intelligibility types* created by Lim and Dey for an investigation into the information demanded by users from context-aware intelligent systems [22]. These intelligibility types describe different types of questions and answers that help explain the inner-workings of black box systems to outside observers. The types include *why* (e.g "Why did the system do x?") and *what* (e.g. "What did the system do?") questions and answers. Lim and Dey show that users primarily ask *why* questions and that participants who receive *why* answers are best able to predict the future output

of black-box systems [23]. However, while analyzing professional *StarCraft II* broadcast commentary, Dodge et al. discovered that expert humans primarily answer implicit *what* questions when explaining game events and they answer *why*-type questions the least frequently. Dodge et al. pose three alternatives to explain this contradiction between shoutcaster utterances and Lim and Dey's findings: 1.) well-informed audiences understand game events and do not need *why* explanations, 2.) consuming *why* explanations is too cognitively demanding for audiences during live matches when time and attention resources are scarce, or 3.) producing *why* explanations is too difficult for human shoutcasters. Investigating these three possibilities is the central contribution of this paper.

Understanding these divergent explanation strategies and their effects on esports audiences is useful for explaining both human and black-box AI behavior. Our work will be especially helpful for explanation approaches that utilize multimedia, interactivity, and natural language explanations for intelligibility in live situations where time and attention are scarce. Ehsan et al. validated an approach to generating natural language *rationales*, direct explanations presented from the acting agent's perspective, to explain black-box agent actions [13]. The rationales were created from a corpus of human players performing a think-aloud activity while playing an arcade game. The think-aloud utterances were annotated with state and action information, then a neural translation model was trained on the corpus to generate novel natural language rationale utterances from game states. The generated natural language rationales outperformed a baseline on ratings of *Confidence*, *Human-Like*, *Adequately Justified*, and *Understandable* with human judges when observing AI play traces of the arcade game alongside different text rationales. In general, many XAI studies test for user trust in the model to make correct decisions [18, 40] and user confidence in the decision-making process [1, 17]. Conversely, our study measures intelligibility to an outside observer of a complex, multi-agent system where players act and react to each other while pursuing adversarial goals in real time. We use recall, which has been used to test for comprehension [9] in the context of competitive sports, to measure intelligibility. While other work explores explanation strategies for game players, this paper focuses on strategies to explain game events to informed viewers.

3 STUDY DESIGN

This section presents an overview of our investigation, the online tools we created to perform the study, how match and explanation content was curated, participant demographic and recruitment information, and the procedure used to perform our study.

3.1 Overview

Our study investigates the three possibilities posed by Dodge et al. to explain the contradiction between their expectations given Lim and Dey's work on *intelligibility types* [22, 23] and actual observations of utterances made by shoutcasters during live *StarCraft II* broadcasts [12]. Dodge et al. expected human experts to use *why* explanations, but found shoutcasters primarily answer *what* questions and answer *why*-type questions the least often. The researchers offer three possible explanations: 1.) well-informed audiences are capable of predicting game actions and do not need *why* explanations, 2.) time limits the audience's consumption of complex *why* explanations during live games, or 3.) it is too difficult for shoutcasters to synthesize *why* explanations in real time. The purpose of our study is to answer this open question. To this end, we create a second-screen companion application to augment an existing esports broadcast with additional text explanations. We recruit active *Dota 2* players through the crowdsourcing platform Prolific and split them into three groups that correspond to the three possibilities. **Group 1** receives a 10 minute excerpt from the start of a *Dota 2* broadcast. **Group 2** receives the broadcast along with an interactive map that displays text explanations. Icons appear periodically on the map and display short *what* event summaries when clicked. **Group 3** also receives the broadcast and map, but are given novel *why* explanations curated from expert *Dota 2* players. After their session, each group is asked to gauge the cognitive effort they expended, answer screener questions to determine whether they were paying attention, and then take part in a short multiple-answer quiz to measure their recall of game events.

3.2 Online Tools

Our study was conducted online. We used the crowdsourcing platform Prolific to recruit our participants. Survey and quiz data was collected with Qualtrics. The test environments were engineered with HTML and JavaScript. Figure 1 shows the test environment. The environment has three main components: a *Dota 2* esports broadcast video, an interactive map, and an explanation overlay window. The video is played start to finish with broadcast audio of two shoutcasters commenting on game events. The broadcast was provided by an embedded HTML 5 video with disabled controls so users could not pause or scrub the video without modifying their settings. The video is initially paused and an HTML button below the video begins the session and hides itself when clicked. JavaScript is used to track the video's current time mark and make decisions based on its position. A second HTML button appears 10 seconds before the video ends that allows the user to return to the Qualtrics survey. All three groups receive the same broadcast video, start button, and end button. In addition to the broadcast video, **Groups 2** and **3** also receive a map that displays interactive explanations. Certain events during the match trigger a marker icon to appear on the map. Each icon's location on the interactive map corresponds to the location of the game event it explains.

There are 12 game events with markers that are spaced at roughly 30 second to 1 minute intervals throughout the broadcast. Figure 3 shows the distribution of explanation events over the 10 minute broadcast clip. Markers can be clicked by the user and remain on screen for 1 minute after they first appear. When a marker is clicked, the icon fades away and an overlay window appears on the map. The window contains a title, thumbnail image, and text explanation of a game world event. **Group 2** participants are given a *what* explanation that mirrors shoutcaster commentary. **Group 3** participants are given novel *why* explanations curated from expert *Dota 2* players. Figure 2a shows a *what* explanation and Figure 2b shows a corresponding *why* explanation. The two explanations appear at the same time, describe the same sequence of events,

(a) A Group 2 *what* explanation.

(b) A Group 3 *why* explanation.

Figure 2: Two images of explanation windows overlayed on the interactive map. Figure 2a shows a *what* event summary explanation given to Group 2. Figure 2b shows a corresponding curated *why* explanation of the same event given to Group 3. Each explanation appears in a popup overlay when an icon is clicked and includes a title and image. The image and title for each event are shared between Group 2 and 3, but the explanation text is different.

have the same title and thumbnail image, but contain different text interventions.

3.3 Match Selection

Online adversarial multiplayer games, like *Dota 2*, regularly undergo rule updates, game balances, and the addition of new content. These updates are called *patches* and can drastically alter the way an esport is played professionally in a short period of time. The dominant strategies for a particular patch are called its *meta-game* or *meta*. When selecting a professional match to use in the study, we wanted to make sure the patch was as recent as possible so that our participants, who are all *Dota 2* players, were familiar with the patch rules and meta strategies. We also wanted to ensure the shoutcaster commentary was as high quality as possible. To strike a balance between these two concerns, we used the first game from the championship series of the most recent professional *Dota 2* tournament. The game features the teams *Alliance* and *Team Secret*. We chose to show the first 10 minutes of the match so that the audience could both watch from the beginning of match play and a number of important events could take place without our participants having to view a full 45 minute match.

During the 10 minute clip, we provide explanations on the interactive map for 12 game events. We wanted participants to have plenty of time to read the explanations without feeling rushed, but also to provide a consistent stream of new events. We aimed to provide an explanation every 30 seconds to 1 minute. We allow the event icons to remain on screen for 1 minute after they first appear. Once an icon is clicked, the explanation text remains on screen until the overlay window is closed by the participant. We record which icons are clicked and how long text box overlays remain open. The distribution of events over the 10 minute video is shown in Figure 3. The average time between explanation events is 46.42 seconds, the minimum time is 11 seconds, and the max is 90 seconds. We chose to explain events with game importance that the shoutcasters directly focus and comment on in the broadcast. We highlight three types of events: 1.) **skirmishes** where players on two teams trade blows and take damage; 2.) **lane updates** where unfolding strategies, the status of duels, or unexpected game actions are covered; and 3.) **deaths** where one player is killed by a member of the other team. In total, we explain 3 skirmishes, 4 lane updates, and 5 deaths.

3.4 Explanation Curation

To create **Group 2**'s explanations we summarized shoutcaster commentary about the event, focusing on *what* information. The summary explanations are one or two sentences long and average 21.75 words per explanation. The longest *what* summary is 26 words long and the shortest is 18 words. **Group 3**'s *why* explanations are over three times longer than the *what* explanations, at 75.33 words on average. The longest *why* explanation is 95 words long and the shortest is 48 words. A recent meta-analysis of reading rate suggests that average silent English reading takes place at 238 words per minute [7], so both *what* and *why* explanation sets are

Figure 3: The distribution of 12 explanation events over the 10 minute video. The grey horizontal bar represents the video's timeline and the 12 red vertical bars represent the relative positioning of explanation events on the timeline.

well under the minimum threshold needed to read each explanation, given the icon timeout window of 60 seconds and average buffer time of 46.42 seconds between events. Even though all explanations can be comfortably read in the allocated time, this difference in explanation length likely contributes to the increased cognitive load reported by the *why* explanation group.

To generate *why* explanations we devised a curation task and recruited 3 expert *Dota 2* players to participate. *Dota 2* rates skill and matches players together in online games based on a metric called MatchMaking Rating (MMR), similar to the ELO system used to rate Chess players [14]. The system awards or deducts a small amount of MMR based on whether a player wins or loses each game. Our experts were rated 7830, 6100, and 5690 MMR at the time of the task. Their average MMR of 6540 places them in the highest MMR tier, *Immortal*, which is the top 0.78% of Season 4 *Dota 2* players according to a nearly 4 million player sample [25]. Figure 4 shows the distribution of players per tier during Season 4 of *Dota 2*. The average MMR of our expert reviewers falls in the top bucket.

Dodge et al. describe natural language *why* explanations as those that connect two different time slices together and report the effect of an action at a particular time on an outcome. In our task, we asked expert *Dota 2* players to create *why* explanations for each of our 12 game events. We gave them written instructions that defined an *action* as something a player does, a *state* as a configuration of the game world at a particular time, and a *why* explanation as one that connects two or more actions or states in order to explain a causal sequence. The instructions included timestamps and short one sentence descriptions for each of the 12 game events. The participants were given the replay ID for our game to use in *Dota 2*'s replay viewer, which all three participants were familiar with. *Dota 2*'s replay viewer does not contain broadcast commentary and allows users to freely navigate game time and space. Participants were instructed to use the replay viewer to provide *why* explanations for each of the game event timestamp and description pairs provided. Once the *why* explanations were collected from the experts, we coded each sentence of each explanation according to what information the sentence was communicating. For each event, we compared the three explanations according to their coding and accepted the explanation with the most common features between the three experts. We accepted full explanations instead of synthesizing new ones to maintain coherence and a consistent author voice.

3.5 Participants

131 participants were recruited on the Prolific crowdsourcing platform. We choose to use a crowd-sourced sample because online administration of standard psychology tests have been shown to produce results comparable to traditional paper-and-pencil questionnaires [10] and our experimental tool is inherently online and technology-driven. We choose Prolific as our crowdsourcing platform because Prolific participants have been shown to be more diverse, less dishonest, and provide higher quality data when compared to Amazon Turk [29]. Potential Prolific participants needed three qualities to be considered for our study: recent experience playing *Dota 2*, English proficiency, and a desktop or laptop computer. We recruited potential participants with a demographic survey that asked if they had played *Dota 2* at least twice in the last six months and requested their *Dota 2* rank and MMR. We filtered any participant who gave inconsistent answers between their rank and MMR in the initial survey. We wanted all participants to be active *Dota 2* players but did not control the population for skill, so a range of MMRs would be randomly assigned to each group. All participants were required to be proficient in English to control for comprehension of the text and audio explanations. Finally, participants were required to use a desktop or laptop computer in order to use the online experiment tool. Participants were randomly assigned to three groups (no, *what*, and *why* explanations) at the start of the experiment. Table 1 shows the low, high, average, and standard deviation of MMR for each of the three groups.

The MMR distributions of **Groups 1** and **3** are similar, but **Group 2** is shifted downward and has higher variance. This means that players in **Group 2** are less skilled on average and less densely distributed around the mean than the other two groups. If this randomly-assigned MMR distribution across groups biases results in any way, we would expect a lower recall score from **Group 2** due to lower average skill compared to **Groups 1** and **3**. Fortunately, this negative difference in skill reinforces our eventual result when **Group 2** outperforms **Groups 1** and **3** on the recall task.

3.6 Procedure

Participants are automatically directed from Prolific to Qualtrics where they consent to participate, verify their Prolific ID, and are randomly assigned to one of the three groups. The participant is shown an introduction screen where their task is described. All three groups are introduced to the *Dota 2* game they will watch, including its tournament context and the two teams. Participants assigned to **Groups 2** and **3** are additionally introduced to the interactive map and provided with instructions about how event icons appear, how to activate them, and how to close them once the text commentary has been read. All three groups are then given a link to the external tool. Each group receives a custom link to their version of the experiment environment. The participants then perform their task and are redirected back to Qualtrics for the concluding survey. The survey's cognitive load and multiple choice recall tasks are the same for all three groups.

Upon arriving at Qualtrics, participants are asked to rate the mental effort they expended while watching the match on a nine point scale adapted from Paas and Van Merriënboer [28]. The scale

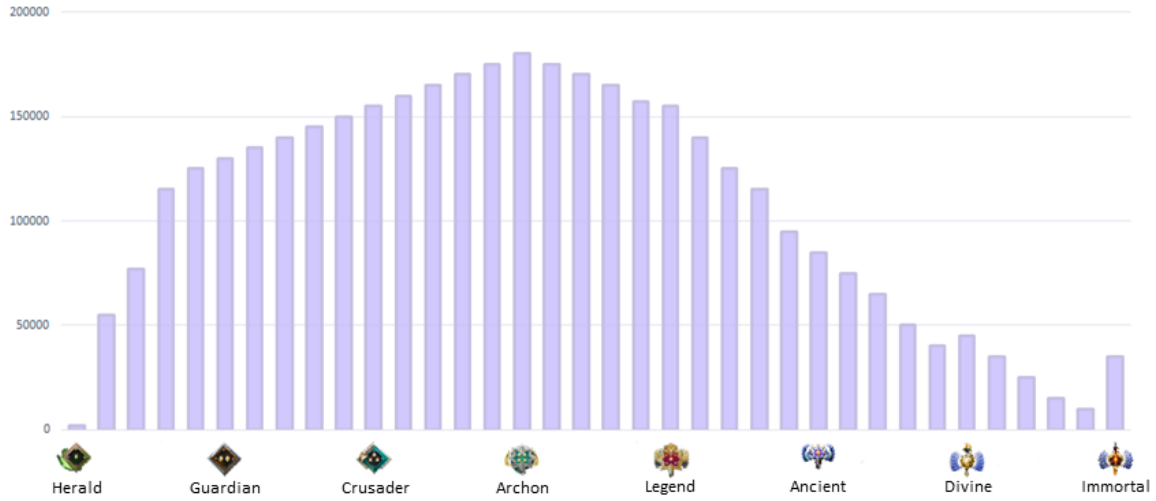


Figure 4: Approximate distribution of 3,938,325 players by MMR tier in Season 4 of *Dota 2* [25]. Each icon at the bottom of the graph represents the start of a tier group. Each tier has five levels before the next tier begins. Each bar in the graph corresponds to a single tier level. The height of the bar corresponds to the number of players in that tier level. The tier on the far left is *Herald 1*, which contains players between 0 and 720 MMR. The final tier is *Immortal*, which contains players above 6000 MMR.

	Low MMR	Low Level	High MMR	High Level	Avg. MMR	Avg. Level	St. Dev.
Group 1	924	*Guardian 2*	7800	*Immortal*	3217	*Legend 1*	1485.78
Group 2	306	*Herald 2*	5720	*Divine 5*	2933	*Archon 5*	1547.87
Group 3	780	*Guardian 1*	7810	*Immortal*	3366	*Legend 2*	1499.07

Table 1: The low, high, average, and standard deviation of MMR for each of the three participant groups. Beside every MMR is the tier and level that MMR belongs to in the distribution shown in Figure 4.

is labeled from *Very, Very Low* mental effort up to *Very, Very High*. Participants are then asked two general recall questions to assess whether they were paying attention during the experimental session. First, they are asked what two teams participated in the *Dota 2* match. This is a multiple choice question with 10 possible answers. Each of the possible answers contains two current professional *Dota 2* teams who are competing in tournaments. Ten total teams are used in the answers and each team is paired in two possible combinations. Only one answer is correct. The second question asks how long the video lasted. It is a multiple choice question with 5, 10, 15, and 20 minutes as possible answers. Users must answer both screening questions correctly and click on at least half of their text events to have their results accepted. This is to ensure they were paying attention and received the interventions we are testing for. Of our 131 participants, 111 met our acceptance criteria.

Finally, we ask each participant to answer seven multiple choice recall questions about the match they watched. Each of the questions corresponds to an event highlighted by an explanation intervention. The seven questions focus on the most unique and memorable of the 12 original events. A question is asked about each of the four lane update events, the first skirmish to take place, the first kill to take place, and a memorable group kill that occurs late in the clip. Each question asks the participant to identify the player at the center of the event. For example, *What player landed the first attack?* Each of the ten players who participated in the game along with their character's name are listed as a possible answer. For example, *MATUMBAMAN (Wraith King)*. The full set of questions is given in Table 2. We chose to ask *what* questions because only one group was given *why* information. Once the participant has answered all seven questions, they are thanked for their participation and automatically directed back to Prolific.

4 ANALYSIS AND RESULTS

We want to answer whether *why* explanations are avoided by human shoutcasters because they are: 1.) not needed by informed audiences, 2.) too cognitively expensive for audiences to consume, or 3.) too difficult for shoutcasters to produce. These three explanations map onto our three groups: **Group 1** no additional explanation interventions, **Group 2** interactive *what* explanations, and **Group 3** interactive *why* explanations. Given these three possibilities and groups, we expect to observe one of three possible outcomes:

Possibility 1 Informed audiences do not benefit from additional explanations. We expect to observe either: **Possibility 1.1** no difference in recall between the three groups or **Possibility 1.2** higher performance in **Group 1**. In the second case, we expect higher reports of cognitive effort from **Groups 2** and **3**.

Explanation Event	Multiple-Choice Question
Event 1	What player landed the first attack?
Event 3	What mid-lane player was shown with a 9-6 last hit advantage over their rival?
Event 4	What Dire player was shown free farming in Radiant territory?
Event 6	What player scored the first kill?
Event 8	What core player is shown teleporting to lane after collecting runes?
Event 9	What player is shown killing a courier?
Event 10	What player is ganked by three enemies in the bottom lane?

Table 2: The seven multiple-choice questions used to test participant recall of game events. Each question had 10 choice options, one for each player participating in the match. We only asked questions about easily-identifiable unique events that receive an explanation. We chose to only ask *what* questions since *why* information is only given to one group of participants.

We decompose this first outcome into two possibilities because there is a chance the interactive map distracts well-informed viewers leading to worse results than if they just watched the broadcast. If this is the case, we expect to see reports of higher cognitive effort in the two groups with distracted participants.

> **Possibility 2** *Why* answers are too expensive to consume during a real-time match. We expect that **Group 2** will outperform the other two groups on the recall task. We also expect higher cognitive effort from **Group 3**.

In this case, informed audiences benefit from explanations but the *why* explanation type requires too many cognitive resources to consume during a live match. If this is the case, we expect limited *what* explanations will best aid recall and consuming *why* explanations will result in the highest reported cognitive effort.

> **Possibility 3** *Why* explanations are easy to consume but hard to produce. We expect **Group 3** to outperform the other two groups on the recall task.

In this case, informed audiences benefit from *why* explanations and can consume them during a live match, but they are hard for human commentators to produce on demand. If this is the case, we expect consuming *why* explanations will best aid recall. *Why* explanations may also increase cognitive effort, but only because other explanation types underutilize audience attention resources.

4.1 Results

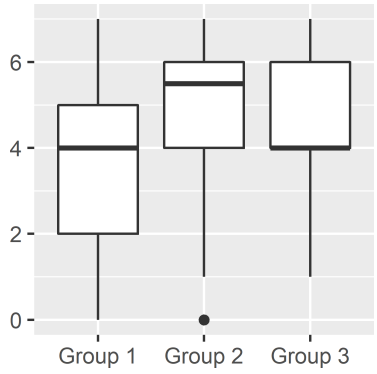
First, we analyze the recall scores. We performed a one-way between subjects ANOVA to compare the effect of interactive text explanation interventions on game event recall for no intervention, *what* explanation, and *why* explanation conditions. We found a significant effect of text interventions on correct recall answers at the $p<.05$ level for the three conditions [$F(2,108)=4.13$, $p=0.0187$]. A post hoc Tukey test showed that the *what* explanation group (M = 5.03, SD = 1.90) is significantly higher than the no intervention group (M = 3.78, SD = 2.16). The *why* explanation group (M = 4.64, SD = 1.69) was not significantly different from the other two, laying in the middle. A boxplot of the recall results is shown in Figure 5a. The results line up with our expectations from **Possibility 2**. The recall test results of **Group 2** are significantly higher than **Group 1**. Meanwhile, **Group 3** is neither significantly lower than **Group 2** or significantly higher than **Group 1**.

Next, we analyze the cognitive effort reports. We expect **Group 3** to report the highest cognitive effort to explain the underperformance of participants who received *why* explanations on the recall task relative to those who received *what* explanations. We performed a second one-way between subjects ANOVA to compare the effect of interactive text explanation interventions on reports of cognitive effort for no intervention, *what* explanation, and *why* explanation conditions. We found a significant effect of text interventions on cognitive effort reports at the $p<.05$ level for the three conditions [$F(2,108)=3.86$, $p=0.024$]. A post hoc Tukey test showed that the *why* explanation group (M = 6.20, SD = 1.79) is significantly higher than the no intervention group (M = 5.10, SD = 1.61). The *what* explanation group (M = 5.65, SD = 1.81) was not significantly different from the other two, laying in the middle. A boxplot of the mental effort results is shown in Figure 5b. Again, the results line up with our expectations from **Possibility 2**. The cognitive effort results of **Group 3** are significantly higher than **Group 1**. **Group 2** is now neither significantly lower than **Group 3** or significantly higher than **Group 1**.

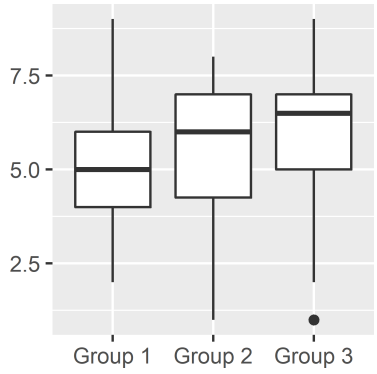
Together, these results support **Possibility 2**, that *why* answers have the potential to be beneficial to informed audiences but are difficult to consume during a real-time match. Dodge et al. discuss this possibility at length, theorizing that expert human shoutcasters use a combination of *what* and *what-could-happen* explanations as a satisficing approximation of *why* answers due to the limited time and attention budget for live audiences. Our results provide evidence that informed audiences indeed benefit from explanations and report *why* explanations as requiring more cognitive effort than the *what* answers given by shoutcasters. These findings can be used by applications that make real-time natural language interventions to explain mechanics, agent motivation, and behavior in complex systems. Providing short, interactive *what* summaries of important events guides audience attention and increases recall without overwhelming users with resource-intensive *why* explanations.

5 DISCUSSION

That informed audiences benefit from interactive explanations when watching an esports broadcast but detailed *why* explanation are cognitively demanding to process in real-time aligns with predictions made by Dodge et al.. These results have implications for automated explanation systems in real-time sequential decision domains with respect to audience intelligibility and recall.



(a) A boxplot of the recall task results.



(b) A boxplot of the mental effort results.

Figure 5: Boxplot graphs of our results. Figure 5a shows the results of the recall task and Figure 5b shows the results of the mental effort reports. The graphs show that Group 2 skews higher than the other two groups on the recall task and Group 3 skews highest on the mental effort reports. These results align with our Possibility 2 expectations.

This section provides a detailed discussion and analysis of these implications.

5.0.1 Second Screen Companion Applications. Our study uses a companion application, situated next to the broadcast video in our testing environment, to display interactive prompts and natural language explanations. Our results will impact how information is generated and presented to audiences in second screen contexts. For example, the Weavr project produced an interactive second screen companion application for *Dota 2* that displays text explanations, real-time statistics, and visualizations for live games on a smart phone device. The application was deployed to 170 people in the context of a real-world professional tournament environment [20]. Participants downloaded the companion application to their smart phone at the start of the tournament, used the application over the course of the weekend, and allowed their usage patterns to be analyzed. 27 participants were further recruited to provide detailed qualitative feedback on their usage over the weekend tournament. The qualitative feedback revealed that learning was a major motivator and many users used the application to observe players they could learn from. This is further evidence that well-informed audiences are not perfect at predicting or understanding complex behaviors from top-tier experts. This matches with our result that providing explanations to well-informed audiences increases intelligibility and recall.

Additionally, participants reported that balancing information density was a major challenge for the application. Similar to our testing environment, the app had a central map that displayed location-based interactive markers. When clicked, these markers presented natural language status updates, similar to shoutcaster *what* commentary, about game state information like item usage, kills, gold collected, etc. Participants reported being overwhelmed by a large number of 'less important' game highlights and would prefer a smaller number of 'important' highlights. Our results indicate that, for intelligibility, these 'important' highlights do not have to be detailed explanations that break apart and analyze game content. If an automated system could replicate our curation method by identifying important events at the same level as broadcast operators and sending a simple *what* summary of that event every 45 seconds, our results indicate it will lead to an increase in average recall of game events when compared to an audience that watches a flat broadcast with no second screen information. Our results also indicate that when designing these real-time systems it is important to make information as easily digestible as possible because overwhelming users with complex explanations in real-time will lead to a decrease in recall back towards the flat broadcast audience.

5.0.2 Automatic Summary Generation. One potential application of this work is to create a second screen companion application capable of generating novel explanations, similar to the *what* summaries curated in our experiment, to increase audience intelligibility. While the explanations employed in our study were human-authored, producing explanations in this way for all games may be too expensive for all but the most well-resourced of applications. One possible solution is to automatically generate *what*-type event summaries. One of the many advances machine learning has made in recent years is in the areas of language modelling and automatic text generation. *Word2Vec* [24] is an algorithm that produces a word embedding model, which is a set of high-dimensional vector representations of words created by analyzing co-occurrence in large data sets. These models are powerful because they embed semantic similarity and can be used for analogical reasoning [15]. For example, many models encode representations that support common sense conceptual arithmetic like *king - man + woman = queen*. These representations have helped drive the development of modern neural language models, like BERT [11] and GPT-3 [6], which are used for a range of language understanding and generation tasks. One of the language tasks these models have been applied to is *machine summarization*. Machine summarization is the task of automatically generating a natural language summary of source content.

There are two main approaches to summarization, *extractive* and *abstractive* summaries. Extractive summarization is the easier approach and the most common. These approaches remove text directly from the summarization source and selectively add it to the summary. Abstractive summarization, on the other hand, generates

novel summary text that does not appear in the summarization source. Work in this area is still developing and deep neural networks have made recent advances [33, 34]. These types of deep neural summarization systems could be used to generate abstractive summaries of shoutcaster commentary around an event selected to be shown to the audience. If live shoutcaster commentary is not available, it could be simulated by a language model trained for *Dota 2* contexts. Another possibility would be to train a system to map short video clips directly to summary explanations, similar to video captioning [8], or produce summaries from game state vector representations, called *moments* [41], and represent interesting narrative properties, like cognitive interest [4]. If trained to summarize directly from image or state sequence information, the model would not need real or synthetic shoutcaster commentary to generate live explanation interventions.

5.0.3 Multi-Modal Explanations. Another line of future work is to generate and assess explanations with types of content other than natural language. Aside from the standard title and icon image, our explanations were natural language text-only. However, there are many ways to explain complex behavior in sequential decision systems other than natural language. Sports broadcasts often use visualizations with statistics and illustrations when introducing or updating storylines. It is likely that intermixing statistics, visualizations, and natural language explanations can create a coherent and low attention-cost framework to convey information to the audience. Recent work by Dodge et al. shifts in this direction, from a model agnostic natural language approach to experiments that measure the impact of explaining reinforcement learning agents with model-specific white box statistics and visualizations [2]. They find that participants shown both saliency maps and reward-decomposition bars, which are visualizations of internal model input and output metrics, attain a higher understanding of the AI model. However, these explanations come at a cost as participants with access to these metrics also report higher levels of cognitive load and predict the agent's next move at a lower rate than participants with no explanations at all. These results again line up with our findings that more intensive explanations are useful but can be hard to process, especially in real-time. Providing metrics and images in addition to natural language explanations can be useful, but they will likely be restricted by the same cognitive resource scarcity that governs real-time natural language explanations. When natural language is paired with statistics or graphics, they must not overwhelm the audience. The Weavr application [20], for example, could be used to provide multimodal explanation interventions that are more effective at indirectly conveying *why* information without overwhelming its audience. Additional types of interventions may be possible in VR environments [32]. Further experiments could test how different pairings or statistics could boost intelligibility in different sequences and circumstances.

5.0.4 Offline Why Explanations and Additional Metrics. We test *why* explanations in the context of real-time sequential decision systems and find they are not as effective as short *what* explanations at increasing user recall because they require more mental effort to process. However, previous work [22, 23] shows that *why* explanations are the most demanded and effective in learning to predict future decisions in offline, non-sequential environments. This is similar to sports contexts where in-depth causal chains are often explored in off-line segments between matches or competitive segments. When building agents that automatically explain behavior, it could still be useful to provide interactive *why* explanations as long as they can be provided during dips in action when the user has more attention and can bring to bear more cognitive effort in processing the detailed explanation. Additionally, there may be other reasons why an audience wants explanations of expert or AI-driven behavior aside from intelligibility. For example, respondents in the Weavr [20] qualitative study indicated they watch in-depth explanation segments because they want to improve as *Dota 2* players. Further experiments could test the effect of different explanation types on metrics beyond comprehension, recall, and intelligibility, like learning or task performance outcomes.

6 CONCLUSION

In this paper we answered an open question in AI-based explanation systems of why expert esports commentators primarily answer *what* questions when explaining complex behavior in real-time strategy games. Using an interactive companion application we tested for recall in participants who receive no additional explanations, summary *what* explanations, and *why* explanations. We find that while informed audiences benefit from additional explanations, *why* explanations are too cognitively demanding and lower performance on the recall task. The broader implications of our work are that providing interactive explanations can increase intelligibility and recall of complex behaviors and events over flat video and audio commentary, but providing explanations that are too cognitively demanding can diminish the effect. Our findings will inform future work in the contexts of explainability and explainable AI in sequential decision making systems to direct audience attention without overwhelming them with high-value but costly *why* explanations.

REFERENCES

[1] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 275–285.

[2] 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 Transactions on Interactive Intelligent Systems* 10, 2 (2020), 15:1–15:37.

[3] Camille Barot, Michael Branon, Rogelio E. Cardona-Rivera, Markus Eger, Michelle Glatz, Nancy Green, James Mattice, Colin M. Potts, Justus Robertson, Makiko Shukonobe, Laura Tateosian, Brandon R. Thorne, and R. Michael Young. 2017. Bardic: Generating Multimedia Narratives for Game Logs. In *Proceedings of the The AIIDE-17 Workshop on Intelligent Narrative Technologies at the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment.* AAAI, Palo Alto, CA, USA, 154–161.

[4] Morteza Behrooz, Justus Robertson, and Arnav Jhala. 2019. Story Quality as a Matter of Perception: Using Word Embeddings to Estimate Cognitive Interest. In *Proceedings of the 15th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment.* AAAI, Palo Alto, CA, USA, 3–9.

[5] Blizzard. 2010. StarCraft II: Wings of Liberty.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[7] Marc Brysbaert. 2019. How Many Words Do We Read Per Minute? A Review and Meta-Analysis of Reading Rate. *Journal of Memory and Language* 109 (Dec. 2019), 94.

[8] Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. 2019. Deep Learning for Video Captioning: A Review. In *International Joint Conference on Artificial Intelligence* (Macao, China) *(IJCAI '19)*. IJCAI, CA, USA, 6283–6290.

[9] Harry L. Chiesi, George J. Spilich, and James F. Voss. 1979. Acquisition of Domain-Related Information in Relation to High and Low Domain Knowledge. *Journal of Verbal Learning and Verbal Behavior* 18, 3 (1979), 257–273.

[10] Robert N. Davis. 1999. Web-Based Administration of a Personality Questionnaire: Comparison with Traditional Methods. *Behavior Research Methods, Instruments, & Computers* 31, 4 (Nov. 1999), 572–577.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[12] 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 *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, 562:1–562:12.

[13] 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 ACM International Conference on Intelligent User Interfaces* (Los Angeles, CA, USA) *(IUI '19)*. ACM, New York, NY, USA, 263–275.

[14] Arpad E. Elo. 1978. *The Rating of Chess Players, Past and Present.* Arco Publications Limited, London, UK.

[15] Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn't. In *Proceedings of the NAACL Student Research Workshop* (San Diego, CA, USA) *(NAACL '16)*. ACL, Stroudsburg, PA, USA, 8–15.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27* (Montreal, CA) *(NIPS 2014)*. Neural Information Processing Systems Foundation, San Diego, CA, 2672–2680.

[17] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, USA) *(CVPR 2018)*. IEEE, New York, NY, USA, 8779–8788.

[18] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. 2018. To Trust Or Not To Trust A Classifier. In *Advances in Neural Information Processing Systems 31* (Montreal, CA) *(NIPS 2018)*. Neural Information Processing Systems Foundation, San Diego, CA, USA, 5541–5552.

[19] Greg Kohs (Director). 2017. AlphaGo. Moxie Pictures.

[20] Athanasios Kokkinakis, Simon Peter Demediuk, Isabelle Nölle, Olu Olarewaju, Sagarika Patra, Justus Robertson, Peter York, Alan Pedrassoli Chitayat, Alistair Coates, Daniel Slawson, Peter Hughes, Nicolas Hardie, Ben Kirman, Jonathan Hook, Anders Drachen, Marian F. Ursu, and Florian Block. 2020. DAX: Data-Driven Audience Experiences in Esports. In *Proceedings of ACM International Conference on Interactive Media Experience (IMX) 2020* (Barcelona, Spain) *(IMX '20)*. ACM, New York, NY, USA, 1–12.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25* (Lake Tahoe, NV) *(NIPS 2012)*. Neural Information Processing Systems Foundation, San Diego, CA, 1097–1105.

[22] Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (Orlando, FL, USA) *(UbiComp '09)*. dl.acm.org, New York, NY, USA, 195–204.

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

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26* (Lake Tahoe, NV) *(NIPS 2013)*. Neural Information Processing Systems Foundation, San Diego, CA, 3111–3119.

[25] Vincenzo "Skulz" Milella. 2020. *Dota Seasonal Rank distribution and Medals - 2020.* Esports Tales. https://www.esportstales.com/dota-2/seasonal-rank-distribution-and-mmr-medals

[26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-Level Control Through Deep Reinforcement Learning. *Nature* 518, 7540 (Feb. 2015), 529–533.

[27] OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. arXiv:1912.06680 [cs.LG]

[28] Fred G. W. C. Paas and Jeroen J. G. Van Merriënboer. 1993. The Efficiency of Instructional Conditions: An Approach to Combine Mental Effort and Performance Measures. *Human factors* 35, 4 (Dec. 1993), 737–743.

[29] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research. *Journal of Experimental Social Psychology* 70 (May 2017), 153–163.

[30] Shaun Prescott. 2020. *Dota 2's The International 2020 has broken its prize pool record, but still no dates for the tournament.* PC Gamer. https://www.pcgamer.com/dota-2s-the-international-2020-has-broken-its-prize-pool-record-but-still-no-dates-for-the-tournament/

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. In *ICML Workshop on Human Interpretability in Machine Learning* (New York, NY, USA) *(WHI 2016)*. IMLS, New York, NY, USA, 91–95.

[32] Justus Robertson, Rogelio E. Cardona-Rivera, and R. Michael Young. 2020. Invisible Dynamic Mechanic Adjustment in Virtual Reality Games. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, New York, NY, USA, 282–289.

[33] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 379–389.

[34] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '17)*. Association for Computational Linguistics, Vancouver, Canada, 1073–1083.

[35] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the Game of Go Without Human Knowledge. *Nature* 550, 7676 (Oct. 2017), 354–359.

[36] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Herve Jegou. 2019. Fixing the Train-Test Resolution Discrepancy. In *Advances in Neural Information Processing Systems 32* (Vancouver, CA) *(NeurIPs 2019)*. Neural Information Processing Systems Foundation, San Diego, CA, 8252–8262.

[37] Valve. 2013. Dota 2.

[38] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning. *Nature* 575, 7782 (Nov. 2019), 350–354.

[39] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In *Advances in Neural Information Processing Systems 29* (Barcelona, Spain) *(NIPS 2016)*. Neural Information Processing Systems Foundation, San Diego, CA, USA, 82–90.

[40] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201.

[41] Xiaoxuan Zhang, Zeping Zhan, Misha Holtz, and Adam M Smith. 2018. Crawling, Indexing, and Retrieving Moments in Videogames. In *Proceedings of the 13th International Conference on the Foundations of Digital Games* (Malmö, Sweden) *(FDG '18)*. ACM, New York, NY, USA, 1–10.

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision* (Venice, Italy) *(ICCV'17)*. IEEE, New York, NY, USA, 2223–2232.