Proceedings of the Nineteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2023)

Distribution Fairness in Multiplayer AI Using Shapley Constraints

Robert C. Gray[1], **Jichen Zhu**[2], **Santiago Ontañón**[3]
[1] Drexel University
[2] IT University of Copenhagen
[3] Google Research / Drexel University
robert.c.gray@drexel.edu, jichen.zhu@gmail.com, santiontanon@google.com

Abstract

Experience management (EM) agents in multiplayer serious games face unique challenges and responsibilities regarding the fair treatment of players. One such challenge is the *Greedy Bandit Problem* that arises when using traditional Multi-Armed Bandits (MABs) as EM agents, which results in some players routinely prioritized while others may be ignored. We will show that this problem can be a cause of player non-adherence in a multiplayer serious game played by human users. To mitigate this effect, we propose a new bandit strategy, the Shapley Bandit, which enforces fairness constraints in its treatment of players based on the Shapley Value. We evaluate our approach via simulation with virtual players, finding that the Shapley Bandit can be effective in providing more uniform treatment of players while incurring only a slight cost in overall performance to a typical greedy approach. Our findings highlight the importance of fair treatment among players as a goal of multiplayer EM agents and discuss how addressing this issue may lead to more effective agent operation overall. The study contributes to the understanding of player modeling and EM in serious games and provides a promising approach for balancing fairness and engagement in multiplayer environments.

1 Introduction

Player modeling and experience management (EM) are critical aspects of modern digital game design, as they enable developers to create engaging and personalized experiences. Multi-Armed Bandits (MABs), a class of reinforcement learning algorithms, have emerged as an effective tool for this purpose (Gray, Zhu, and Ontañón 2020, 2021). There is also a growing body of research on the application of MABs in multiplayer modeling, where a single agent is charged with managing the experience of multiple players simultaneously (Zhu and Ontañón 2019) and which offers a more complex problem due to the need to consider interactions among multiple players.

When dealing with multiplayer EM in serious games, where the primary goal is to facilitate interventions toward specific goals, the focus shifts from simply balancing the game to accommodate player skill to ensuring that the EM agent caters to individual needs within the shared experience. This introduces unique challenges and responsibilities for the agent, particularly where the agent must make decisions that affect multiple players simultaneously. Even beyond the ethical considerations for why such an agent holds a responsibility to provide fair treatments across the player base, doing so in the case of interventional applications may ultimately lead to more effective agent operation overall. For instance, consider the element of player non-adherence, where a player that does not feel they are receiving a worthwhile experience within an activity may reduce their participation, thereby reducing the opportunities through which the intervention can operate with those players and nullifying even the greatest of innovations it may present.

In this study, we examine a natural pitfall in the use of traditional MAB strategies as EM agents, where the agent may fail to properly administer its intervention across the player group in its pursuit of maximizing a particular outcome metric. For instance, a traditional strategy aiming to maximize player physical activity in an exercise game may prioritize its motivational techniques for only the highest-performing players. We call this the *Greedy Bandit Problem*, and we demonstrate its potential to promote player non-adherence in a multiplayer serious game. We then propose a solution for mitigating the effects of this problem via a new bandit strategy, the Shapley Bandit, that enforces fairness constraints in its treatment of players based on the Shapley Value.

It is worth noting that this paper relates to a contemporary publication (Gray et al. 2022) where we validate this approach with human users. Where that paper discusses the human factors and user study evaluation, this paper discusses the rationale and details for the Shapley Bandit AI design and our preliminary validation via simulation prior to human user studies. The remainder of this paper is structured as follows. Section 2 presents literature related to this problem space, followed by our motivating scenario in Section 3. Section 4 analyzes data from a Pretest User Study with human participants, enabling us to identify the Greedy Bandit Problem and offer the Shapley Bandit as a solution in Section 5. We then describe our simulation environment in Section 6, which we use to evaluate our approach in Section 7.

2 Background and Related Work

The following discusses necessary background regarding Multi-Armed Bandits, Player Modeling and Experience Management, and the Shapley Value.

2.1 Multi-Armed Bandits

The Multi-Armed Bandit (MAB) problem is a sequential decision problem (Robbins 1952; Lai and Robbins 1985) where an agent must repeatedly decide which option of a set of available options (called "arms") to select (or "pull") to maximize their expected reward over a series of such pulls. The agent does not know the true reward distribution of each arm and, therefore, must explore by pulling different arms to learn more about the reward probabilities (*exploration*) while spending as many opportunities as possible to pull the arms it believes to yield the highest rewards (*exploitation*).

Several strategies have been developed to address the MAB problem, including epsilon-greedy (Kuleshov and Precup 2000; Lattimore and Szepesvári 2020), upper confidence bound (UCB) (Auer, Cesa-Bianchi, and Fischer 2002), and Thompson sampling (Thompson 1933). Epsilon-greedy is a simple and intuitive approach that balances exploration and exploitation by randomly selecting either the "greedy" action (i.e., the arm that so far has yielded the greatest rewards) or exploring randomly, based on probability determined by the ε parameter. UCB and Thompson sampling are more sophisticated approaches that use Bayesian inference and uncertainty estimates (i.e., confidence intervals) to guide exploration and exploitation.

2.2 Player Modeling & Experience Management

In the domain of player modeling and Experience Management (EM), several studies have explored the use of AI agents to create adaptive games (Bates 1992) that cater to individual player needs, preferences, and skill levels. Yannakakis and Togelius provide a comprehensive overview (Yannakakis and Togelius 2018) of the various techniques employed in AI-driven game design, including player modeling through observations of metrics that predict aspects of the user (Yannakakis et al. 2013; Drachen, Canossa, and Yannakakis 2009). Toward this task, a variety of AI techniques can be deployed toward building player models, such as support vector machines (Missura and Gärtner 2009), neural networks (Min et al. 2016), or MABs (Gray et al. 2020; Vinogradov and Harrison 2022).

EM agents construct a model using the AI for a player and, in the case of software-based interventions like serious games, leverage this understanding of the player to tailor the gaming experience for the player toward its particular goals, such as health outcomes (Fujiki et al. 2008) or learning objectives (Valls-Vargas et al. 2015). The EM agent is able to perform this tailoring through the use of *EM levers* (Gray, Zhu, and Ontañón 2021), or elements within the game environment that provide the agent with opportunities to affect the game state. Essential to our own study scenario and method (Section 3), other studies have found success in using additional characters in the game environment as EM levers (Feltz et al. 2014; Samendinger et al. 2017).

2.3 Shapley Value

Proposed in 1953 by Lloyd S. Shapley (Shapley 1953, 1997), the Shapley Value addresses the problem of attributing individual utility (most often to determine commensurate rewards) for a set of participants operating in cooperation toward a common goal (Winter 2002). Due to overlapping and unseen effects of influence and synergy, it may be difficult to know precisely how much each participant contributed to the performances of others or to what degree others contributed to their own individual performance within the joint endeavor. However, Shapley aimed to find an estimate for this value, proposing that an optimal assessment would meet the following criteria, or axioms (Ma and Tourani 2020; Roth 1988): (1) Symmetry, where completely interchangeable participants should be interpreted as having equal contributions to the team, (2) Nullity, where a participant that adds no value to any sub-division of the overall group is assigned no contribution beyond their individual achievement, (3) Additivity, where no contribution is lost if we were to separate the game into sub-events or rounds, and (4) Efficiency, where the sum of the contribution of all participants equals the total contribution for the full group.

More formally, in a coalition N containing a participant i (among others), considering every possible way that N could form to include participant i would be to consider every permutation of the sub-coalitions (S) of participants in N other than i (i.e., $N \setminus \{i\}$). If the increase in the coalition's utility provided by participant i for each of these sub-coalitions is averaged, it results in the expected marginal contribution that i provides to the coalition as a whole. Shapley summarized this operation with the following equation:

$$\varphi_i(N, v) = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)![v(S \cup \{i\}) - v(S)] \tag{1}$$

Additionally, Shapley provided proofs that this formulation not only achieves all four of the axioms but provides the *only* solution that does so (Roth 1988). Importantly, this axiomatic approach, facilitated by Equation 1, empowers any solution employing the equation to imbue its results with a notion of stability defined by these qualities. As such, these axioms have been interpreted by some as rules defining a "fair" division, as if they served as an impartial "arbitrator" or "referee" separate from the participants seeking evaluation (Hart 1989). This is in following other approaches using the Shapley Value to satisfy specific fairness properties in games (van den Brink 2002; Balkanski and Singer 2015), specifically those regarding *distribution fairness* (Cohen 1987; Alexander and Ruderman 1987) as a principle of Organizational Justice Theory (Greenberg 1990; Greenberg and Colquitt 2013) founded on J.S. Adams' equity theory (Adams 1963, 1965). In this work, we aim to use the Shapley Value to estimate the relative contribution of individual players toward the achievement of a team within a video game for the purpose of attributing the reward they should receive within the game system.

3 Study Scenario

Our formulation of the Greedy Bandit Problem and our solution proposed in this work were the result of an analysis of a previous user study conducted by our research team (Zhu et al. 2020). We will refer to this study as the Pretest User

Figure 1: Two participant teammates engage remotely in a daily software activity in which their steps from the previous day are presented. An artificially constructed third teammate is added to the game by the AI, presenting as a genuine participant alongside the others. The MAB strategy determines the steps of this third teammate to be either above, between, or below those of the others (i.e., option A, B, or C), offering additional comparison opportunities that will ideally target the SCO preferences of the two real participants.

Study (or simply Pretest) throughout this work, with a summary provided for convenience here.

Our ongoing research centers around an intervention that models and leverages the psychological concept of social comparison theory to motivate players in an exercise game (or *exergame*) to improve their daily habits around physical activity (PA). The field of social comparison examines the personality traits that govern a person's propensity to compare themselves to others, consciously or subconsciously, when assessing their own abilities or beliefs (Festinger 1954; Gibbons and Buunk 1999). Depending on a person's social comparison orientation (SCO), they will experience positive or negative motivation in a task when exposed to others who are perceived as doing better (i.e., "upward" comparisons) or worse (i.e., "downward" comparisons) than themselves.

The intervention embedded in our exergames is facilitated by an AI agent that both models individual SCO by observing player behavior and adapts the intervention to best accommodate the individual's predicted preferences. In our scenario, participants are grouped into pairs that engage in a PA-related web activity with a third teammate controlled by the AI agent but presented to the participants as real. The agent, governed by our MAB strategy, chooses each day among three potential arms to determine the performance of this third teammate, as depicted in Figure 1. The agent can choose to report the fabricated teammate's steps as above, between, or below those of the two real participants, which will create additional comparison opportunities that can leverage the SCO preferences it has modeled for each of the two genuine participants. When engaging in their daily session, each of the participants will be shown the step achievements of their two other teammates as they report their motivation to personally engage in PA (via a 5-point Likert scale survey question) and examine data regarding daily lifestyle health habits.

The Pretest deployed this scenario among a group of 55 human user participants (37 women, 18 men) recruited from undergraduate courses at a university in a large U.S. city, where they were asked to participate in one session per day with the web-based activity and wear a Fitbit device to track their daily steps. Our findings demonstrated that an MAB strategy built around an ε-greedy policy based on a regression model predictor (i.e., experimental condition) yielded statistically higher motivation than a control group with a random policy. Additional details regarding results from the Pretest can be found in its own publication (Zhu et al. 2020), but continuing analysis of our results indicated interesting trends regarding player participation, discussed in the next section.

4 Analysis of Pretest User Study Data

We analyzed our user data from the Pretest to look beyond player motivation and performance and instead explore a proxy for player satisfaction—namely, player adherence—relative to their treatment by the Pretest bandit strategy. In this analysis, we wished to examine both a metric relating to the user's *treatment* by the system and a metric relating to the user's *engagement*.

Specifically, we analyzed all sessions from the Pretest for players in the (greedy) experimental condition. For each of these sessions, the analysis considered both the choice made by the bandit strategy and that choice's predicted success with each participant. From this, we determined how often the bandit strategy chose the arm most likely and least likely to benefit each player ("top" and "bottom" arms, respectively). We then aggregated a count of these for each player, which served as a metric reflecting the *treatment* of the player in terms of the degree to which the AI "catered" (i.e., prioritized positive treatment) to them. A second metric we collected was the number of days the player missed their session during the experimental phase. This served as our marker for *engagement*, a proxy for the player's interest in continuing to participate in the exercise.

We then compared these values, beginning with our graph in Figure 2. Here we graph our participants in descending order of their "Top" counts (grey bars). On a secondary axis, we graph the percentage of missed days in the experimental phase (orange line). With a Pearson's R of -0.40, we observe a large (Gignac and Szodorai 2016) negative correlation between a player's top treatment counts and their likelihood to miss a session.

The above considers how often a player received positive preferential treatment, but it does not consider how often players received treatment against their preferences. For this reason, the next analysis considers the *net* top treatments for a player (T_i), calculated as the difference between the number of times the player received their top and bottom treatments from the bandit strategy.

We also wanted to include an element of player performance (i.e., daily steps achieved) to examine the degree to which each player's treatment was commensurate with their effort. We predicted that one cause for player disengagement (i.e., missed days) might be due to an innate sense of receiving a "bad deal" in terms of how well their comparison experiences aligned with their individual SCO versus the

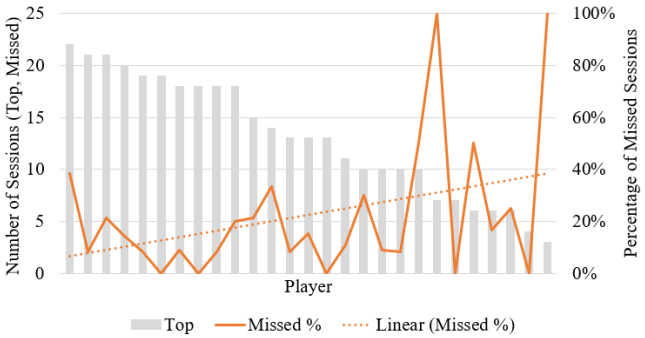

Figure 2: Individual participants in the Pretest are sorted by the number of times they received their top treatment for either steps or motivation (grey). On a secondary axis, we chart the likelihood that the player would miss a day in the experimental phase (orange). We see a large (Gignac and Szodorai 2016) negative correlation between the number of times a player received their top treatment and the likelihood they would miss a given session (Pearson's R = -0.40).

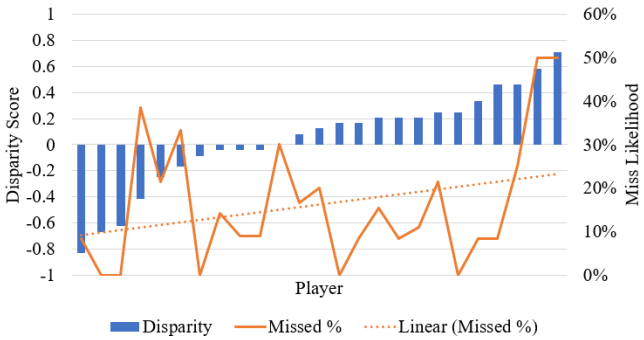

Figure 3: Disparity scores (blue) for participants in the Pretest, where participants on the left represent those who received very good treatment relative to their step performance, and players on the right received poor treatment relative to their efforts. On a secondary axis, we chart the likelihood that the player would miss a day in the experimental phase (orange) to find a large (Gignac and Szodorai 2016) positive correlation (Pearson's R = 0.36).

amount of effort (i.e., daily steps) they had been committing to the activity. This would follow the concept of *psychological meaningfulness*, in which humans are more likely to engage or disengage in activities based on the perceived reward received for their physical, emotional, and cognitive investments (Kahn 1990). Though the precise knowledge of their step performance versus others and the degree to which the MAB catered to them would not be known to the players, we posited that an impression of the relationship between these factors might have been sensed by players over the course of the study. Therefore, we created a new derived metric called the *disparity score* (D) that measures the degree to which these two values may be misaligned for players.

Computing the disparity score values requires $\{E_1, ..., E_n\}$, the effort achieved by the n players (i.e., average steps) throughout the experimental phase as well as $\{T_1, ..., T_n\}$, the treatments given to each of the n players (i.e., the net top treatment counts) over the same period. The *disparity score* (D_i) for a particular player (i) is computed as follows, where the $PR(V, V_i)$ function denotes the *percentile rank* of value for a player (V_i) among a set of values for all players (V):

$$D_i = PR(E, E_i) - PR(T, T_i). \quad (2)$$

In other words, the D_i value for player i is the difference between their percentile rank of step performance and their percentile rank of net top treatments, and the results for all players are illustrated in Figure 3. Here we graph the disparity scores for each player, sorted from lowest to highest. Players near the center of the graph (i.e., with D_i near zero) represent those for whom treatment was commensurate with their efforts. Players closer to the left side of the graph were given much better treatment relative to their performance, and players on the right were given top treatment much less frequently compared to that player's efforts. On a secondary graph, we chart the percentage of missed days observed for that player in the experimental phase. This figure helps us to visualize the large (Gignac and Szodorai 2016) positive correlation (Pearson's R = 0.36) between a player's disparity score and the percentage of sessions they missed.

We then perform regression analysis on the correlation between the disparity score and miss likelihood to find a slope and intercept ($y = 0.1394x + 0.157$, $R^2 = 0.13$) to help us define a model for the relationship between them. With the disparity score serving as our estimate of the *psychological meaningfulness* the activity may hold for a player, we take this result to serve as a method for estimating the likelihood a user will miss a session based on their disparity score.

5 The Greedy Bandit Problem

We believe this analysis reveals a potential issue that arises in the case where an AI must make decisions that potentially benefit some players more than others, where there is a risk that certain players may be routinely prioritized and others neglected. This may result in some players becoming disenfranchised and growing less engaged over time, lessening the AI intervention's overall potential impact and efficacy. We argue that the AI has a responsibility to be aware of this potential risk and would benefit from mitigating designs.

We refer to this dilemma as the *Greedy Bandit Problem*, where the term "greedy" is often used in algorithmic fairness literature as the opposite of something that is designed with a notion of "fairness" (Balkanski and Singer 2015; D'Amour et al. 2020). To address this problem, we propose a two-component approach. In the first, where typical bandit design may compare player output metrics directly, we propose a more nuanced comparison technique (i.e., the Shapley Value) that better assesses the contributions that individuals make when they work in a team (Section 5.1). In the second, where a "greedy bandit" may pursue a single metric related to performance, we consider alternative, player-focused metrics by which bandit strategy success should be determined.

5.1 Application of the Shapley Value

The Shapley Value provides a method by which the contributions of individuals can be assessed from the sum of their joint effort in a way that achieves fairness as defined by four axioms. Given a coalition of players (N) containing player i, we are interested to know the true value that this player contributes (φ_i) to the coalition when engaged in a joint endeavor pursued (or game played) by that group, or (N, v).

To do so, we consider every possible permutation by which the coalition could be constructed and average the marginal increase brought to the group by player i at the point in which player i is added during construction. Specifically, for this three-member coalition, we consider all possible permutation formation sequences (F) that could construct coalition $N = \{A, B, C\}$. This set would include the following six items, with double arrows ($\Rightarrow$) denoting where player A is added to the formation sequence:

$$\begin{aligned}
F_1 &: \emptyset \Rightarrow \{A\} \rightarrow \{A, B\} \rightarrow \{A, B, C\} \\
F_2 &: \emptyset \Rightarrow \{A\} \rightarrow \{A, C\} \rightarrow \{A, B, C\} \\
F_3 &: \emptyset \rightarrow \{B\} \Rightarrow \{A, B\} \rightarrow \{A, B, C\} \\
F_4 &: \emptyset \rightarrow \{B\} \rightarrow \{B, C\} \Rightarrow \{A, B, C\} \\
F_5 &: \emptyset \rightarrow \{C\} \Rightarrow \{A, C\} \rightarrow \{A, B, C\} \\
F_6 &: \emptyset \rightarrow \{C\} \rightarrow \{B, C\} \Rightarrow \{A, B, C\}. \qquad (3)
\end{aligned}$$

To find the average marginal increase that player A brings to the efforts of the full coalition N, we find the average of the values player A added to each of these six hypothetical formations at the time they were added. To provide a value for all elements of F, we require the following variables: $\boldsymbol{v(A_0)}$, $\boldsymbol{v(B_0)}$, and $\boldsymbol{v(C_0)}$ refer to the utility that the players achieve each when working alone, which in our case is the baseline step performance demonstrated by each player prior to the game as well as the step value assigned to the virtual player (C) by the AI agent. $\boldsymbol{v(AB)}$, $\boldsymbol{v(AC)}$, and $\boldsymbol{v(BC)}$ refer to the utility achieved by players working in their respective pairs, which we estimate as the sum of the steps of the players in each pair. $\boldsymbol{v(ABC)}$ refers to the utility achieved by the entire coalition N.

With these variables, we are able to define a value for each element of F for any player i, examining our definitions described in Equation 3, to yield the corresponding marginal utility values for that player $v(F_{i,n})$. We present F_A as our example and derive values for the utility brought to the team by player A in each of the six formation sequences:

$$\begin{aligned}
v(F_{A,1}) &= v(A_0) \\
v(F_{A,2}) &= v(A_0) \\
v(F_{A,3}) &= v(AB) - v(B_0) \\
v(F_{A,4}) &= v(ABC) - v(BC) \\
v(F_{A,5}) &= v(AC) - v(C_0) \\
v(F_{A,6}) &= v(ABC) - v(BC). \qquad (4)
\end{aligned}$$

According to the Shapley approach, the attribution of utility $\varphi_i(N, v)$ assigned to player i will be the average of the values in this set (F_i):

$$\varphi_i(N, v) = \frac{1}{6} \sum_{n=1}^{6} v(F_{i,n}). \qquad (5)$$

5.2 The Shapley Bandit

As discussed in Section 5, the typical approach to bandit strategy design will aim to maximize a particular value, regardless of how its operation may or may not prioritize certain arms above others. However, when these arm selections correspond to outcomes for human participants, extra care should be taken to ensure that players are not repeatedly excluded over time, leading to their potential disenfranchisement. Therefore, we address the Greedy Bandit Problem by adding a constraint to our bandit strategy design that aims to align the favorable attention that each player receives to their individual efforts.

Shapley Disparity: Given a set of players N, a Shapley Bandit estimates the *Shapley Disparity* for each player i at each time step t. When a reward is received for an arm pull, the Shapley Value (Equation 5) is computed for that player and added to a *Cumulative Shapley Value*, CSV_i. From this, we define the player's *CSV Ratio*, $CSVR_i$, as their Cumulative Shapley Value divided by the total of that of all players. Additionally, the bandit maintains a *Treatment Counter*, TC_i, for each player that tracks how often the agent chose the arm predicted to most benefit that player. From this, we similarly define the player's *TC Ratio*, TCR_i, as their Treatment Counter divided by the sum of that of all players. Finally, we define each player's *Shapley Disparity* as $SD_i = |CSVR_i - TCR_i|$, measuring the difference between the player's proportional contribution and the proportion of pulls in which the agent has prioritized the player.

Shapley Bandit: Instead of a greedy bandit strategy that would simply select the arm predicted to maximize total steps, the Shapley Bandit aims to reduce the *Shapley Disparity* among players. Specifically, it will select the arm predicted to most benefit the player who, if catered to, would result in the lowest total *Shapley Disparity* among all players. As with conventional (ε-greedy) bandit strategies, this prediction is based on previous rewards observed, with a small probability that the strategy will choose a random arm.

By prioritizing alignment of the proportion of top treatments for each individual to their proportion of total team contribution, we anticipate that this strategy will empower the agent to prioritize high performers while still ensuring that it never fully ignores any member of the group. Of course, we expect that this will result in an incurred cost to total utility (i.e., rewards will no longer be fully maximized as they would with a greedy bandit) but that the cost of this fairness constraint will provide a better experience to all participants. We will examine this further via simulation experiments, first evaluating the impact on user experience (via adherence metrics) in Section 7.2 and then examining the cost of this tradeoff (via total utility) in Section 7.3.

6 Simulation Environment

We aim to create a simulation in which we can evaluate the potential efficacy of the Shapley Bandit prior to conducting human user studies. To do so, we employ *virtual players* provisioned with models that adjust their behavior according to their exposure to different *social comparisons* (discussed in Section 3) and described in further detail in our

previous simulation study (Gray, Zhu, and Ontañón 2021). In short, virtual players will either be motivated or demotivated (in terms of daily step activity) based on their exposure to upward or downward comparisons and depending on their (randomized) internal preferences for such comparisons.

Beyond this, we also include a virtual player behavior for skipping their daily sessions based on their perception of the misalignment between their effort and the treatment they are receiving in the activity. We create a model that allows a player to predict these two values (effort and treatment relative to other players) to better mimic the subconscious sense a human player might build regarding their perception of psychological meaningfulness in this activity.

The first component of this subsystem is a notion of the player's own step performance relative to others. Though no virtual player will have access to any other player's steps, it's reasonable to believe that a human participant will have some knowledge of average human walking behavior. Therefore, we allow the virtual players to understand the distribution of steps expected from players like them, from which they can find their position within that general expectation. Specifically, virtual players were allowed to know that daily steps were generally distributed with a mean of 11832 and a standard deviation of 2369, which are values derived from a study on human walking behavior conducted by Furberg et al. (Furberg et al. 2016). Therefore, virtual players were able to use the average of their daily steps so far ($\bar{\rho}_i$) to estimate their effort relative to the group ($\hat{E}_i$) using the following, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution:

$$\hat{E}_i = \Phi(\frac{\bar{\rho}_i - 11832}{2369}). \tag{6}$$

The second component of our behavioral subsystem is a notion of the treatment the player feels they have received from the system. This may not be known or even explicitly considered by human participants, but it may be sensed through their implicit satisfaction with the system based on how well the selection targets have been catered to accommodate their individual SCO preferences. Therefore, we allow virtual players to track how often their comparison targets have been in favor or against their preference. An internal *treatment counter* increments by one whenever both of their teammates compare to them in the virtual player's preferred direction, and it decrements whenever both teammates compare against their preference.

Human players would have insight neither into the individual SCO of the other players nor into the MAB strategy's operation and choices among arms. Similarly, virtual players do not have access to this higher-level information; however, for the sake of the simulator, we allow virtual players to understand the expected distribution of such scores in a game lasting two weeks. It could be reasoned that in a large sample, the mean of such treatment counter values would rest on zero, and it would present in a normal distribution between the extreme values of ± 14 (i.e., if a player received their best or worst treatment exclusively). Using the intuition of the range rule (Wan et al. 2014), the virtual player could guess a standard deviation of treatment counter values to be $S \approx {}^{14*2}/_{6} = 4.7$ over a large group. Therefore, virtual players were able to use the value of their own treatment counter at any point (H_i) to estimate their treatment relative to the group ($\hat{T}_i$) using the following:

$$\hat{T}_i = \Phi(\frac{H_i}{4.7}). \tag{7}$$

With both $\hat{E}_i$ and $\hat{T}_i$ values for a virtual player, we can compute an estimate of their *disparity score* ($\hat{D}_i$) at any time, following the definition provided in Equation 2:

$$\hat{D}_i = \hat{E}_i - \hat{T}_i, \tag{8}$$

where $\hat{D}_i$ is a value within $[-1, 1]$. This value can be generated at any time step (i.e., $\hat{D}_{i,t}$) by using the $\hat{E}_i$ and $\hat{T}_i$ values generated at that same time step. As an update to our virtual player behavioral model, we can introduce a new behavior in which the virtual player may choose to miss one of their sessions based on their disparity score estimate at that time, using insight from human participant data discussed in Section 4. Specifically, we apply the coefficient and intercept of the regression analysis to model the virtual player's likelihood of missing a given session ($M_{i,t}$) as follows:

$$M_{i,t} = 0.1394 * \hat{D}_{i,t} + 0.157. \tag{9}$$

Virtual players keep track of the total number of times their non-adherence behavior is triggered and report it at the end of the experiment along with the other metrics we collected regarding human players in the Pretest, such that we can perform the same analysis on our simulation results.

7 Evaluation

Our evaluation consists of three parts. First, in Section 7.1 we aim to validate our simulation such that we can verify the manifestation of the Greedy Bandit Problem in the same degree that we observed in the human user study environment of the Pretest. Second, once this is established, in Section 7.2 we can deploy the Shapley Bandit within this environment to gain an understanding of how our strategy may help to mitigate the effects of the Greedy Bandit Problem. Finally, in Section 7.3 we run a third experiment to examine the cost incurred by the Shapley Bandit relative to the performance of the Greedy Bandit.

7.1 Simulation Validation

To compare our simulation results with our Pretest analysis above, we wish to run a *batch* of experiments. For example, the only way for us to generate values for E_i that are meaningful in a larger set (E) for use in analysis by Equation 8, we must perform our analysis on larger groups of players ($\{p_1, ..., p_n\}$). Therefore, we set up our experiment to run batches of 12 such teams (i.e., 24 total virtual players) to mimic the Pretest so that we can repeat the same analysis on the results of the virtual players in our simulation.

With results of one such batch presented in Figure 4, we see many promising similarities and some differences when comparing the graphs resulting from our analysis of simulation data against our same analysis of real human behavior data in Section 4. We first see that our simulation is capable of creating the same pattern of disparity across the user base,

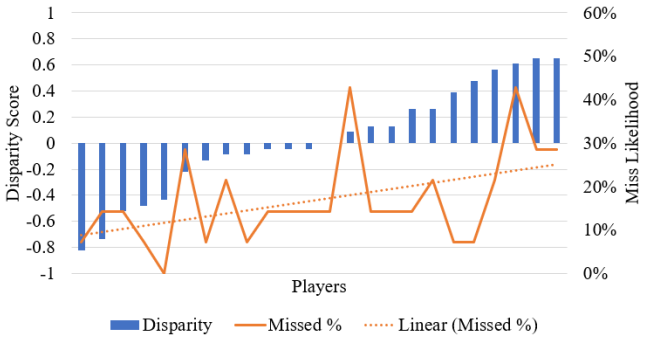

Figure 4: Disparity scores for a batch of virtual players, intended for comparison to Pretest data (Figure 3). The simulation under a greedy bandit can create a similar disparity spread as human user data. Non-adherence rates, based on Equation 9, are also reported for each virtual player (orange) to again reflect real human behavior patterns.

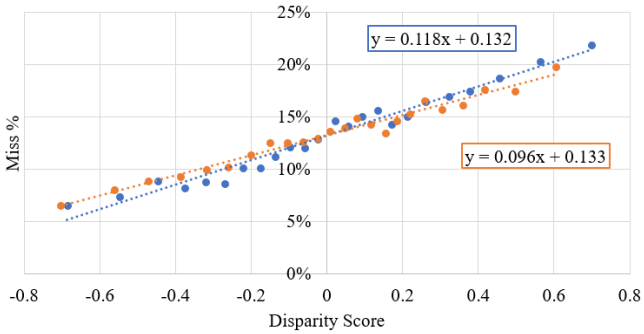

Figure 5: Scatter plots of the disparity score / non-adherence relationship in the Greedy Bandit experiment (blue) vs. the Shapley Bandit experiment (orange). The smaller slope of the Shapley condition regression (0.096 vs. 0.118 in the Greedy condition) indicates a reduced likelihood of non-adherence relative to disparity ($p < 0.001$).

which indicates that the simulator can manifest the issues of disparity that are symptoms of the Greedy Bandit Problem. Similarly, the regression analysis finds a coefficient and intercept that is not too dissimilar to that of the human data (i.e., coefficient of 0.122 vs. 0.139 and intercept of 0.167 vs. 0.157 in the simulation and Pretest data, respectively), where the overall trend and shape of the data are similar.

We extend this experiment to reduce the impact of stochastic elements in our simulation. First, we increase the number of days in each game to 100 to allow the agent more of an opportunity to understand the players. Second, we increase our test from a single batch of games to a set of 100 batches, where the results are aggregated for all 24 players across batches based on their rank in disparity among those in their batch cohort. In other words, the key metrics (i.e., top treatments and miss %) for the players with the lowest disparity score in each batch were averaged to construct the first *aggregate* player. Metrics for the second-lowest players were then averaged in the same way, and so on.

The scatter plot for these players, each reflecting *aggregate* players across 100 batches, is illustrated in blue in Figure 5. Regression analysis over this set of data again finds a coefficient and intercept that is similar to that of the human data (i.e., coefficient of 0.117 vs. 0.139 and an intercept of 0.132 vs. 0.157), this time reflecting the trend to an even greater degree ($R^2 = 0.97$, vs. $R^2 = 0.13$ in the human data) due to the synthetic nature of the simulation and averaging of significantly more points of data (2400 virtual players in simulation vs. 24 players in the human data).

From these results, we see that across a larger sample size, we can expect virtual player metrics to manifest the intent of our model fairly closely. We determined this performance to be satisfactory to support our simulation as a test bed for exploring the potential of the Shapley Bandit.

7.2 Shapley Bandit Evaluation

We leverage our simulation environment to explore the efficacy of the Shapley Bandit strategy, described in Section 5.2, to assess the degree to which it may help alleviate the effects of the Greedy Bandit Problem. We repeat the last experiment described above with 100 batches, except we replace the default Greedy Bandit with a Shapley Bandit for each game.

Our first analysis compares the miss rates of the aggregate players in the Shapley Bandit experiment to those of the Greedy Bandit experiment, which are shown in Figure 5. These results allow us to compare the performance between conditions of MAB strategy effect on participant non-adherence. First, we identify that there is a tighter horizontal grouping among the Shapley Bandit disparity values, suggesting a smaller variance in the disparity provided to the players in that condition. Because the purpose of introducing the Shapley Value was to better align bandit rewards to individuals based on their performance, we would consider this an intended effect of the strategy. F-Test analysis on the full data sets (Greedy vs. Shapley conditions, n=2400 players in each) indeed demonstrates a mean disparity score of zero with a statistically significant smaller variance ($p < 0.05$) in the Shapley condition.

Second, regression analysis on both data sets yields a smaller coefficient in the Shapley condition ($\beta_S < \beta_G$), found to be statistically significant ($p<0.001$) via a z-test for the difference between two regression coefficients (Paternoster et al. 1998). This suggests that players attended to by this strategy exhibited a smaller likelihood for non-adherence even at the same disparity level.

In general, we do not expect the Shapley Bandit to outperform the greedy bandit on performance metrics, and indeed we observed that virtual players in the Greedy Bandit condition achieved more steps per day per player (average of 12366.0 vs. 12224.6, $p < 0.001$) due to the fact that exploitation pull targets were unconstrained. However, we endure this lower performance in a willing exchange, where we expect the Shapley Bandit to instead achieve a more uniform *distribution* of metrics (of both treatment and performance) due to the fairness constraints it applies. Indeed, again we observe via F-test that players in the Shapley condition experienced lower variance among their step achievements ($p < 0.001$). Similarly, though we observe a non-

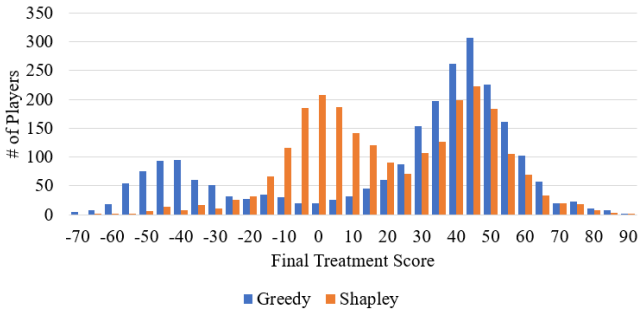

Figure 6: Distributions of final internal Treatment Counter values (TC_i) perceived by the virtual players in both Greedy Bandit (avg. = 19.84) and Shapley Bandit (avg. = 20.26) experiments (not statistically significant, $n = 2400$ in each condition). The bimodal nature in both graphs is likely due to both high and low-performing players in both simulations; however, because the Shapley strategy ensures that even the lower-performing players are catered to, it raises the treatment of those players, thereby providing smaller variance in treatment among all players ($p < 0.001$).

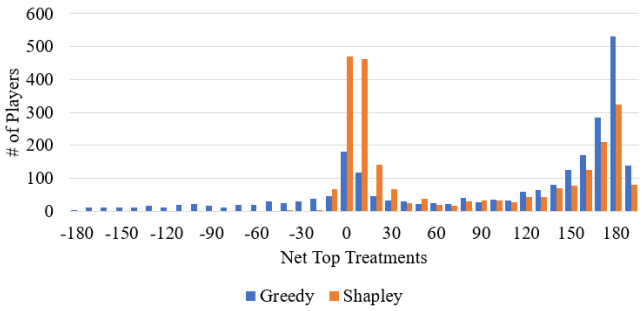

Figure 7: Distributions of net top treatment counts (TC_i) values in both Greedy Bandit (avg. = 95.77) and Shapley Bandit (avg. = 77.6) experiments ($n = 2400$ in each condition). Though Greedy outperforms ($p < 0.001$) in this metric due to its unconstrained targeting of only high-performing players, the Shapley Bandit achieves lower variance ($p < 0.001$) in net treatments across all players. Additionally, significantly ($p < 0.001$) fewer players in the Shapley Bandit condition (12.6%) receive more bottom treatments than top compared to the Greedy condition (17.6%).

statistically significant indication that players in the Shapley Bandit condition may achieve overall lower miss likelihood than those in the Greedy condition (11.97% in the Shapley condition vs. 12.04% in the Greedy condition, $p = 0.328$), our focus returns once again to the smaller variance of this metric across players in the Shapley condition ($p < 0.001$), further supporting the Shapley Bandit's aim of providing a more uniform treatment of players among the group.

Additional F-Test analysis reveals that players in the Shapley condition indeed observed statistically significant ($p < 0.001$) lower variance in overall treatment scores perceived by the virtual players (Figure 6) and net top treatments given by the agent (Figure 7). These results are promising because they not only demonstrate a solution that improves the overall experience for players but that also treats players more uniformly. In both conditions, the virtual players operated with the same parameters, behavior sets, and models for non-adherence based on the players' performance and disparity scores (Equation 9). These results show that the Shapley Bandit succeeded in reducing variance in disparity, where its focus on the TC Ratio (TCR_i) required that even lower-performing players still received attention.

This effect is especially evident in Figure 6, which visualizes the virtual players' internal perceptions of treatment. While both conditions present a bimodal graph with high- and low-performing players clustered, the Greedy Bandit condition (blue) shows much more inequality between the two groups, where the bandit catered to higher-performing players at the expense of ignoring the remainder to a significant degree. In contrast, the Shapley condition (orange) ensured that even the lower-performing players were still given proper attention. Though the Shapley strategy sacrificed some performance in not maximizing its exploitation of the higher players, we see that the overall scores among its *lower* grouping are seated higher than that of the Greedy Bandit's lower group. The results for treatment average in our simulation were inconclusive, yielding an average treatment score of 19.84 for the Greedy Bandit and 20.26 for the Shapley Bandit ($p = 0.329$). However, and again more important to our objective, this tightening of the treatments for both player groups results in a closer grouping of all players overall, yielding a statistically significantly lower variance of treatment scores ($p < 0.001$) in the Shapley group.

A separate but related metric can further explain this operational difference between the strategies, illustrated in Figure 7. In this analysis, we observe not the virtual players' estimation for their treatment but the actual degree to which they were catered to as tracked by the agent. We review all decisions in the trial and note the number of times a player received the arm that was predicted to be that player's top arm, and we also record the number of times each player received their worst arm. The difference between these values (i.e., top count subtracting worst count) for each player yields that player's *net top treatments* metric.

We see in this distribution of player net top treatments where the Greedy Bandit exploited top-performing players (i.e., at x>110) to a greater degree than the Shapley Bandit. However, without a fairness constraint, we also see that many players experienced negative top treatment scores, some greatly so. For example, seven players received their top treatment as few as twice in 100 pulls in the Greedy condition. In contrast, the Shapley strategy aimed to correct this pattern and instead set up a resistance to negative net top treatments that is visible in the graph around x=0. While this resulted in a slightly lower average than that achieved by the Greedy Bandit, the distribution illustrates how the Shapley strategy very clearly ($p < 0.001$) achieved a smaller variance of player experience with regard to this metric. Further, we observe that the Shapley Bandit condition resulted in significantly fewer players with a net top treatment score below zero (12.6% in the Shapley vs. 17.6% in the Greedy condition, $p < 0.001$).

Bandit Strategy	Overall Avg. Reward	Pull #100
Greedy Bandit	12431.9 ($\pm$0.068)	12530.3
Shapley Bandit	12338.6 ($\pm$0.061)	12435.0
UCB1	11927.3 ($\pm$0.049)	12011.0
ε-greedy	11928.8 ($\pm$0.049)	12031.2
Random	11617.5 ($\pm$0.015)	11621.2

Table 1: Average rewards of Greedy vs. Shapley Bandit strategies in multiplayer scenario ($h = 100$, $\pm$99.9% CI)

7.3 Overall Performance in Study Scenario

Consideration of the lower step results in the Shapley condition prompts a final experiment in which we aim to assess more precisely what the Shapley approach loses in performance in order to support this more uniform experience for its players. Specifically, we wish to run a larger simulation measuring the average step performance of teammates in our sample scenario when under the control of various MAB strategies. We set up an experiment of one million games (i.e., a total of two million players), each lasting 100 time steps, and observed the average step performance of those players when governed by a battery of MAB strategies.

We examine the following set of strategies: Random (random pulls), UCB1 ($C = 800$), and ε-greedy ($\varepsilon = 0.01$) along with the Greedy and Shapley Bandits, both set up as they were in the previous experiments. It's worth noting that the distinction between the Greedy Bandit and the standard ε-greedy strategy is the use of the regression-based model for predictions, mentioned in Section 3 and described in previous work (Zhu et al. 2020). We began with a pre-test in which values for C and ε parameters were determined via parameter sweep (Gray et al. 2020). Results of this experiment are presented in Table 1, where we observe merely a 0.75% loss in average reward for the Shapley Bandit.

8 Discussion

Results from our simulation evaluation confirmed that the environment was capable of reproducing similar metrics with a Greedy Bandit as those we observed in the Pretest study with human users, providing a capable tool for exploring the potential of the Shapley Bandit. With that in place, our experiments involving the Shapley Bandit demonstrated an ability to better homogenize user experiences with respect to the agent's decisions, if not always in significantly raising the expected value of virtual player metrics (e.g., treatment scores or missed sessions), then decisively with respect to the variance among those metrics. In particular, the results illustrated in Figures 6 and 7 demonstrate the strategy's ability to significantly alter the distribution of player experience in terms of how they were catered to by the agent. Our simulations indicated a modest but statistically significant reduction in the degree to which disparity score influences player non-adherence likelihood both in regression coefficient and variance among players, which we expect to be reflected in higher player adherence when we deploy this strategy in human user studies.

Regarding overall MAB strategy efficacy, we see from the results of the final experiment that although the Greedy Bandit does yield better performance overall, the Shapley Bandit incurs a relatively low cost (0.75%) for all the benefits demonstrated in the previous experiments. That is, in our simulation, the Shapley Bandit approach yields a strategy that is 99.2% as effective as a Greedy Bandit but achieves this outcome with smaller variances across a number of metrics important for player satisfaction and fairness. By considering the distribution of attention it gives to all players in the group, the Shapley Bandit demonstrates capability in making worthwhile tradeoffs around metrics that are meaningful when serving as an AI agent working with human players. From these results, we were confident in further evaluating this approach with human participants.

9 Conclusion

Our primary contributions in this work include (1) the identification of the Greedy Bandit Problem as it was observed in a human user study guided by a greedy MAB strategy, (2) the construction of a simulation and virtual player behavior strategy that embodies the concept of *psychological meaningfulness*, by which player non-adherence (i.e., a virtual player choosing not to participate) could be used as a metric for player satisfaction, and (3) an alternative bandit strategy called the Shapley Bandit that leverages the Shapley Value to better align AI agent treatments to player contributions in a multiplayer environment.

Following analysis of data from a Pretest User Study with which we formalized the Greedy Bandit Problem, we designed a simulation scenario (including virtual players) that we confirmed to be capable of generating observations of the Greedy Bandit Problem to a similar degree as that which manifested in the human user data. We then deployed our Shapley Bandit strategy in that same environment and determined that it was able to mitigate many of the issues of the Greedy Bandit Problem. In our final experiment, we evaluated the direct performance of both strategies to verify a relatively small difference in performance (less than 1%) in exchange for the benefits it was shown to provide.

We identify some limitations in this work, most notably that our approach discussed here is only so far validated in a simulation based on human user data; however, this investigation provided the insight needed before we could validate the approach with human users (Gray et al. 2022). Additionally, our solution may face scaling issues as the groupings of players grow larger, but only to the extent that the Shapley Value itself faces such scaling issues. Overall, we believe that this approach provides a good foundation for moving forward with MAB-based EM strategies involving human players, where considerations like fairness, player satisfaction, and adherence are a concern.

Acknowledgments

This work is partially supported by the National Science Foundation (NSF) under Grant Number IIS-1816470. The authors would like to thank all past and current members of the projects.

References

Adams, J. S. 1963. Towards an understanding of inequity. *The journal of abnormal and social psychology*, 67(5): 422.

Adams, J. S. 1965. Inequity in social exchange. In *Advances in experimental social psychology*, volume 2, 267–299. Elsevier.

Alexander, S.; and Ruderman, M. 1987. The role of procedural and distributive justice in organizational behavior. *Social justice research*, 1(2): 177–198.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3): 235–256.

Balkanski, E.; and Singer, Y. 2015. Mechanisms for fair attribution. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 529–546.

Bates, J. 1992. Virtual reality, art, and entertainment. *Presence: Teleoperators & Virtual Environments*, 1(1): 133–138.

Cohen, R. L. 1987. Distributive justice: Theory and research. *Social justice research*, 1(1): 19–40.

D'Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; and Halpern, Y. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534.

Drachen, A.; Canossa, A.; and Yannakakis, G. N. 2009. Player modeling using self-organization in Tomb Raider: Underworld. In *2009 IEEE symposium on computational intelligence and games*, 1–8. IEEE.

Feltz, D. L.; Forlenza, S. T.; Winn, B.; and Kerr, N. L. 2014. Cyber buddy is better than no buddy: A test of the Köhler motivation effect in exergames. *GAMES FOR HEALTH: Research, Development, and Clinical Applications*, 3(2): 98–105.

Festinger, L. 1954. A Theory of Social Comparison Processes. *Hum. Relations*, 7(2): 117–140.

Fujiki, Y.; Kazakos, K.; Puri, C.; Buddharaju, P.; Pavlidis, I.; and Levine, J. 2008. NEAT-o-Games: blending physical activity and fun in the daily routine. *Computers in Entertainment (CIE)*, 6(2): 21.

Furberg, R.; Brinton, J.; Keating, M.; and Ortiz, A. 2016. Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016. https://doi.org/10.5281/zenodo.53894. Accessed: 2019-09-11.

Gibbons, F. X.; and Buunk, B. P. 1999. Individual differences in social comparison: Development of a scale of social comparison orientation. *J. Pers. Soc. Psychol.*, 76(1): 129–142.

Gignac, G. E.; and Szodorai, E. T. 2016. Effect size guidelines for individual differences researchers. *Personality and individual differences*, 102: 74–78.

Gray, R. C.; Villareale, J.; Fox, T.; Dallal, D. H.; Ontañón, S.; Arigo, D.; Jabbari, S.; and Zhu, J. 2022. Improving Fairness in Group-Based Social Exergames via Shapley Bandits. In *Proceedings of the 28th ACM Conference on Intelligent User Interfaces (IUI'21)*.

Gray, R. C.; Zhu, J.; Arigo, D.; Forman, E.; and Ontañón, S. 2020. Player Modeling via Multi-armed Bandits. In *in Proceedings of the 15th International Conference on the Foundations of Digital Games*.

Gray, R. C.; Zhu, J.; and Ontañón, S. 2020. Regression Oracles and Exploration Strategies for Short-Horizon Multi-Armed Bandits. In *Proceedings of the 2020 IEEE Conference on Games*.

Gray, R. C.; Zhu, J.; and Ontañón, S. 2021. Multiplayer Modeling via Multi-Armed Bandits. In *Proceedings of the 2021 IEEE Conference on Games*.

Greenberg, J. 1990. Organizational justice: Yesterday, today, and tomorrow. *Journal of management*, 16(2): 399–432.

Greenberg, J.; and Colquitt, J. A. 2013. *Handbook of organizational justice*. Psychology Press.

Hart, S. 1989. Shapley value. In *Game theory*, 210–216. Springer.

Kahn, W. A. 1990. Psychological conditions of personal engagement and disengagement at work. *Academy of management journal*, 33(4): 692–724.

Kuleshov, V.; and Precup, D. 2000. Algorithms for the multi-armed bandit problem. In *Journal of Machine Learning Research*, volume 1, 1–48.

Lai, T. L.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Ma, S.; and Tourani, R. 2020. Predictive and causal implications of using shapley value for model interpretation. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, 23–38. PMLR.

Min, W.; Mott, B. W.; Rowe, J. P.; Liu, B.; and Lester, J. C. 2016. Player Goal Recognition in Open-World Digital Games with Long Short-Term Memory Networks. In *IJCAI*, 2590–2596.

Missura, O.; and Gärtner, T. 2009. Player modeling for intelligent difficulty adjustment. In *Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12*, 197–211. Springer.

Paternoster, R.; Brame, R.; Mazerolle, P.; and Piquero, A. 1998. Using the correct statistical test for the equality of regression coefficients. *Criminology*, 36(4): 859–866.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5): 527–535.

Roth, A. E. 1988. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

Samendinger, S.; Forlenza, S. T.; Winn, B.; Max, E. J.; Kerr, N. L.; Pfeiffer, K. A.; and Feltz, D. L. 2017. Introductory dialogue and the Köhler effect in software-generated workout partners. *Psychology of sport and exercise*, 32: 131–137.

Shapley, L. S. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28): 307–317.

Shapley, L. S. 1997. *A value for n-person games*, 69–79. Princeton University Press. ISBN 9780691011929.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4): 285–294.

Valls-Vargas, J.; Kahl, A.; Patterson, J.; Muschio, G.; Foster, A.; and Zhu, J. 2015. Designing and tracking play styles in solving the incognitum. In *Proceedings of the Games+ Learning+ Society Conference, to appear. Games+ Learning+ Society*.

van den Brink, R. 2002. An axiomatization of the Shapley value using a fairness property. *International Journal of Game Theory*, 30(3): 309–319.

Vinogradov, A.; and Harrison, B. 2022. Using multi-armed bandits to dynamically update player models in an experience managed environment. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, 207–214.

Wan, X.; Wang, W.; Liu, J.; and Tong, T. 2014. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology*, 14(1): 1–13.

Winter, E. 2002. The shapley value. *Handbook of game theory with economic applications*, 3: 2025–2054.

Yannakakis, G. N.; Spronck, P.; Loiacono, D.; and André, E. 2013. Player modeling. *Artificial and Computational Intelligence in Games*.

Yannakakis, G. N.; and Togelius, J. 2018. *Artificial intelligence and games*, volume 2. Springer.

Zhu, J.; Dallal, D. H.; Gray, R. C.; Villareale, J.; Ontañón, S.; Forman, E. M.; and Arigo, D. 2020. Personalization Paradox in Behavior Change Apps: Lessons from a Social Comparison-Based Personalized App for Physical Activity. In *ACM PACM on Human Computer Interaction*.

Zhu, J.; and Ontañón, S. 2019. Experience Management in Multi-player Games. In *Proceedings of the IEEE Conference on Games*.