Proceedings of the Sixteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-20)

Studying General Agents in Video Games from the Perspective of Player Experience

Cristina Guerrero-Romero,[*1] **Shringi Kumari,**[*2]
Diego Perez-Liebana,[1] **Sebastian Deterding**[2]
[1]Queen Mary University of London
[2]University of York
{c.guerreroromero, diego.perez}@qmul.ac.uk, {sk1382, sebastian.deterding}@york.ac.uk

Abstract

Research actively explores and advances the play strength of general agents, which are able to play video games without having specific knowledge about them. However, how general agents impact player experience and motivation when implemented in commercially viable games is largely unexplored. In this paper, we investigate this relationship as initial work towards linking general agent behaviour and player experience as a step towards making general agents applicable to commercial video games. Specifically, we created two versions of a simple competitive human-versus-agent game having two general Monte Carlo Tree Search (MCTS) agents with different behaviours. These agents, without having specific knowledge about the game, have two unique goals: i) maximising score; and ii) exploring (more suitable for the game we chose). We integrated these agents into a 'capture the flag' game and conducted a study to investigate the effects on several player motivation components of the Intrinsic Motivation Inventory (IMI) and Player Experience of Need Satisfaction (PENS) scale. Enquiry in this direction opens up the possibilities to start analysing general agents from the perspective of the player's journey.

Introduction

There is an active body of research creating and exploring the improvement of General Video Game Playing (GVGP) agents. These agents have been shown to play a wide range of video games at competitive strength without specific knowledge about them (Perez-Liebana et al. 2019a). Their performance evaluation is usually limited to winning rates and scores (Perez-Liebana et al. 2019b). Put simply: Research is concerned with how strongly agents play, assuming that higher play strength is better. This leaves us with little knowledge about how different agents uniquely impact player experience (PX): is playing against them actually enjoyable? Furthermore, often game designers want to elicit a specific player experience or emotion that has nothing to do with the strength of play, for instance in the game *Journey* (Thatgamecompany 2012), the whole game was designed around the player experience arc that the designers wanted (Chen 2013).

In this paper, we investigate the question of how GVGP agents can be designed keeping PX in mind. This is as an important initial step towards making general agents viable for commercial use as research towards such integration is practically non-existent. We think that PX and Player Motivation (PM) can be important parameters for designing GVGP agents, rather than just focusing on the AI's win rate. An AI could have a high winning rate but still be incapable of eliciting a desired PX. With our work we hope to initiate this line of research. We particularly compare two versions of a simple competitive player-versus-agent 'capture the flag' General Video Game AI (GVGAI) (Perez-Liebana et al. 2016) game, each with a general Monte Carlo Tree Search (MCTS) agent with a different general behaviour (without specific knowledge about the game). One agent aimed to just win and maximise their score while the other one was encouraged to also explore the game space. We studied how these distinct behaviours affected PM, namely 'tension', 'perceived competence' and 'enjoyment', using the Intrinsic Motivation Inventory (IMI) (Deci and Ryan 2003) and Player Experience of Need Satisfaction (PENS) (Ryan, Rigby, and Przybylski 2006) scales. We discuss how such PX evaluation of agents is pivotal for making GVGP agents ultimately applicable in commercial games.

Background

General Video Game Playing

General Video Game Playing (GVGP) (Levine et al. 2013) refers to Artificial Intelligence (AI) algorithms that can play video games without having prior or specific knowledge about the game, their rules or environment. The entities that are driven towards a goal are called agents and, in the context of GVGP, they are given the name of General Agents (GA). There are different approaches followed in the creation of GA, distinguishing between two main groups: search and learning algorithms. Search algorithms play in real-time and have a forward model available to them, which allows to simulate future states by providing state-action pairs. This way, the algorithms can estimate how the game changes as different actions are carried out in the game. Learning algo-

[*]Both authors contributed equally to this research.

rithms, on the other hand, require offline training to learn a behavioural policy to be used when playing the game, as they do not use a forward model. The GVGAI competition has featured planning (Perez-Liebana et al. 2016; Gaina et al. 2017) and learning (Torrado et al. 2018) tracks where searching and learning algorithms have been used, respectively.

The level of generality of the agents depends on the heuristics included in the algorithm and, in most cases, these are focused on winning and maximising the score in the game. The idea behind improving GA is creating high-quality algorithms that perform well in different games, making sure the strength comes from the approach and not from a game-tailored hand-crafted heuristic. However, merely changing the heuristics of these general algorithms to a different general goal (as exploring the map) also affects their performance (Guerrero-Romero, Louis, and Perez-Liebana 2017). Competitions and frameworks are available and used as a benchmark to measure the quality of the general algorithms and approaches (Perez-Liebana et al. 2019a).

One tool available for the study and development of general agents is the General Video Game AI (GVGAI) Framework (Perez-Liebana et al. 2016). It provides a simple structure to research both learning and search algorithms, providing a simple Video Game Description Language (VGDL) (Schaul 2013) for game development. One of the sample algorithms available with the framework is Monte Carlo Tree Search (MCTS). MCTS is a search algorithm that builds a tree incrementally and asymmetrically, balancing between exploitation and exploration of its nodes. It has different variations and enhancements, their strengths depending on the situation under consideration (Browne et al. 2012).

AI and General Video Game Playing in Commercial Games There is an increasing interest in using AI in commercial games: Procedural Content Generation (PCG) has become popular (e.g. *No Man's Sky* (Hello Games 2016)), along with the use of player modeling and analytics (El-Nasr, Drachen, and Canossa 2016).

Commercial game AI programmers are still frequently using behaviour selection algorithms like Finite-State Machines (FSM), Behaviour Trees (BT), Utility Systems, Goal-Oriented Action Planners (GOAP) or Hierarchical Task Networks (HTN) (Rabin 2013). These types of AIs are hand-designed, given specific steps and goals based on the details of the particular game's design. Sturtevant in (Rabin 2015) introduces the adoption of search algorithms in commercial games, stating that path planning techniques are extensively used and the use of other search approaches appears to be growing. Amongst others, they present the successful application of MCTS in strategic games like *Total War: Rome II* (Creative Assembly 2013). This is a good example of the extension and use of search algorithms in video games, but their heuristics are heavily tailored towards the game (Thompson 2018). Despite the research interest, the use of GVGP agents (with heuristics that do not have specific information about the game) haven't found their way into commercial applications yet.

Player Experience

The field of player experience (PX) concerns itself with people's perceptions of and responses to the use of games (Bernhaupt 2015). Work in this field focuses on a relatively small cluster of (somewhat overlapping) constructs such as enjoyment, engagement, fun, presence, flow and immersion (Caroux et al. 2015; Mekler et al. 2014). It is driven by the tenet that people buy and play games for the sake of the experience playing them provides. Hence, one key PX constructs is intrinsic motivation, the psychological processes that energise and direct behaviour toward an activity which are generated by the activity itself - put simply, the activity is done 'for its own sake' (Ryan and Deci 2017). By far the most well-established theory for intrinsic motivation is self-determination theory (SDT) (Ryan and Deci 2017). According to SDT, just like the physiological needs of hunger or thirst, people have innate psychological needs for experiences of competence (successfully affecting one's environment), autonomy (acting with volition, willingness, and in congruence with one's self), and relatedness (mutual connection and support between self and others). Activities like gameplay are intrinsically motivating because and when they give rise to experiences that satisfy these basic psychological needs. SDT has become an established major theory in studying and evaluating PX (Tyack and Mekler 2020). Specifically, PX researchers routinely use two SDT-based survey instruments to operationalise and assess positive PX: the general Intrinsic Motivation Inventory (IMI) (Deci and Ryan 2003) and the game-specific Player Experience of Need Satisfaction (PENS) scale (Ryan, Rigby, and Przybylski 2006).

PX research largely assumes that play strength has a U-shaped relation to enjoyment mediated by competence: if competitors are significantly stronger than the player, the player will mostly lose, thwarting their sense of competence. If competitors are significantly weaker, the player will win without exerting much effort, thus feeling little competence. Recent work suggests that suspense or 'tension' related to outcome, not competence, may mediate the relation between difficulty/competitor play strength and enjoyment (Abuhamdeh, Csikszentmihalyi, and Jalal 2015).

Player Experience and General Video Game Playing There is existing work linking AI with PX; Guckelsberger et al. (2017) predicted player experience in a PCG game by using computational models of intrinsic motivation. Emmerich and Masuch (2016) studied if the existence of a virtual AI in a game had an impact on its experience. Furthermore, there is strong work on player profiling to support PX when it comes to matchmaking, difficulty adjustment and AI directors (Yannakakis and Togelius 2018).

Contrasting with this, existing research linking PX and general AI is limited to either automatic game design or Experience Driven Procedural Content Generation (EDPCG) techniques to some degree (Yannakakis and Togelius 2011). To give an example of automatic game design, Kunanusont, Lucas, and Perez-Liebana (2018) used N-Tuple Bandit Evolutionary Algorithms (NTBEA) to tune parameters of multiple GVGAI games. The objective of this work was to adjust the experience of the AI players so the distribution of the scores obtained would fit certain curves. However, human players were not involved in the generation nor evaluation of the final games. In the tapestry of game AI and PX, we clearly see a void regarding work that investigates GVGP agents from a PX point of view.

Case Study

This work is a first study of the relation between general agents implemented in commercial-like games and PX. We define commercial-like games as games eventually meant for selling versus games primarily used for research with no intention to face players. The objective is to present a plausible line of research whose ultimate goal is having a successful integration of GVGP agents, with general goals, in commercial games. This is a complex problem to solve, as games come in myriad genres often trying to elicit very different PX. We take a first step in tackling this problem by choosing a popular game type - ’capture the flag’. The player’s goal is taking control of an area, element or group of elements for a longer period of time than the opposite player, or team. The symbol of what is being controlled is called *the flag*.

Hypothesis

We expected that a general agent whose behaviour is more suited to the game (*Exp.*), in this case being able to explore the map to take control of the *flags*, would create more ’tension’ in the player than an agent that focuses on winning alone (*Std.*). From our casual playtests, we could see that the players playing against the exploration type agent were feeling more ’under pressure/tensed’ yet enjoying the game equally; some actually preferring the tenser game experience. To test these relations, we measure pressure/tension and enjoyment as PX constructs (see below). We hypothesised that (1) pressure/tension will be significantly higher for *Exp.* (2) and that there will be no significant difference in enjoyment between *Exp.* and *Std.* Since the two chosen agents have different behaviors, we also do an open exploration on how they impact player’s perceived competence with (3) no specific hypothesis regarding competence.

The Game

A player-versus-NPC (agent) competitive game called *Skulls and Tombstones* was created in VGDL. The ’capture the flag’ type of game was chosen because it is simple, short and there is no complex strategy needed to win it. The core mechanic of controlling game elements is representative for a large number of existing game features across genres.

As seen in Figure 1, the game consists of two avatars, one assigned to the player and the other to the agent; trees, which act as walls and conform the limits of the map; and a series of skulls and tombstones spread around. The goal of the game is capturing tombs (turning them into the colour representing the player) by bringing a skull to a tomb. The skulls need to be collected by colliding with them. Picking a skull does not affect the score – points are not awarded or removed until a tomb has been captured. The player with more tombs captured when the time runs out is considered the winner. Only one skull is allowed to be carried at a time and it is possible to re-capture tombs that have been already coloured by the other player. The colour blue is assigned to Player0 (human); the colour red to Player1 (AI) and the time limit is set to 30s. The level was designed so the elements are distributed around the map. This kind of layout was aligned with an exploratory behaviour we were looking for the game to elicit.

Figure 1: Screenshot of *Skulls and Tombstones*.

General Agents

Thanks to the two-player GVGAI competition (Gaina et al. 2017), a series of search algorithms suitable for two-player games had already been created for the GVGAI Framework. In the study, we use Monte-Carlo Tree Search (MCTS) (Browne et al. 2012) for both agents as it shows good performance in the two-player competition (Perez-Liebana et al. 2019a). We built two versions of *Skulls and Tombstones*, each integrating the MCTS with a different goal. Both goals are general, so no details of the game are provided to the agent, but they are distinct enough to generate different behaviours. These two versions of the game were considered the two conditions for the study; more details of the agent used in each of them is given below. Pseudo-codes are also included, where H is considered an arbitrary high value.

Sample MCTS (Std.) This is a vanilla implementation of the algorithm as described in (Browne et al. 2012). This agent is provided in the GVGAI framework and has been used without making any modifications to it. Its heuristic is general, with the goal to win by maximising the score. Algorithm 1 is the pseudocode of the value function.

Exploring-Encouraged MCTS (Exp.) The Sample MCTS was updated inspired by (Guerrero-Romero, Louis, and Perez-Liebana 2017). They took a series of sample agents from the GVGAI Framework, isolated the state evaluation and plugged in a series of heuristics that go beyond just winning or maximising the score. The same approach has been followed in the *Exp.* MCTS, so the core of the algorithm or its design parameters have not been modified. The main difference comes from its goal (heuristic) that allows the agent to get rewards by visiting new positions in the map. The agent used in this experiment is also rewarded by the conventional rewards (winning and score) but encouraged to visit those tiles of the map where the agent has been fewer number of times. Also, the score is used as the difference between the current score of the game and the score resulting in the forward model, versus just the latest as in *Std.*. For balancing the exploratory heuristic, the final tuning was reached by carrying out trial and error and analysing the trace of the rewards. We gave high priority to exploring the map and visiting new positions while allowing the agent to score or win if it gets the chance. Algorithms 2 and 3 shows the pseudocode of the value function implemented. $C = 10$ is a constant that scales the heuristic value.

Algorithm 1 Pseudocode of the value function in *Std.*

if $is\_EndOfTheGame()$ **and** $is\_Loser()$ **then**
  return H^-
else if $is\_EndOfTheGame()$ **and** $is\_Winner()$ **then**
  return H^+
return $get\_StateGameScore()$

Algorithm 2 Pseudocode of the value function in *Exp.*

if $is\_EndOfTheGame()$ **and** $is\_Loser()$ **then**
  return H^-
else if $is\_EndOfTheGame()$ **and** $is\_Winner()$ **then**
  return H^+
else if $is\_OutOfBounds()$ **then**
  return H^-
return $EncourageExplH()$ + (new_score - $game\_score$)*C

Algorithm 3 Pseudocode of the heuristic calculation by $EncourageExplH()$, used in the value function for *Exp.*

if $nTimesVisitedPosition() > 0$ **then**
  return $-C * nTimesVisitedPosition()$
return C

Agent Behaviour To score in *Skulls and Tombstones* it is needed to perform two actions: (a) get the skulls (which does not affect the score) and (b) carry them to the tombs. *Std.* exclusively focuses on winning and maximising score so when the roll-out is not long enough to discover the potential score rewards, it is unable to find them, facing a large, flat reward landscape. As a result, the agent's decision-making becomes highly arbitrary, resulting in being idle or circling randomly on the map. In contrast, while *Exp.* also takes winning and maximising the score into consideration, it is driven by an exploratory heuristic. This heuristic makes the agent to move all over the map, visiting positions (tiles) it has not yet visited or visited the least number of times. Therefore, the agent is more likely to collect the skulls scattered around the map and to see the reward given by the score change, acquiring a dynamic behaviour.

Method

Participants We recruited 50 participants of which we had to exclude 15 due to some minor errors by the researchers or the fact that they did not fill the entire questionnaire form. 35 participants - 15 female and 20 male - were recruited in person. All participants were 18 or above years of age. We collected play behaviour data and made sure that all participants had played games before and were familiar with the studied game genre and controls. Players were picked from a diverse pool and were not limited to just students or any one category. They were spread across demographics and gaming abilities. None of the participants had previously played the chosen game. The players were provided with the game information sheet[1] before they consented to take part in the study.

Material The game (*Skulls and Tombstones*) described above was a standalone executable played at university computers. The aim of the game was simple and visual enough for players to be clear about the game goals and rewards. The game had no music or sound effects as feedback in any of the conditions. Game sessions were made to quit after 30 seconds each. Players played the entire game for 3 rounds, 90 seconds in total.

Procedure The study was a between participant setup (Shaughnessy, Zechmeister, and Zechmeister 2000), where two different groups of participants played the two different versions (game with *Std. agent* and game with *Exp. agent*). Participants were recruited in person alternating between the conditions with the conducting researchers being present throughout the process. Participants were given an information sheet along with instructions how to play the game, and asked for their consent and demographic details (age, gender, gaming experience). Players played the game three times and then answered questions as explained in the information sheet provided at the start of the experiment. (In playtests, during the game's development, we learned that the game and its controls were simple enough to understand in one playthrough. Playing three times gave players enough gameplay experience to answer PX related questions). All players were asked to play on an iMac using the keyboard. The game was played with the four arrows where only corresponding directional actions were allowed ($\leftarrow$, $\rightarrow$, $\uparrow$ and $\downarrow$). Players were made to play in a quiet zone for the duration of the study. Their game log was stored in real-time, while IMI and PENS questionnaires were filled out after they finished the game. The questionnaires were filled one after the other. Players were debriefed at the end.

Questionnaires We use IMI and PENS (7 point Likert) scales which are frequently used and validated questionnaires for PX. IMI is based on SDT and PENS is designed to capture SDT components in players. These

[1] https://osf.io/tmc6x/

scales are most directly related to the motivational model we use. We chose particular sub-scales from each to assess specific PX we were interested in. IMI comprises of seven sub-scales: interest/enjoyment, perceived competence, effort, value/usefulness, pressure/tension, and perceived choice. The interest/enjoyment sub-scale is considered the self-report measure of intrinsic motivation, while the pressure/tension sub-scale is considered to capture outer pressures to perform an activity. The PENS scale has subscales for the three basic needs (competence, autonomy, relatedness), as well as presence and intuitive controls. For the present study, we adopted the ’pressure/tension’ (5 questions), ’perceived competence’ (6 questions), and ’enjoyment’ (7 questions) IMI sub-scales. PENS ’competence’ (3 questions) sub-scale scale was used in addition only because it is specifically designed for video games, unlike the general IMI, however it did not have all the components like ’enjoyment’ and ’tension’ that we were interested in. We chose these particular sub-scales to suit our study design.

Logs Gameplay data was logged of each session containing: name of the AI integrated in the game, human and AI score per game tick, number of total game ticks (400), final scores and winner (*human*, *AI* or *draw*).

Pre-Study

A preliminary study was conducted with 38 participants with the same hypothesis but the experimental settings weren’t as clean. We found promising results: (1) ’pressure/tension’ was significantly higher when players were playing against *Exp.* and (2) players felt significantly more competent when playing against *Std.*. Based on this we decided to conduct the study again with more rigour which we present in this paper.

Results

We examine whether the difference in agents’ behaviour impacted ’pressure/tension’, ’perceived competence’ and ’enjoyment’. Our primary hypothesis was that the *Exp.* agent, as its behaviour fits the characteristics of the game created, will feel harder to beat, leading to higher pressure/tension. We further hypothesised that *Exp.* will not be less enjoyable than *Std.*, even if players feel higher ’tension’ against it. Lastly, we wanted to explore in which version of the game do players perceive themselves as more competent.

Tension

The players feel more tense while competing against *Exp.* (see Figure 2). This is demonstrated by a nearly significant two tailed t-test (t=2.02, df=33, p=0.051) in support of the hypothesis with an effect size of Cohen’s d=0.68. The effect size lying between a medium (0.5) and large (0.8) based on Cohen’s suggestions makes us consider the results in the direction of the hypothesis. The higher tension is expected for players who feel more nervous while competing against an agent which can explore the game map better. This follows the same tendency as we saw in the pre-study (much more significantly).

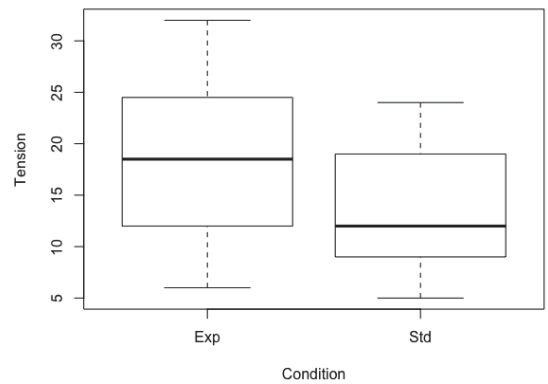

Figure 2: Tension in the game with *Exp.* and *Std.* agents.

Perceived Competence

We had no particular hypothesis regarding ’perceived competence’ yet we wanted to explore this component. As measured by IMI, players feel more competent when competing against *Std.*, demonstrated by a two tailed t-test (t=-3.93, df=33, p$<$ 0.001) with a large effect size of Cohen’s d=-1.33. Interestingly, competence as measured by the PENS sub-scale does not show a significant difference between the two conditions (t=-1.7, df=33, p=0.099).

Enjoyment

We found no significant difference between how enjoyable player’s found the two game versions (t=0.99, df=33, p=0.329). We expected that the *Exp.* version of the game would not be less enjoyable than the *Std.* version.

Discussion

With this work we start building a bridge between GVGP agents and PX. We hope to make it possible for game designers looking at PX to integrate GVGP agents in their games. To this end, we evaluate GVGP agents in how they impact player motivation rather than their ’raw’ performance as measured in scores or winning rates. This paper presents a novel approach where we apply PX measures to games with GVGP agents. As a first step, we specifically use the GVGAI framework, tackle general agents as NPCs and focus on a search algorithm: MCTS.

One of the main contributions of this work is that we demonstrate how in short term researchers and designers can make general agents with varying heuristics, integrate them in small games and then test how that effects PX. Using GVGAI allowed us to make a simple game, however visual feedback for players was limited. We iterated over the game design and play-tested until it was self-explanatory and engaging. Using changing colors on the tombs as a feedback of score allowed us to overcome UI limits of the framework. We chose specific PX measures like that of ’tension’ since *Exp.* was expected to behave in a fashion that would be perceived as more purposeful, making the gameplay more tense (pressure/tension). We expected that in this case such suspenseful tension would not make the game less enjoyable than playing against *Std.* measured with 'enjoyment'. We suggest researchers to pick PX based on the kind of game they choose and the experiences they are interested in studying. Using PX questionnaires is only one method to do this; more open ended investigations can be done with qualitative research.

The study described in this paper was not conducted to show if one agent was *better* than the other, but to show how two general agents with different behaviour sets have an effect on specific components of PX. We found that the *Exp.* created more tension in comparison to *Std.*. While 'tension' is a negative indicator of intrinsic motivation, it still constitutes an important experiential factor in many games (Seif El-Nasr et al. 2006). They are designed to make players feel tension, for example horror games or 'tight' arcade-style games like *Super Meat Boy* (Team Meat 2010). We would like to add that more tension does not necessarily equal a better or worse PX, also shown by 'enjoyment' results that did not show any significant difference between the two conditions. We merely argue that a specific degree of experienced tension is often an important design goal for game designers. From the IMI 'perceived competence' results, we found that the players felt more competent when playing against the *Std.* agent. However, according to PENS, we found no significant difference in perceived competence of the players whether they played against *Std.* or *Exp.*. If we examine the 'competence' sub-scale of PENS, it entails items that assess whether the game's challenge was perceived to *match* the player's skills. The underlying rationale here is that a better match should result in higher perceived competence, as players perceive successes as 'well-earned' and failure as 'near-miss'. In contrast, the IMI items of 'perceived competence' sub-scale focus on how well a person can perform the given task regardless of the difficulty *match*. Based on PENS results, the two versions don't differ significantly in perceived 'competence' in terms of difficulty-skill *match*. Designers and researchers need to choose what they are interested in as there is a large range in which challenge can be presented to players. For some games interaction difficulty match could be an important aspect (e.g. *Super Mario Bros*. (Nintendo 1985)) while it could be inconsequential for other games that do not pose interaction challenges but focus on emotional ones (e.g. *Walking Simulators* (Irwin 2017)).

Whenever game designers want to use a general AI to elicit aimed PX, we suggest that they are more likely to succeed if they judge the general AI behaviour from the perspective of the player. Existing work is primarily looking at general AI from the perspective of the game and how well they perform in it, excluding the player out of the equation where we believe that the player is central to such discussions. We believe that our proposed line of research would eventually extend the option for game designers to use general agents, especially as NPCs, for successfully eliciting desired PX in commercial games.

The approach proposed in this paper allows us to inspect agents in terms of the experience that the designers want their players to have: for instance, if a designer wanted the game to have more 'tension' they could use *Exp.* while if they wanted the players to feel more competent (in IMI sense) they could use *Std.*. It is possible to reduce the skill level of the AI by reducing the number of the iterations of the algorithm, as this is directly proportional to the playing skill up to a certain degree (Nelson 2016). This kind of tweaking becomes possible if the designers know what experience they want the players to have. They can then evaluate the agent for that experience and then adjust the heuristics to fine-tune the player's journey. Furthermore, if AI researches also choose to adopt this approach of evaluating agents by using player reactions, we would see the field broaden in terms of diversely behaved general AI, which suits the currently growing landscape of games.

Limitations and Future Work

The GVGAI Framework was originally created to compare general agents, not as a video game creation tool. Hence, the game controllers and interface were not ideal and could have an effect on the experience. It is possible to replicate our approach with more polished game prototypes built in other frameworks more suitable for game development. The work demonstrated in the paper could be extended to use other search algorithms or even learning algorithms. As a next step, we suggest conducting a similar study using *Unity* as they provide the *ML-agents* toolkit that can be used to train agents with learning techniques (Juliani et al. 2018).

We would like to underline that the general heuristics showcased in this paper are quite favourable to the type of game under consideration. We would like to encourage researchers to create more nuanced and diverse general heuristics that can be evaluated for different PX and emotions when they are accommodated to the needs of the type of game under consideration. Some examples of such general heuristics are proposed in (Guerrero-Romero, Lucas, and Perez-Liebana 2018).

We have used a quantitative method to evaluate PX, in which we would like to flag that our sample size could have been larger. There are non-aligning results between the two measures of competence and the pre-study results suggest that a larger scale replication would make this measure clearer. 'Tension' from IMI is not designed to capture game-like pleasurable tenseness but the item has face validity for capturing data to test our hypothesis. Given these initial results, we suggest using other research methods in addition to quantitative analysis for more diverse investigations. Qualitative methods for an in depth analysis or even eye tracking if the researchers or designers are interested in real time reactions.

In summary, as future work we would like to transfer the GVGP agents from GVGAI framework to more commercial game engines like Unity. We recommend testing the AI on other games to demonstrate that its PX patterns carry across games. We would also like to create a wider variety of games and heuristics to be studied from the perspective of player experience and player motivation.

Acknowledgments

We thank Andrew Martin and Sokol Murturi for their help. This work was funded by the EPSRC CDT in Intelligent

Games and Game Intelligence (IGGI) EP/L015846/1.

References

Abuhamdeh, S.; Csikszentmihalyi, M.; and Jalal, B. 2015. Enjoying the Possibility of Defeat: Outcome Uncertainty, Suspense, and Intrinsic Motivation. *Motivation and Emotion* 39(1):1–10.

Bernhaupt, R. 2015. User Experience Evaluation Methods in the Games Development Life Cycle. In *Game User Experience Evaluation*. Springer. 1–8.

Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; and Colton, S. 2012. A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in games* 4(1):1–43.

Caroux, L.; Isbister, K.; Le Bigot, L.; and Vibert, N. 2015. Player-Video Game Interaction: A Systematic Review of Current Concepts. *Computers in Human Behavior* 48:366–381.

Chen, J. 2013. Designing Journey. https://www.gdcvault.com/play/1017700/Designing. Accessed: 2020-01-31.

Creative Assembly. 2013. *Total War: Rome II*.

Deci, E. L., and Ryan, R. M. 2003. Intrinsic Motivation Inventory. *Self-Determination Theory* 267.

El-Nasr, M. S.; Drachen, A.; and Canossa, A. 2016. *Game Analytics*. Springer.

Emmerich, K., and Masuch, M. 2016. The Influence of Virtual Agents on Player Experience and Performance. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, 10–21.

Gaina, R. D.; Couëtoux, A.; Soemers, D. J.; Winands, M. H.; Vodopivec, T.; Kirchgeßner, F.; Liu, J.; Lucas, S. M.; and Perez-Liebana, D. 2017. The 2016 Two-Player GVGAI Competition. *IEEE Transactions on Games* 10(2):209–220.

Guckelsberger, C.; Salge, C.; Gow, J.; and Cairns, P. 2017. Predicting Player Experience without the Player. An Exploratory Study. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 305–315.

Guerrero-Romero, C.; Louis, A.; and Perez-Liebana, D. 2017. Beyond Playing to Win: Diversifying Heuristics for GVGAI. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, 118–125. IEEE.

Guerrero-Romero, C.; Lucas, S. M.; and Perez-Liebana, D. 2018. Using a Team of General AI Algorithms to Assist Game Design and Testing. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8. IEEE.

Hello Games. 2016. *No Man's Sky*.

Irwin, J. 2017. Devs Discuss the History and the Future of So-Called 'Walking Sims'. https://www.gamasutra.com/view/news/291307/Devs_discuss_the_history_and_the_future_of_socalled_walking_sims.php. Accessed: 2020-1-27.

Juliani, A.; Berges, V.-P.; Vckay, E.; Gao, Y.; Henry, H.; Mattar, M.; and Lange, D. 2018. Unity: A General Platform for Intelligent Agents. *arXiv preprint arXiv:1809.02627*.

Kunanusont, K.; Lucas, S. M.; and Perez-Liebana, D. 2018. Modeling Player Experience with the N-Tuple Bandit Evolutionary Algorithm. In *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Levine, J.; Bates Congdon, C.; Ebner, M.; Kendall, G.; Lucas, S. M.; Miikkulainen, R.; Schaul, T.; and Thompson, T. 2013. General Video Game Playing. In *Artificial and Computational Intelligence in Games*, volume 6. Dagstuhl Follow-Ups. 77–83.

Mekler, E. D.; Bopp, J. A.; Tuch, A. N.; and Opwis, K. 2014. A Systematic Review of Quantitative Studies on the Enjoyment of Digital Entertainment Games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 927–936.

Nelson, M. J. 2016. Investigating Vanilla MCTS Scaling on the GVG-AI Game Corpus. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–7. IEEE.

Nintendo. 1985. *Super Mario Bros.*

Perez-Liebana, D.; Samothrakis, S.; Togelius, J.; Schaul, T.; Lucas, S. M.; Couëtoux, A.; Lee, J.; Lim, C.-U.; and Thompson, T. 2016. The 2014 General Video Game Playing Competition. *IEEE Transactions on Computational Intelligence and AI in Games* 8(3):229–243.

Perez-Liebana, D.; Liu, J.; Khalifa, A.; Gaina, R. D.; Togelius, J.; and Lucas, S. M. 2019a. General Video Game AI: A Multitrack Framework for Evaluating Agents, Games, and Content Generation Algorithms. *IEEE Transactions on Games* 11(3):195–214.

Perez-Liebana, D.; Lucas, S. M.; Gaina, R. D.; Togelius, J.; Khalifa, A.; and Liu, J. 2019b. General Video Game Artificial Intelligence. *Synthesis Lectures on Games and Computational Intelligence* 3(2):1–191.

Rabin, S. 2013. *Game AI Pro: Collected Wisdom of Game AI Professionals*. CRC Press.

Rabin, S. 2015. *Game AI Pro 2: Collected Wisdom of Game AI Professionals*. CRC Press.

Ryan, R. M., and Deci, E. L. 2017. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York: Guilford Press.

Ryan, R. M.; Rigby, C. S.; and Przybylski, A. 2006. The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion* 30(4):344–360.

Schaul, T. 2013. A Video Game Description Language for Model-Based or Interactive Learning. In *2013 IEEE Conference on Computational Inteligence in Games (CIG)*, 1–8. IEEE.

Seif El-Nasr, M.; Niedenthal, S.; Kenz, I.; Almeida, P.; and Zupko, J. 2006. Dynamic Lighting for Tension in Games. *Game Studies Journal* 7(1).

Shaughnessy, J. J.; Zechmeister, E. B.; and Zechmeister, J. S. 2000. *Research Methods in Psychology*. McGraw-Hill.

Team Meat. 2010. *Super Meat Boy*.

Thatgamecompany. 2012. *Journey*.

Thompson, T. 2018. Revolutionary Warfare The AI of Total War (Part 3). https://www.gamasutra.com/blogs/TommyThompson/20180212/314399/Revolutionary_Warfare__The_AI_of_Total_War_Part_3.php. Accessed: 2020-01-24.

Torrado, R. R.; Bontrager, P.; Togelius, J.; Liu, J.; and Perez-Liebana, D. 2018. Deep Reinforcement Learning for General Video Game AI. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8. IEEE.

Tyack, A., and Mekler, E. D. 2020. Self-Determination Theory in HCI Games Research: Current Uses and Open Questions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–22.

Yannakakis, G. N., and Togelius, J. 2011. Experience-Driven Procedural Content Generation. *IEEE Transactions on Affective Computing* 2(3):147–161.

Yannakakis, G. N., and Togelius, J. 2018. *Artificial Intelligence and Games*. Springer.