Evaluating Planning-Based Experience Managers for Agency and Fun in Text-based Interactive Narrative

Alejandro Ramirez\(^1\) and Vadim Bulitko\(^1\) and Marcia Spetch\(^2\)
Departments of Computing Science\(^1\) and Psychology\(^2\)
University of Alberta
Edmonton, Alberta, T6G 2E8, CANADA
{ramirez|bulitko|mspetch}@ualberta.ca

Abstract

Artificial intelligence (AI) techniques have been applied to video games to make the overall experience more enjoyable. In games with interactive storytelling (IS), player actions can substantially affect plot events and plot characters. Therefore, AI planning techniques have been used to shape the plot in response to player actions that conflict with authorial goals. While such methods are poised to increase player fun and agency, two recent implementations (ASD and PAST) have not been formally evaluated to date. In this paper we do so via a series of user studies for the first time. We show that ASD significantly enhances fun and agency, whereas PAST gets mixed results with an interaction between effects of the experience manager and player prior gaming experience in one user study, and marginally significant results for increased agency in a study with a constrained story domain.

Introduction

Research on interactive storytelling (IS) is at the intersection of artificial intelligence (AI), interactive entertainment, and traditional storytelling. It is important both because storytelling is deemed to be cognitively rich and because implementing such a storyteller with AI is likely to make video games more interactive, narratively complex and customized for a particular audience.

IS systems have been usually implemented around the concept of a drama manager: “an intelligent, omniscient, and disembodied agent that monitors the virtual world” in order to intervene and improve “some model of quality of experience” (Riedl and Bulitko 2013). A generalization of this idea is the experience manager, which extrapolates this concept to non-dramatic situations in a wide range of interactive environments (Riedl 2012). In the context of interactive narrative, fun (i.e., how entertaining and amusing an interactive experience is) and agency (i.e., the ability of the player to affect the virtual world) are frequently deemed important goals targeted by IS systems.

While some experience managers can improve fun and agency in an interactive narrative, they do so by relying heavily on story annotations and extra game content, therefore increasing the authorial burden on the developers.

To alleviate the authorial burden, AI planning techniques have been used to automatically anticipate ruptures in the original narrative due to player actions and repair it, aiming to increase players’ agency. Of particular interest is the system called Automated Story Director (ASD), which represents stories as formal plans and computes narrative repairs to retain story goals a priori specified by the author (e.g., an evil wolf eats granny in The Little Red Riding Hood folk story) (Riedl et al. 2008).

More recently, this approach was extended with a player model which enabled automatic customization of such repairs to the particular player type as detected during gameplay (Ramirez and Bulitko 2012). The resulting system, called Player-Specific Automated Storytelling (PAST), requires the author to additionally specify a mapping between narrative events and player types (e.g., killing the wolf likely indicates a fighter inclination).

Despite their promise, neither of the two experience managers (ASD and PAST) has been formally evaluated with a user study. In this paper we do so in the context of text-based narratives centered on The Little Red Riding Hood folk story. The results indicate that ASD can increase player fun and agency compared to a fixed-narrative baseline. On the other hand, the use of player modelling within PAST did not consistently improve fun or agency in one evaluation with a broad story, but improved agency in another evaluation with a more constrained story.

The rest of the paper is organized as follows. We first define the problem formally and then review related work. We then describe the two experience managers in detail and present the empirical test bed employed in the user studies. Finally, we present and discuss the results and conclude the paper with future work directions.

Problem Formulation

The problem tackled in this paper is to determine if planning-based experience managers can deliver on their promise of increasing player fun and agency in the context of an interactive narrative experience. Specifically, the scope is set around planning-based managers that allow player actions outside the exemplar narrative, and that use automated planning to repair the narrative to a consistent state. This can be measured with a post hoc instrument (a psychologically validated questionnaire) in the context of a user study.

The standard setup is to have an experimental and a control condition which are different only in the use of the experience manager being evaluated. One important con-
sideration here is the balance of story content across these conditions. This is fundamental to prevent fun and agency measured in the experimental condition from being affected merely by player’s experience of a particular influential story content, absent from the control condition. Such balance is usually achieved by having players in the control condition play content experienced by random players from the experimental condition (known as “yoking” in psychology). Consequently, players in both conditions cover similar “story ground” in approximately the same proportion.

Related Work
While several IS approaches have been proposed in the last forty years, from choose-your-own-adventure books by Edward Packard (Kraft 1981) to modern experience and drama managers (Riedl and Bulitko 2013), formal evaluations of their performance have been scarce.

One notable exception is PaSSAGE (Thue et al. 2007; 2010; 2011) which was evaluated via a series of user studies accumulating over one thousand subject hours. While significant effects of using a player model to select content were demonstrated, the experience manager within PaSSAGE was manually scripted and did not use automated content generation techniques.

User studies have also been carried out for performance metrics other than fun and agency. For instance, a suspense manager was evaluated in a user study (Cheong and Young 2008) and significant benefits were found. Data-driven approaches (Yu and Riedl 2012; 2013) and “interestingness” ratings (Sharma et al. 2007) have been used for player modelling. The results indicate that it is possible for an experience manager to model a player cognitively. However, none of these studies evaluated planning-based experience managers for improving player fun and agency.

Planning-Based Experience Management
In this section, we briefly present the foundation of the two experience managers evaluated in the paper. Both ASD and PAST are based on representing narratives as formal plans. To illustrate, we will use an interactive extension of the classical narrative of The Little Red Riding Hood in line with the previously published work on ASD and PAST. In the story, a child (red) travels through a forest to deliver food to her grandmother (grandy). She is confronted by an evil wolf (wolf) and the player decides in an interactive fashion how the narrative unfolds from there on. Any event in this interactive experience is a planning operator, denoted here with a Lisp-like notation. For instance, the event of the wolf eating the granny is encoded as the planning operator (eat X Y) which has the preconditions (alive Y) and (hungry X) and the post-condition (eaten Y) with X bound to the wolf and Y to the granny.

In this approach, the author specifies the so-called exemplar narrative. This is done by listing some game events and causal and temporal dependencies among them. Note that the exemplar narrative needs not be linear. For instance, with the original exemplar narrative of ASD (Riedl et al. 2008), the player can experience the events in different temporal orders: the wolf can eat Little Red first or he can eat her granny first (Figure 1).

The agency the player has in the interactive story can, however, lead to ruptures where a precondition of an event is invalidated. For instance, the player could choose to (kill wolf) thereby establishing the post-condition of (:not (alive wolf)) and thus invalidating the authorial goal (eaten granny). Such ruptures can be foreseen by the AI planner and repairs to the narrative can be precomputed in response. The rupture above can be repaired by having a fairy resurrect the wolf. The authorial goals are shown as narrative “islands” and the outcome in the figure.

Frequently, a given rupture can be repaired in multiple ways necessitating, a method to select among them. Whereas ASD used an ad hoc method of selecting between contingency narratives, PAST uses a player model to do so. For instance, if the player has demonstrated a fighting inclination then PAST could fix the rupture resulting from (kill player wolf) by introducing a fiery pack of wolves. On the other hand, if the player is deemed to be more of a method actor then PAST could have a magic fairy peacefully resurrect the wolf.

In PAST, the player model is a vector of five numbers, each indicating an inclination towards a certain style of play: fighter, power gamer, storyteller, method actor, and tactician, based on Robin Laws’ player types (Laws 2001). This approach was previously found effective within PaSSAGE although without any automated planning (Thue et al. 2007). The model is initialized with all components being equal. Afterwards, the vector is adjusted based on annotations of

---

**Figure 1:** An interactive version of The Little Red Riding Hood employed in the studies: solid lines are causal links that can be threatened by ruptures; dotted lines are temporal links. Events: H (greet red wolf), T (tell-about red wolf granny), G (eat wolf red), R (eat wolf granny), K (kill hunter wolf), E (rescue hunter red), S (rescue hunter granite), C (give granite cake). Reproduced from (Ramirez and Bulitko 2012).
the actions the player takes (e.g., killing the wolf when other options are available indicates a fighter inclination). Each action and rupture in the virtual domain has to be annotated by the author; PAST then uses these to calculate a global annotation for each contingency narrative.

We made minimal changes to PAST from the original design (Ramirez and Bulitko 2012) for our user studies. Specifically, the original PAST precomputed all contingency narratives for all possible ruptures during an off-line phase based on player archetypes. The version of PAST evaluated in this paper forgoes the off-line precomputation. As a result, it saves memory and time off-line and allows to select a narrative repair more closely fitting the player model at hand (instead of one of the five generic model archetypes). Additionally, the planning algorithm is invoked ad hoc only when a rupture takes place.

Evaluation

To evaluate ASD and PAST we constructed a text-based interactive adventure–inspired partly by RPG text-based adventure games–around the exemplar narrative of The Little Red Riding Hood, in line with the literature (Riedl et al. 2008; Ramirez and Bulitko 2012). Three different versions were developed. The version used to evaluate ASD had three possible ruptures and enough content for the AI planner to repair any of them. The version used to evaluate PAST had five ruptures and three to five possible contingency narratives for each so that a repair suitable for different player types can be chosen. The latter also had a prologue with enough interaction for the system to populate the player model before the first rupture would take place. The third version was also created to evaluate PAST, and had only one rupture and five different repairs available.

Methodology

Three user studies were conducted. Participants gave written consent and received a university course credit for their participation. These participants were tested in groups of ten to fifteen people, with sessions lasting approximately 30 minutes. The participants werebriefed before the start, and then interacted with the story, selecting from on-screen choices (Figure 2). Upon completion, an evaluation survey was administered and a debriefing form was provided.

For the survey, we used a previously validated instrument (Vermeulen et al. 2010). The instrument had twelve questions aimed at measuring player fun and six questions aimed at their agency. Each question was a 5-point Likert scale from completely disagree to completely agree. The scores on all fun questions were added together and then normalized using the maximum and minimum possible scores. The result is referred to as the fun score. The same was done for the agency score. The instrument has been previously used to evaluate experience managers (Thue et al. 2011).

Because our experiment consisted of a reading assignment, there was an inherent risk that some participants might skim through the text without reading it. To reduce noise caused by such skimming we used task completion time to estimate the reading speed for each participant and compared it to average reading speeds reported in the literature. First-year college students are reported to skim text at 450 words per minute (Carver 1993). Given the 10% slow down due to reading off the screen (Zeffle 1998), we eliminated data from all subjects that read above 405 words per minute. We then additionally removed data from all subjects whose fun and/or agency score was an outlier according to Tukey’s outlier filter (also known as interquartile outlier detection). The survey also measured participant gaming skills, in order to classify them into gamers (one hour or more of gameplay per week) or non-gamers (less than one hour per week).

The data for each metric were analyzed using a two-way ANOVA ($\alpha = 0.05$) with the factors being experience management and gaming skills. The data did not violate the assumptions of independent observations, normality (assessed by the Kolmogorov-Smirnov test) or unequal variance (assessed by the Levine’s and Welch’s tests). Post hoc comparisons used Tukey’s HSD test.
Results

User Study 1: ASD. The goal of the first user study was to compare ASD to an exemplar narrative. The control condition were participants who experienced the Little Red Riding Hood interactive narrative with choices limited to changing the temporal order of events. The experimental condition was additionally allowed to rupture the narrative at up to three plot points. The number of participants was $n = 81$ ($n = 98$ before removing outliers). There were 42 subjects (21 gamers, 21 non-gamers) in the experimental condition and 39 (23 gamers, 16 non-gamers) in the control condition.

The experimental condition showed significantly higher scores for both fun ($F(1,76) = 5.46, p = 0.02$) and agency ($F(1,77) = 20.46, p = 0.001$) compared to the control condition. The interaction between gaming skills and experience management was not statistically significant. In the first row of Figure 3 means and standard errors are shown.

User Study 2: PAST. In the second user study, we evaluated the effects of player modelling in PAST. We had three different conditions for the main factor of experience management: control, experimental and exemplar; players had the same set of choices in all three. In the experimental condition, narrative ruptures were repaired according to the player model learnt for the actual player. In the control condition, the player model was taken from a randomly drawn player in the experimental condition (as a part of the yoking
procedure). If random model coincided with the player’s actual model then the subject would be counted in the experimental condition. The exemplar condition contained all players who never deviated from the exemplar narrative (i.e., never caused a rupture).

In total, \( n = 72 \) participants \( (n = 98 \text{ before outlier removal}) \) were evaluated. The experimental condition had 23 subjects \( (8 \text{ gamers}, 15 \text{ non-gamers}) \), the control condition had 22 subjects \( (6 \text{ gamers}, 16 \text{ non-gamers}) \) and the exemplar condition had 27 subjects \( (8 \text{ gamers}, 19 \text{ non-gamers}) \).

No statistically significant results were observed in the main factors for either fun or agency. However, a significant interaction between these was found for fun \( (F(2, 66) = 4.73, p = 0.012) \). Agency showed a similar interaction but it failed to reach significance \( (F(2, 67) = 1.93, p = 0.1538) \). Means and standard error bars are depicted in the bottom row of Figure 3. Tukey’s HSD post hoc test showed that non-gamers in the control condition had significantly higher fun ratings than non-gamers in the experimental condition; no other comparisons reached significance.

We attempted to replicate user study 2 later in the academic term with \( n = 63 \) participants \( (n = 83 \text{ before outlier removal}) \). The subjects were distributed over the three conditions as follows: experimental: 19 \( (2 \text{ gamers}, 17 \text{ non-gamers}) \); control: 19 \( (5 \text{ gamers}, 14 \text{ non-gamers}) \); exemplar: 25 \( (10 \text{ gamers}, 15 \text{ non-gamers}) \). This time, no significant effects were found, which we speculatively attribute to the lack of motivation in end-of-the-term university students who signed up for the study late and likely to simply get course credit.

User Study 3: PAST with Constrained Domain. In the third user study, we also studied the effects of player modelling and planning; however, we constrained the story space to only 5 \( (\text{down from } 31 \text{ stories in user study } 2) \) possible different stories by permitting only one rupture at a fixed time for all participants. Furthermore, only two conditions were present for the experience management factor: control (ASD) and experimental (PAST).

The total number of participants was \( n = 34 \ (n = 41 \text{ before outliers were removed}) \), with 17 subjects in control \( (5 \text{ gamers}, 12 \text{ non-gamers}) \) and 17 subjects in experimental condition \( (4 \text{ gamers and } 13 \text{ non-gamers}) \). No statistically significant results were found for fun \( (p < 0.05) \); however, a marginally significant result \( (p < 0.10) \) for agency \( (F(1, 67) = 5.46, p = 0.097) \) was found. No statistically significant interactions were found.

The Pull-back Effect. Some participants appear to have experienced a pull-back effect wherein they felt that upon breaking the narrative through their choices they were pulled back into the original plot. To illustrate, here are two comments collected from the user studies:

*The plot of events seems predetermined, anything the user chooses may eventually loop back to the static plot.*

*(...) It was disappointing when I would try to go away from the typical Red Riding Hood story and then it would just inevitably go back to that (it made me feel like my input in what happened next didn’t matter).*

Discussion

The results for the first user study indicate that ASD increases player fun and agency over the exemplar narrative. This is despite the pull-back effect reported by some participants in the comment’s section.

The results of the second user study did not show significant effects of player modelling in PAST for either player
fun or agency. This is surprising since PAST is a combination of ASD and a player model from PaSSAGE, which in isolation showed an effect in increasing fun and/or agency. We believe several factors may cause this.

Firstly, the PAST user study had fewer participants than the ASD study split into more conditions, thus causing a loss of statistical power. Second, the latter part of the study was run at the end of the academic term, possibly affected by a lack of motivation in “last minute” sign ups, potentially adding noise. Third, the richer narrative space (i.e., more ruptures and more diverse repairs) could have induced a higher variance. To illustrate, Figure 5 shows the uneven distribution of players over different stories (i.e., each story being a collection of different events). Our attempt to control variance by randomly assigning player models from the experimental subjects to control condition subjects (yoking) accounted only for the types of repairs, hence not controlling for the actual ruptures taken. Lastly, perception might differ between gamers and non-gamers. Since there is no statistical significance, it is impossible to tell if this is a fluke or an indicative result. If the latter is the case then PAST may increase fun for gamers but not for non-gamers in this domain.

The third user study, conducted in a constrained story domain, obtained marginally significant results that may indicate that agency was increased by using PAST when compared to ASD. A higher significance level might have been obtained with more participants.

**Conclusions and Future Work**

This paper presents the first formal evaluation of two recent planning-based experience managers in the context of IS. Both were evaluated with subjects playing a text adventure based on a folk story on the basis of their self-reports of fun and agency. It appears that planning-based experience management can improve fun and agency over even a non-linear exemplar narrative. Player modelling shows mixed results: in one large story domain, positive results are seen in gamers, whereas agency is increased in a constrained story domain. This highlights that the balance of content is of particular importance when evaluating experience managers.

Results presented in this paper support previous conjectures that treating IS as AI planning is an effective way to improve player’s experience while affording them more choice within a narrative. Adding player modelling on top seems promising, but further evaluations are required. We feel that existing systems can be deployed in the field to automatically foresee narrative ruptures and automatically compute repairs. The resulting narratives can then be fleshed out with multimedia elements by game developers.

**Acknowledgements**

We would like to gratefully acknowledge the funding provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the help and consultations by David Thue, and the great work by Wayne DeFehr and Shane Riczu with different writing tasks for the prototypes.
References


