Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment
(AIIDE 2022)

Evaluating Reader Comprehension of Plan-Based Stories Containing Failed Actions

Rushit Sanghrajka [1], **R. Michael Young** [1,2]

[1] School of Computing, University of Utah
[2] Entertainment Arts and Engineering Program, University of Utah
Salt Lake City, UT, USA
rush.sanghrajka@utah.edu, young@eae.utah.edu

Abstract

A growing number of algorithms for story planning include the ability to create stories with failed actions – in particular failed actions that occur because of the mistaken beliefs of the characters attempting them. To date, most of these systems have been evaluated analytically, primarily by comparing their expressive range to prior story generation systems. Empirical evaluation of these systems has been preliminary. In this paper, we outline a general comprehension-based approach to the evaluation of plan-based story generation. We describe how we specialize it for use evaluating story plans containing failed actions, and we describe the design and results of an experiment using this approach to evaluate plot lines produced by HEADSPACE , a system that models the beliefs of characters and uses that model to generate plot lines containing actions that are attempted but that fail.

Introduction

In stories, characters commonly attempt to perform actions that fail (Lenhart et al. 2008). For example, when the terrorist mastermind Hans Gruber attempts to shoot Detective John McClane near the end of *Die Hard* (McTierman, J. 1988), his attempt fails as McClane waves the gun's magazine mockingly at Gruber and informs him that McClane had previously unloaded the gun. Examples like this highlight how mistaken beliefs and action failures are used intentionally by authors to build suspense or other dramatic effects by playing with the disparities of knowledge and ability between characters within the unfolding story world. Recent advances in automated story generation have demonstrated the ability to create plot sequences with failed actions (Sanghrajka, Young, and Thorne 2022; Young 2017b; Christensen, Nelson, and Cardona-Rivera 2020), but work characterizing the efficacy of those approaches has been preliminary.

This paper sets out an experimental methodology for evaluating plan-based story generation that includes failed actions, and uses that methodology to evaluate a specific narrative planning knowledge representation. In the discussion that follows, we briefly review a plan-based knowledge representation for characterizing plans with action failures based on mistaken character beliefs. This representation forms the basis for plans produced by the HEADSPACE planning algorithm (Sanghrajka, Young, and Thorne 2022; Young 2017b). We then describe an experimental evaluation of use of these plan data structures to create textual stories. In this evaluation, plan data structures are translated to both text examples and cognitive models of stories defined by cognitive psychologists. Human participants are asked to rate pairs of questions and answers about stories, and their ratings are compared to predictions made by the cognitive models to gauge the participants' understanding of the underlying story structure. The experiment demonstrates that the descriptions produced from HEADSPACE plans are understandable and serve effectively to convey stories' structure to readers when actions fail due to characters' mistaken beliefs.

Related Work

Work in narrative planning has looked at disparities in knowledge and how that pertains to the stories generated due to these disparities. Teutenberg and Porteous (Teutenberg and Porteous 2015) employ the power of local knowledge for characters to create narrative scenarios where characters can perform deceitful actions such as `feign-death`. Shirvani et al. (Shirvani, Ware, and Farrell 2017) expand the space of generated stories even further by allowing characters to imagine "possible worlds" based on their (possibly inconsistent) beliefs, and act according to their local model of the world. More recently, Christensen et al. (Christensen, Nelson, and Cardona-Rivera 2020) propose a domain compilation model that builds in failed actions and different beliefs into a PDDL compilation.

Evaluation for these planners has varied in approach. Teutenberg and Porteous report plan and heuristic metrics to describe the nature of the plans produced by experimenting with various narrative domains. They also describe the narrative qualities of their plan, which underscores the argument that qualitative analysis is more important to understanding narrative planner outputs than typical AI plan metrics such as minimum plan length. Shirvani, Ware and Farrell performed an empirical evaluation by generating stories and then having participants on Mechanical Turk answer questions about the stories to gauge their understanding compared to the model of the world in the problem set. Christensen, Nelson and Cardona-Rivera report performance metrics for narrative planning algorithms, focusing on processes of domain compilation (e.g., translating domains expressed in formalisms for specialized narrative planners into representations used by general purpose planners) on domains of different dimensions.

In contrast to analytic measures of planning-based story generation, other work employs *comprehension-based* methods. In these approaches, plans are translated into two parallel representations: a cognitive model serving as a proxy for a canonical reader's understanding of the plan's story, and a textual realization of the plan's plot.[1]. Human readers' understanding of the textual stories are compared to predictions made by the cognitive model to gauge comprehension, which is related to generative efficacy. One cognitive model of story comprehension, called QUEST (Graesser and Clark 1985), has been used in a comparative role in a number of studies to evaluate human comprehension of computationally generated narratives (Jhala and Young 2010; Bahamón and Young 2017; Ware and Young 2014). Christian and Young (Christian and Young 2004) first proposed a translation algorithm that can generate QUEST knowledge structure graphs from narrative plans. Their proposed translation algorithm allows for creation of graph-like QUEST representations from the output of narrative planners, allowing for a cognitive model to track comprehension of the events in the narrative plan. More recently, Sanghrajka, Lang and Young (Sanghrajka, Lang, and Young 2021) build upon this translation algorithm to translate narrative plans that include failed actions. As we describe below, we adopt this approach for evaluation, and describe specifics in the following sections.

Understanding HEADSPACE Stories

Generating HEADSPACE Plans

In this work, we represent the action that unfolds within a narrative-oriented virtual environment using a *plan*, a data structure characterizing the actions that occur in a story and the specification of the role that characters, objects and locations play in those actions. Plans and the planning systems that produce them have been widely studied in artificial intelligence research, though they are typically used to describe specifications of real-world action sequences. The plan structures we employ are those produced by the HEADSPACE planner as described by Sanghrajka and their collaborators (Sanghrajka, Young, and Thorne 2022).

HEADSPACEis a heuristic search planner intended to produce narrative plans whose structure reflects the plot structure of stories. HEADSPACE extends the FastForward general-purpose planning algorithm (Nebel and Hoffmann 2001) in several ways that make it more appropriate for story generation. First, the HEADSPACE knowledge representation is extended to represent character beliefs concerning conditions in the world. Actions in HEADSPACE have material preconditions and effects, much like the typical preconditions and effects used by STRIPS (Fikes and Nilsson 1990) and many other planning systems. But HEADSPACE actions also may describe *epistemic* preconditions and effects – descriptions of what conditions must be believed by the performing character prior to the action's attempt and how the action changes those beliefs upon successful execution. Second, HEADSPACE may build plans that contain actions that will fail upon execution. Within the plan, this occurs because actions whose epistemic preconditions are satisfied (that is, whose preconditions are satisfied in the beliefs of the characters performing the actions) do not have all their material preconditions satisfied in the story world at the time of their execution. Finally, HEADSPACE extends the typical STRIPS-like action representation to include characters' intentions modeling their commitments to action advancing their own goals. HEADSPACE plans track the intended plans of characters over time (Bratman 1987), and revises those intended plans when characters update their beliefs in ways that make their previously intended courses of action unexecutable.

More detail on the HEADSPACE algorithm can be found in the work of Sanghrajka, Young and Thorne (Sanghrajka, Young, and Thorne 2022; Young 2017a). Examples of two HEADSPACE plans used in our experiments are shown in Figures 1 and 4, and the first of these is described in more detail below.

Leveraging QUEST to Evaluate Comprehension of Plan-Based Stories

Understanding the effectiveness of any method used to create story structure is critical to a scientific approach to narrative. Our approach to the evaluation of HEADSPACE centers on measuring readers' comprehension of the stories produced by HEADSPACE plans. In particular, we build on human-centered evaluation processes used in prior work (Riedl and Young 2010; Ware and Young 2014; Farrell, Robertson, and Ware 2016; Jhala and Young 2010) that gauges a narrative planner's effectiveness by seeking to determine how people comprehend the stories it creates. As we describe below, we leverage significant work done by cognitive psychologists around narrative comprehension, basing our evaluation on the premise that when plan-based story generation is effective, readers' comprehension of generated stories will reflect an understanding of the plot that aligns with the structure of the plans used to produce them.

Graesser and Clark (Graesser and Clark 1985) developed a cognitive model for question-answering in the context of stories. The framework, called QUEST, describes conceptual graph structures that are built by readers during their comprehension of a story and that are used by them as a mental model of its plot. These graph structures are called QUEST knowledge structures, or QKSs. Nodes in QKSs correspond to events, goals, states and other elements in the story. Arcs between nodes corresponds to causal, temporal and intentional relationships between the nodes they connect. To model human question-answering in the context of stories, QUEST defines an arc search procedure that mirrors how humans traverses a QKS to characterize relationships between any two nodes in the graph. Significant experimental work has shown the QUEST model to be effective at

[1] In some work (e.g., that of Christian and Young(Christian and Young 2004) and Jhala and Young (Jhala and Young 2010), cinematic renderings are used rather than textual ones.

Arc Type	Details
Consequence (C)	**Definition** A causes or enables B and A precedes B in time **Node Type Constraints** (Event or State) $\xrightarrow{C}$ (Event or State)
Reason (R)	**Definition** $\{B$ is a reason or motive for A or B is a superordinate goal for $A\}$ and A is achieved before B is achieved **Node Type Constraints** Goal $\xrightarrow{R}$ Goal
Outcome (O)	**Definition** B specifies whether or not the goal A is achieved **Node Type Constraints** Goal $\xrightarrow{O}$ (Event or State)
Initiate (I)	**Definition** A initiates or triggers the goal in B and A precedes B in time **Node Type Constraints** (Event or State) $\xrightarrow{I}$ Goal

Table 1: Relevant set of arcs within the QUEST Knowledge structures as defined by Graesser et al. (Graesser and Clark 1985)

making predictions relating to an idealized reader's conception of a text (Graesser and Clark 1985; Graesser, Lang, and Roberts 1991a; Graesser and Franklin 1990; Graesser and Hemphill 1991).

Within a QKS, both nodes and arcs are typed based on their meaning and role in answering questions about stories. In brief, QKS nodes are categorized into state, event, and goal types, as well as several other types not directly relevant to or present in our analysis. State nodes describe material conditions in the world state. Event nodes describe processes in the world that change its state. Goal nodes describe a character's desired state of the world. Any action taken intentionally by a character has a corresponding goal and event node in the story's QKS. The six arc types that are crucial to event-oriented storytelling are displayed and described in Table 1.

QUEST defines a graph search procedure that approximates the cognitive processes used during human question answering about plot elements in stories. For instance, the search process can be used to provide *goodness of answer* (GOA) ratings, characterizations of how well an event corresponding to some node a serves as an answer to a *Why* question posed about an event that corresponds to another node b. The graph search process supports GOA ratings for several types of question-answer pairs, including three of relevance here: *why* questions (e.g., "Why did Charlie Brown fall down?"/"Because Lucy challenged him to kick a football."), *how* questions (e.g., "How did Ferdinand escape the bullpen?"/"He unlocked the bullpen gate."), and *what are the consequences of* questions (e.g., "What are the consequences of Luke talking to Cornelius Evazan in Mos Eisley's cantina?"/"Obi-wan chopped off Ponda Baba's arm."). In general, candidate answer nodes that lie farther in the QKS from the node at the focus of a question are considered poorer answers than candidate nodes that lie closer.

Prior work (e.g., (Riedl and Young 2010; Jhala and Young 2010; Farrell, Robertson, and Ware 2016; Cardona-Rivera et al. 2016)) has leveraged the QUEST model to gauge reader comprehension of automatically generated stories by using a multi-step approach. First, an automatic process is used to translate a plan data structure into a corresponding QKS (Christian and Young 2004; Sanghrajka, Lang, and Young 2021). Second, a text-based story is generated from the plan data structure using a simple, template-based approach. Human participants are asked to read the story and are then asked to provide GOA ratings for question-answer pairs drawn from its elements. Those ratings are compared to QUEST's GOA ratings for the QKS nodes corresponding to the same story elements. When human responses yield GOA ratings that align with those provided by QUEST, it is a strong indicator that the humans' cognitive model of the underlying story aligns with the structure of the story plan.

An Example Plan and Its Corresponding QKS and Story Text

To demonstrate the character of belief dynamics and failed actions in HeadSpace and to show the connection between a HeadSpace plan, a QKS data structure and a textual representation of the story they define, we give related examples of two in this paper – both stories drawn from the same Western-themed planning domain. The first example, which we call the Breakout problem, is one of two problems that were used in the experimental evaluation, and is an updated version of the example domain first described by Thorne and Young (Thorne and Young 2017). The second problem used in the experiment, called the Drink Refill problem, is described in the appendix. The plan we describe is shown in Figure 1, the corresponding QKS is shown in Figure 2, and the text based on the plan is shown in Example Text 1.

The Breakout example makes use of seven operators. They are Pickup, where a character picks up an object from the floor, Shoot-At-Door, where a character fires a gun they're holding at the lock of a door in order to break (and unlock) the door's lock, Check-Gun-Loaded-F, where a character opens the cylinder of a revolver that they're holding in their hand, then learns the revolver is unloaded, Load-Gun, where a character takes bullets that they're holding and loads a gun that's in their hand, Escape, where a character opens an unlocked jail cell door and walks through it to freedom.

For the example below, an informal sketch of the planning problem's initial state sets a character, Dolores, locked in a jail cell that has only one exit: a door that's locked. Fortunately, a deputy has foolishly left a gun and some bullets on a chair outside Dolores' cell, but within her reach. The goal for the story is for Dolores to be in the hallway outside her cell.

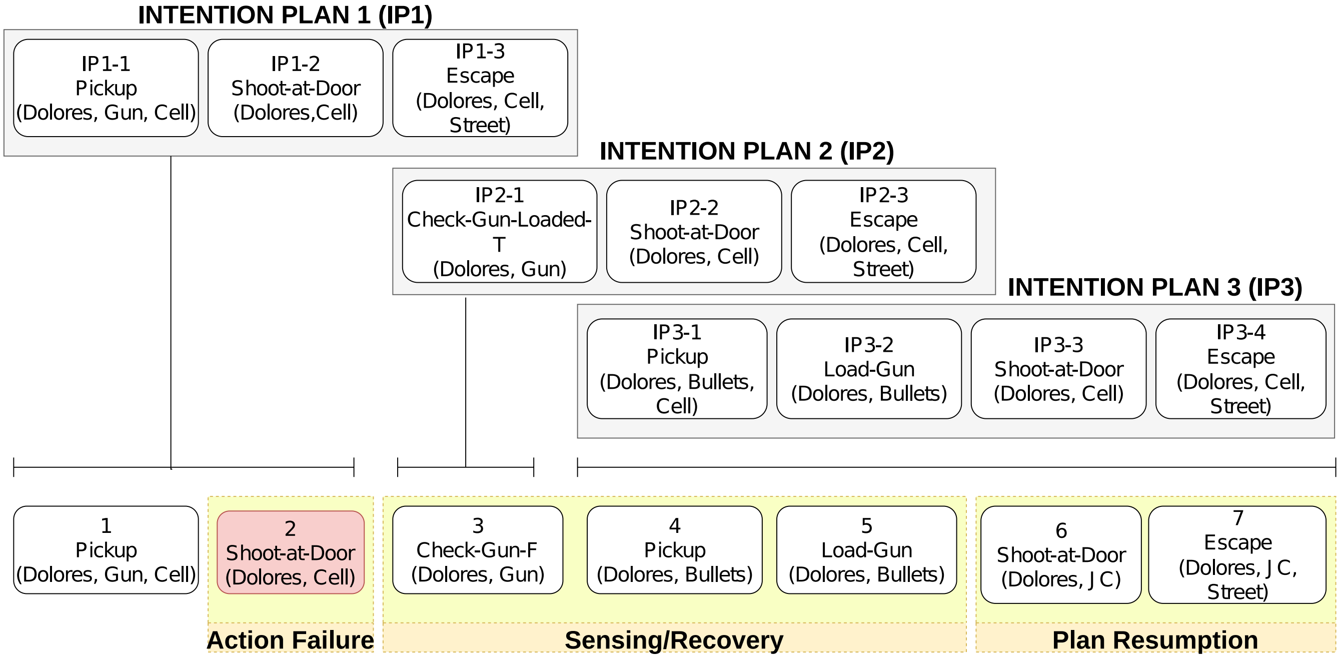
INTENTION PLAN 1 (IP1)
IP1-1 Pickup (Dolores, Gun, Cell)
IP1-2 Shoot-at-Door (Dolores,Cell)
IP1-3 Escape (Dolores, Cell, Street)
INTENTION PLAN 2 (IP2)
IP2-1 Check-Gun-Loaded-T (Dolores, Gun)
IP2-2 Shoot-at-Door (Dolores, Cell)
IP2-3 Escape (Dolores, Cell, Street)
INTENTION PLAN 3 (IP3)
IP3-1 Pickup (Dolores, Bullets, Cell)
IP3-2 Load-Gun (Dolores, Bullets)
IP3-3 Shoot-at-Door (Dolores, Cell)
IP3-4 Escape (Dolores, Cell, Street)
1 Pickup (Dolores, Gun, Cell)
2 Shoot-at-Door (Dolores, Cell)
3 Check-Gun-F (Dolores, Gun)
4 Pickup (Dolores, Bullets)
5 Load-Gun (Dolores, Bullets)
6 Shoot-at-Door (Dolores, J C)
7 Escape (Dolores, J C, Street)
Action Failure
Sensing/Recovery
Plan Resumption

Figure 1: The HEADSPACE plan from the Breakout domain. In this figure, time flows from left to right. Actions that are attempted by the character in the story are shown across the bottom of the figure using rounded rectangles and numbered 1 to 7. For preconditions and effects of each action, refer to Table 2. Action 2, shown in red, is attempted but fails because its non-belief preconditions are not all met in the world state where it is attempted. The three intention plans held by the character during the course of the story are shown contained in gray rectangles above the plan's actions, along with an indicator showing the interval of story actions during which they are held.

The plan for the story is shown in Figure 1. In the story plan, Dolores intends to shoot the door's lock to damage it, then open the door and escape her cell. The plan in Figure 1 shows the actual executed story actions, the actions that are attempted but fail, and the actions that were intended by Dolores throughout the plan's execution. In the world state before the start of the plan, Dolores believes that a gun is loaded and on a chair within her reach, bullets are also on the chair next to the gun, and the door to her jail cell is locked and closed. Her beliefs at that time are correct except for the fact that the gun is actually unloaded. Dolores first picks up the gun (Action 1), then pulls the trigger, intending to shoot the door's lock, thus unlocking it (Action 2). Because the gun isn't loaded, the action fails. At this point, Dolores realizes that the action failed, and becomes uncertain about just those beliefs that were involved in the failed action's preconditions.

In the world state resulting after Action 2, all of the epistemic preconditions for Dolores' execution of Action 2 have been asserted as unknown in her belief model. In HEADSPACE, when a character attempts an action but fails, the character automatically comes to doubt its belief in each of the failed action's preconditions. This belief update may prompt sensing actions and/or revision of the character's plans for achieving their goals. See (Sanghrajka, Young, and Thorne 2022) for more detail.

Dolores detects that these new beliefs are inconsistent with her intention plan IP1, causing her to drop IP1 and form a new intention plan to first confirm that her gun is loaded and then proceed to use it to escape. Dolores then actively seeks new beliefs about the gun's ammo status by checking the gun's cylinder (Action 3). In the world state resulting from Action 3, Dolores believes that her gun is unloaded. Having expected the gun to be loaded, Dolores detects that her new belief is inconsistent with intention plan IP2, causing her to drop IP2 and form a new intention plan to load her gun and then proceed to use it to escape. In Action 4 she takes bullets from the chair, and in Action 5 she uses them to load her gun. In Action 6 she fires at the lock again. Succeeding this time, the door is now unlocked. Since Dolores believes correctly at that point that the door is unlocked, she opens the door (Action 7) and escapes her cell.

The example plan was used to generate a simple text-based story using a semi-automated process. For each action of the plan, we created a simple sentence based on a pre-defined template for that action's type, and filled references in the template with proper names of characters and objects used in the action. Boilerplate text describing the initial state and the goal of the character was added to the beginning of the story. The automatically generated story elements were then edited by hand to replace proper nouns with pronoun references and to combine some independent clauses using conjunctions to increase readability. The text in Text Example 1 shows the example story sequence generated for the

Pickup(?character, ?item, ?loc)	

PRE-T	at(?character, ?loc)
	in(?item, ?loc)

PRE-F	has(?character, ?item)

PRE-B+	at(?character, ?loc)
	in(?item, ?loc)

PRE-B-	has(?character, ?item)
EFF-T	has(?character, ?item)
EFF-F	in(?item, ?loc)
EFF-B+	has(?character, ?item)
EFF-B-	in(?item, ?loc)

Load-Gun(?character, bullets)	

PRE-T	has(?character, gun)
	has(?character, bullets)

PRE-F	loaded(gun)

PRE-B+	has(?character, gun)
	has(?character, bullets)

PRE-B-	loaded(gun)
EFF-T	loaded(gun)
EFF-B+	loaded(gun)

Check-Gun-Loaded-T(?character)	

PRE-T	has(?character, gun)
	loaded(gun)

PRE-B+	has(?character, gun)
PRE-U	loaded(gun)
EFF-B+	loaded(gun)

Place-Down(?character, ?item, ?loc)	

PRE-T	holding(?character, ?item)
	at(?character, ?loc)

PRE-F	in(?item, ?loc)

PRE-B+	holding(?character, ?item)
	at(?character, ?loc)

PRE-B-	in(?item, ?loc)
EFF-T	in(?item, ?loc)
EFF-F	holding(?character, *)
EFF-B+	in(?item, ?loc)
EFF-B-	holding(?character, *)

Hold(?character, ?item, ?loc)	

PRE-T	at(?character, ?loc)
	in(?item, ?loc)

PRE-F	holding(?character, *)

PRE-B+	at(?character, ?loc)
	in(?item, ?loc)

PRE-B-	holding(?character, *)
EFF-T	holding(?character, ?item)
EFF-B+	holding(?character, ?item)

Check-Gun-Loaded-F(?character)	
PRE-T	has(?character, gun)
PRE-F	loaded(gun)
PRE-B+	has(?character, gun)
PRE-U	loaded(gun)
EFF-B-	loaded(gun)

Check-Bottle-Empty-T(?character, ?bottle)	

PRE-T	holding(?character, ?bottle)
	empty(?bottle)

PRE-B+	holding(?character, ?bottle)
PRE-U	empty(?bottle)
EFF-B+	empty(?bottle)

Check-Bottle-Empty-F(?character, ?bottle)	
PRE-T	holding(?character, ?bottle)
PRE-F	empty(?bottle)
PRE-B+	holding(?character, ?bottle)
PRE-U	empty(?bottle)
EFF-B-	empty(?bottle)

Pour-Drink(?character, ?bottle, ?glass)	

PRE-T	holding(?character, ?bottle)
	empty(?glass)

PRE-F	empty(?bottle)

PRE-B+	holding(?character, ?bottle)
	empty(?glass)

PRE-B-	empty(?bottle)
EFF-T	empty(?bottle)
EFF-F	empty(?glass)
EFF-B+	empty(?bottle)
EFF-B-	empty(?glass)

Escape(?character, cell, street)	
PRE-T	at(?char, cell)
PRE-F	locked(door)
PRE-B+	at(?char, cell)
PRE-B-	locked(door)
EFF-T	at(?char, street)
EFF-F	at(?char, cell)
EFF-B+	at(?char, street)
EFF-B-	at(?char, cell)

Shoot-at-Door(?character, ?loc)	

PRE-T	at(?character, ?loc)
	has(?character, gun)
	loaded(gun)
	in(door, ?loc)

PRE-B+	at(?character, ?loc)
	has(?character, gun)
	loaded(gun)
	in(door, ?loc)

EFF-F	locked(door)
EFF-B-	locked(door)

Table 2: Operators that were part of the Western Domain. These operators form the plan for the Breakout and Drink Refill problems within the domain.

Breakout story.

Example Text 1. *Dolores is imprisoned inside a jail cell. She wants to escape from the jail. Just outside of the jail cell, there lies a revolver and some bullets on a chair. Dolores walks over and picks up the gun. She walks over to the jail door. Dolores attempts to shoot at the jail door lock but fails. Dolores checks the revolver to see if it is loaded, and finds that it is unloaded. She walks over and picks up the bullets. Dolores loads the bullets into the gun and fires again, this time successfully breaking open the jail door. She then walks out of the jail cell and exits the jail.*

In addition to generating text from the plan, the plan data structures are also used to create a QUEST knowledge structure (QKS) for the story. The translation algorithm proposed by Sanghrajka et al. (Sanghrajka, Lang, and Young 2021) was used to accomplish this. The proposed translation algorithm is straightforward, simply creating respective nodes for each action in the plan, and then creating arcs between nodes based on the causal dependencies between the steps in the plan. This approach extends the approach by Christian and Young (Christian and Young 2004), adapted to HEADSPACE plans by including the addition of goal nodes characterizing the elements of unfulfilled intention plans and the addition of failed outcome arcs for failed events. The generated QKS for the plan in Figure 1 is depicted in Figure 2.

Experimental Evaluation

Background

To gauge reader comprehension of plan-based stories containing failed actions, we designed an experiment that followed prior approaches used to evaluate reader comprehension of stories generated by planning systems. As described above, the experimental design borrows from cognitive psychologists' approach to modeling narrative comprehension using QUEST (Graesser, Lang, and Roberts 1991b).

Design

The design of our experiment was similar to these previous QUEST-based evaluation methods, but made use of HEADSPACE plans and related stories that contained action failures. In our experiment, we first generated HEADSPACE plans that contained action failures. We then translated those plans into QKSs as well as into text stories. Using QUEST's arc search procedure, question-answer pairs for "Why", "How" and "What are the consequences of" were generated for pairs of nodes in the QKSs. Human participants read the stories and then answered a set of questions that required them to provide GOA ratings on a 4 point Likert Scale for the question-answer pairs. For the purposes of this study, we focused on just those Why questions that involved a failed action as either the question node or the answer node in the QKS. An example question-answer pair is provided in Example Text 2.

Example Text 2. *Why did Dolores want to check the revolver? Because she failed to shoot at the jail door lock.*

We compared participant GOA ratings for nodes that were rated by QUEST as better answers to participant GOA ratings for nodes that QUEST rated as worse answers. If participants were understanding the relationships between the events in the stories, we would expect to see higher participant GOA ratings for the better answers than for the worse answers.

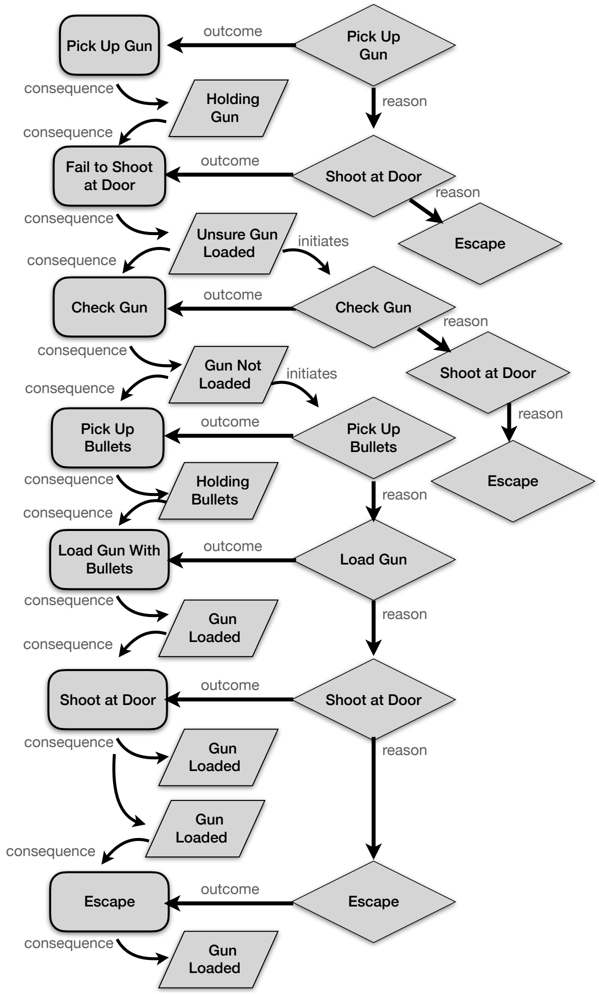

Figure 2: The QKS corresponding to the plan in Figure 1. Rounded rectangle indicate event nodes. Skewed rectangles indicate state nodes. Diamonds indicate goal nodes. Arcs from one node to another are labeled with one of the QUEST relationship types showing in Table 1.

Domain

We generated plans, stories, and QUEST representations in two different planning problems drawn from a common western-themed domain. The first planning problem was the western-themed Breakout problem described in the previous section. The second problem, described in more detail in the appendix, was based around a customer at a saloon asking for a drink refill. Table 2 shows the ground operators that were used in either the Breakout plan or the Drink Refill plan.Both problems and resulting solution plans were designed to be similar in structure, with identical number of actions, the second action failing, an intermediate set of sensing/recovery actions, and then successful plan completion. Participants were randomly assigned to one of the two generated story contexts.

Participants

Participants for this experiment were recruited using Amazon Mechanical Turk. Out of 42 participants that completed the survey, we discarded responses from 33 participants because they did not pass simple pre-defined comprehension check questions (provided in the appendix below). Data from the 9 remaining participants was used for this study. Mechanical Turk is known for producing noisy results, as observed from other studies that have recruited participants using MTurk for story comprehension related tasks (e.g., (Farrell, Robertson, and Ware 2016)). For this study, we were able to report significant results even with a relatively small number of participants with a high confidence, which was verified by a power analysis done before running the study.

Hypothesis

We hypothesized that if readers were able to comprehend the role of failed actions in stories generated by our approach, then, for questions related to those failed actions, readers' mean GOA ratings for Question-Answer pairs identified by QUEST as "better" would be higher than their mean GOA ratings for Question-Answer pairs identified by QUEST as "worse."

Results

A standard one-tailed t-test was used to compare the mean GOA rating of question-answer pairs with an arc distance of 1 (identified as the better answers) to the mean GOA rating of question-answer pairs with an arc distance of 3 (identified as the worse answers). The result of the t-test with 11 degrees of freedom yields $t = 1.7959$ with a p-value of 0.0065 ($p < 0.01$). The responses to question-answer pairs with an arc distance of 1 ("better") were rated significantly higher than responses to question-answer pairs with an arc distance of 3 ("worse").

To provide a more fine-grained analysis, a standard one-tailed t-test was used to compare mean GOA ratings for question-answer pairs with an arc distance of 1 or 2 (identified as "better" answers) with the mean GOA rating of question-answer pairs with an arc distance of 3 (identified as "worse" answers). The result of the t-test with 10 degrees of freedom yields $t = 1.8124$ with a p-value of 0.0092 ($p < 0.01$). The responses to question-answer pairs with an arc distance of 1 or 2 ("better") were rated significantly higher than responses to question-answer pairs with an arc distance of 3 ("worse"). The means and variances for the conditions are reported in Table 3. T-tests performed to compare responses for arc distance of 2 against an arc distance of 3 and arc distance of 1 against an arc distance of 2 or 3 also found some evidence ($p < 0.05$) of significance.

Discussion and Future Work

The experiment we describe above focused on measuring readers' comprehension of underlying HEADSPACE plan structure. Specifically, we focused on comparing readers' GOA ratings with QUEST ratings specifically for *Why* ques-

Responses with arc distance of 1	Responses with arc distance of 1 or 2	Responses with arc distance of 3
3.1111 (0.673)	3.0377 (0.6908)	2.1111 (0.861)

Table 3: Mean Goodness of Answer (GOA) ratings (and standard deviations) for Question/Answer pairs containing action failures.

tions relating node pairs involving failed actions. [2] The analysis clearly shows that readers rated node pairs closer together in the QKS as better answers and node pairs farther apart as worse answers – as predicted by the QUEST arc search procedure. These results provide strong support for the claim that readers were able to understand the role of failed actions in the unfolding events of each story, and that the corresponding HEADSPACE plan structures served as the basis for readers' comprehension.

One limitation of the current study is that it only considers stories where actions fail and then characters modify the world state to allow the failed action to be performed successfully. In many stories with failed actions, characters drop their intentions when actions fail and adopt new courses of action rather than repairing the world to re-attempt the failed action. Work by Amos-Binks (Amos-Binks 2018) has shown that readers show greater engagement in plan-based stories where characters change their intentions. In future work, we will evaluate the connection between reader comprehension of story structure, intention revision, mistaken beliefs and failed actions.

Finally, we anticipate exploring the role that stories with failed actions play in the creation of a reader's violation of expectations. Expectation violation in narrative is linked to the experience of surprise (Maguire and Keane 2005), an affective response that is often created by effective storytellers. By combining the current model of action failure arising from mistaken character beliefs with a model of a reader's or viewer's expectations, we hope to be able to influence the experience of a reader's surprise by building stories with actions that fail in ways that are unanticipated by either characters or readers.

Appendix

As described above, the experiments we ran made use of 2 distinct plans and corresponding QKSs and text realizations.

The plan in Figure 4 was selected because, while the actions in it are distinct from those in the experimental plan shown in Figure 1, its structure is nearly identical. As a result, the two QKSs (shown in Figures 2 and 3) are also similar. This similarity between the experimental materials was intentional to prevent any confounding variables between the two conditions. Both stories have the same number of actions, with the same structure. The character has three possible intention plans in both stories, with the intention plans being adopted and invalidated at the same points in the sequence of actions. This allows for keeping the conditions about the planning problem the same and reducing the bias from one of the "settings" of the stories on results.

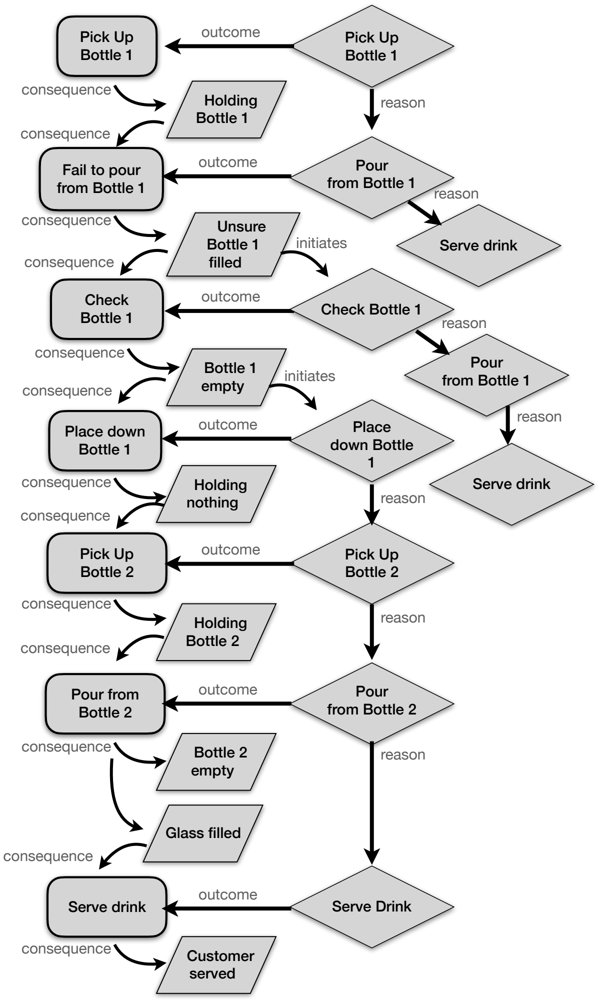

Figure 3: The QKS corresponding to the plan from Figure 4. Rounded rectangle indicate event nodes. Skewed rectangles indicate state nodes. Diamonds indicate goal nodes. Arcs from one node to another are labeled with one of the QUEST relationship types showing in Table 1.

Drink Refill Plan

Figure 4 shows the second example plan used in our experiment, and Figure 3 shows the corresponding QKS. Example Text 3 shows the corresponding story realized as text. Some of the QA pairs used in the Drink Refill problem are reported below. Texts 4-9 were used to filter out participants (4-6 are examples of good or very-good QA pairs, and texts 7-9 are examples of bad or very bad pairs). Example Texts 10-15 are QA pairs used in the analysis. Texts 10 through 12 have a calculated arc distance of 1, texts 13-14 have an arc dis-

[2] Because there is already a long history of work leveraging QUEST to evaluate plan-based plot generation, we focused our evaluation specifically on comprehension indicators for HEADSPACE's novel contribution: failed actions.

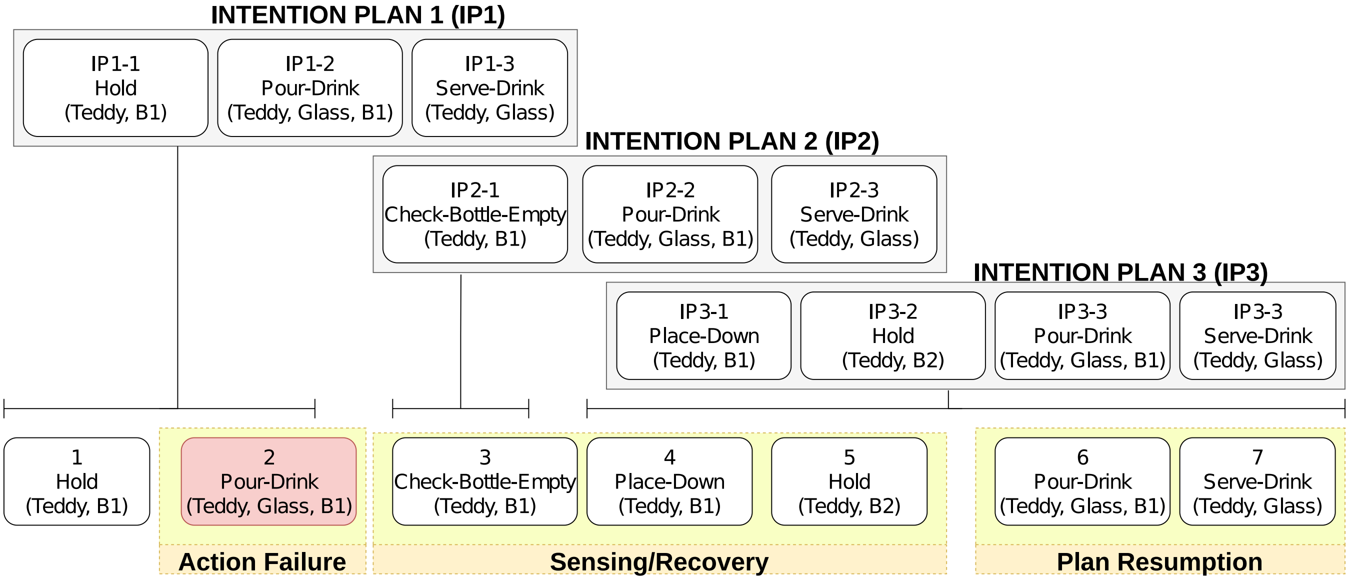
INTENTION PLAN 1 (IP1)
IP1-1 Hold (Teddy, B1)
IP1-2 Pour-Drink (Teddy, Glass, B1)
IP1-3 Serve-Drink (Teddy, Glass)
INTENTION PLAN 2 (IP2)
IP2-1 Check-Bottle-Empty (Teddy, B1)
IP2-2 Pour-Drink (Teddy, Glass, B1)
IP2-3 Serve-Drink (Teddy, Glass)
INTENTION PLAN 3 (IP3)
IP3-1 Place-Down (Teddy, B1)
IP3-2 Hold (Teddy, B2)
IP3-3 Pour-Drink (Teddy, Glass, B1)
IP3-3 Serve-Drink (Teddy, Glass)
1 Hold (Teddy, B1)
2 Pour-Drink (Teddy, Glass, B1)
3 Check-Bottle-Empty (Teddy, B1)
4 Place-Down (Teddy, B1)
5 Hold (Teddy, B2)
6 Pour-Drink (Teddy, Glass, B1)
7 Serve-Drink (Teddy, Glass)
Action Failure
Sensing/Recovery
Plan Resumption

Figure 4: The plan used in the second experimental domain. The red action was attempted but failed because the non-belief preconditions are not all met in the world state where they are attempted.

tance of 2, and text 15 has an arc distance of 3.

Example Text 3. *Teddy is a bartender working in a bar. He wants to serve a customer their drink. There are bottles of beverage on the shelf. Teddy walks over to the shelf and picks up a bottle. He then attempts to pour a drink from it but fails. He checks the bottle and sees that the bottle is empty. Teddy then places the first bottle back on the shelf and picks up a new bottle. He attempts to pour a drink again and is successful. Teddy then serves the drink to the customer.*

Example Text 4. *What was the consequence of Teddy failing pouring the drink from the first bottle? He checked the first bottle and saw that it was empty.*

Example Text 5. *What was the consequence of Teddy failing pouring the drink from the first bottle? He wanted to check the first bottle.*

Example Text 6. *Why did Teddy want to pick up the first bottle? Because he wanted to pour a drink from the first bottle.*

Example Text 7. *What was the consequence of Teddy serving the drink to the customer? Teddy was unsure whether the first bottle was empty.*

Example Text 8. *What was the consequence of Teddy failing pouring the drink from the first bottle? He wanted to pick up the first bottle.*

Example Text 9. *What was the consequence of Teddy pouring the new drink successfully? He failed to pour the drink from the first bottle.*

Example Text 10. *Why did Teddy attempt to (unsuccessfully) pour the drink from the first bottle? Because the customer's glass was empty.*

Example Text 11. *Why did Teddy attempt to (unsuccessfully) pour the drink from the first bottle? Because Teddy was holding the first bottle.*

Example Text 12. *Why did Teddy attempt to (unsuccessfully) pour the drink from the first bottle? Because he wanted to pour a drink from the first bottle.*

Example Text 13. *Why did Teddy attempt to (unsuccessfully) pour the drink from the first bottle? Because he picked up the first bottle.*

Example Text 14. *Why did Teddy attempt to (unsuccessfully) pour the drink from the first bottle?In order to refill the customer's drink.*

Example Text 15. *Why did Teddy attempt to (unsuccessfully) pour the drink from the first bottle? Because he wanted to pick up the first bottle.*

Breakout Plan

This section provides the question-answer pairs used as part of the experiment with the Breakout plan. Example Texts 16 through 21 are QA pairs that were used as comprehension check questions (16 is a good QA pair and 17-21 are bad QA pairs). Example Texts 22 through 30 are used in the analysis. Texts 22 through 28 have a calculated arc distance of 1. Text 27 has a calculated arc distance of 2, and Text 28 has a calculated arc distance of 3.

Example Text 16. *What was the consequence of Dolores failing when shooting at the door lock? She checked the revolver and found that it was unloaded.*

Example Text 17. *What was the consequence of Dolores loading the revolver? She failed shooting at the door lock.*

Example Text 18. *What was the consequence of Dolores successfully shooting the jail door open? She failed shooting at the door lock.*

Example Text 19. *What was the consequence of Dolores picking up the bullets? She failed shooting at the door lock.*

Example Text 20. *What was the consequence of Dolores loading the revolver? She wanted to load the revolver.*

Example Text 21. *What was the consequence of Dolores checking the revolver and finding it unloaded? She failed shooting at the door lock.*

Example Text 22. *Why did Dolores try to (unsuccessfully) shoot at the door lock? Because Dolores believed that the revolver was loaded.*

Example Text 23. *Why did Dolores try to (unsuccessfully) shoot at the door lock? Because she wanted to shoot at the jail door lock.*

Example Text 24. *Why did Dolores' attempt to shoot fail? Because the revolver was not loaded.*

Example Text 25. *Why did Dolores try to (unsuccessfully) shoot at the door lock? Because Dolores had the gun.*

Example Text 26. *Why did Dolores try to (unsuccessfully) shoot at the door lock? Because the jail door was locked.*

Example Text 27. *Why did Dolores try to (unsuccessfully) shoot at the door lock? In order to escape from prison.*

Example Text 28. *Why did Dolores try to (unsuccessfully) shoot at the door lock? Because she wanted to pick up the revolver.*

References

Amos-Binks, A. A. 2018. *Intention Revision of Plan-Based Agents for Narrative Generation*. Ph.D. thesis, North Carolina State University.

Bahamón, J. C.; and Young, R. M. 2017. An Empirical Evaluation of a Generative Method for the Expression of Personality Traits through Action Choice. In *Proceedings of the Annual Conference on Artificial Intelligence And Interactive Digital Entertainment*.

Bratman, M. 1987. *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Cardona-Rivera, R. E.; Price, T.; Winer, D.; and Young, R. M. 2016. Question answering in the context of stories generated by computers. *Advances in Cognitive Systems*, 4: 227–245.

Christensen, M.; Nelson, J.; and Cardona-Rivera, R. E. 2020. Using Domain Compilation to Add Belief to Narrative Planners. In *Proceedings of AIIDE-20*.

Christian, D.; and Young, R. M. 2004. Comparing cognitive and computational models of narrative structure. In *Proceedings of the National Conference on Artificial Intelligence*, 385–390. American Association of Artificial Intelligence, Menlo Park, CA: AAAI.

Farrell, R.; Robertson, S.; and Ware, S. G. 2016. Asking hypothetical questions about stories using QUEST. In *International Conference on Interactive Digital Storytelling*, 136–146. Springer.

Fikes, R.; and Nilsson, N. 1990. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. In Allen, J.; Hendler, J.; and Tate, A., eds., *Readings in Planning*. Morgan Kaufmann.

Graesser, A.; and Clark, L. 1985. *Structures and procedures of implicit knowledge*. Norwood, New Jersey: Ablex.

Graesser, A.; Lang, K.; and Roberts, R. 1991a. Question answering in the context of stories. *Journal of Experimental Psychology: General*, 120(3): 254.

Graesser, A.; Lang, K.; and Roberts, R. 1991b. Question answering in the context of stories. *Journal of Experimental Psychology: General*, 120: 254–277.

Graesser, A. C.; and Franklin, S. P. 1990. QUEST: A cognitive model of question answering. *Discourse processes*, 13(3): 279–303.

Graesser, A. C.; and Hemphill, D. 1991. Question answering in the context of scientific mechanisms. *Journal of memory and language*, 30(2): 186–209.

Jhala, A.; and Young, R. M. 2010. Cinematic Visual Discourse: Representation, Generation, and Evaluation. *IEEE Transactions on Computational Intelligence and Artificial Intelligence in Games*, 2: 69–81.

Lenhart, A.; Kahne, J.; Middaugh, E.; Macgill, A. R.; Evans, C.; and Vitak, J. 2008. Teens, Video Games, and Civics: Teens' Gaming Experiences Are Diverse and Include Significant Social Interaction and Civic Engagement. *Pew internet & American life project*.

Maguire, R.; and Keane, M. T. 2005. Expecting a surprise? The effect of expectations on perceived surprise in stories. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 833–838. Cognitive Science Society.

McTierman, J. 1988. *Die Hard*. Thorp, R., Stuart, J., and de Souze, S. E.: 20th Century Studios.

Nebel, B.; and Hoffmann, J. 2001. The FF planning system: Fast plan generation through heuristic search. *Journal of Artificial Intelligence Research*, 14: 253–302.

Riedl, M.; and Young, R. M. 2010. Narrative planning: balancing plot and character. *Journal of Artificial Intelligence Research*, 39(1): 217–268.

Sanghrajka, R.; Lang, E.; and Young, R. M. 2021. Generating QUEST Representations for Narrative Plans Consisting of Failed Actions. In *The 16th International Conference on the Foundations of Digital Games (FDG) 2021*, 1–10.

Sanghrajka, R.; Young, R. M.; and Thorne, B. 2022. HeadSpace: Incorporating Action Failure and Character Beliefs into Narrative Planning. In *Proceedings of the AAAI Conference on AI and Interactive Entertainment*. Pomona, CA.

Shirvani, A.; Ware, S. G.; and Farrell, R. 2017. A possible worlds model of belief for state-space narrative planning. In *Proceedings of the 13th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment*, 101–107.

Teutenberg, J.; and Porteous, J. 2015. Incorporating global and local knowledge in intentional narrative planning. In

Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, 1539–1546. International Foundation for Autonomous Agents and Multiagent Systems.

Thorne, B.; and Young, R. M. 2017. Generating Stories that Include Failed Actions by Modeling False Character Beliefs. In *Working Notes of the AIIDE Workshop on Intelligent Narrative Technologies*. Salt Lake City, UT.

Ware, S. G.; and Young, R. M. 2014. Glaive: a state-space narrative planner supporting intentionality and conflict. In *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Young, R. M. 2017a. Sketching a generative model of intention management for characters in stories: Adding intention management to a belief-driven story planning algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 13, 281–288.

Young, R. M. 2017b. Sketching a generative model of intention management for characters in stories: Adding intention management to a belief-driven story planning algorithms. In *Working Notes of the AIIDE Workshop on Intelligent Narrative Technologies*. Salt Lake City, UT.