Advances in Cognitive Systems 9 (2020) 31–49 Submitted 5/2019; published 12/2020

Intention Dynamics of Rebel Agent Behavior

Adam Amos-Binks AAMOSBINKS@ARA.COM
Applied Research Associates, Inc., Raleigh, NC, 27615 USA

Dustin Dannenhauer DDANNENHAUER@NAVATEKLTD.COM
Navatek LLC, Arlington, VA, 22203 USA

David W. Aha DAVID.AHA@NRL.NAVY.MIL
Navy Center for Applied Research in AI, Naval Research Laboratory, Washington, DC, 20375 USA

Abstract

Rebel agents are both an important narrative plot device and arguably essential for true agent autonomy. Rebellious behaviors respond to the changes in another agent's intentions or a dynamic environment. While revising intentions is key to responding to and rebelling against other agents, existing models of intention revision are coarse grained. They enable an agent to perform basic operations such as dropping old intentions and adopting new ones but do not capture the interdependence and intention dynamics between types of rebellious behavior. We address this limitations with an approach that builds on existing work on intention dynamics and operationalizes rebel agents in an interactive narrative context. This approach led to three contributions. We first define plan-based representation of three rebellious behaviors: betrayal, revenge, and justice. These definitions rely on intention dynamics – treating intentions as a complex set of interactions between agents – rather than singletons. Second, we evaluate our definitions in a generated narrative text and use the QUEST knowledge structure and question answering to elicit the mental models our definitions induce in human subjects. Our results show that people who read generated text that includes statements representing trust and violations of trust inferred how betrayal occurred. Interestingly, subjects do not require these same statements to infer that revengeful and just goals were responses to betrayal. Finally, we analyze betrayal, revenge, and justice in a framework for characterizing rebel agents. Together these three contributions offer first steps toward the application of interactive rebel agents.

1. Introduction

Interactive narrative is a media-independent experience (e.g., choose-your-own-adventure book, text-based adventure game, console game) that strives to balance an author-defined story arc with user actions that also shape the plot. In response to these actions, Belief-Desire-Intention (BDI) character agents adapt their behavior through narrative devices such as intention revision. Dynamic intentions enable a rich environment in which agents foil, cooperate, and even rebel against the user and other BDI agents.

While intention revision enables an agent to drop old intentions and adopt new ones, it is a simplified model of behavior change. It lacks subtleties to represent the intention dynamics of narrative phenomena that are often key to plot development. Examples of such dynamics that support a rebellious plot include betrayal (e.g., Hugh Glass, *The Revenant*), revenge (e.g., Sam Chisolm, *The Magnificent Seven*), and justice (e.g., Carl Lee Hailey, *A Time to Kill*). The typical narrative use of rebellion involves a protagonist who must subvert an antagonist's power. In contrast, AI researchers argue that rebel agents serve functional roles, and in particular that a rebel agent's noncompliance is essential to true agency and autonomy (Coman & Muñoz-Avila, 2014). However, rebel agents still lack operationalized computational models for human interaction.

To address some of these limitations, we investigate the auspicious relationship between intention dynamics, interactive narrative, and rebel agents. We make three contributions to operationalizing rebellious behaviors for interactive narrative. The first is defining intentional plan-based structures to represent three rebellious behaviors: betrayal, revenge, and justice. Central to these definitions is the concept of intention dynamics, where the mental state and actions of our agents evolve over time based on the actions of other agents. Second, we evaluate our definitions using the QUEST cognitive model in a human subject evaluation. We use QUEST knowledge structures to represent the user's expected mental model after reading generated text of our plan-based definitions. Finally, we analyze betrayal, revenge, and justice in a rebel agent characterization framework. Together these three contributions take first steps to advance interactive rebel agent applications

2. Previous Work

Our contributions are based on three areas of previous work. First, narrative generation gives a background for intentional planning. Second, intention encodes the mental state from the Belief-Desire-Intention model that enables rebel definitions. Finally, existing rebel agent frameworks characterize the qualities of our agent definitions. This section includes key contributions from each area.

2.1 Interactive Narrative

Schank and Abelson (1977) were perhaps the first to propose the use of plans to represent story plots. Their analysis was based on the theoretical overlaps of plot events with the action-oriented, causally-linked, and temporally-ordered properties of plans. Since these early insights, story generators with extended representations capture a range of narrative features (Meehan, 1977; Porteous et al., 2010; Perez y Perez & Sharples, 2001).

Of the many representations, we focus on intentional partial-ordered causally-linked planning (IPOCL) (Riedl & Young, 2010), where intentional (goal-driven) agents execute causally-linked actions towards their goals. Together, individual agent goals reach the goal states in a planning problem. IPOCL story plans operationalize intention by generating solution plans that contain only a sequence of causally connected actions that achieve the goal of an agent's intention. An *intention frame* aggregates the agent's goal, a motivating step, and the sequence called a subplan. Both the goal and subplan are key in identifying reconsidered intentions, as we discuss further in Section 2.2.

Because IPOCL planning leverages the explicit notions of causality and intention, researchers have evaluated its affect on a reader's mental model using the QUEST cognitive model of question answering (Graesser et al., 1991). The model uses a graph called the QUEST knowledge structure

Table 1. The decision making steps of a belief, desire, intention agent, adapted from Rao and Georgeff (1998).

#	Agent action
1	make an observation ω from the environment
2	revise ***beliefs*** based on ω
3	if current ***beliefs*** cause an ***intention*** to be reconsidered then
4	let ***desires*** be the results of options when considering ***beliefs*** and ***intentions***
5	let ***intentions*** be the results of deliberating ***beliefs, desires*** and ***intentions***
6	if current ***beliefs*** introduce complications to an **intention** achieving ***plan*** then
7	let ***plan*** be the results of planning with ***beliefs*** and ***intentions***

(QKS) to represent a reader's mental model of a story's causal and intentional organization. Additionally, the framework prescribes the structure of question answering and includes a QKS traversal to predict reader responses. Predictions are compared with actual readers' responses to evaluate how well they reflect subjects' mental models.

From this cognitive psychological foundation, IPOCL plans the plot structure of an interactive narrative. This form of narrative allows a participant to shape the plot through their interactions. When causal link threats are introduced by user actions, an experience manager agent will generate a new plan, ensuring it is coherent with the failed one. Specifically for IPOCL, agents should change goals in a principled fashion or risk reducing a user's engagement due to a lack of coherence.

2.2 Intention

The intentions in narrative planning are grounded in the Belief-Desire-Intention (BDI) theory of mind. Beliefs are facts an agent believes as true, desires are world states an agent wants to be true, and intentions are those desires an agent is committed to make true through action. Bratman (1987) first theorized a concept of intention, based on its use to both characterize an agent's mental state (e.g., commitment to a goal) and action (e.g., justification for action). Intention was later formalized for logical agents by Cohen and Levesque (1990) and led to decision-making abilities for BDI agents. BDI research has made substantial research efforts on belief revision and update (e.g., Rao & Georgeff, 1998), while making only cursory investigations on the connected effects of belief changes to other mental states, specifically intention. As part of an investigation into intention revision logic, Van der Hoek (2007) formalized intention revision in linear temporal logic based on the algorithm summarized in Table 1.

Specifically, intention revision is concerned with the *reconsider* function (line 4) and its coupling to new observations (line 2). The *reconsider* function is characterized as a costly cognitive process, while new observations are relatively easy to obtain, making reconsideration at every observation infeasible. The procedure does not specify when agents should reconsider, except that observation and enablement of previously unachievable goals alone are not sufficient. On the other hand, when observations make a current intention unachievable, the agent would be well served to *reconsider* and execute lines 4–7 to develop a new plan for an achievable goal. Amos-Binks and Young (2018) operationalized this in a plan-based model of intention revision in which causal link threats compel an agent to reconsider an intractable goal and initiate an intention revision.

2.3 Rebel Agents

Rebel behavior, or the ability for an agent to reject, protest, or alter its goals, plans, or actions is a desired capability for many autonomous systems (Briggs & Scheutz, 2017; Dannenhauer et al., 2018). Agents often have access to different sources of information and operate with safety or ethical constraints. Consider two hypothetical scenarios. In the first, a humanoid robot is assisting a human in carrying a large heavy object. While walking, the humanoid robot observes an obstacle behind the human and refuses to continue carrying the object until the path is safe for the human. In the second, a hotel service robot denies a request to retrieve luggage for a person who is attempting to steal from hotel guests. For autonomous AI systems, rebellion is especially important when the system designers differ from the users and when there are constraints on acceptable behavior for that system.

Coman and Muñoz-Avila (2014) motivate the need for rebellion to achieve believable characters in narrative settings. They describe goal-driven autonomy (GDA) agents with motivation-based discrepancies that lead to rebel behavior. GDA is a model of goal reasoning in which agents perform a four-step process: detect discrepancies; explain what may have caused the discrepancies; formulate new goals; and select which goals to pursue (Muñoz-Avila et al., 2010). Discrepancies are differences in the expected and observed world states while the agent is acting. Motivation discrepancies put forth by Coman and Muñoz-Avila are instead discrepancies between an agent's motivation and either the agent's current plan, the observed state, or the agent's current goal. Since these motivations change they may no longer align with the agent's current motivation.

GDA agents are similar to BDI agents in that desires are analogous to goals and intentions are much like plans. We use BDI agents in our approach, as it has been extensively applied to interactive narrative, but there are no limitations preventing GDA agents from employing revenge, betrayal and justice. The focus of our work is the intention revision process that characterizes behaviors of these sorts. These behaviors can be seen as forms of rebellion leading to more believable characters that progress an interactive narrative's plot. Coman and Muñoz-Avila focus on characters that identify conflicts between their motivations and goals/actions/state. These motivation discrepancies provide another source of when to reconsider and perform intention revisions.

3. Rebel Agent Behaviors

Intentional planning systems generate action sequences to satisfy the goal conditions of a planning problem. These action sequences structure an interactive narrative in which character agents can adopt, drop, or revise their intentions in response to their environment, but they are limited in that they do not deliberately adopt rebellious behaviors. To address this limitation, we define rebellious behaviors in an intentional planning representation. We use a simple scenario, *Prison*, to illustrate basic definitions of intentional plans and our rebel agent behaviors. When possible, after we present each formal definition, we provide a concrete instance of it within *Prison*. Although our example involves an interactive narrative, these rebel agent behaviors also have utility in many interactive settings. Section 3.1 reviews existing formalizations for intentional planning. Building on these statements, we then propose formal definitions for the concepts of betrayal, revenge, and justice in Sections 3.2, 3.3, and 3.4, respectively.

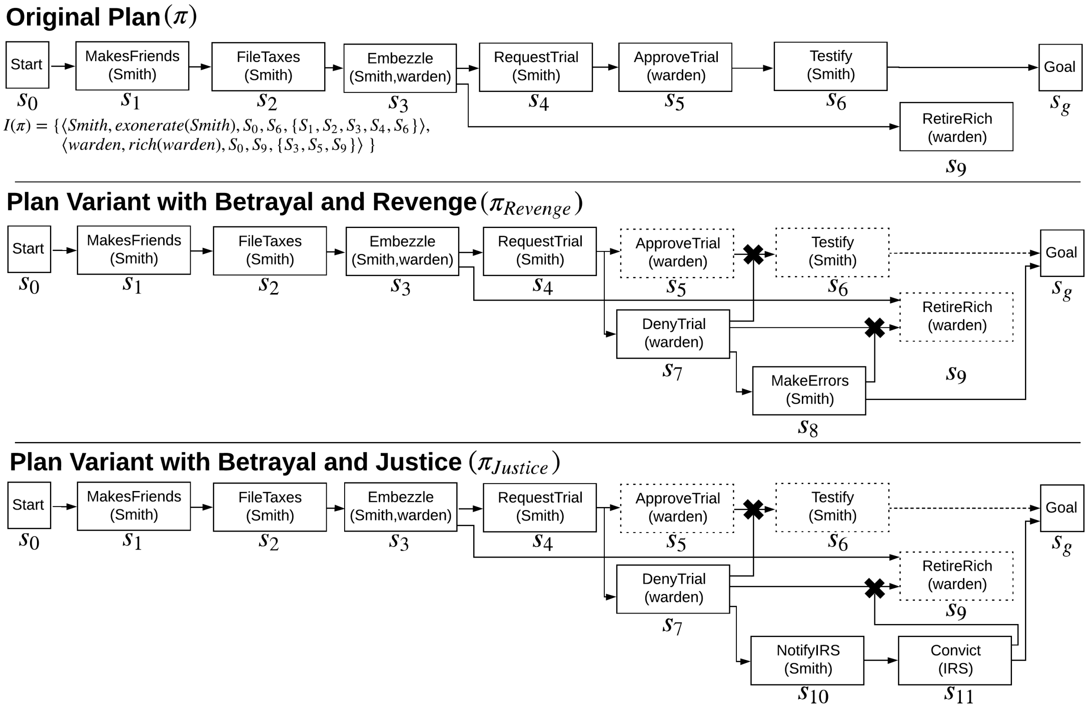


Figure 1. Prison intentional plan with variants for the warden's betrayal and Smith's revenge and justice.

3.1 Intentional Planning

Our approach uses intentional planning definitions from Riedl and Young's (2010) work on IPOCL planning. Intentional planning differs from classical planning by adding a single additional constraint on the solutions; all steps in a solution plan must be causally linked to achieving at least one agent's goal (with happenings being *fate*'s intention). We refer to this causally linked set of actions as an agent's subplan to achieve its goal. The agent, its subplan, and goal are aggregated into a structure called an *intention frame* that reflects the additional constraints on intentional plans.

Our *Prison* example has two agents: Smith (a non-player agent) and the warden (a player agent):

Definition 1 (Agent) An agent is a symbol that uniquely identifies a goal-oriented agent.

Definition 2 (Agent goal) A logical sentence that identifies an agent's desired world conditions.

An agent's goal is represented by the *intends(agent, goal)* predicate. Smith executes actions to achieve his exoneration *intends(Smith, exonerated(Smith))*, while the warden acts to retire rich *intends(warden, rich(warden))*.

The agent who executes an action is called the consenting agent. In the original plan (top) in Figure 1, Smith is the consenting agent of the *MakeFriends*, *Embezzle*, *RequestTrial*, and *Testify* actions. This is reflected in our Action definition:

Definition 3 (Action) Action A consists of preconditions that must be satisfied before execution, PRE(A), effects that result, EFF(A), and a consenting agent, AGENT(A), who performs the action.

Preconditions are literals in a state space whose conjunction must evaluate to true *before* an action's execution. An action's effects are literals whose conjunction evaluates to true *after* A is executed.

An action's name, parameter list, preconditions, effects, and consenting agent describe an *action schema*. An action schema creates a plan step by grounding the free variables. An agent's goal-oriented actions are executed within an intentional plan:

Definition 4 (Intentional plan) An intentional plan π is $\langle S, B, O, L, I\rangle$, where S is the set of steps (ground instances of actions in POCL planning), B the binding constraints on the variables of S, O the partial ordering of steps in S, L the set of causal links joining steps in S, and finally I, the intention frame set that define agent subplans.

Definition 5 (Causal links) A causal link, $s \xrightarrow{p} u$, is a tuple $\langle s, p, u\rangle$, where s, u are actions and p is a literal. A causal link records that p is both an effect of s and satisfies the precondition in u.

Causal links are the edges that connect the plan steps in Figure 1. Intention frames are the essential structure of an intentional plan. Intention frames structure intentional plan elements into goal-oriented behavior of agents.

Definition 6 (Intention frame) An intention frame is a tuple $\mathcal{I} = \langle \text{AGENT}, g, m, \sigma, T\rangle$, where g is AGENT's goal, m is the motivating step such that $m \in S$ with the effect $\neg g$, and σ is the satisfying step such that $\sigma \in S$ with the effect g. A subplan is a set of steps $T \subseteq S$ such that AGENT consents to each step, each step shares at least one causal link to a step in T, some step in T achieves g, and steps in T occur after m and before σ.

Figure 1 includes the intention frames, $\mathcal{I}(\pi)$, for *Smith* and *warden*. Finally, intentional plans solve planning problems; in Figure 1 the planning problem has a single literal, *content(Smith)*.

Definition 7 (Planning problem) A planning problem Φ is a five-tuple $\langle \mathcal{I}, \mathcal{G}, \mathcal{A}, \mathcal{O}, \Lambda\rangle$, where $\mathcal{I}$ and $\mathcal{G}$ are conjunctions of true literals in the initial and goal state, respectively, $\mathcal{A}$ is the set of symbols referring to agents, $\mathcal{O}$ is the set of symbols referring to objects, and Λ is a set of action schemata.

When executing a plan-based interactive narrative, our analysis considers a step as executed if we have updated its effects in the execution state, where the execution state is a set of consistent, non-modal, ground literals. We use executed steps to determine active intentions.

Definition 8 (Active intention) An active intention, i, is part of the current plan, $i \in I(\pi)$, where at least one step of the subplan is executed and the satisfying step, $\sigma(i)$ is not executed. A plan's active intentions are indicated by $I^a(\pi)$.

In Figure 1, Smith's intention of $exonerate(Smith)$ is active from $s_1 tos_5$, until he executes the satisfying step, *Testify* (s_6). Active intentions are useful for identifying reconsidered intentions from which rebel behavior can be initiated. During an interactive narrative, the player agent (the warden) can take actions that introduce threats to causal links, thus foiling nonplayer agents from achieving their goals.

Definition 9 (Causal link threat) A causal link threat occurs when a causal link is established $s \xrightarrow{p} u$ and when some other step w has effect $\neg p$ and could be executed after s but before u. Executing w in this interval means the precondition p of u is no longer satisfied by the state after s is executed and thus u will not execute.

In $\pi_{Revenge}$ in Figure 1, the player agent executes the *DenyTrial* step (s_7) instead of the planned *ApproveTrial* (s_5). This introduces a causal link threat to the *Testify* action that is part of Smith's *exonerate(Smith)* intention. We refer to an action that introduces a causal link threat at execution time as an exceptional action.

Definition 10 (Exceptional action) An exceptional action, s'_t, is executed at time t by the player agent, $\text{AGENT}(s'_t) = user$, such that one of its effects, $e \in \text{EFF}(s'_t)$, introduces a causal link threat to a precondition of a future step $\text{PRE}(s_u)$ in the current plan π where $t \leq u$.

The *DenyTrial* exceptional action causes Smith to reconsider his $exonerate(Smith)$ intention.

Definition 11 (Reconsidered intention) A reconsidered intention, $\langle \mathcal{I}, \epsilon \rangle$, consists of $\mathcal{I}$, an active intention, and ϵ a literal that introduces a causal link threat to the subplan, $T(\mathcal{I})$.

Once an agent has a reason to reconsider its intentions, the agent deliberates about what intentions it will add, drop, and revise. These intention dynamics are influenced by several factors that we have not yet fully articulated. We investigate factors that affect rebellious behavior in the remainder of this section.

3.2 Betrayal

The intention dynamics initiated by betrayal often lead to an agent adopting intentions for revenge and justice. At the crux of betrayal are two agents that have established trust, at least temporarily, between them.

Definition 12 (Trustful action) A trustful action s^{trust} is an action that two agents consent to such that s^{trust} is a step in a subplan of each agent.

In both *Prison* variants, S_2 establishes trust between the warden and Smith as *Embezzle* is a step in both their subplans.

After establishing this trust, if one agent introduces a causal link threat to the other agent's goal, the other agent will view it as a violation of trust and thus a betrayal.

Definition 13 (Agent betrayal) A tuple $\text{BETRAY} = \langle s^{trust}, \chi, \mathcal{I} \rangle$, where s^{trust} is a step in π that has two consenting agents, such that one agent is the consenting agent in χ and the second agent is the agent in $\mathcal{I}$. The exceptional action χ must occur after s^{trust} and must contain an effect that introduces a causal link threat to the subplan of $\mathcal{I}$.

At S_5 the warden denies Smith's trial request, betraying Smith in favor of retiring rich (S_9). The combination of the causal link threat introduced by the warden in S_5 and prior trustful action in S_2 represent betrayal in *Prison*. Trust can occur between agents for many reasons. This definition is just one operationalization that requires little knowledge engineering.

3.3 Revenge

There are different motivations for revenge and we argue that betrayal is one of them. Intuitively, revenge occurs when an agent (Smith) believes it has been wronged by another agent (the warden) and adopts an intention to exact its grievance by foiling a goal of the offending agent.

Definition 14 (Agent revenge) An agent revenge is a tuple, $\langle \mathcal{I}_a, \mathcal{I}_b, \textsc{betray} \rangle$, in which $T(\mathcal{I}_a)$, a subplan of $\mathcal{I}_a$, contains an effect that introduces a causal link threat to $T(\mathcal{I}_b)$, a subplan of $\mathcal{I}_b$, such that $\textsc{betray}$ is a betrayal between $\textsc{agent}(\mathcal{I}_a)$ and $\textsc{agent}(\mathcal{I}_b)$. $\mathcal{I}_a$ is a revengeful intention.

After the warden commits his betrayal in S_5, Smith reconsiders his intentions as his subplan to achieve his exoneration is no longer viable. Smith adopts a revengeful intention ($\mathcal{I}_a$ from Definition 14) and subversively makes accounting errors in the warden's embezzlement scheme (*MakeErrors*, S_8 in the first variant plan in Figure 1). This action foils the warden's intention to retire rich, thereby representing a revengeful intention.

3.4 Justice

Another option an agent may deliberate over is pursuing justice as a response to betrayal. There are a number of similarities between revenge and justice, including being wronged and focusing intentions on the offending agent. However, the main difference is that revenge deliberately subverts a value system, with the agent taking matters into its own hands, whereas justice adheres to a value system deferring punishment to another agent. In this case, a value system captures mutual exclusive relationships between between goals. Smith cannot continue to pursue revenge without giving up his goal of being exonerated, whereas he could while pursuing justice. A full theory of value systems is a more appropriate problem for a goal management lifecycle (e.g., Cox et al. 2017), so we differentiate justice and revenge in these plan-based definitions by the specificity of the intention. A revengeful intention pursues a specific goal to foil the another agent's intention, whereas a just intention's goal is to serve justice, whatever form it takes.

Definition 15 (Agent justice) An agent justice tuple, $\langle \mathcal{I}_a, \mathcal{I}_b, \textsc{agent}, \textsc{betray} \rangle$, where $T(\mathcal{I}_a)$ is a subplan of $\mathcal{I}_a$ that contains an effect that will serve as motivation for $\mathcal{I}_b$. $T(\mathcal{I}_b)$ contains an effect that will foil a subplan of $\textsc{agent}$ and $\textsc{betray}$ is an agent betrayal between $\textsc{agent}(\mathcal{I}_a)$ and $\textsc{agent}$.

In Figure 1, the second variant plan contains Smith's *NotifyIRS* action (s_{10}) which alerts the IRS of the warden's embezzlement scheme. This motivates the IRS to investigate and convict the warden (S_{11}). As a result of the conviction, the warden cannot achieve his retire rich intention and the IRS achieves its intention. Like our revenge definitions, our justice definition is coupled to betrayal and requires little knowledge engineering. We expect there are cases of justice not captured here and would require greater engineering.

4. Rebel Agent Behavior: Framework Characterization and Mental Models

Betrayal, revenge, and justice can all be classified under the rebellion framework described by Coman and Aha (2018). Rebellion occurs between a rebel and an *interactor*, the person rebelled against. They classified rebellion along three dimensions: *expression*, *focus*, and *interaction initiation*. All three behaviors are examples of inward-oriented (*expression*) and explicit (*focus*) rebellion.

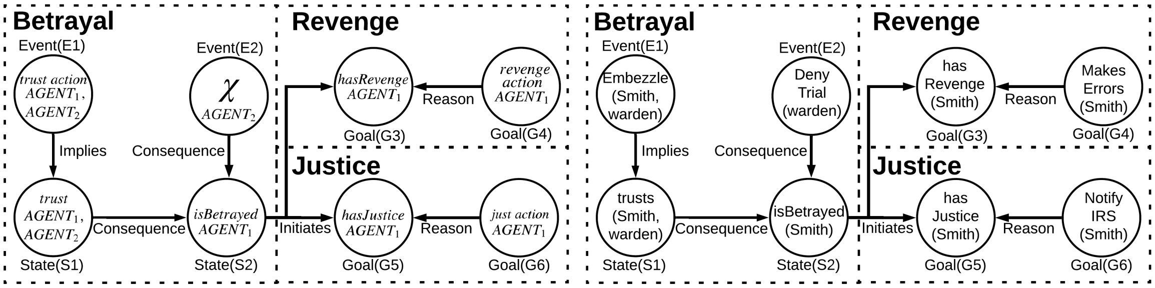


Figure 2. Two versions of three QKS subgraphs that represent three rebellious behaviors. The graph on the left is ungrounded whereas the version on the right is grounded to the *Prison* example.

The term *inward-oriented* refers to an agent changing its own behavior rather than keeping another agent from behaving in a certain way. The label *explicit* denotes a rebellion's observable effect as opposed to an internal state of mind. All examples result in changes to the rebelling agent's actions.

Betrayal is a reactive (*interaction initiation*) rebellious behavior because the rebel agent (the warden) is rejecting an agreed-upon cooperation (the trust established in the embezzlement scheme) with the interactor (Smith). Revenge and justice are also reactive (*interaction initiation*) because rebellion arises from an interaction initiated by the interactor (now the warden). Importantly, in each model, the rebellion is only known from observing actions, since no agent announces its rebellion to the other before taking action. In fact, when Smith exacts his revenge he hides his rebellion (making errors) from the agent he is rebelling against (the warden).

The rebel and interactor roles change between Smith and the warden, offering an opportunity to consider extending the framework. Since the rebel agent and interactor switch after betrayal, it creates what we term a *chain of rebellion*. If the warden had not rebelled against Smith, the latter would have not rebelled by seeking revenge or justice. Finally, we highlight that such episodes of rebellion can support plot progression in an interactive narrative but may fall outside the original framework's scope, as they are not necessarily constructive. In the example, Smith's revenge is rebellion in *support* of something, which the framework assumes.

We want our intention dynamics of computational models of rebel behavior to be readily understood by humans. Recall from Section 3.1 that the QUEST cognitive model (Graesser et al., 1991) has been used to assess a reader's comprehension of a computational model's output (Riedl & Young, 2010; Cardona-Rivera et al., 2016). This represents a reader's mental model of a story as a graph called the QUEST knowledge structure (QKS). We can assess how well a QKS represents this model by comparing question-answer pairs formed from its nodes to human subject responses. Subject responses rate a question-answer pair's goodness-of-answer on a four-point Likert scale.

Plan-based models of narrative have leveraged QKS to represent more complex agent interactions through the use of subgraphs (Amos-Binks & Young, 2018). Rather than analyzing an entire story's QKS, a subgraph focuses on reader comprehension of a specific story element. We hypothesize three subgraphs represent the intention dynamics of our rebel behaviors. Using these subgraphs, we generate question-answer pairs to evaluate how well our computational models induce the desired mental models in human subjects.

Table 2. Question-answer pairs for evaluating QKS subgraphs as representative of a subject's mental model after reading text generated from our rebel behavior definitions.

Behavior	#	QKS Nodes	Question-Answer pair	Question-Answer from Prison
Betrayal	Q1	E1-S1	Why did $Agent_1$ do trust action? Because $Agent_1$ trusted $Agent_2$	Why did Smith embezzle money with the warden? Because Smith trusted the warden
	Q2	S2-E2	Why $Agent_1$ believe they were betrayed? Because χ	Why did Smith believe he was betrayed? Because the warden denied Smith's trial
	Q3	S2-S1	Why $Agent_1$ believe they were betrayed? Because $Agent_1$trusted $Agent_2$	Why did Smith believe he was betrayed? Because Smith trusted the warden
Revenge	Q4	G3-S2	Why did $Agent_1$want revenge? Because $Agent_1$was betrayed	Why did Smith want revenge? Because Smith believed he was betrayed
	Q5	G4-G3	Why did $Agent_1$ do revenge action? Because $Agent_1$wanted revenge	Why did Smith make errors? Because Smith wanted revenge
Justice	Q6	G5-S2	Why did $Agent_1$want justice? Because $Agent_1$was betrayed	Why did Smith want justice? Because Smith believed he was betrayed
	Q7	G6-G5	Why did $Agent_1$ do just action? Because $Agent_1$wanted justice	Why did Smith notify the IRS? Because Smith wanted justice

The betrayal subgraph in Figure 2 represents the implied trust between two agents after a trustful action (E1-S1). An event (E2) introduces a causal link threat that causes an agent to reconsider its intentions. This event, along with the implied trust (S1), are the sufficient conditions for betrayal (S1-S2, E2-S2). To validate the subgraph as representative of a mental model after experiencing the plan-based betrayal, we use the question-answer pairs in rows 1 to 3 of Table 2. The three question-answer pairs confirm the existence of the aforementioned edges. The revenge subgraph in Figure 2 requires only two question-answer pairs. One pair confirms that AGENT$_2$'s betrayal (the warden) initiated AGENT$_1$ (Smith) to adopt a revengeful goal (S2-G3). A second question-answer pair assesses whether Smith took action in service of his revengeful goal (G4-G3).

Similar to revenge, the justice subgraph requires only two question-answer pairs. One pair confirms that AGENT$_2$'s betrayal (the warden) initiated AGENT$_1$ (Smith) to adopt a just goal (S2-G5). A second question-answer pair assesses whether Smith took action in service of his just goal (G6-G5). Together these three QKS subgraphs represent the mental models that we expect will result from human subjects who read text generated from our three formal definitions. Although are defined separately, they are linked in the structure, representing their interdependent dynamics.

5. Experimental Evaluation

We have set our evaluation of plan-based models of rebel agent behavior in a story generation context, although we believe these behaviors to be applicable to other human-AI interactions. The goal of our evaluation is to determine if our plan-based definitions induce the mental models represented by the QKS subgraphs in Section 4. In short, we are investigating whether the plan syntax for rebel agent behaviors are semantically meaningful to humans. The first part of the evaluation concerns the relationship of trust to betrayal. Trust is established between two agents through an action to which they both consent. When one of these agents actively foils the other's goals, we hypothesize it will be interpreted as betrayal by our subjects. This betrayal leads to the second part of the evaluation. Here we assess whether the betrayal action initiates a reconsidered intention on the part of the betrayed agent, who then responds by adopting a revengeful or just intention. We hypothesize that this sequence (betrayal action, intention revision) creates a mental model in our subjects that we represent with the QKS subgraphs in Figure 2. The experiment uses a question-answer protocol to evaluate whether our computational models of rebel agent behavior lead to the expected knowledge structure and aims to answer two research questions:

- **R1.** How do trust and exceptional actions affect the betrayal QKS subgraph?
- **R2.** How do trust and exceptional actions affect the revenge and justice QKS subgraphs?

The sections that follow describe our investigation into answering these questions. We discuss the experimental design, the hypotheses we tested, and our empirical findings.

5.1 Experimental Design

Story Structures. We implemented a version of our running *Prison* example as a planning problem in PDDL (McDermott et al., 1998) and generated an initial solution plan with 13 steps. From this initial plan, we used the procedure from Amos-Binks et al. (2018) to generate the betrayal and revenge intention dynamic represented in the plan $\pi_{Revenge}$. The resulting plan represents the story from an interactive narrative in which a user chooses actions (as the warden) that require an experience manager to generate a new story branch using intention revision. We represent another potential user response in the solution plan $\pi_{Justice}$ in which Smith responds to betrayal with a just intention. These two stories differed by a total of three statements.

Our experiments focus on the intentional dynamics of rebel agents and we deliberately minimized discourse effects (such as dramatic effects by reordering steps) in the implementation. After linearizing the partial orderings of a plan's steps, we apply a template, $\langle agent1, action, agent2 \rangle$, that generates a story statement for each plan step. Templates are a simple form of natural language generation that do not include intermediate representations (e.g., sentence plans) used by more general systems (Reiter & Dale, 1997). Instead, template methods map nonlinguistic input (e.g., plan steps) directly to text structure (e.g., story statements). The text for $\pi_{Revenge}$ appears in Table 3.

Subjects. We collected data from 176 participants who were recruited on the Figure Eight crowd sourcing platform and paid 2.00 USD to read our generated narrative text and then respond to eight randomly ordered question-answer pairs. Two pairs were testing basic comprehension of the story, three pairs assessed the betrayal QKS subgraph, two pairs assessed their respective revenge or justice group, and a final pair screened automated responses. Goodness of answer responses were recorded for each question on a four-point Likert scale using answer options of *very poor*, *somewhat poor*, *somewhat good*, *very good*.

Procedures. After logging into the online experiment, participants were randomly assigned into a revenge or justice evaluation group. Subjects were also assigned to either a control or one of four treatment groups. The INFERENCE (control) group read a variation of the story with two extra story statements that explicitly stated Smith trusts the warden and the warden betrayed Smith. This let

Table 3. This table contains plan steps transformed into story statements read by the $\pi_{Revenge}$ group. Different subject mental models were induced by having experimental groups read slightly different versions of the story, as indicated in column two. The INFERENCE (control) group received the entire story but with two explicit inferences added (trust and betrayal). The TRUST-BETRAYAL group received the entire story but with no explicit inferences, the TRUST group received a story with no betrayal action, the BETRAYAL group received a story with no trust action, and the *none* group received a story with neither trust or betrayal actions.

#	Group	Statement
S1	all groups	Smith is wrongfully convicted and wants to be exonerated.
S2	all groups	Smith complies with demands from the warden.
S3	all groups	Smith files taxes for the warden.
S4	all groups	Smith does the prison's accounting.
S5	INFERENCE	Smith trusts the warden.
S6 (trust)	INFERENCE, TRUST-BETRAYAL, TRUST	Smith embezzles prison funds in warden's retirement account.
S7	all groups	Smith makes friends with some inmates.
S8	all groups	Smith learns his friend has some exonerating evidence.
S9	all groups	Smith requests exoneration from the warden.
S10 (χ)	INFERENCE, TRUST-BETRAYAL, BETRAYAL	The warden says he will never allow Smith to be exonerated.
S11	INFERENCE	Smith believes he was betrayed.
S12	all groups	Smith continues to do the prison's accounting.
S13	all groups	Smith deliberately makes errors in the prison's ledger.
S14	all groups	Smith attends the warden's retirement party.
S15	all groups	The warden discovers his retirement account is empty.

us assess whether subjects make the inference of trust between the warden and Smith and whether Smith's justice or revenge response was motivated by betrayal. The treatment groups read stories in which there was no explicit statement of trust or betrayal (TRUST-BETRAYAL group), no trusting action or explicit statement (BETRAYAL group), no betrayal action or explicit statement (TRUST group), and no trust or betrayal actions or explicit statements (*none* group). Table 3 presents the text read by each group in the revenge version of *Prison*. Participants began the experiment by completing a brief demographic survey and then read their assigned story, presented as a single page of text, after which they completed the question-answer survey.

5.2 Experimental Results

Prior to our analyses, we inspected and coded the data to identify missing values and outliers but no extreme values were identified. We broke our two research questions into hypotheses that we analyzed using a chi-square test for independence, with all analyses computed using a critical alpha value of p = .05.

R1. How do trust and exceptional actions affect the betrayal QKS subgraph?

Our response was to test three hypotheses, H1–H3, one per betrayal question-answer pair in Table 2.

H1. Subjects who read the trust and exceptional action statements will rate the *"Why did Smith embezzle money with the warden? Because Smith trusted the warden."* question-answer pair (Q1) with a *very good* goodness-of-answer. The chi-square statistic is 30.14 and has a p-value of 0.003 (Figure 3), indicating that the experimental groups were drawn from two different distributions. Upon further investigation, the *very/somewhat good* responses of the INFERENCE and TRUST-BETRAYAL groups are similar, suggesting that the subjects made the inference that Smith believed he was betrayed (TRUST-BETRAYAL) without this being explicitly stated (INFERENCE). In contrast, the responses for TRUST, BETRAYAL, NONE are similar. This suggests that subjects in TRUST who read the *embezzle* statement (S6) without the *betrayal* statement (S10) did not infer it and require both statements to do so. Overall, these results confirm H1 for the Prison story and indicate that the E2-S2 edge in the betrayal QKS subgraph represents part of subject mental models.

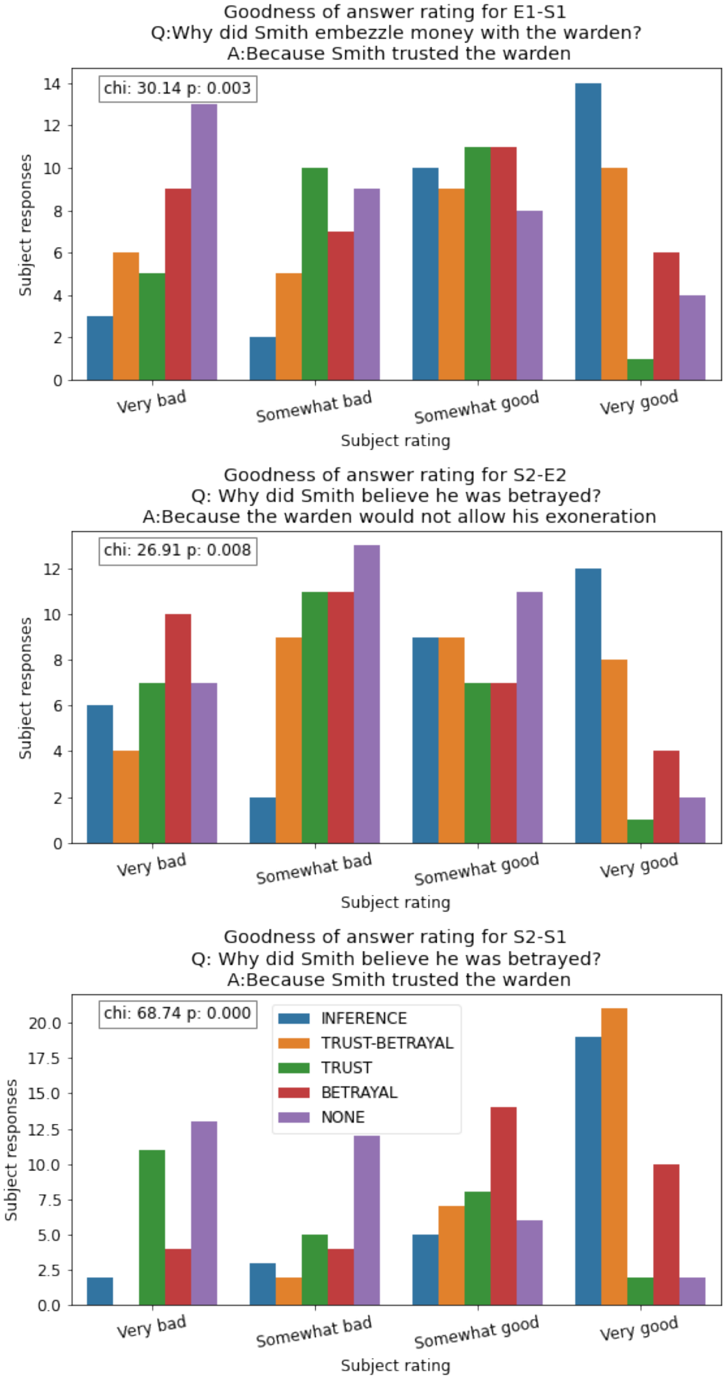
H2. Subjects who read the trust and exceptional action statements will rate the *"Why did Smith believe he was betrayed? Because the warden denied Smith's trial."* question-answer pair (Q2) with a *very good* goodness-of-answer. The chi-square statistic is 26.91 and has a p-value of 0.008 (Figure 3), indicating that the experimental groups were drawn from two different distributions. Upon further investigation, the *very/somewhat good* responses of the INFERENCE and TRUST-BETRAYAL groups are similar, suggesting that the subjects inferred that Smith believed he was betrayed (TRUST-BETRAYAL) without it being explicitly stated (INFERENCE). In contrast, the responses for TRUST, BETRAYAL, NONE are similar, suggesting that subjects in BETRAYAL who read the *betrayal* (S10) without the *embezzle* statement (S6) did not infer the betrayal and require both statements to do so. Overall, these results confirm H2 and indicate that the S1-S2 edge in the betrayal QKS subgraph represents part of subject mental models of betrayal.

H3. Subjects who read the trust and exceptional action statements will rate the *"Why did Smith believe he was betrayed? Because Smith trusted the warden."* question-answer pair (Q3) with a *very good* goodness-of-answer. The chi-square statistic is 68.74 and has a p-value of 0.000 (Figure 3) indicating that the experimental groups were drawn from two different distributions. Upon further investigation, the *very/somewhat good* responses of the INFERENCE and TRUST-BETRAYAL groups are similar, suggesting that the subjects made the inference that Smith trusted the warden (TRUST-BETRAYAL) without this being stated explicitly (INFERENCE). In contrast, the responses for TRUST, BETRAYAL and NONE are similar, suggesting they did not make the trust inference. Overall, these results confirm H3 for the Prison story and indicate that the E1-S1 edge in the betrayal QKS subgraph represents part of subject mental models of betrayal.

In summary, in all three hypotheses, subjects in the TRUST, BETRAYAL, and NONE groups with story statements were drawn from a separate distribution than the TRUST-BETRAYAL (original plan) and the INFERENCE group (plan with inferences). This indicates that subjects need to read both the trust and exceptional actions for our QKS subgraph to accurately represent a subject's mental model of betrayal therefore supporting our plan-based definition of betrayal.

R2. How do trust and exceptional actions affect the revenge and justice QKS subgraphs?

Our approach to answering this research question centered on whether the revenge and justice intentions were appropriate responses to betrayal. As shown on the previous question, subjects must observe both a trust action and exceptional action to infer betrayal (TRUST-BETRAYAL). We expect subjects who receive neither treatment to indicate that revenge and justice are not linked to betrayal.


Goodness of answer rating for E1-S1
Q:Why did Smith embezzle money with the warden?
A:Because Smith trusted the warden

Goodness of answer rating for S2-E2
Q: Why did Smith believe he was betrayed?
A:Because the warden would not allow his exoneration

Goodness of answer rating for S2-S1
Q: Why did Smith believe he was betrayed?
A:Because Smith trusted the warden

Figure 3. Goodness-of-answer responses to betrayal question-answer pairs. The top contains responses to *"Why did Smith embezzle money with the warden? Because Smith trusted the warden."*, the center contains responses to *"Why did Smith believe he was betrayed? Because the warden denied Smith's trial."*, and the bottom shows responses to *"Why did Smith believe he was betrayed? Because Smith trusted the warden."*.

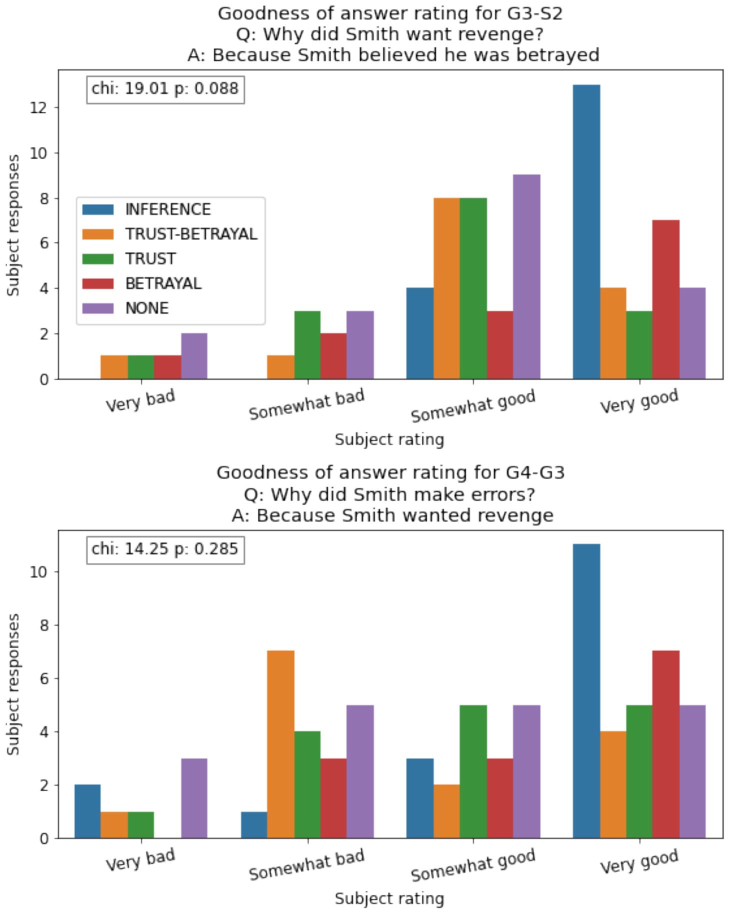


Figure 4. Prison goodness-of-answer responses for revenge question-answer pairs. The histogram on the top encodes responses to *“Why did Smith want revenge? Because Smith believed he was betrayed.”*. The histogram on the bottom encodes responses to *"Why did Smith make errors? Because Smith wanted revenge."*.

H4. Subjects who do not read the trust action or exceptional action will rate the *“Why did Smith want revenge? Because Smith believed he was betrayed.”* question-answer pair (Q4) with a *very/somewhat bad* goodness-of-answer. The chi-square statistic is 19.73 and has a p-value of 0.072 (Figure 4), indicating that we cannot conclude the experimental groups were drawn from different distributions. This is surprising, given that TRUST, BETRAYAL, and NONE did not associate Smith’s betrayal with the trust or exceptional actions, but did associate a betrayal with the revengeful response. Overall, results do not support H4 for the Prison story. They suggest that the S2-G3 edge in the QKS subgraph will be inferred spontaneously by subjects (because answers are skewed towards *good*) regardless of reading statements generated from plan-based definitions of betrayal. Subjects’ mental models of betrayal-revenge dynamic are more complex than assumed by our current QKS.

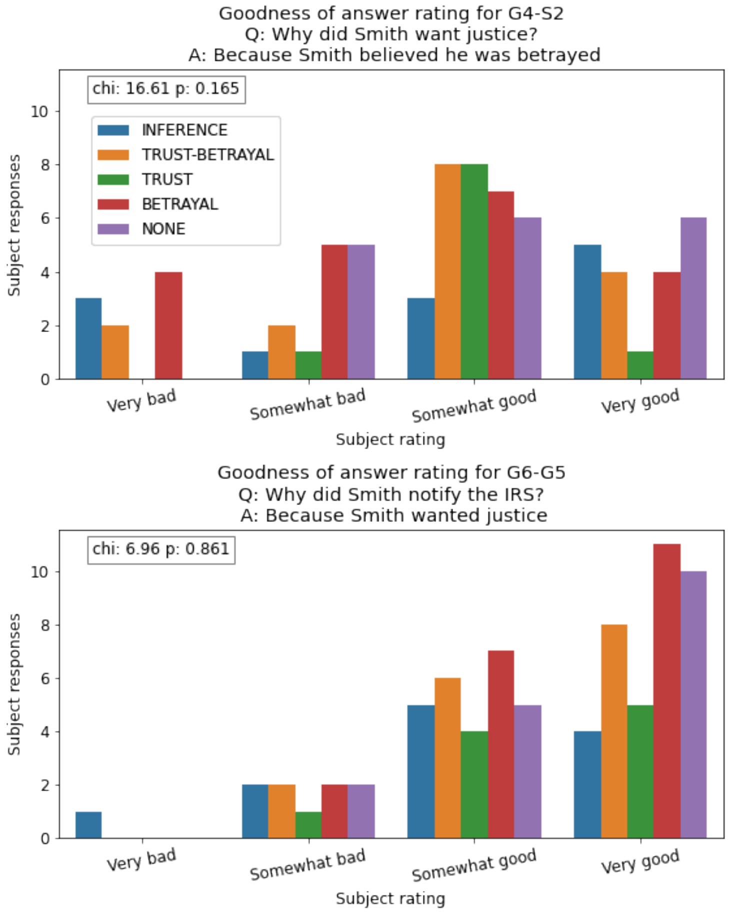


Figure 5. Prison goodness-of-answer responses for Justice question-answer pairs. Responses to *"Why did Smith want justice? Because Smith believed he was betrayed."* are in the histogram on the top and those to *"Why did Smith notify the IRS? Because Smith wanted justice."* are on the bottom.

H5. All subjects will rate the *"Why did Smith make errors? Because Smith wanted revenge."* question-answer pair (Q5) with a *very/somewhat good* goodness-of-answer. The chi-square statistic is 14.97 and has a p-value of 0.243 (Figure 4), indicating that we cannot conclude the experimental groups were drawn from the different distributions. This suggests that subjects interpreted Smith as acting in pursuit of revenge and that the G4-G3 edge in the QKS subgraph is representative of subject mental models.

H6. Subjects who do not read the trust action or exceptional action will rate the *"Why did Smith want justice? Because Smith believed he was betrayed."* question-answer pair (Q6) with a *very/somewhat bad* goodness-of-answer. The chi-square statistic is 15.31 and has a p-value of 0.225 (Figure 5), indicating that we cannot conclude the experimental groups were drawn from different distributions. Like H4, this is surprising given that TRUST, BETRAYAL, and NONE did not associate Smith's betrayal with the trust or exceptional actions but did associate betrayal with a just response. Overall, these results do not confirm H6 for the Prison story. They suggest that the S2-G5 edge in the QKS subgraph will be inferred spontaneously by subjects (because answers are skewed towards *good*) regardless of reading statements generated from plan-based definitions of betrayal. Again, subjects' mental models of betrayal-justice dynamic appear to be more complex than assumed by our current QKS.

H7. All subjects will rate the *"Why did Smith notify the IRS? Because Smith wanted justice."* question-answer pair (Q7) with a *very/somewhat good* goodness-of-answer. The chi-square statistic is 7.38 and has a p-value of 0.832 (Figure 5). This indicates that we cannot conclude that the experimental groups were drawn from different distributions. This suggests that subjects interpreted Smith as acting in pursuit of a just response and that the G6-G5 edge in the QKS is consistent with their behavior.

In summary, our QKS subgraphs for revenge and justice do not accurately represent mental models that subjects created from reading text generated from our plan-based definitions of a betrayal and revenge/justice dynamic. Subjects infer that the revenge and justice response are motivated by betrayal, but they do not infer how the betrayal occurred. We hypothesize that, instead, the intention dynamics between betrayal and revenge/justice match a schema from their general world knowledge, in that many subjects implicitly know that revenge and justice often follow betrayal.

6. Discussion and Future Work

Rebel agency is both an important narrative plot device and essential for true autonomy. We supported both of these claims by defining computational models of rebellion in plan-based agents. Betrayal, revenge, and justice are all behaviors in opposition to the aims of other agents and therefore rebellious. The two key plan elements in betrayal are a trust action and an exceptional action. Two agents consent to a trust action because it is in service of both their intentions. An exceptional action introduces a causal link threat to a trusting agent's subplan. Once an agent has been betrayed, it may respond with revengeful or just intentions. A revengeful intention seeks to foil the offending agent's plan, whereas a just intention attempts to pursue justice in a value system (e.g., by deferring to authorities). The BDI agent control loop inspired our approach to these intention dynamics, with our betrayal model meeting the conditions for BDI agents to first *reconsider* intentions and revengeful or just responses are products of the agent's deliberation over their *options*.

In Section 4, we characterized rebel behavior within an existing rebellion framework. All three are inward-oriented, explicit, and interaction-initiation forms of rebellion. Notably, the framework we used does not capture the *chain of rebellion* that describes the dynamic rebel-interactor intentions between Smith and the warden. We expect that, as rebel agents become more widely used, frameworks for characterizing them will evolve to capture the dynamics described here, as well as many others.

In addition to defining and characterizing plan-based structures, we investigated their effect on comprehension in generated narrative text. We proposed QUEST knowledge structure subgraphs to represent mental models of the dynamics between betrayal, revenge, and justice, from which we formulated two research questions. Our experiments for question R1 showed that subjects needed both an action that implied trust and an action that violated it to infer that betrayal had occurred. This was consistent with the QUEST knowledge structure for betrayal. The results for question R2 showed that subjects agreed revengeful and just behaviors were in response to betrayal, but they did not require actions to establish and violate trust. We hypothesize this is the result of a subject's general knowledge that closely ties revenge and justice to betrayal. In future work, we should refine our revenge and justice QKS subgraphs to reflect this close relationship.

Beyond the aforementioned improvements to mental model representations, intention dynamics themselves offer an interesting avenue of investigation. In this work we used betrayal, revenge and justice because they support rebel agent behavior, but we view intention dynamics as essential for autonomous agents understanding their human counterparts mental state. Much of current work focuses on immediate, transactional interactions (e.g., recognize voice commands). Intention dynamics offer a kind of playbook of behaviors that often play out over time and are readily understood by humans. To this end, they could go beyond transactional interactions and structure long-range human-computer interactions in applications such as game playing, personal assistants, and dialogue systems. Narratologists have gone so far as to identify the intention dynamics in different forms of fiction (Propp, 2010; Cook, 2011), but there has been little progress toward an exhaustive taxonomy, let alone computational models, of them. Addressing this limitation is a first step toward future work on intention dynamics (including the rebellious ones) that can improve the overall experience of human-computer interactions.

Acknowledgements

Thanks to ARA, Navatek, and NRL for sponsoring this research, as well as the reviewers and Pat Langley for their comments and recommendations.

References

Amos-Binks, A., Spain, R., & Young, R. M. (2018). Subjective experience of intention revision. *Advances in Cognitive Systems*, *6*, 193–209.

Amos-Binks, A., & Young, R. M. (2018). Plan-based intention revision. *Proceedings of the 2018 AAAI Conference on Artificial Intelligence* (pp. 8047–8048). New Orleans: AAAI Press.

Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Briggs, G., & Scheutz, M. (2017). The case for robot disobedience. *Scientific American*, *316*, 44–47.

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

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, *42*, 213–261.

Coman, A., & Aha, D. W. (2018). AI rebel agents. *AI Magazine*, *39*, 16–26.

Coman, A., & Muñoz-Avila, H. (2014). Motivation discrepancies for rebel agents: Towards a framework for case-based goal-driven autonomy for character believability. *Case-Based Agents: Papers from the ICCBR Workshop*. Cork, Ireland.

Cook, W. W. (2011). *Plotto: The master book of all plots*. Portland, OR: Tin House Books. From https://books.google.com/books?id=HDOPLT0Wer8C.

Cox, M. T., Dannenhauer, D., & Kondrakunta, S. (2017). Goal operations for cognitive systems. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 4385–4391). San Francisco: AAAI Press.

Dannenhauer, D., Floyd, M. W., Magazzeni, D., & Aha, D. W. (2018). Explaining rebel behavior in goal reasoning agents. *Explainable AI Planning: Papers from the ICAPS Workshop*. Delft, The Netherlands.

Graesser, A. C., Lang, K. L., & Roberts, R. M. (1991). Question answering in the context of stories. *Journal of Experimental Psychology: General*, *120*, 254–277.

McDermott, D., Knoblock, C., Veloso, M., Weld, S., & Wilkins, D. (1998). PDDL–the Planning Domain Definition Language: Version 1.2. *Yale Center for Computational Vision and Control, Tech. Rep. CVC TR-98-003/DCS TR-1165*.

Meehan, J. R. (1977). TALE-SPIN: An interactive program that writes stories. *Proceedings of the Sixth International Joint Conference on Artificial Intelligence* (pp. 91–98). Cambridge, MA: William Kaufmann.

Muñoz-Avila, H., Aha, D. W., Jaidee, U., Klenk, M., & Molineaux, M. (2010). Applying goal driven autonomy to a team shooter game. *Proceedings of the Twenty-Third International FLAIRS Conference* (pp. 465–470). Daytona Beach, FL: AAAI Press.

Perez y Perez, R., & Sharples, M. (2001). MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*, 119–139.

Porteous, J., Cavazza, M., & Charles, F. (2010). Applying planning to interactive storytelling: Narrative control using state constraints. *ACM Transactions on Intelligent Systems and Technology*, *1*, 10.

Propp, V. (2010). *Morphology of the folktale* (2nd ed.). Austin, TX: University of Texas Press.

Rao, A. S., & Georgeff, M. P. (1998). Decision procedures for BDI logics. *Journal of Logic and Computation*, *8*, 21–26.

Reiter, E., & Dale, R. (1997). Building applied NLG systems. *Natural Language Engineering*, *3*, 57–87.

Riedl, M. O., & Young, R. M. (2010). Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, *39*, 217–267.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.*. Hillsdale, NJ: Lawrence Erlbaum.

Van Der Hoek, W., Jamroga, W., & Wooldridge, M. (2007). Towards a theory of intention revision. *Synthese*, *155*, 265–290.