Computational Models of Rebel Agent Behavior for Interactive Narrative

Adam Amos-Binks
Applied Research Associates, Inc.
Decision Systems Group
Raleigh, NC, USA
aamosbinks@ara.com

Dustin Dannenhauer
NRC Postdoctoral Fellow
Naval Research Laboratory
Washington, DC, USA
dustin.dannenhauer.ctr@nrl.navy.mil

David W. Aha
Navy Center for Applied Research
in Artificial Intelligence
Naval Research Laboratory
Washington, DC, USA
david.aha@nrl.navy.mil

Abstract

The emerging interest in rebel agents for autonomy and connection with the intention dynamics of rebellious behavior has yet to be made. To address this limitation we make three contributions. First, we define plan-based computational models of betrayal, revenge, and justice as rebel agent behavior for interactive narratives. Second, we use the QUEST knowledge structure to develop representations of the desired mental models created by the rebellious behaviors and propose a method to evaluate them. Lastly, we characterize the behaviors within an existing rebel agent framework. These contributions operationalize rebel agents in a strong application context with cognitive psychological foundations.

Introduction

Interactive narrative strives to balance an author defined story arc with user actions that also shape the plot. In response to these actions, belief-desire-intention character agents may adapt their behavior through narrative devices such as intention revision. Dynamic intentions enable these agents to support a rich virtual environment where they foil, co-operate, and even rebel against the user agent.

While intention revision enables an agent to drop old intentions and adopt new ones, it is a coarse-grained model of behavior change. It lacks fine-grained detail to represent more specific intention dynamics of narrative phenomena that are often key to plot development. Finer-grained examples of intention dynamics include revenge[e.g.], betrayal[e.g.], and justice[e.g.] that support a rebellion story arc. In contrast to the typical narrative use of rebellion where a protagonist must subvert an antagonist's power, an emerging concept from the AI community has rebel agents serving functional roles. Namely that a rebel agent's non-compliance is essential to true agency and autonomy (Coman and Muñoz-Avila 2014), however it has yet to be operationalized in computational models of narrative.

To address some of these limitations, we examine the auspicious relationship between intention dynamics, interactive narrative, and rebel agents. With our approach, we make three contributions to operationalizing rebellious behaviors for plan-based intentional agents. The first is intentional plan-based definitions of 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 leverage the QUEST cognitive model in a proposed evaluation. We use QUEST knowledge structures to represent the user's expected mental model after experiencing the plan-based definitions. Finally, we classify betrayal, revenge, and justice behaviors under the rebellion framework (Aha and Coman 2017). Together these three contributions take the first steps in advancing rebel agent applications.

Previous Work

Our contributions are based on three areas of previous work. First, narrative generation, gives a background for intentional planning. Second, intention, is the mental state and mechanism that enables rebel definitions. Lastly, rebel agent frameworks characterize the qualities of our definitions.

Interactive Narrative

Schank and Abelson 1977 were perhaps the first to publish the concept of classical planning representing story plots. It was based on the theoretical overlaps of plot events with the action-oriented, causally-linked, and temporally-ordered properties of plans. Since these first insights, story generators have extended representations to capture a range of narrative features (Meehan 1977; Porteous, Cavazza, and Charles 2010; Perez y Perez and Sharples 2001).

We focus on intentional partial ordered causally linked planning (Riedl and 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 only generating solution plans that contain actions on a sequence of causally connected actions that achieve the goal of an agent's intention. An *intention frame* aggregates an agent's goal, a motivating plan step and a causally connected action sequence called a subplan. Both the goal and subplan are key in identifying reconsidered intentions and discussed further the next section.

Because IPOCL planning leverages the explicit notion of causality and intentionality, researcher's have evaluated its affect on a reader's mental model using the QUEST cognitive model of question-answering (Q-A) in the context of

Algorithm 1 BDI control loop excerpt (Rao and Georgeff 1998)

```
1: ... B(beliefs),D(desires),I(intentions), π(plan)
2: get next observation ω
3: revise B on the basis of ω
4: if (reconsider(B,I)) then
5:     D = options(B,I)
6:     I = filter(B,D,I)
7:     if not sound(π,B,I) then
8:         π = plan(B,I)
9: ...
```

stories (Graesser, Lang, and Roberts 1991). The model uses a graph called the QUEST knowledge structure (QKS) to represent a reader's mental model of a storys causal and intentional structure. Additionally, QUEST defines the structure of Q-A and includes a QKS traversal to predict reader responses. QUEST predictions are compared with actual readers responses to evaluate how well the QKS represents a mental model.

From this firm cognitive psychological foundation, IPOCL plans have been used to structure the plot of an interactive narrative (IN). An IN allows a participant to shape the plot through their interactions. When causal link threats are introduced by user actions, an experience manager agent (EM) will generate a new plan, ensuring it is coherence with the failed one. Specifically for IPOCL, any agent who changes goals must do so in a principled fashion, or risk reducing a user's engagement due to a lack of coherence.

Intention

The use of intentions in narrative planning is 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. the justification for action). Intention was later formalized for logical agents by Cohen and Levesque 1990 and lead to the development of decision making abilities for BDI agents (Rao and Georgeff 1998).

The BDI research community has made substantial research efforts on belief revision and update (e.g. (Rao and Georgeff 1998)), while only making 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 time logic based on Alg. 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 unfeasible. It is not specified exactly when agents should reconsider, but that observation and the enabling of previously unachievable goals alone are not sufficient. On the other hand, when observations are made that make a current intention unachievable, the agent would be well served to *reconsider* and execute lines 5-8 to develop a new plan for an achievable goal. This was operationalized in a plan-based model of intention revision by Amos-Binks and Young 2018 where causal link threats cause an agent to reconsider an intractable goal and initiate a revision.

Rebel Agents

Rebel behavior, or the ability for an agent to reject, protest, or alter it's goals, plans, or actions is a desired capability for many autonomous systems (Briggs and Scheutz 2017; Dannenhauer et al. 2018). Agents often have access to different sources of information and operate with safety or ethical constraints. Consider the following hypothetical scenarios: (1) 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. (2) A hotel service robot denies a request to retrieve luggage for a person who is attempting to steal from other hotel guests. For autonomous AI systems, rebellion is especially important when the designers of a system are different than the users of the system, 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 which lead to rebel behavior. GDA is a model of goal reasoning where agents perform a four-step process: detect discrepancies, explain what may have caused the discrepancies, formulate new goals, and select which goals to pursue (Munoz-Avila et al. 2010). Normally 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 (A) the agent's current plan, (B) the observed state, or (C) the agent's current goal. Since motivations change, A, B, or C may no longer align with the agent's current motivation.

GDA agents perform similar processes to BDI agents where desires (BDI) are similar to goals (GDA) and intentions (BDI) are similar to plans (GDA). We describe our approach using a BDI perspective, although GDA agents could also use such an approach. The focus here of this work is on the intention revision process that characterizes behaviors such as betrayal, revenge, and justice. Each of these behaviors can be seen as a form of rebellion that leads to more believable characters. Coman and Muñoz-Avila focus on characters that identify conflicts between their motivations and goals/actions/state while in this work we focus on finer-grained intention revision for behaviors including betrayal, revenge, and justice. Motivation discrepancies described by Coman and Muñoz-Avila could be used to identify when to perform the intention revision we describe here.

Computational Models for Rebel Agents

Intentional planning systems generate action sequences that reach the goal conditions of a planning problem. These plans scaffold the plot of an interactive narrative where intentional agents can adopt, drop, or revise their intentions in response to their interactive narrative environment but are limited as they do not deliberately adopt rebellious behaviors. To address this limitation, we provide intentional planning definitions that characterize different types of rebellious behavior. We use a simple example, *Prison*, to both provide examples of basic definitions of intentional plans and capture how a rebellious non-player character agent reacts to an interactive narrative player agent. Second, we construct the desired QUEST knowledge structures that represent the mental models resulting from the rebellious behavior. Finally, we outline how these rebellious behaviors are characterized by an existing rebel agent framework.

Intentional Planning

Our approach uses intentional planning definitions from Riedl and Young's work on IPOCL 2010. Intentional planning differs from classical planning 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 (happenings are *fate*'s intention). We refer to this causally linked set of actions as an agent's subplan to achieve their goal. The agent, their subplan and goal are aggregated into a structure called an intention frame that reflects the additional constraints on intentional plans.

Our *Prison* example in Figure 1 has two agents, Smith (a non-player character 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) Is a logical sentence that identifies a desired world-state of an agent.

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 help Smith, *intends (warden, hasTrial(Smith))*.

The agent who executes any given action is called the consenting agent. In the original plan (top) in Figure 1, Smith is the consenting agent of the *MakeFriends*, *SharedPlan*, *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, $\textsc{Pre}(A)$, effects that result, $\textsc{Eff}(A)$, and a consenting agent, $\textsc{agent}(A)$, 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 steps by grounding the free variables and result in plan steps $s_1 - s_6$ in the original plan. 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 the set of steps (a step is a ground instance of an action in POCL planning) is S, 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 connecting the steps in Fig. 1. Finally, intention frames are the essential element 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 $\mathcal{I} = \langle \textsc{agent}, g, m, \sigma, T \rangle$ where g is $\textsc{agent}$'s goal, motivating step $m \in S$ with the effect $\neg g$, the satisfying step $\sigma \in S$ with g as an effect. A subplan for a to achieve g is a set of steps $T \subseteq S$ that $\textsc{agent}$ consents to, each step shares at least one causal link to another step in T, and achieves g. Steps in T occur after m and before σ.

The original plan in Figure 1 includes the intention frames for *Smith* and *warden*. Finally, intentional plans solve planning problems, the plan in Figure 1 solves a planning problem with a single condition, *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, $\mathcal{A}$ the set of symbols referring to agents, $\mathcal{O}$ the set of symbols referring to objects, and Λ a set of action schemata.

While executing a plan-based interactive narrative, we label 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 $exonerated(Smith)$ is active from $s_1 - s_5$, until he executes the satisfying step, *Testify* (s_6). Active intentions are useful for identifying reconsidered intentions and support our definition of intention revision. During a plan-based interactive narrative, the player agent (the warden) can take actions introducing causal link threats, preventing non-player agents from achieving goals.

Definition 9 (Causal link threat) A causal link threat occurs when a causal link is established $s \xrightarrow{p} u$, and 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 q of u is no longer satisfied by the state after s is executed and thus u will not execute.

In the betryal-revenge variant 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.

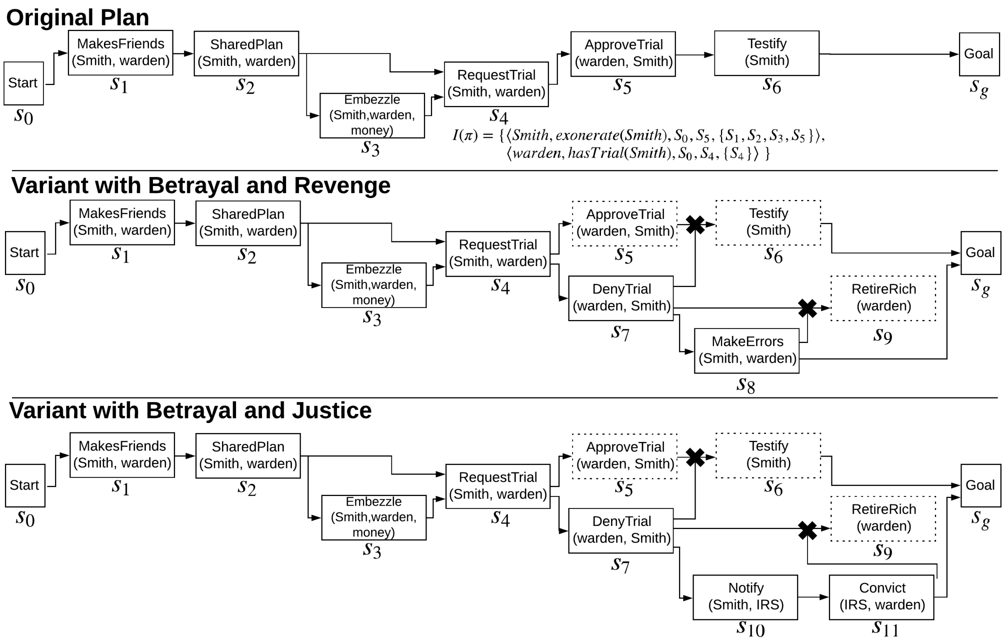

Figure 1: *Prison* intentional plan (π) with two variants capturing the warden's betrayal and Smith's options; revenge or justice

Definition 10 (Exceptional Action) An exceptional action s'_t executed at time t by the user agent, $\text{AGENT}(s'_t) = user$, where 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 $t \leq u$.

This exceptional action causes Smith to reconsider (as in line 4, Alg. 1) his $exonerate(Smith)$ intention.

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

If an intention is reconsidered, an agent deliberates whether the goal is worth pursuing. Cohen and Levesque (Cohen and Levesque 1990) prescribe that an agent should only drop a goal after achieving it or when the agent believes the goal is unachievable. We are interested in the latter:

Definition 12 (Unachievable Goal) A goal is unachievable, g_u, if using a agent's belief state as the initial state, no subplan to achieve $g(\mathcal{I}_R)$ exists.

Agents maintain a belief state of their environment represented as sets of consistent, non-modal, ground literals. They update their belief state by observing the effects of actions. After the *DenyTrial* action, Smith believes exoneration is unachievable as no subplan exists to achieve exoneration. This belief leads Smith to drop this goal and because there was a shared plan with the warden to achieve it, he believes he was betrayed. Smith must now consider his options (line 5 in Alg. 1) a new goal (revenge or justice) that also solves the problem, which we characterize as an intention revision:

Definition 13 (Intention Revision) An intention revision is $\langle \mathcal{I}_R, \mathcal{I}' \rangle$ where $\mathcal{I}_R$ is an active intention $g(\mathcal{I}_R)$ is unachievable and $\mathcal{I}'$ is an intention frame where $g(\mathcal{I}') \neq g(\mathcal{I}_R)$, and $\text{AGENT}(\mathcal{I}') = \text{AGENT}(\mathcal{I}_R)$.

Betrayal

Betrayal is an intention dynamic closely associated with, and often leads to, the intention revisions of revenge and justice. At the crux of betrayal is two agents with common or closely aligning intentions. From this mutual interest, the two agents develop a shared plan that requires, at least temporarily, trust between them. If an agent chooses to drop the shared intention by introducing a causal-link threat in pursuit of another goal, the other agent will view it as a violation of trust.

A shared plan can emerge for a variety of reasons. However, we avoid exhaustively defining it and instead indicate it with a simple operator. We use an action who's preconditions are that both agents are pursuing the same goal. Its lone effect is $sharedPlan$ that we use to define betrayal.

Definition 14 (Betrayed Intention) A betrayed intention, $\langle \mathcal{I}_a, \mathcal{I}_b \rangle$, where the subplan of $\mathcal{I}_a$, $T(\mathcal{I}_a)$, contains an effect that will introduce a causal link threat to the subplan of $\mathcal{I}_b$, $T(\mathcal{I}_b)$, such that the goal of $\mathcal{I}_b$ has a shared plan, indicated by $sharedPlan(g)$, and $\text{AGENT}(\mathcal{I}_a) = \text{AGENT}(\mathcal{I}_b)$.

In both *Prison* variants, S_2 is when the shared plan is created. From $S_2 - S_4$, Smith and the warden execute actions towards their shared goal. However, at S_5 the warden denies Smith's trial request at which point he drops his goal of helping Smith in favor of retiring rich. The combination of the causal link threat introduced by the warden in S_5 and the $sharedPlan$ represent betrayal in *Prison*.

Revenge

Revenge can be motivated by different reasons and we argue that betrayal is one of them. Intuitively, the concept of revenge is when an agent (Smith) who believes they have been wronged by another agent (the warden) adopts an intention to exact their grievance by foiling a goal of the offending agent . A plan-based definition is as follows:

Definition 15 (Revengeful Intention) A revengeful intention, $\langle \mathcal{I}_a, \mathcal{I}_b \rangle$, where the subplan of $\mathcal{I}_a$, $T(\mathcal{I}_a)$, contains an effect that will introduce a causal link threat to the subplan of $\mathcal{I}_b$, $T(\mathcal{I}_b)$, such that $\text{AGENT}(\mathcal{I}_b)$ had previously executed an exceptional action with effect ϵ that led to $\text{AGENT}(\mathcal{I}_a)$ reconsidering their intentions.

After the warden commits his betrayal in S_5 by adopting an intention to retire rich (($\mathcal{I}_b$) from Def. 15), Smith reconsiders his intentions as his subplan to achieve his exoneration is no longer viable. He drops his exoneration intention as their is no subplan to achieve it. In its place, Smith adopts a revengeful intention (($\mathcal{I}_a$) from Def. 15) and subversively makes accounting errors in the warden's embezzlement scheme (*MakeErrors*, S_8 in the first variant plan in Fig. 1). This action foils the warden's intention to retire rich, thereby representing a revengeful intention.

Justice

Pursuing justice is another response to betrayal an agent may deliberate over. There are a number of similarities between revenge and justice. However the main difference is that revenge deliberately subverts a legal or value system, whereas justice adheres to the system. We differentiate them in our plan-based definitions by the specificity of the intention. A revengeful intention pursues a specific goal to foil of another agent where the goal of a just intention is to bring about justice, whatever form it takes in the value system.

Definition 16 (Just Intention) A just intention, $\mathcal{I}$, is when the goal of $\mathcal{I}$ is *servedJustice(*AGENT_2*)*, such that AGENT_2 had previously executed an exceptional action with effect ϵ that had caused $\text{AGENT}(\mathcal{I})$ to reconsider their intentions.

In Figure 1, the second variant plan contains Smith's $Notify$ action (s_{10}) that alerts the IRS of the warden's embezzlement scheme. This compels the IRS to investigate and convict the warden. As a result of the conviction, the warden cannot achieve his retire rich intention and Smith achieves his justice intention. These intention dynamics represent our definition of a just intention.

Rebel Agent Framework Characterization

Betrayal, revenge, and justice can all be classified under the rebellion framework in (Aha and Coman 2017). Rebellion occurs between a rebel and an *interactor* which is the person rebelled against. Rebellion is classified under three dimensions: *expression*, *focus*, and *interaction initiation*. All three behaviors are examples of inward-oriented (*expression*) and explicit (*focus*) rebellion. Inward-oriented refers to an agent changing its own behavior rather than preventing another agent from behaving in a certain way. Explicit refers to a rebellion's observable effect as opposed to an agent expressing it in their own state of mind. All of the examples here result in a change in the actions of the rebelling agent.

Betrayal is a reactive (*interaction initiation*) rebellious behavior because the rebel agent (i.e. warden) is rejecting an agreed-upon cooperation (e.g. *SharedPlan*) with the interactor (i.e. smith). Revenge and justice are also reactive (*interaction initiation*) because rebellion arises from an interaction initiated by the interactor. In the case of revenge and justice, the rebel and interactor are flipped. In betrayal, the warden is the rebel and smith the interactor. However, in revenge and justice, the warden is the interactor and smith is the rebel. Importantly, in each model, the rebellion is only known from observing actions since no agent announces their rebellion to the other before taking action. In the revenge case, Smith attempts to hide their rebellion (making errors) from the agent they are rebelling (the warden).

It may seem that revenge and justice are outward-oriented forms of rebellion since they focus on altering the end state of another agent. The framework from Aha and Coman uses the initial interaction to define reactive vs. proactive, rather than the desired outcome. Thus in all of these examples, the rebellion is reactive. However, the rebel agent and interactor flip after betrayal, creating a chain of rebellion. If the warden had not rebelled against smith, smith would have not rebelled seeking revenge or justice. Finally, these episodes of rebellion may fall outside the scope of the framework because they are not necessarily constructive (i.e. Smith taking revenge may not be an example of rebellion in *support* of something, the kind of rebellion the framework classifies).

Proposed evaluation

To evaluate our computational models of rebel behavior, we turn to the QUEST cognitive model. QUEST represents a reader's mental model of a story with a graph structure called the QUEST knowledge structure (QKS). How well a QKS represents a mental model can be assessed by forming Question-Answer pairs from QKS nodes, and then comparing QUEST's prediction to human subject responses. Subject responses are asked to rate a pair's Goodness-Of-Answer (GOA) 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 and Young 2018). Rather than analyzing the whole story structure, the QKS subgraph approach focuses on reader comprehension of a specific point in a story. Using the QKS subgraph approach, we develop three hypothesized QKS subgraphs for our rebel behaviors. To validate them as appropriate representations for human subject evaluation, we identify Q-A pairs from the subgraph that will confirm the existence of the edges when they are essential and the absence when appropriate.

Betrayal

Our betrayal QKS subgraph in Figure 2 captures the concept of a shared plan between two agents (G1-E1) along with the event that through the causal link threat indicates a dropped goal (E2-G2) and resulting betrayal (E2-S1, E2-S2). To validate the subgraph as representative of a mental model after experiencing the plan-based betrayal, we would use the Q-A pairs in rows 1-6 in Table 1. Their are four Q-A pairs to confirm the existence of the aforementioned edges and two Q-A pairs to ensure that reader's separate each agent's goals into

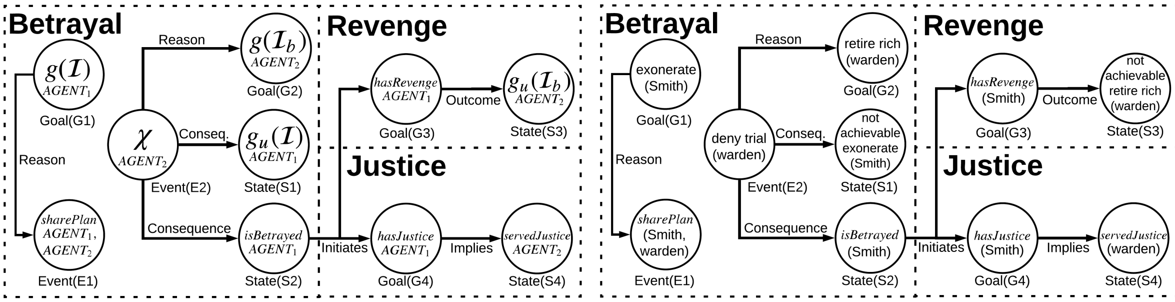

Figure 2: This figure contains three QKS subgraphs representing three rebellious behaviors from *Prison*.

Q-A	Type	GOA	Question	Answer	Question from Prison	Answer from Prison
				Betrayal		
E2-G2	Why	Good	Why did $Agent_2$ do χ?	Because $Agent_2$ wanted $g(I_b)$	Why did the warden deny Smith's trial?	Because the warden wanted to retire rich
E2-S1	Cons.	Good	What was a cons. of χ?	That $g_u(I)$	What was a con. of the warden denying the trial?	That Smith could not achieve exoneration
E2-S2	Cons.	Good	What was a cons. of χ?	That $Agent_1$ was betrayed	What was a cons. of the warden denying Smith's trial?	That Smith was betrayed
E2-G1	Why	Bad	Why did $Agent_2$ do χ?	Because $Agent_2$ wanted $g(I)$	Why did the warden deny Smith's trial?	Because he wanted to exonerate Smith
E2-E1	Why	Bad	Why did $Agent_2$ do χ?	Because $Agent_2$ shared a plan	Why did the warden deny Smith's trial?	Because he had a shared plan with Smith
E1-G1	Why	Good	Why did $Agent_{1,2}$ shared a plan?	Because $Agent_{1,2}$ wanted $g(I)$	Why did warden and Smith share a plan?	Because they wanted to exonerate Smith
				Revenge		
G3-S2	Why.	Good	Why did $Agent_1$ want revenge?	Because $Agent_1$ was betrayed	Why did Smith want revenge?	Because Smith was betrayed
G3-S3	Cons.	Good	What was a cons. of revenge?	That $g_u(I_b)$	What was a cons. of Smith's revenge?	That the warden could not retire rich
G3-G1	Why	bad	Why did $Agent_1$ want revenge?	Because $Agent_1$ wanted $g(I)$	Why did Smith want revenge?	Because he wanted to be exonerated
				Justice		
G4-S2	Why	Good	Why did $Agent_1$ want justice?	Because $Agent_1$ was betrayed	Why did Smith want justice?	Because Smith was betrayed
G4-S4	Cons.	Good	What was a cons. of justice?	That justice was served	What was a cons. of Smith's justice?	That the warden was served justice
G4-G1	Why	Bad	Why did $Agent_1$ want justice?	Because $Agent_1$ wanted $g(I)$	Why did Smith want justice?	Because he wanted to be exonerated

Table 1: Question-answer pairs for evaluating QKS subgraphs as representative of mental models from reading rebel behavior.

separate goal hierarchies, This separation means the agents' new goals are not subgoals of previous goals, implying the agent was required to deliberate.

Revenge

Our revenge QKS subgraph from Figure 2 requires only 3 Q-A pairs. Two pairs confirm that AGENT_2's betrayal (the warden) initiated AGENT_1 (Smith) adopting a revengeful goal (S2-G3) and that an outcome of pursuing this goal was AGENT_2 was unable to achieve their goal. A final Q-A pair assesses whether the revengeful goal was a subgoal of the shared plan goal between AGENT_1 and AGENT_2.

Justice

Similar to revenge, the justice QKS subgraph requires only 3 Q-A pairs. Two pairs confirm that AGENT_2's betrayal (the warden) initiated AGENT_1 (Smith) adopting a just goal (S2-G4) and that achieving this goal implies the warden received his justice. A third Q-A pair assesses whether the just goal was a subgoal of the original goal with a shared plan.

Conclusion

Rebel agents are both an important narrative plot device and arguably essential for true agent autonomy. Our approach has made first steps towards both these goals by defining computational models of rebellious behavior for plan-based agents. An integral part of our models is our use of the BDI agent control loop. Specifically, our model of betrayal meets the sufficient conditions for BDI agents to first *reconsider* their intentions while responding with revengeful or just behavior is part of the agent's deliberation over their *options*.

In addition to defining the models, we have also hypothesized their affect on comprehension. Using the QUEST cognitive model, we have proposed a QUEST knowledge structure subgraph to capture the intended effects for each of betrayal, revenge, and justice. The resulting Question-Answer pairs also provide a convenient mechanism to produce explanations of behavior. Lastly, we have characterized the models as part of existing rebel frameworks.

Our future work in the immediate term is to implement this and use the output in a human subject experiment to validate the proposed QKS subgraphs.

References

Aha, D. W., and Coman, A. 2017. The AI Rebellion: Changing the Narrative. In *Thirty-First AAAI Conference on Artificial Intelligence*, 4826–4830.

Amos-Binks, A., and Young, R. M. 2018. Plan-based Intention Revision. In *AAAI Conference on Artificial Intelligence*. New Orleans: AAAI.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Palo Alto: Center for the Study of Language and Information.

Briggs, G., and Scheutz, M. 2017. The Case for Robot Disobedience. *Scientific American* 316(1):44–47.

Cohen, P. R., and Levesque, H. J. 1990. Intention is Choice with Commitment. *Artificial Intelligence* 42(2-3):213–261.

Coman, A., and Muñoz-Avila, H. 2014. Motivation discrepancies for rebel agents: Towards a framework for case-based goal-driven autonomy for character believability. In *In Proceedings of the 22nd International Conference on Case-Based Reasoning (ICCBR)*.

Dannenhauer, D.; Floyd, M. W.; Magazzeni, D.; and Aha, D. W. 2018. Explaining Rebel Behavior in Goal Reasoning Agents. In *Workshop on Explainable Planning at ICAPS-18*.

Graesser, A. C.; Lang, K. L.; and Roberts, R. M. 1991. Question answering in the context of stories. *Journal of Experimental Psychology: General* 120(3):254–277.

Meehan, J. R. 1977. TALE-SPIN: An Interactive Program that Writes Stories. In *International Joint Conference on Artificial Intelligence*, volume 77.

Munoz-Avila, H.; Aha, D. W.; Jaidee, U.; Klenk, M.; and Molineaux, M. 2010. Applying Goal Driven Autonomy to a Team Shooter Game. In *Proceedings of the Twenty-Third Florida Artificial Intelligence Research Society Conference*. Daytona Beach, FL: AAAI Press.

Perez y Perez, R., and Sharples, M. 2001. MEXICA: A Computer Model of a Cognitive Account of Creative Writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Porteous, J.; Cavazza, M.; and Charles, F. 2010. Applying Planning to Interactive Storytelling: Narrative Control Using State Constraints. *Transactions on Intelligent Systems and Technology* 1(2):1–21.

Rao, A. S., and Georgeff, M. P. 1998. Decision procedures for BDI logics. *Journal of Logic and Computation* 8(3):21–26.

Riedl, M. O., and Young, R. M. 2010. Narrative Planning : Balancing Plot and Character. *Journal of Artificial Intelligence Research* 39:217–267.

Schank, R. C., and Abelson, R. P. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry Into Human Knowledge Structures.* Oxford, England: Lawrence Erlbaum.

Van Der Hoek, W.; Jamroga, W.; and Wooldridge, M. 2007. Towards a theory of intention revision. *Synthese* 155(2):265–290.