Explaining Rebel Behavior in Goal Reasoning Agents

Dustin Dannenhauer[1] and **Michael W. Floyd**[2] and **Daniele Magazzeni**[3] and **David W. Aha**[4]
[1]NRC Postdoctoral Fellow; NRL; Navy Center for Applied Research in AI; Washington, DC
[2]Knexus Research Corporation; Springfield, VA
[3]King's College London, Strand, London WC2R 2LS
[4]Naval Research Laboratory; Navy Center for Applied Research in AI; Washington DC

{dustin.dannenhauer.ctr, david.aha}@nrl.navy.mil, michael.floyd@knexusresearch.com, daniele.magazzeni@kcl.ac.uk

Abstract

Generating human-comprehensible explanations is an important requirement for autonomous systems in human-agent teaming environments. Humans and agents often have their own knowledge of the world, knowledge of objectives being pursued and tasks being performed, and their own constraints. Given these differences, an agent may be issued goals that violate its own constraints or preferences, or are undesirable for the team's task. Numerous situations may arise where rebellion by dropping or changing goals leads to a more beneficial outcome. Agents with goal reasoning capabilities may rebel by rejecting or altering the goals and plans expected of them by human teammates. Explanations help build trust and understanding between the human and agent, leading to greater overall effectiveness. In this paper we outline motivating examples for explainable rebellious behavior in goal reasoning systems and identify open research questions.

Introduction

In recent years there has been increased interest in autonomous agents capable of rebellion (Briggs and Scheutz 2016). Rebellious agents are agents that may reject, revise, or in some form protest a goal issued to them by another agent (including humans). This ability to rebel is necessary for any agent that receives goals from multiple sources (including self-provided goals) that may conflict. Consider service robots that give tours or deliver goods (e.g., a hotel room service robot); these agents should reject goals that could lead to any number of undesirable situations, such as those that damage the robot. Many positive reasons for rebellion have been described (Aha and Coman 2017; Briggs and Scheutz 2015), including:

- **Differential Information Access**: The agent may have access to information the human does not have. A simple example is when an agent is helping a human to move a box in a warehouse and, while carrying the object, the agent observes a harmful obstacle behind the human. The agent stops moving and informs the human, rebelling against the given goal of moving the box to the target destination.
- **Oversubscription**: The agent is tasked with goals from two teammates, and thus may reject goals from one of them. For example, an agent on a team is tasked by a supervisor with obtaining information (by mapping an environment and taking pictures and video) while other team members are responsible for moving obstacles to clear a path. When another agent asks the surveillance agent to help move an obstacle, the agent may reject that goal since it does not align with its current surveillance goal.
- **Ethical Conflict**: The agent has a conflict with the ethics of a given goal. For example, an agent may be asked to take a harmful action against another human, violating the agent's ethical code of not harming any humans. This may involve hard constraints, such as not taking any action that could have a harmful effect (Briggs and Scheutz 2015), or a preference for avoiding states with low ethical scores.
- **Impasse**: A provided goal may not be achievable due to resources (e.g., battery level) or obstacles.
- **Task Violation**: A provided goal may violate some global constraints or preferences of the agent, such as delivering a package to a destination outside the country or safe area. These constraints may be similar to ethics, but may be more task-specific.
- **Safety**: An autonomous vehicle may be given a goal to reach a destination in a short period of time. However, the plan to reach the destination could become dangerous. An autonomous car may need to violate traffic laws or drive off-road, or a drone may need to travel too fast to avoid obstacles (i.e., if flying in a forest or indoor environments such as in urban search and rescue settings).

While there are many motivations for positive rebellion, it makes a rebel agent less predictable to other interacting agents. Thus, any agent that can rebel should also be able to explain its behavior. Indeed, legal measures are being adopted to provide individuals affected by automated decision-making with a "right to explanation", as referred to in the recent EU General Data Protection Regulation (GDPR), in place from May 2018 (European Parliament and Council 2016).

The interpretability of AI systems has been a popular topic of workshops and related events since 2016, and in 2017 DARPA launched the Explainable AI (XAI) Program. Most of these efforts have focused on providing transparency to the decision making of machine learning (ML) systems in general, and deep networks more specifically[1].

[1]Exceptions, for example, include the broader intent of the

While XAI research on data-driven ML is well-motivated, AI Planning is well placed to address the challenges of transparency and explainability in a broad range of interactive AI systems. For example, research on Explainable Planning has focused on helping humans to understand a plan produced by the planner (e.g., (Sohrabi, Baier, and McIlraith 2011; Bidot et al. 2010)), on reconciling the models of agents and humans (e.g., (Chakraborti et al. 2017)), and on explaining why a particular action was chosen by a planner rather than a different one (e.g., (Smith 2012; Langley et al. 2017; Fox, Long, and Magazzeni 2017).

Most prior work on rebel agents considers how robots can avoid hard ethical constraints, such as actions that may harm self or others (Briggs and Scheutz 2015), or that violate a constraint in a task specification when working alongside a human (Gregg-Smith and Mayol-Cuevas 2015). These agents have rebelled against actions that have harmful effects. Instead, we are concerned with agents equipped with automated planners that may execute complex sequences of actions to achieve goals.

Since rebellious goal reasoning agents are concerned with decisions regarding which goals to pursue in addition to planning decisions, explanations will also need to consider goal reasoning decisions. Explaining such planning and goal related decisions is the focus of this paper. More specifically, we motivate why the intersection of rebel agents and explainable planning is an important research area, outline future problems for explainable rebellious planning and goal reasoning agents, and discuss several research questions that require attention.

Related Work

Briggs et al. (2015) present an approach to determine the speech directives a robot should utter when rebelling. They describe five felicity conditions that must be met before a robotic agent adopts a goal issued by a human. These conditions may not be sufficient for automated planning agents to accept or reject goals. For example, metrics are not considered, so there is no ability for an agent to reject a goal because it would be much more expensive/risky/unsafe than what the human may have assumed when issuing the goal.

To our knowledge, no prior work describes automated planning agents that can reject goals and explain their decision for rebellion. Research on goal reasoning (Vattam et al. 2013) considers agents that can perform operations on their goals, such as changing or dropping a goal via various operations (Cox, Dannenhauer, and Kondrakunta 2017; Cox 2016) and, in the formalism of the Goal Lifecycle, by applying goal refinement strategies (Roberts et al. 2016). However, in prior work these goal changes occur due to encountering unexpected external events or opportunistic situations, rather than rebelling against a provided goal. Cox and Veloso (1998) describe goal transformations that occur during plan generation in PRODIGY, where one of the transformations is the dropping of a goal. Although rebellious goal changes could be encoded in such an architecture (e.g., rebellion by dropping a goal could be considered a *retraction* operation as described in Table 1 in (Cox and Veloso 1998)), rebellious goal changes have not been considered or implemented.

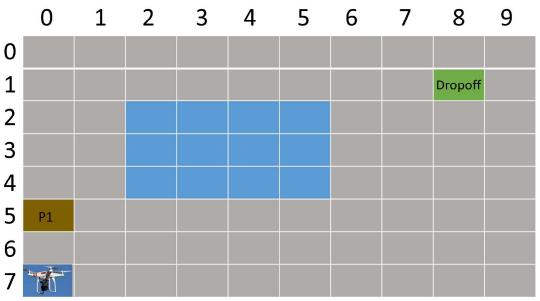

Figure 1: Top-down environment view for an autonomous delivery drone. The environment contains the drone (Row 7, Column 0), a package (Row 5, Column 0), a dropoff location (Row 1, Column 8), and a body of water (rectangular region between Row 2, Column 2 and Row 4, Column 5).

Real-world planning systems operate on large amounts of data and can generate plans that overwhelm a human operator, especially when they are subject to severe time constraints for decision making (e.g., whether to abort a mission). Thus, methods for explaining decisions (e.g., by comparing multiple plans) must generate explanations that can be quickly digested.

Motivating Example

We now provide a motivating example of a rebel agent operating in a dynamic and uncertain environment. Figure 1 displays a top-down view of a state in an environment in which an autonomous delivery drone operates (Row 7, Column 0). In this example, the agent controls the drone and we use the two terms interchangeably. This state contains a package (Row 5, Column 0), a dropoff location (Row 1, Column 8), and a large body of water (a rectangular region between Row 2, Column 2 and Row 4, Column 5). The drone's human operator provides it with goals and can communicate with it, but the operator is not physically located in the area depicted in Figure 1. For this example, we assume that the initial goal utterance provided by the operator is *"deliver the package to the dropoff location and return to your initial position"*.

In the simplest scenario, the drone would take the provided goal and generate the following plan (Figure 2): fly to the package, pick up the package, fly directly to the dropoff location, drop off the package, and fly directly to the initial location. In this version, there would be little need for explanation since the drone would achieve the specified goal using an expected plan (i.e., flying directly between all locations). Additionally, since no unexpected events or environment changes occur, the drone would have no new information to provide its operator.

Consider a variant of this scenario where the drone considers flying over water to be dangerous but the operator is

Workshops on XAI at IJCAI-17 and IJCAI-18

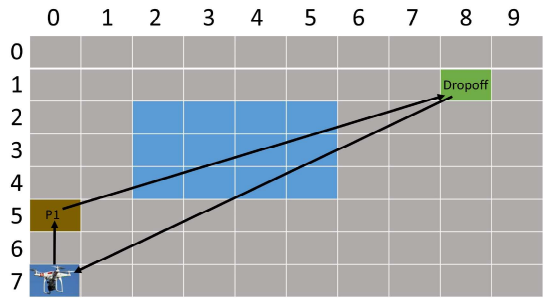

Figure 2: Visualization of drone's plan to fly directly to the package, pick up the package, fly directly to the dropoff, drop off the package, and fly directly to its initial position.

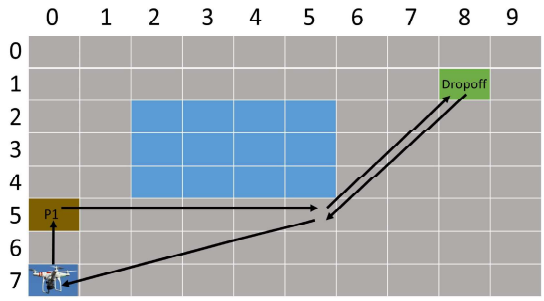

Figure 3: Variant of the drone's delivery plan that avoids flying over the water.

not aware of this water-avoidance preference. The plan generated by the drone will still deliver the package, but will take a slightly longer route that avoids the water (Figure 3). If no pre-execution communication occurred (e.g., making the operator aware of the preference or getting feedback on the generated plan), the drone would execute the plan without knowing that it was rebelling against the operator's desire for a plan that uses only direct flight routes.

Either during plan execution, if communication is possible, or during a post-mission debriefing, the operator may ask questions to improve their understanding of why the drone's plan differed from their expectations:

- **Operator**: *"Were you pursuing another goal?"*
  Agent: *"No, only the goal you provided me"*
- **Operator**: *"Was that the most efficient plan to achieve the goal?"*
  Agent: *"Yes, it was the most direct route I could take to deliver the package given my preferences"*
- **Operator**: *"Why didn't you just fly over the water'*
  Agent: *"Because flying over the water is too dangerous"*

Based on these questions and the provided explanations, the drone can convey that its act of rebellion was due to a divergent set of planning preferences, thereby leading to differences in the drone's generated plan and the plan that the operator expected to see executed.

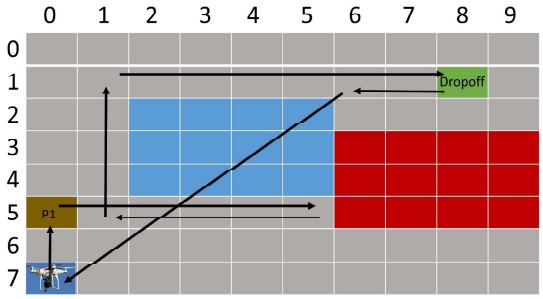

Figure 4: Variant of the drone's delivery plan that requires replanning when the drone detects the fire (rectangular region between Row 3, Column 6 and Row 5, Column 9) and when the drone detects it is low on fuel.

Consider a third variant where the operator is aware that the drone prefers to avoid flying over water. In this variant, the operator would expect the drone to fly around the water but, because they are not physically located in the environment, would be unaware of any environment changes that occur. As shown in Figure 4, when the drone reaches the water's edge it notices a large fire blocking its path to the dropoff location. This would cause the drone to replan to fly clockwise around the water. However, after delivering the package to the dropoff location, the drone realizes its fuel is lower than expected and its planner finds that the only feasible plan, due to the limited fuel, is to fly over the water.

To the operator, who is unaware of the fire, it may appear that the drone rebelled (against the expected plan or provided goal) or opportunistically modified its goals. For example, to an outside observer it would be reasonable to assume that the drone retrieved another package at the location where it detected the fire and delivered that package to a location near the top of the water body. As in the previous scenario, and most situations involving a rebel agent capable of goal reasoning, the operator may wish to query the agent about its goals and plans. For example, these questions, and possible associated explanations from the agent, include:

- **Operator**: *"Why did you stop flying towards the dropoff location and fly clockwise around the water?"*
  Agent: *"Because there was a large fire blocking my path"*
- **Operator**: *"Why did you return by flying over the water?"*
  Agent: *"Because I was low on fuel and my desire to return home outweighed my preference to avoid flying over water"*
- **Operator**: *"Why didn't you take a shorter path over the water by flying directly north after detecting the fire?*
  Agent: *"Because my original plan was to avoid the water, but I used more fuel than expected and needed to replan"*

As a final scenario, consider when the drone executes its initial plan but an unexpected opportunity presents itself. Figure 5 shows that the drone, while attempting to deliver the package, observes a serious car accident. The drone may

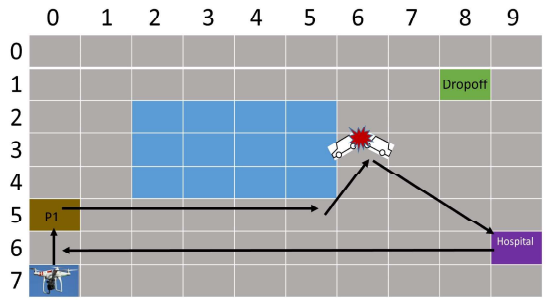

Figure 5: Variant of the drone's plan when it encounters a car crash (Row 3, Column 6) and generates a goal to assist the victims by airlifting them to the hospital (Row 6, Column 9).

have a strong internal motivation to help preserve life whenever possible, so it would rebel against its delivery goal, abandoning or suspending it, and instead formulate a new goal to assist the crash victims. The drone would generate a plan to achieve its new goal, and perform actions to airlift the car's driver to a nearby hospital[2]. Its new plan would involve airlifting the car's driver to a nearby hospital. After completing that plan and achieving the goal, the extra fuel drain from carrying a human would make the delivery goal impossible, so the drone would return to its initial position. The operator may have the following questions for the drone:

- **Operator**: *"Were you pursuing another goal?"*
  Agent: *"Yes, I formulated a new goal to assist a car crash victim"*
- **Operator**: *"Why did you help the victim instead of completing the delivery?"*
  Agent: *"I am programmed to prevent the loss of human life and the injuries looked serious."*
- **Operator**: *"Why didn't you complete the delivery afterwards?'*
  Agent: *"Because the weight of the victim drained my fuel faster than normal; I was unable to generate a plan to complete the delivery given my fuel level"*

As this motivating example demonstrates, when an agent operates in a sufficiently complex environment where it interacts with a teammate, many opportunities/needs for explanation arise. These increase when the agent and its teammate may have different (or unknown) planning preferences (e.g., water-avoidance), methods to evaluate plan quality (e.g., plan duration vs. plan safety), or observable information (e.g., an agent located in the environment vs. an operator externally located). Additionally, the knowledge that an agent can rebel or modify its own goals may make an operator more likely to assume goal changes or rebellion are the cause of unexpected behavior. Thus, the agent needs to answer such questions by explaining its plans and goals, as well as any additional information related to the reason it behaved unexpectedly.

We simplified the presentation and discussion of our examples by including assumptions, namely a lack of online communication or time-sensitive tasks preventing rapid communication, that restricted questions and explanations to occur post-run. However, the need for explanation also exists during the course of operation. For example, upon seeing a real-time report of the drone's GPS location, the operator may ask questions about why the drone appears to be deviating from expectations. Similarly, the drone could realize that its changing behavior is unexpected and provide proactive explanations that it believes will provide necessary context to the operator. Although such real-time questions and explanations could lead to unnecessary supervision of the agent (e.g., questioning every minor variance from expectations, even if the differences have no impact on overall success), they also provide the ability to incrementally correct misunderstandings and prevent the operator from becoming overwhelmed by how much the agent's behavior deviates from expectations. This is important in situations where disuse is possible, since a human may just label a robot as untrustworthy and stop using it rather than ask the tens or hundreds of post-run questions necessary to understand its behavior.

Explaining Goal Reasoning Decisions

The field of goal reasoning has seen at least three general frameworks emerge: Goal-Driven Autonomy (GDA) (Molineaux, Klenk, and Aha 2010), the Goal Lifecycle (Roberts et al. 2016), and Goal Operations and Transformations (Cox, Dannenhauer, and Kondrakunta 2017; Cox and Veloso 1998). Explainable goal reasoning is an open problem which we hope to motivate others in the community to pursue, and has been identified as an important factor for goal reasoning agents that are members of human-robot teams (Molineaux et al. 2018). Although there have been recent examples of goal reasoning agents being deployed as members of human-agent teams in complex domains (Floyd et al. 2017; Gillespie et al. 2015), these agents do not explicitly explain their behavior to human teammates. We highlight the challenges and issues that may arise to create explainable goal reasoning agents in each of these frameworks.

Explaining Goal-Driven Autonomy

- Explainable GDA agents will likely require not only a general trace, but also explanations of components that affect the pursuit of goals besides planning, including: motivator functions (Coddington et al. 2005; Muñoz-Avila, Wilson, and Aha 2015), discrepancy detection (Dannenhauer, Munoz-Avila, and Cox 2016; Karneeb et al. 2018), internal explanation (Molineaux, Kuter, and Klenk 2012), and goal selection. A goal may be abandoned because discrepancy detection (using informed expectations or goal regression expectations) notices an anomaly. A logical follow-up question is *"what was the anomaly?"* and *"why is the anomaly important?"*. When it comes to explaining discrepancy detection and why an action failed,

[2]This example assumes that a drone exists that could perform such an airlift maneuver. This may be unreasonable for current-generation drones, but is used for illustrative purposes.

and why that action is important (especially if using goal regression expectations) we may want to use a dependency graph (Ayan et al. 2007) to demonstrate that a certain part of our goal cannot be reached if this action is not performed.

- Since a human may ask why an agent formulated a new goal g', the agent may need to backtrack to the internal explanation component (e.g., that the explanation is that there is an obstacle), and then backtrack to the anomaly detection (e.g., that action 7 of the current plan failed), and then backtrack to the planning system to say *"I (the agent) chose this plan instead of others because it was more efficient"*.

Explaining the Goal Lifecycle

- The Goal Lifecycle is built on Goal-Task Network (GTN) Planning and an open question remains about how to generate explanations from GTNs. Another open question is whether the questions asked about GTNs differ from those considers by existing work of explainable planning (Fox, Long, and Magazzeni 2017).
- As goals are refined toward completion or backtracking for later processing, a number of metrics (either domain-dependent or domain-independent, like inertia) are maintained. An open question is how to use these metrics in explanations? Another open question is whether the information maintained by the Goal Lifecycle during goal refinement is enough to handle all the questions a user would ask the system. If not, what other information would be needed to answer these questions and how would the formalism change to support this?

Explaining Goal Operations and Transformations

- Performing goal operations and transformations to resolve an unachievable goal to an achievable goal is an open problem. Since goal operations are formulated similarly to planning operations on world states, the trace of goal operations could be stored. Only some goal operations have been formalized completely into operators, and they make use of an ontology. Explanations of why the agent chose to pursue goal g' when the current goal g is unachievable or undesirable may rely on explaining relationships between goal predicates that are in the ontology. The example from (Cox, Dannenhauer, and Kondrakunta 2017), where the goal of *on(B,A)* is generated when *stable-on(B,A)* is unachievable, has to do with the fact that the *on* predicate is closely related to the *stable-on* predicate in the goal predicate ontology.

Common Themes for Explainable GR Agents

- A trace of behavior will be needed to rewind decision making in all three of these frameworks. The formalism of the Goal Lifecycle contains some trace information for explanations and the MIDCA cognitive architecture from (Cox, Dannenhauer, and Kondrakunta 2017) maintains a trace of the cognitive level decision making which includes some goal reasoning decision making: goal selection, goal-related discrepancy detection, and internal explanation. However, some goal operations, such as goal transformation (one method of goal formulation) could occur at the metacognitive level, thus warranting explanations of the meta-level processes. Therefore, a trace of the metalevel could identify the meta-level reasoning processes performed by the agent and provide metalevel explanations.

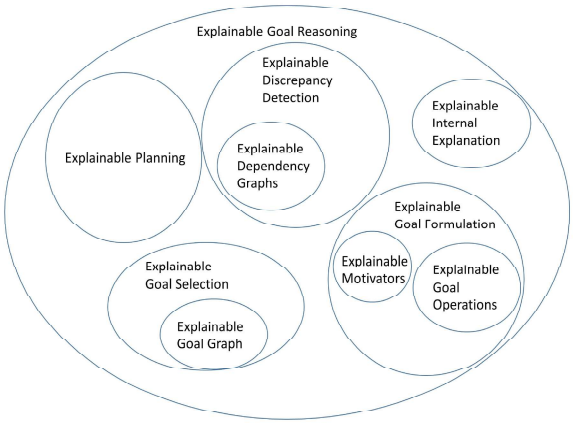

Figure 6: Multi-level overview of goal reasoning processes.

- Complex situations will call for human operators to desire multi-level explanation mechanisms that at the top level answer questions by first identifying which goal reasoning mechanism(s) were involved and then exploring the details of those specific mechanisms. The space of explanations for goal reasoning systems is likely much larger than planning systems without goal reasoning components. A preliminary multi-level overview of goal reasoning components is presented in Figure 6.

  The larger bubbles in Figure 6 represent more general explanation requests, and smaller bubbles represent more narrow requests. In attempting to learn why an agent abandoned a goal, the operator may first ask a high-level question of: *"What caused you to abandon your goal?"*. The agent may give the high-level answer: *"Because I detected an unresolvable discrepancy"*. To answer this question, the agent may first start within the largest bubble, *Explainable Goal Reasoning*, and answer with a response tied to a bubble within that bubble: *Explainable Discrepancy Detection*. When the operator asks a follow up question *"Why was the discrepancy unresolvable?"* the agent may first identify this question falls within the *Explainable Discrepancy Detection* bubble and generate an answer related to the inner bubble *Explainable Dependency Graphs*: *"I could not achieve my goal because I could not execute a specific action that achieves a specific goal condition"*. This example illustrates that explanations may be relevant to different goal reasoning processes. Before answering a question effectively, an agent may need to identify which goal reasoning components are relevant to the question.

Discussion

In the case where the agent rejects a goal, the agent needs to show (to the extent that is possible) that it could not find any way to achieve that goal without violating one of the conditions. And this is where explainable planning, and in particular the use of planning, comes into play. Given that, in general, it is not feasible to formally prove that there are no plans, the agent and the user can use the explainable planning framework (XAIP) (Fox, Long, and Magazzeni 2017) to provide a justification/understanding of why the agent rebelled.

The XAIP framework in this context can be used in two ways:

1. The rebel agent uses the XAIP framework itself, by questioning the current plans (that is not valid) and exploring alternative plans.
2. The agent and the human (or the other agent) use XAIP in cooperation, where the human questions the initial plan (trying to find feasible ways to achieve the goal).

In both scenarios, we envision XAIP being used both during and after scenarios (i.e., online vs. post-run debriefing).

Open Questions for Explainable Rebel Agents

1. Explainable rebellion: How can rebel agents explain why the rejected actions violated their ethical models (e.g., for example rejecting actions that cause harm)? Also, how can rebel agent explain why they chose one goal over another when using a metric-based evaluation (e.g., this may involve reporting that states needed to reach the goal or goal states themselves have low ethical scores)?
2. While there has been some work on implementing goal operations in various frameworks, an agent with all possible goal operations and goal strategies has not yet been realized. Two open problems are: *How can each goal operation / strategy be explained?* and in the case when multiple goal reasoning operations / strategies have been applied *How can multiple explanations be composed in a digestible way?*
3. Will explainable goal reasoning agents in complex domains with complex reasoning components run into potential space problems? Might some questions warrant explanations that are just too expensive to compute? For example, suppose an agent has executed hundreds or thousands of actions since the last time the operator interacted with the agent to give it a goal. During that time, the agent's goal may have changed multiple times due to anomalies (e.g, the goal was suspended, another goal chosen, then later the original goal was re-adopted but the old plan was invalid so agent did re-planning). Whenever the agent changes goals or plans, how many alternative plans should it store to provide evidence of its decisions? Or should the agent save computational and storage time by only computing information needed for explanations when necessary (e.g., when a question is asked)? How does the tradeoff between maintaining complete explanation knowledge and lazy on-demand explanation generation impact the amount of time necessary to generate explanations (i.e., precomputed explanations vs. dynamically generated explanations)? Does the particular domain impact these choices (e.g., the agent's available onboard hardware resources, time-sensitive nature of explanations, expected explanation needs of a particular user)?

Conclusion

An agent's behavior may be the result of changes to both the agent's goals and plans. Thus, an agent that can modify its plans or goals, and also reject plans or goals from teammates, requires increasingly sophisticated methods of explanation. In this paper, we have discussed why explanation is an important factor for rebellious goal reasoning agents and why explainable planning is a key consideration for such agents. In addition to a motivating example illustrating the potential explanations that may be required of such an agent, we have also examined the explanation needs of existing goal reasoning frameworks and many of the open questions that remain. As goal reasoning agents are deploying in increasing complex domains as part of human-agent teams, we hope this discussion will motivate the development of explanation capabilities in goal reasoning agents.

References

Aha, D. W., and Coman, A. 2017. The AI Rebellion: Changing the Narrative. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 4826–4830.

Ayan, N. F.; Kuter, U.; Yaman, F.; and Goldman, R. P. 2007. Hotride: Hierarchical ordered task replanning in dynamic environments. In *Planning and Plan Execution for Real-World Systems–Principles and Practices for Planning in Execution: Papers from the ICAPS Workshop*.

Bidot, J.; Biundo, S.; Heinroth, T.; Minker, W.; Nothdurft, F.; and Schattenberg, B. 2010. Verbal plan explanations for hybrid planning. In *Proceedings of Multikonferenz Wirtschaftsinformatik*, 2309–2320.

Briggs, G., and Scheutz, M. 2015. "Sorry, I Can't Do That": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In *Proceedings of the AAAI Fall Symposium on AI for Human-Robot Interaction*, 32–36. AAAI Press.

Briggs, G., and Scheutz, M. 2016. The Case for Robot Disobedience. *Scientific American* 316(1):44–47.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 156–163.

Coddington, A.; Fox, M.; Gough, J.; Long, D.; and Serina, I. 2005. MADbot: A motivated and goal directed robot. In *Proceedings of the 20th National Conference on Artificial Intelligence*, 1680–1681. AAAI Press.

Cox, M., and Veloso, M. 1998. Goal transformations in continuous planning. In *Proceedings of the AAAI Fall Symposium on Distributed Continual Planning*, 23–30. AAAI Press.

Cox, M. T.; Dannenhauer, D.; and Kondrakunta, S. 2017. Goal operations for cognitive systems. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 4385–4391. AAAI Press.

Cox, M. T. 2016. A model of planning, action, and interpretation with goal reasoning. In *Proceedings of the 4th Annual Conference on Advances in Cognitive Systems*, 48–63. Cognitive Systems Foundation.

Dannenhauer, D.; Munoz-Avila, H.; and Cox, M. T. 2016. Informed expectations to guide GDA agents in partially observable environments. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2493–2499.

European Parliament and Council. 2016. Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119:1–88.

Floyd, M. W.; Karneeb, J.; Moore, P.; and Aha, D. W. 2017. A goal reasoning agent for controlling UAVs in beyond-visual-range air combat. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4714–4721. AAAI Press.

Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable planning. In *Explainable Artificial Intelligence: Papers from the IJCAI Workshop*.

Gillespie, K.; Molineaux, M.; Floyd, M. W.; Vattam, S. S.; and Aha, D. W. 2015. Goal reasoning for an autonomous squad member. In *Goal Reasoning: Papers from the ACS Workshop*, 52–67.

Gregg-Smith, A., and Mayol-Cuevas, W. W. 2015. The design and evaluation of a cooperative handheld robot. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 1968–1975. IEEE.

Karneeb, J.; Floyd, M. W.; Moore, P.; and Aha, D. W. 2018. Distributed discrepancy detection for beyond-visual-range air combat. *AI Communications* 31(2):181–195.

Langley, P.; Meadows, B.; Sridharan, M.; and Choi, D. 2017. Explainable agency for intelligent autonomous systems. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 4762–4764. AAAI Press.

Molineaux, M.; Floyd, M. W.; Dannenhauer, D.; and Aha, D. W. 2018. Human-agent teaming as a common problem for goal reasoning. In *Proceedings of the AAAI Spring Symposium on Integrating Representation, Reasoning, Learning, and Execution for Goal Directed Autonomy*. AAAI Press.

Molineaux, M.; Klenk, M.; and Aha, D. W. 2010. Goal-driven autonomy in a Navy strategy simulation. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*. AAAI Press.

Molineaux, M.; Kuter, U.; and Klenk, M. 2012. DiscoverHistory: Understanding the past in planning and execution. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multi-Agent Systems*, 989–996. IFAAMAS.

Muñoz-Avila, H.; Wilson, M. A.; and Aha, D. W. 2015. Guiding the ass with goal motivation weights. In *Goal Reasoning: Papers from the ACS Workshop*, 133–145.

Roberts, M.; Shivashankar, V.; Alford, R.; Leece, M.; Gupta, S.; and Aha, D. 2016. Goal reasoning, planning, and acting with ActorSim, the actor simulator. In *Poster Proceedings of the 4th Annual Conference on Advances in Cognitive Systems*.

Smith, D. E. 2012. Planning as an iterative process. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2180–2185. AAAI Press.

Sohrabi, S.; Baier, J. A.; and McIlraith, S. A. 2011. Preferred explanations: Theory and generation via planning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. AAAI Press.

Vattam, S.; Klenk, M.; Molineaux, M.; and Aha, D. W. 2013. Breadth of approaches to goal reasoning: A research survey. In *Goal Reasoning: Papers from the ACS Workshop*.