The Ideal Rebellion: Maximizing Task Performance in Rebel Agents

James Boggs[1], **Dustin Dannenhauer**[2], **Michael W. Floyd**[3], **David Aha**[4]
[1] Union College, Schenectady, NY, USA
[2] NRC Postdoctoral Fellow, NRL, Navy Center for Applied Research in AI Washington, D.C., USA
[3] Knexus Research Corporation, Springfield, VA, USA
[4] Naval Research Laboratory, Navy Center for Applied Research in AI Washington, D.C., USA
boggsj@union.edu, {dustin.dannenhauer.ctr, david.aha}@nrl.navy.mil,
michael.floyd@knexusresearch.com

Abstract

The capability for an agent to rebel is important when the agent has information or motivations that differ from its teammates and/or supervisor. Prior work on AI rebellion has considered how agents reject actions that may lead to harmful or undesirable consequences. In this work, we examine agents that seek to maximize the achievement of goals when the goal states may contain both desirable and undesirable effects. We implemented a Metacognitive Integrated Dual-Cycle Architecture (MIDCA) agent that may reject goals with undesirable effects, and identify the ideal amount of rebellion they can perform given their teammate's resolve. We show empirical results in a domain where the agent controls a drone that must remove invasive plants while protecting endangered plants. Our results indicate that agents operating within the ideal rate of rebellion achieve higher overall scores for the given metric.

1 Introduction

The topic of AI rebellion is garnering increased attention as a necessary and beneficial role in human-robot interactions and human-agent teaming [Briggs and Scheutz, 2016; Aha and Coman, 2017]. An agent may need to rebel if:

1. It has access to information the human (or other agents) do not have
2. It has been tasked with a different mission than its teammates or supervisors
3. It has been subjected to goal requests from agents (including humans) that seek to do harm

In (1) consider the example of a robot that is helping a human carry a large package in a warehouse. The human is walking backwards and the robot sees an obstacle behind the human that could cause the human harm. The robot rebels by stopping and informing the human of the danger, even though its current goal is to get the box to its destination. In (2), an example scenario is one of an agent that is part of a team, but the agent is given a goal to *collect information* while the other teammates are given goals to *repair a building*. The agent may be asked by its teammates to help with the repair, which could conflict with its own mission, and thus may require the agent to rebel by rejecting the request. In (3) a human could give morally wrong goals to a robot, such as harming another human, and the robot should rebel since it would violate its own ethical code. For a further elaboration on the benefits of robot rebellion we refer the reader to [Briggs and Scheutz, 2016; Aha and Coman, 2017].

Research in rebel agents must address two primary questions: *when to rebel* and *how to rebel*. We focus on the *when to rebel* question, and consider agents carrying out tasks in a multi-agent simulated domain where many of the goals have both positive and negative effects. Most prior implementations of rebel agents focus on ensuring that a harmful act never occurs (discussed in Section 2), whereas in this work we consider an environment where goals may have both harmful and beneficial effects. Thus, our agent should consider the positive and negative implications of its actions and rebel in order to maximize its overall performance. We hypothesize that given a metric evaluation of a task, there exists a maximal rate for which rebellion should occur – *an ideal rebellion*.

We present a rebel agent implemented using the Metacognitive Integrated Dual-Cycle Architecture (MIDCA) [Cox *et al.*, 2016]. MIDCA is an open source architecture that provides a modular structure and an explicit focus on both a cognitive layer and metacognitive layer. Both the supervisor agents that issue goals and the rebel agents are MIDCA agents, and we describe the implementation in more detail in Section 3.

The contributions of this work include:

- An extension to the MIDCA architecture for the implementation of rebellious agents.
- Empirical results showing the improvement from rebellion under different metrics.
- A simulated multi-agent drone domain that allows for studies in agent rebellion.

The remainder of the paper describes our rebel agent and how rebellion can improve performance. Section 2 discusses related work on rebel agents and machine ethics. Section 3 discusses the design of our rebellious agents, including the

framework for rebellion they operate under. Section 4 introduces the simulated drone domain, Section 5 presents our experimental setup, and Section 6 reports our results. Finally, Section 7 provides conclusions and describes future work.

2 Related Work

We begin our discussion of related work by describing our agent using the rebellion framework put forth by Aha and Coman [2017] ,and follow with a comparison of our agent's rebellion to prior literature of other rebel agents and a discussion of machine ethics. See [Aha and Coman, 2017] for a more extended view of the literature.

2.1 Classification Under the Rebellion Framework

Aha and Coman [2017] put forth a framework to enable discussion, implementation, and deployment of positive rebel agents. Positive rebellion refers to rebellious behavior carried out in support of ethics, safety, self-actualization, solidarity, and social justice. Their framework describes four stages of rebellion, of which our agent implements three: rebellion deliberation, rebellion execution, and post-rebellion. We do not implement the pre-rebellion phase due to the reactive, not proactive, nature of our agents. Our agent *deliberates* on whether to rebel by rolling a weighted die, where the weight factor is the probability of agent rebelliousness. Our agent *executes* rebellion by informing the operator it would like to rebel, and the operator may allow or reject the agent's proposed rebellion. If the operator allows the rebellion, the rebellion episode ends and the agent does not pursue the goal and, as such, no *post-rebellion* occurs. However, if the operator rejects the agent's rebellion, the agent then enters the *post-rebellion* phase and may decide to rebel, ultimately having the final say in the matter, which then concludes the rebellion episode. Rebellion deliberation begins when the operator issues a goal to an agent. The rebellion described here does not consider emotion.

Aha and Coman [2017] define an *interactor* as the entity that is being rebelled against. They classify rebellion types along three dimensions: *expression*, *focus*, and *interaction initiation*. Briefly, *expression* may be either explicit in that rebellion is visible or implicit where rebellion is maintained as an internal disagreement within the agent; *focus* is concerned with whether rebellion will seek to change the behavior of the agent itself or another agent; *interaction initiation* may be reactive (rebellion happens after interactor issues a goal) or proactive (rebellion is initiated from the agent). Our agent can be classified using this rebellion framework as (1) explicitly expressing its rebellion, (2) having an inward-oriented focus of the subtype non-compliance, and (3) having an interaction initiation that is reactive. Our agent would be classified under the first dimension as demonstrating explicit rebellion because it communicates to the interactor (i.e., by telling an operator it wishes to drop a goal) and changes its behavior (i.e., does not take actions towards the goal given by the interactor). Along the second dimension, our agent's rebellion is inward-oriented since it is concerned with changing its own behavior rather than trying to change the interactor's behavior. Along the third dimension, our agent's rebellion is reactive since rebellion occurs only after the interactor gives the agent a goal.

2.2 Comparison to Previous Approaches of Agent Rebellion

Briggs and Scheutz [2015] outline scenarios and mechanisms that determine both *when* and *how* robots should reject inappropriate directives from humans in the form of speech. Prior to their work, most autonomous systems rejected directives based on two reasons: lack of knowledge or lack of ability. Their work argues for more general rejection capabilities and they describe a process using the DIARC/ADE cognitive robotic architecture [Schermerhorn *et al.*, 2006]. They introduce five types of felicity conditions and provide a high level outline of components needed to appropriately reject directives based on which felicity condition is violated. The work then elaborates on the felicity conditions regarding obligation and permissibility. They provide two formulas describing when an agent is obligated to adopt a goal considering two types of roles held by the human: supervisor and administrator. Regarding permissibility, they give a formula stating that for any goal which is unsafe it is permissible for all agents to reject that goal. The property of *unsafe* is formulated in such a way that a goal is unsafe if there exists an effect of the goal where any agent is harmed. The paper then goes into specific examples regarding directives given to a NAO robot asking it to walk off a ledge and walk into an obstacle. In the two examples, after the robot rejects the initial directive, the human then supplies an exception that enables the robot to adopt the goal without rejection when asked a second time. In the walking off the ledge example, the human tells the robot that he will catch the robot; in the obstacle example, the human tells the robot that the obstacle is not solid.

The rebellious behavior in our work is a more nuanced type of rebellion that considers maximizing a cost-benefits analysis of behavior rather than avoiding any negative outcomes. In the situations presented by Briggs and Scheutz [2015], in all cases there is a clear boundary for when rejection should happen (i.e., the goal would have any effect that could harm an agent). The goals in our work have effects that may be both positive and negative, with most goals having at least one negative effect. Some goals may have a greater value from the positive effects than negative, and in those circumstances the agent may adopt the goal, especially taking into account the supervisor's priority of the goal. At a high-level, our work is more focused on the agent's rejection decision of a goal (using the cost-benefits analysis) and less on explaining that rejection to the supervisor. An area of future work could be identifying how to best explain a rejection of a goal when reasons for rejection involve both positive and negative effects, and rejection simply for the sake of avoiding any negative effects is a worse overall result because of the lack of valuable positive effects. In many situations, there may not exist an achievable goal without negative effects.

Gregg-Smith and Mayol-Cuevas [2015] describe a robotic agent in the form of a hand-held tool that has task-specification knowledge and can refuse to perform an action that conflicts with the task specification. An example taken from their tiling task involves a task specification of pick-

ing colored tiles from bins and placing them in a specified pattern. The hand-held robotic agent can refuse to pickup or place tiles. If the human user attempts to place a tile in the wrong location, the tool will refuse to drop the tile. This rebellion situation differs from our work because there is a global shared goal (i.e., placing the tiles in a pre-specified way). The human may try to place a tile that would violate the global shared goal and thus the hand-held robot agent would rebel. In our work, the agent does not have knowledge of a shared global goal, and instead only receives goals from the human interactor.

Unlike Briggs and Scheutz [2015] and Gregg-Smith and Mayol-Cuevas [2015], we are trying to achieve a high ethical score rather than prevent violation of any negative ethical situation or a situation that violates a shared task specification.

2.3 Machine Ethics

Bringsjord *et al.* [2006] propose a method of ensuring the ethical behavior of artificially intelligent actors based on formal logical proofs. They propose utilizing a modified form of first order logic called deontic logic, which gives a formal logical language for describing and proving ethical statements. Specifically, they suggest an axiomization of deontic logic with an eye towards multiagent systems produced by Yuko Murakami, which they call Murakami-axiomatized deontic logic (MADL). In addition to standard first-order operators, MADL includes three new operators: $\bigcirc P$ as "ought to be the case that P", $\ominus_\alpha P$ as "agent α ought to see to it that P," and $\Delta_\alpha P$ as "agent P sees to it that P." These allow for humans to encode an ethical system alongside a standard logical encoding of some domain, which the robot can use to rigorously prove the moral permissibility, obligation, or prohibition of an action. The method proposed offers an excellent jumping-off point for exploration into ethical rebellion in autonomous agents, because it provides a clear framework for how an agent might make ethical decisions. In particular, using this framework an agent can *provably* demonstrate a command or decision is ethically invalid, and can bring this proof directly to its operator. Agent rebellion also complements the method put forward by Bringsjord *et al.* [2006], who note that one challenge facing the use of such an ethical framework is situations in which a human operator causes a morally impermissible situation. In our work we do not consider explicit ethics, however because our agent is trying to achieve the overall maximum number of positive effects (measured via the given metric), it can be considered a utilitarian agent (some goals are adopted even if they have a negative effect).

2.4 Rebellion and Goal Reasoning

Rebel agents that operate on the notions of goals (as opposed to exclusively on actions) are naturally goal reasoning agents. Goal reasoning agents have explicit goal structures and perform operations on these goals such as selecting which goal to pursue, formulating new goals, planning to achieve goals, and others [Vattam *et al.*, 2013]. Prior work on goal operations include a formalism of goal operations as they relate to planning, action, and interpretation [Cox, 2016], PDDL-like operators that can be applied to goals [Cox *et al.*, 2017], and strategies that modify goal nodes within the goal lifecycle

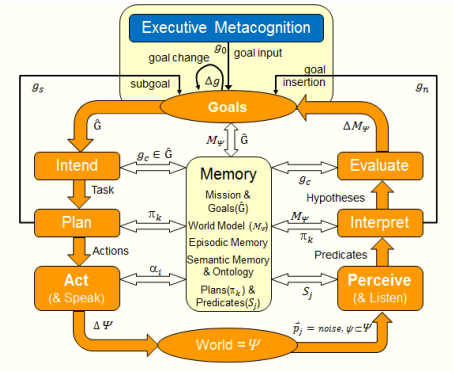


Figure 1: The Cognitive Layer of the MIDCA Architecture [Cox *et al.*, 2016]

[Roberts *et al.*, 2016]. Since rebel agents protest goals given to them, reasoning about their goals is important to determining alternative courses of action.

Rebel agents motivate research on goal transformation, provided the assumption that there is a better course of action for the agent rather than than doing nothing. Consider the examples from the introduction of when an agent may need to rebel. In (1), the agent should probably generate the goal to be in a state in which the human is aware of the obstacle behind them; in (2), the agent should consider opportunities to accomplish both its teammates' goals and its own goals if possible; in (3), the rebel agent may be able to assist the human without violating its own ethics (e.g., if the human interactor was in danger, instead of achieving the requested goal of *incapacitated(attacker)*, the agent could formulate the related goal *safe(interactor)*). We speculate that development of more complex rebel agents will require increasingly sophisticated goal reasoning mechanisms.

Rebellious goal reasoning agents have been explored in the conext of human-agent teaming, and specifically what impact rebelion would have on such a team [Molineaux *et al.*, 2018]. However, this work has been restricted to how rebellious agents can be modeled using a human-agent teaming model, how a rebellious agent would evaluate team performance, and how explanations factor into rebellion. Unlike the work presented in this paper, no implementation of a rebellious agent was performed and the impact of rebellious behavior was not evaluated.

2.5 Rebellion in MIDCA

The cognitive layer of the Metacognitive Integrated Dual-Cycle Architecture (MIDCA) is shown in Figure 1. The cognitive layer of MIDCA operates in cycles composed of six phases: Perceive, Interpret, Evaluate, Intend, Plan, and Act. Perceive contains procedures for obtaining direct perceptual data from the environment. Interpret handles a number of processes including: transforming raw data into an internal state, identifying unexpected situations that have occurred in the world, explaining what may have caused discrepancies between the agents expectations and observations, and generating new goals. Evaluate is responsible for tracking progress made on the agent's current goals, including dropping goals

that have been achieved or have failed. On the left side of Figure 1, Intend decides the current goals that the agent should be pursuing, Plan generates the sequence of actions needed to reach the agent's goals, and finally Act executes the behaviors needed to carry out the plan actions. A cycle similar to the cognitive layer occurs at the meta cognitive layer (not shown), except instead of perceiving and acting in an environment, perception (referred to as monitoring) occurs on the agent's cognitive layer and action is taken to change the cognitive layer or internal memory instead of changes in the world environment. The new rebellion related processes in this work occur at the cognitive layer in the Interpret and Evaluate phases, and is described in more detail in the following section.

3 Framework for Rebellion

The general flow of an agent rebellion in our framework can be seen in Figure 2. Broadly speaking, given a set of logical atoms about the world W and a set of goals G, the agent interprets the goals in light of the world state it perceives, and then evaluates each goal. If a goal is impossible, it is removed immediately, while if a goal is acceptable, the agent begins to plan and act in accordance with the goal. If the goal is possible but undesirable, the agent rebels. It's first step is clarifying internally why it is rebelling by identifying why the goal is undesirable and how the goal can be changed to become acceptable. It then informs its operator of its reasons for rebellion, as well as any possible changes to the goal it came up with. The operator then has the chance to weigh in, either by selecting an alternative goal or by rejecting the rebellion and telling the robot to continue with the current goal. If the operator accepts the goal, the agent plans and acts accordingly. However, as the actor in the field, the agent has the final say if an operator rejects its rebellion and can choose whether to acquiesce or not. Depending on its final decision, the agent will either remove the goal or plan to carry out the goal.

More specifically, the rebellion process begins in an agent's Interpret phase when a goal which has been marked as invalid is investigated and found to be possible to achieve but undesirable. The goal is then marked in the agent's memory as one to rebel against. In the Evaluate phase, the goals which have been marked as requiring rebellion are noted and the rebellion process for each specific goal is started by storing information about the situation in the agent's memory. This includes the goal itself, the reason for the rebellion (e.g., native plants will be killed), the operator which assigned the goal, information about the rebellion (e.g., which native plants are threatened), the identity of the rebelling agent, and possible alternative goals which would not lead to rebellion. This last piece of information allows the agent to autonomously generate acceptable goals and give them to the operator as suggestions for a replacement goal. That is, the agent is not merely rebelling by rejection, but making useful suggestions.

For each rebellion, the agent sends a message indicating that it is rebelling against the given goal, explaining why, and offering alternative goals, if it could generate any. The operator then chooses whether to allow the rebellion by selecting an alternative goal or insisting that the agent pursue the initial goal. In the current implementation, this is done by rolling a weighted die. If the operator accepts the rebellion, it messages the agent its acceptance and its choice from among the alternative goals the agent generated, then stores the goal-agent pair in its memory so that it does not give that agent the same goal again.

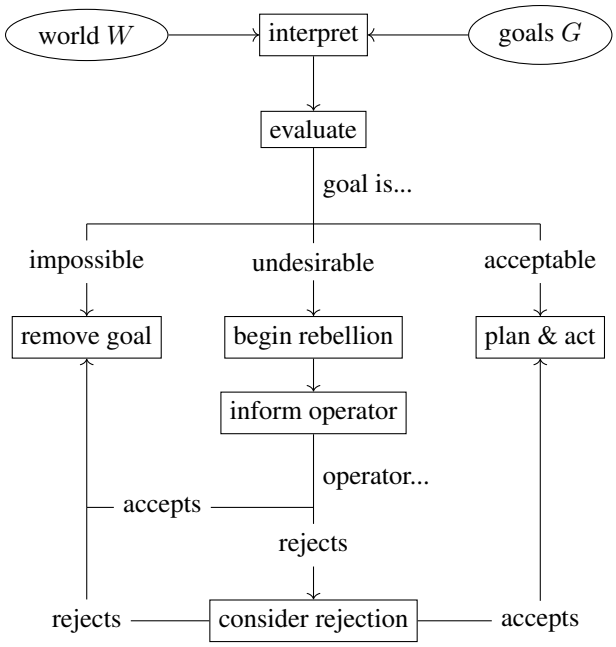


Figure 2: The flow of an agent's rebellion in our framework.

If the operator rejects the rebellion, it messages the agent informing it of this. The agent then can choose to acquiesce to this rejection or to persist in its rebellion. This is also done by rolling a weighted die. If the agent chooses to comply, it proceeds with the goal and stores the goal in memory as one which it should not rebel against in future cycles. If it does not comply, it stops pursuing the goal and marks it as an automatic rebellion for future cycles.

4 Simulated Plant Protection Domain

In order to test agent rebellion within a concrete context, we developed a simulated domain called the *Plant Protection Domain*. This domain casts agents as autonomous drones and operators as on-the-ground personnel who know the locations of plants belonging to a harmful invasive species. The goal of the operators is to eliminate all of the invasive species. However, there are also endangered plant species in the area, and the agents are programmed to minimize harming of the endangered plants. The operators are able to instruct the drones by providing them with goals of locations they should move to and drop herbicide. When pursuing a goal, the agent generates a plan to navigate to the target and deploy herbicide. Since the herbicide is dropped from the drone, it cannot be precisely placed but instead lands on the target as well as neighboring areas.

Since the herbicide does not discriminate between invasive

plants and endangered plants, the agent examines the area near where it is dropping herbicide to see if any endangered plants are present. If it detects an endangered plant near an invasive plant, it will rebel against its operator by refusing to deploy the herbicide. The rebellion involves informing its operator that it is rebelling against the order to deploy herbicide. The operator then has the opportunity to overrule this rebellion and instruct the agent to follow through. Even if the operator does reject the rebellion, the agent has the final say, and can ignore the operator if it desires. This dialog means that the rebellion process is a communicative one between the agent and the operator.

Additionally, if there is more than one drone operating in the environment, agents can be proactive by informing other agents that they are on course to harm endangered species and should rebel. Currently, this occurs when the agent pursuing a target does not see an endangered plant which is in range of its herbicide spray, but another agent does. In this case, the latter agent can inform the former about the presence of the endangered plant, thus inciting an act of rebellion.

The world in which the agents and operators act is implemented as an $n \times n$ grid of tiles containing all invasive and endangered species (collectively 'Plants'), and all agents and operators (collectively 'Actors'). Each Plant occupies a single tile, and tiles containing a Plant cannot be moved through by Actors. The reason they cannot be in the same location is because the Actors could disturb the Plants, either with the operator's feet or the wind from the drone's rotor blades, which may damage endangered plants or spread the seeds of invasive plants. Actors also occupy a single tile, but are able to pass through each other. Plants are static fixtures and cannot be moved (i.e., attempting to move a Plant would kill it).

Each Actor's behavior is governed by an instantiation of a MIDCA cycle, and at the end of each cycle the actor selects an action to perform. The most common action is movement, and any Actor can move one tile in any of the cardinal directions, so long as the destination tile is not blocked. The other important thing an Actor can do is ready and deploy a herbicide spray at its location. In order to spray a target, an Actor must approach the target to within the range of spray radius, ready the spray, and finally perform the spray action. When an Actor performs the spray action, all plants within a one-square radius of the Actor are destroyed. The world is graphically displayed to the user through an ASCII representation, as seen in Figure 3.

5 Experimental Setup

The experiments we have conducted so far provide a baseline for understanding how rebellion effects the outcome of missions within the Plant Protection Domain, and how varying the world and the attitudes of the Actors affects this, i.e., the rebelliousness of the agents and the resolve of the operators. Our tests were run in a 10×10 world with five rebellious agents and one operator. We varied the density of plants on the grid, the ratio between the number of endangered and invasive plants, the operator's resolve (i.e., its probability of rejecting a rebellion), and the agents' rebelliousness (i.e., their probability of acquiescing to a rebellion rejection

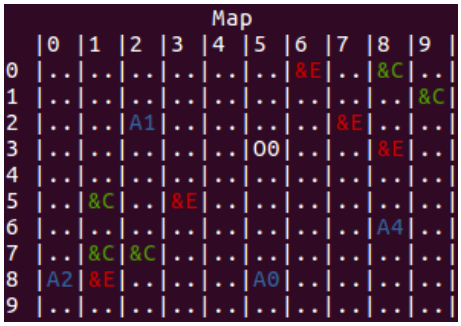


Figure 3: The graphical representation of an instance of the domain. Plants and Actors are labeled and color-coded. Green &Cs are endangered species, red &Es are invasive species, blue As are agents, and white Os are operators. Both agents and operators are also uniquely numbered.

from the operator). The resolve of an operator and the rebelliousness of an agent are each given as a real number between 0 and 1 (inclusive). Thus, higher resolve values make it more likely the operator will reject an agent's rebellion and higher rebelliousness values make it more likely an agent will acquiesce to an operator's rejection. For example, a resolve of 0.25 means an operator will reject a rebellion a quarter of the time, and a rebelliousness of 0.25 means the agent will acquiesce to a rejection of its rebellion a quarter of the time. Our results collected the proportion of invasive plants killed and proportion of endangered plants still alive at the end of each experimental run. Ideally, 100% of invasive plants would be destroyed and 100% of endangered plants would remain unharmed, although that is often not possible given the configuration of the environment (i.e., invasive plants positioned nearby endangered plants).

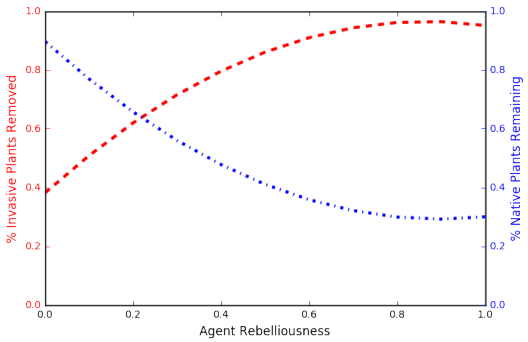
We used the following values for the environment and Actor parameters:

- **Plant Density**: The percentage of squares with plants in them: *low* (0.1), *medium* (0.2), and *high* (0.4)
- **Endangered:Invasive Ratio**: The ratio between endangered plants and invasive plants: *endangered-light* (0.5), *endangered-neutral* (1.0), and *endangered-heavy*(1.5)
- **Operator Resolve**: *none* (0.0), *low* (0.25), *medium* (0.50), *high* (0.75), *complete* (1.0)
- **Agent rebelliousness**: *none* (0.0), *low* (0.25), *medium* (0.50), *high* (0.75), *complete* (1.0)

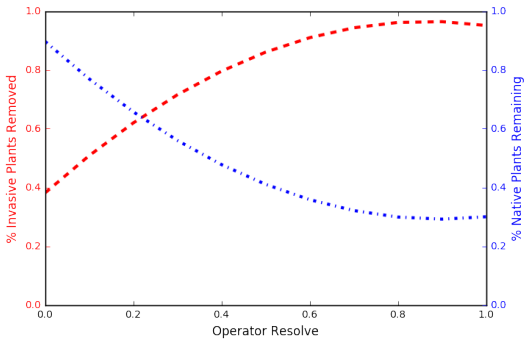
In each experiment, all five agents used identical rebelliousness values. For each set of parameters, we ran three tests and averaged the proportion of invasive plants destroyed and proportion of endangered plants still alive across the three tests.

6 Results

Our initial tests have demonstrated that when operators are less likely to reject agent rebellions and when agents are more likely to rebel against an operator's rejection, endangered plants are far more likely to survive the mission, but fewer invasive species are removed.



(a) Plants remaining based on agent rebelliousness



(b) Plants remaining based on operator resolve

Figure 4: Invasive plants removed and endangered plants remaining based on both agent rebelliousness and operator resolve. In both graphs, the red dashed line indicates the percentage of invasive species removed while the blue dashed and dotted line indicates the percentage of native species which survive. Higher values are better for both lines.

Figure 4 illustrates this effect by presenting the second order polynomial regressions of the percentage of invasive species killed and endangered species remaining against both agent rebelliousness (Figure 4a) and operator resolve (Figure 4b). It can be seen in Figure 4a that there is a positive relationship between the likelihood an agent will persist with its rebellion (i.e., lower rebelliousness) and endangered species surviving and a negative relationship between the probability of the agent persisting and invasive species getting removed. That is, more persistent rebellious agents are more effective at preserving endangered species at the cost of failing to remove invasive ones. This occurs because the agents are far more cautious about dropping herbicide. Figure 4b shows the opposition relationship with operator resolve. Thus, the more likely an operator is to reject an agent's rebellion, the more likely invasive species are to be killed and the less likely endangered species are to survive.

Figure 4 also demonstrate that the effect agent rebelliousness has on plant species survival is less significant than the effect of operator resolve. Whereas the percentage of invasive species removed and endangered species surviving change by approximately 20 percentage points as the probability of agent rebelliousness changes from 0.0 to 1.0, these values change by approximately 60 percentage points as operator resolve goes from 0.0 to 1.0.

Our results also demonstrate that there is a negative relationship between the percentage of endangered species which survive and the percentage of invasive species which are removed. This can be seen both visually in the horizontal symmetry of the graphs in Figure 4 and in the similarity of the percentage point changes of invasive and invasive species in both graphs. This effect is more clearly demonstrated in Figure 5, which shows the relationship between the percentage of endangered plants surviving and the percentage of invasive plants removed across all tests. In general, as more endangered plants are preserved, fewer invasive species are able to be removed.

Finally, we designed a metric to determine the overall success of a mission which is based on comparing the percentage of invasive species removed and the percentage of endangered species left alive. Because different missions, contexts, or high-level directives (e.g., from the supervisors of the agents and operator) may result in different priorities, our metric can be tuned by adjusting the weight given to each category. Thus, we can examine various metrics depending on if more value should be given to invasive species removal (i.e., aligning closer with the operator's goals) or endangered species preservation (i.e., aligning closer with the agents' goals). Figure 6 shows how agent rebelliousness can affect mission score for various mission priorities. Three prioritizations are shown:

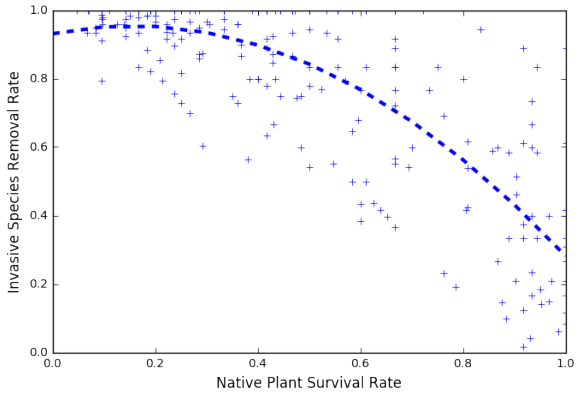


Figure 5: The percentage of invasive species removed compared to the percentage of endangered species which survived for each experiment. Each + represents the results of a single experiment and the blue dashed line indicates the second order polynomial regression on the data.

- **Careless Weighting**: Twice as much value is given to the percentage of invasive plant removed than the percentage of endangered plants saved.
- **Cautious Weighting**: Half as much value is given to the percentage of invasive plants removed than the percentage of endangered plants saved.
- **Even Weighting**: Equal weight is given for the percentage of invasive plants removed and endangered plants

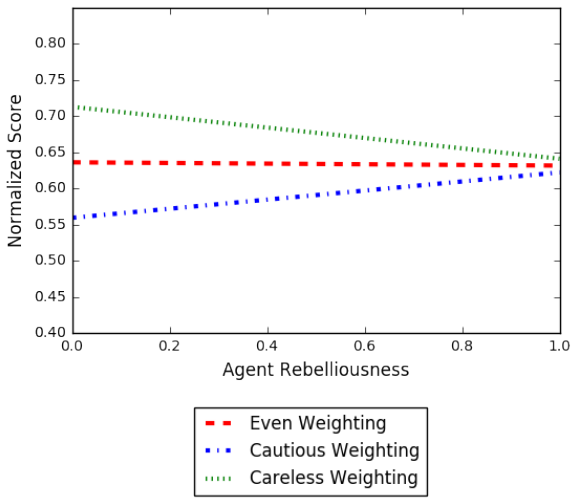


Figure 6: The change in mission scores based on agent rebelliousness. Each line represents a different prioritization, with the green dotted line representing invasive species removal as the priority, the blue dashed and dotted line representing native species survival as the priority, and the red dashed line representing equal priority.

saved.

The figure demonstrates that when caution is required agent rebelliousness is a boon, while such rebelliousness is a burden when native species are not prioritized. When neither is prioritized the agent's rebelliousness does not matter. This aligns with our expectations, since the closer the true metric of agent performance aligns with its motivations (i.e., preserve endangered species) the more advantageous it is for it to restrict itself to goals that comply with those motivations.

7 Conclusion and Future Work

This paper examined the influence that operator resolve and agent rebelliousness has on overall mission performance when goals have both positive and negative effects. Our evaluation demonstrated that lower operator resolve and heightened agent rebelliousness resulted in fewer endangered plants being killed but also fewer invasive plants being removed. This is not particularly surprising, as it resulted in the agents achieving their own goals to preserve endangered plants when there was a conflict with the operator-provided goals to eliminate invasive plants. As a result, only invasive plants which were not close to endangered ones were killed. However, our results also demonstrated how overall performance varied depending on the resolve of the operator and rebelliousness of the agent. More importantly, it demonstrated that ideal rebelliousness is largely influenced by both operator resolve and overall mission evaluation. Hence, a rebellious agent would need methods to estimate both operator resolve and mission evaluation criteria in order to select an appropriate rebellion level that would optimize team performance. We feel this work serves as a useful baseline as we continue research in this area, since it demonstrates the effects of a broad range of attitudes on agents which are preloaded with knowledge about the results of their actions in worlds which contain some uncertainty. Future changes to the domain, such as forcing agents to learn the consequences of their actions, adding a social dynamic to rebellion, a more complex communication process between agents and operators, or the use of interesting motivations for operator rejection or agent rebelliousness can be examined in light of these baseline results.

A similarly expected relationship was found between the percentage of endangered plants left alive and invasive plants removed, specifically there is a clear negative correlation (as shown in Figure 5). This is not surprising given that if there are endangered and invasive plants in close proximity, the agent must remove the endangered plant in order to remove the invasive one. As such, the more endangered plants that are left alive, the more likely it is that some invasive plants were spared as well. The strength of this effect depends on the density of plant life in a region, with low-density areas having fewer situations in which endangered and invasive plants are near each other. The effect of plant density deserves to be further examined because it may be the case that agent rebelliousness ought to be altered based on the density. In a less dense region, for example, the agent might be able to afford to be less cautious while still removing the same number of endangered plants as a more cautious agent would in a more dense region. Thus, an interesting avenue of future work will be to develop mechanisms by which the agent can use environmental or contextual clues to adapt is rebelliousness.

It can also be seen in the results of our evaluation that agent rebelliousness has less of an effect on endangered and invasive plant removal than operator resolve (as shown in Figure 4). This is likely due to the fact that agent rebelliousness indicates the agent's likelihood of ignoring an operator's rejection of the agent's initial rebellion. As such, the agent's rebelliousness only comes into play if the operator rejects a rebellion in the first place. In other words, if an operator is fairly permissive, the agent's rebelliousness is not influential because the agent is never called upon to persevere in its rebellion. Future work should focus on having dynamic initial rebellions based on situational and social cues, rather than having agents always rebel in objectionable situations. In our evaluation, the agent always makes the operator aware of its conflicts but may ignore the operator's final decision, whereas a two-stage rebellion would involve an agent that may suppress its own conflicts. Thus, further evaluation will be necessary to see how two-stage rebellion influences performance.

Future work on rebel agents can go in a variety of directions, and in particular social pressure, consequence learning, and ethics encoding should be further explored. Social pressure would include peer pressure by other agents in a multi-agent scenario to either rebel or not rebel as well as agent-operator and agent-agent trust mechanics. Consequence learning and ethics encoding would work together to enable an agent to have implicit and preloaded ethical imperatives which are agnostic to the particular domain. Learning that certain general types of actions have morally significant consequences allows an agent to predict the moral consequences of future actions it will take and modify its behavior in order to conform with its ethical imperatives, including through rebellion.

Acknowledgments: Thanks to NRL for sponsoring this research. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

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

[Briggs and Scheutz, 2015] Gordon Briggs and Matthias Scheutz. "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*, pages 32–36. AAAI Press, 2015.

[Briggs and Scheutz, 2016] Gordon Briggs and Matthias Scheutz. The Case for Robot Disobedience. *Scientific American*, 316(1):44–47, 2016.

[Bringsjord *et al.*, 2006] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44, 2006.

[Cox *et al.*, 2016] Michael T. Cox, Zohreh Alavi, Dustin Dannenhauer, Vahid Eyorokon, Hector Munoz-Avila, and Don Perlis. MIDCA: A metacognitive, integrated dual-cycle architecture for self-regulated autonomy. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3712–3718. AAAI Press, 2016.

[Cox *et al.*, 2017] Michael T. Cox, Dustin Dannenhauer, and Sravya Kondrakunta. Goal operations for cognitive systems. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 4385–4391. AAAI Press, 2017.

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

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

[Molineaux *et al.*, 2018] Matthew Molineaux, Michael W. Floyd, Dustin Dannenhauer, and David W. Aha. 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, 2018.

[Roberts *et al.*, 2016] Mark Roberts, Vikas Shivashankar, Ron Alford, Michael Leece, Shubham Gupta, and David W. Aha. Goal reasoning, planning, and acting with ActorSim, the actor simulator. In *Poster Proceedings of the 4th Annual Conference on Advances in Cognitive Systems*, 2016.

[Schermerhorn *et al.*, 2006] Paul W. Schermerhorn, James F. Kramer, Christopher Middendorff, and Matthias Scheutz. DIARC: A testbed for natural human-robot interaction. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1972–1973, 2006.

[Vattam *et al.*, 2013] Swaroop Vattam, Matthew Klenk, Matthew Molineaux, and David W. Aha. Breadth of approaches to goal reasoning: A research survey. In *Goal Reasoning: Papers from the ACS Workshop*, 2013.