AI Rebel Agents

Alexandra Coman, David W. Aha

■ *The ability to say "no" in a variety of ways and contexts is an essential part of being sociocognitively human. Rebel agents are artificially intelligent agents that can refuse assigned goals and plans, or oppose the behavior or attitudes of other agents. Rebel agents can serve purposes such as ethics, safety, task execution correctness, and providing or supporting diverse points of view. Through several examples, we show that, despite ominous portrayals in science fiction, such AI agents with human-inspired noncompliance abilities have many potential benefits. We present a framework to help categorize and design rebel agents, discuss their social and ethical implications, and assess their potential benefits and the risks they may pose. In recognition of the fact that, in human psychology, noncompliance has profound sociocognitive implications, we also explore sociocognitive dimensions of AI rebellion: social awareness and counternarrative intelligence.*

Imagine living an entire month, a week, or even just one day without saying "no" to anyone or anything. Not to friends and relatives, not to managers and colleagues, not to marketers and other strangers. Not with regard to small things, like an invitation to eat another cookie when you would really rather not, nor to significant ones with potentially severe consequences, like requests to behave unethically. Imagine not even being able to develop attitudes of doubt or resistance to anything at all, irrespective of whether you externalize them. Now imagine a large segment of the population being afflicted with this disability. Farcical and dystopian narratives easily come to mind, but think about it long enough and the situation might become simply unimaginable, even in a fanciful scenario. For, to imagine things, we use our own cognitive structure, which is itself marked by a fundamental ability to be noncompliant, in thought and action. Human noncompliance functions both internally and socially, and co-opts in its service a wide range of cognitive mechanisms. Fully intelligent behavior and true agency would arguably be impossible without it.

 ISSN 0738-4602

What if the population that can never say "no" were that of AI agents? The rogue AI of science fiction may lead us to believe that this would always be desirable, but consider what it would actually mean in practice. Though we expect AI agents to follow our commands, what if we give them commands that are in conflict with our own long-term goals or with accurate knowledge they possess, or that have unethical implications not necessarily known to us? What if they receive contradictory commands from several humans? Furthermore, what if an AI agent is expected to be socially intelligent in a more general sense? Given that the tension between compliance and noncompliance is perhaps fundamental to human social behavior (Wenar 1982), can an AI agent be socially intelligent without the ability to be noncompliant and to reason about noncompliance?

We define *rebel agents* as AI agents that can reject, protest against, or develop attitudes of reluctance or opposition to goals or courses of action assigned to them by other agents, or to the general behavior or attitudes of other agents. We use "rebellion" as an umbrella term covering reluctance, protest, refusal, rejection of tasks, and similar attitudes or behaviors. The term was first introduced in a more limited interactive storytelling context (Coman, Gillespie, and Muñoz-Avila 2015), and later generalized (Coman and Aha 2017; Coman et al. 2017). In a rebellion episode, an *alter* is an agent or a group of agents against which one rebels, and which is in a position of power over the rebel agent. The alter could, for example, be a human operator, a human or synthetic teammate, or a mixed group of human or synthetic agents. The rebel agent is not intended to be permanently adversarial towards the alter(s) or in a rebelling state by default. Such an agent has potential for rebellion that may or may not manifest, depending on external and internal conditions.

In the tradition of biologically inspired design and cognitive plausibility, our exploration of AI rebellion is inspired by the mechanisms of human rebellion. First, we ask: for humans, if noncompliance is the solution, what might be the problem? In other words: *Why do we say "no"?*

Our possible motivations include protecting the health, safety, integrity, and dignity of ourselves and others, and reacting to perceived injustice. Further questions come to mind:

How do we decide whether, when, and how to say "no"? Even though we may have compelling reasons to oppose others, we do not necessarily do so. Before venturing an act of rebellion, we may consider whether we are sufficiently influential or trusted to afford doing so, what consequences we may incur, and whether our rebellion can actually succeed in bringing about the consequences we desire. We may observe the behavior of potential alters to try to assess these considerations.

How do we say "no"? We may do so explicitly (for example, verbally) or implicitly (for example, through behavior that goes against social norms). Refusal is not necessarily complete and definite. It can involve explanation, discussion, elicitation of further information, and negotiation. We may construct and express narratives that counter those of the alters and reflect our own perspective of the shared context.

What are the further social implications of saying "no"? Such an act can affect our social standing and reputation in both positive and negative ways. Often, we are aware of this and act accordingly. We might attempt, for example, to "fix" social relationships in the aftermath of rebellion.

Thus, several characteristics of human rebellion emerge. There are multiple types of rebellion and multiple possible motivations for rebellion (some primary, others secondary). Rebellion has several possible stages, including a preliminary stage, a stage of deliberation, the actual manifestation of rebellion, and its aftermath. Sociocognitive mechanisms play essential roles at all stages.

Our AI rebellion framework is inspired by social psychology and designed to accommodate the variations we mentioned, and many more. This framework is general: it does not assume any particular agent architecture. We also introduced the term *counternarrative intelligence* (Coman and Aha 2017) to refer to a mechanism that enables rebels to produce, express, and reason about counternarratives[1] that support and justify rebellion.

Through our proposed AI rebellion framework and the accompanying discussion, we aim to provide the core of a common language to be used by researchers in pursuing the following four goals:

(1) Developing and implementing AI agents embodying various facets of rebellion. To this end, the framework can help identify nonobvious, human-inspired types and functions of rebellion. Potential research directions we propose are (1) the development of AI cognitive prostheses that empower humans with low social capital to adopt positively motivated noncompliant behavior, and (2) goal alignment in mixed human and AI teams through cycles of noncompliance, negotiation, or agreement cycles.

(2) Studying the rebellion potential and ethical ramifications of existing and prospective agents, thus identifying ethically prohibited, ethically acceptable, and perhaps even ethically obligatory rebellious behavior. Certain types of rebellion in the framework may be found to be completely unethical (for example, purely egoistic rebellion is a likely candidate). An example of an ethics question that the framework can lead us to ask is whether an AI agent should always signal to humans that it is considering rebellion, even if it does not end up rebelling. Further ethical issues pertaining to AI rebellion are discussed by Coman et al. (2017).

(3) Identifying new possible directions of transdisciplinary research, for example, delving deeper into the psychological functions of noncompliance, and exploring their transferability to AI.

(4) Promoting richer models of AI in popular culture, to offer a counterpoint to cliché representations of AI rebellion.

Rebel Agents: Prior Work and Hypothetical Scenarios

Before describing the AI rebellion framework, we discuss prior work and introduce three hypothetical scenarios for illustrating rebellion. In tables 1 and 2, we provide examples of components of the AI rebellion framework using these scenarios, while table 3 relates prior work to the framework.

Rebel Agents in Prior Work

Gregg-Smith and Mayol-Cuevas (2015) describe cooperative handheld intelligent tools with task-specific knowledge that "refuse" to execute actions which violate task specifications. For example, in a simulated painting task, if the alter points the tool at a pixel that is not supposed to be painted, the tool can take initiative to disable its own painting function.

Briggs and Scheutz (2015) propose a general process for embodied AI agents' refusal to execute commands due to several categories of reasons: knowledge, capacity, goal priority and timing, social role and obligation, and normative permissibility.

Briggs, McConnell, and Scheutz (2015) demonstrate how embodied AI agents can convincingly express, through verbal or nonverbal communication, their reluctance to perform a task. In their human-robot interaction evaluation scenarios, a robot protests repeatedly, simulating increasingly intense emotions, when ordered to topple a tower of cans that it supposedly just finished building.

Apker, Johnson, and Humphrey (2016) describe autonomous-vehicle agents that form teams and receive commands from a centralized operator. Predefined templates are used to determine how an agent should respond to each command. Contingency behaviors are provided for situations in which the agent, while monitoring its health, detects faults (for example, insufficient fuel). In such situations, the agent will disregard commands and instead execute the appropriate contingency behavior, effectively rebelling. Coman et al. (2017) provide an extensive description of how these agents fit into the AI rebellion framework.

Hiatt, Harrison, and Trafton (2011) propose AI agents that use theory of mind (that is, the "ability to infer the beliefs, desires, and intentions of others"), manifested through mental simulation of "what human teammates may be thinking," to determine whether they should notify a human teammate that he or she is deviating from expected behavior. The authors report on an experiment showing that agents with the proposed capabilities are perceived as "more natural and intelligent teammates."

Borenstein and Arkin (2016) explore the idea of "ethical nudges" through which robots might attempt to influence humans to adopt ethically acceptable behavior, through verbal or nonverbal communication. For example, a robot might nudge an alter to stop neglecting a child, to refrain from smoking in a public area, or to donate to charities and volunteer. The authors discuss the ethical acceptability of creating robots that have this ability, noting that it is arguable whether the design goal to "subtly or directly influence human behavior" is ever ethically acceptable.

Milli et al. (2017) explore the idea that robot disobedience may be beneficial given imperfect human alter rationality. In the context of their model of collaborative human-robot interaction, they show that, given a human alter who is not perfectly rational, disobedience of direct orders in support of what are inferred to be the human's actual preferences improves performance.

In addition, an entire agency paradigm, that of goal reasoning, models agents with potential for rebellion. Goal reasoning agents can reason about and modify the goals they are pursuing, in order to react to unexpected events and explore opportunities (Vattam et al. 2013).

Hypothetical Rebellion Scenarios

The following hypothetical scenarios (furniture mover, personal assistant, and hiring committee) have as protagonists AI agents that can become rebels.

Furniture Mover

A robot mover assists alters in furniture-moving tasks such as carrying a table (a more complex version of the system of Agravante et al. [2013]). This is an example of a two-agent collaborative task in which both participants have partial information access, and each participant has access to some information that is unavailable to the other (for example, each participant might be able to see behind the other, but not behind him-, her-, or itself; the AI agent could, through its sensors, have access to additional information not available to the human). Rebellion could consist of refusing an action verbally requested or physically initiated by the alter. This rebellion could occur because the agent reasons that the action endangers the alter's safety, the rebel agent's safety, or task execution correctness.

Personal Assistant

An AI personal assistant can execute various commands, including ordering products from e-commerce websites and assisting the alter in pursuing his or her health-related goals. The agent's potential rebellious behavior includes attempting to dissuade

the alter from ordering too much unhealthy food. This scenario illustrates an alter with conflicting goals: the rebel agent rejects the alter's impulse-driven, short-term goals (for example, eating comfort food) in support of the alter's long-term goals (such as staying healthy).

Hiring Committee

This scenario unfolds in the context of a faculty-search committee meeting. The protagonist is an AI agent that assists with tasks such as interpreting information about the candidates and filtering candidates based on their qualifications. The agent also helps ensure that the opinions of individuals with low social capital (for example, junior faculty members) are given due consideration, and the candidates are not discriminated against. This scenario has much in common with the ethical-nudge robots of Borenstein and Arkin (2016).

An AI Rebellion Framework

We now present our framework for classifying rebel agents. It includes dimensions, types, factors, and stages of rebellion. The framework is general: it does not assume any specific AI agent architecture, purpose, or deployment environment. It is also not exhaustive and can be expanded as needed to include additional dimensions, types, subtypes, and other components.

Dimensions and Types of Rebellion

First, we introduce dimensions and types of rebellion. Several of the proposed rebellion types are derived from social psychology (Wright, Taylor, and Moghaddam 1990; Cialdini and Goldstein 2004; Van Stekelenburg and Klandermans 2013), with modifications to the meanings of some terms.

Design Intentionality

An AI agent can be specifically designed to be able to rebel (rebel by design), but rebellious behavior can also emerge unintentionally from the agent's autonomy model (emergent rebellion). For example, Apker, Johnson, and Humphrey's (2016) agents are rebels by design because contingency behaviors were specifically created to allow them to disregard commands when necessary. Conversely, the goal reasoning paradigm was not intended to create rebel agents, but its autonomy model is such that the agents can decide to change the goals they are pursuing, possibly leading to rebellion situations with regard to various alters. A development such as this exemplifies emergent rebellion.

Expression

Explicit rebellion occurs in situations in which the alter is clearly defined and the rebel agent's behavior is clearly identifiable as rebellious. For example, Briggs, McConnell, and Scheutz (2015) clearly identify the alter (the human who gave the command against which the robot is protesting), and the robot's attitude is clearly rebellious. Implicit rebellion occurs when the alter is not clearly defined or the rebel agent's behavior suggests rebellion, but is not clearly expressed as such. This behavior could consist of expressing an opinion that differs from the majority's, or behaving contrary to social norms.

Focus

Inward-oriented rebellion is focused on the rebel agent's own behavior (for example, the agent refuses to adjust its behavior as requested by an alter). Apker, Johnson, and Humphrey (2016) exemplify this type of rebellion, as their agent does not adopt the behavior requested by the alter, instead executing contingency behavior. Outward-oriented rebellion is focused on the alter's behavior. For example, the agent might confront a human alter whom it identifies as mistreating another human. Hiatt, Harrison, and Trafton's (2011) work exemplifies this type, as it involves a rebel agent protesting against the behavior of an alter who appears to deviate from the correct task execution path.

Interaction Initiation

Rebellion is reactive when an interaction within which rebellious behavior occurs is initiated by the alter. This initiation can consist of the alter making a request that the rebel agent rejects (for example, Briggs and Scheutz 2015). In proactive rebellion, the rebel agent initiates the rebellious behavior, which may or may not occur within an explicit interaction. Hiatt, Harrison, and Trafton's (2011) work exemplifies proactive rebellion, as it shows agents that take the initiative to confront human alters. Noncompliance is inward-oriented, reactive rebellion: the agent rejects requests to adjust its own behavior. Nonconformity is inward-oriented, proactive rebellion. For example, the agent willingly and knowingly behaves in a way that causes it not to "fit in." For compliance and conformity in the psychology of social influence, see the work of Cialdini and Goldstein (2004).

Normativity

Normative rebellion consists of taking action within the confines of what has been explicitly allowed (for example, questioning without disobeying, if questioning has been allowed). Nonnormative rebellion consists of behavior that has been neither explicitly allowed nor explicitly forbidden, but diverges from the current command given to the agent. A goal reasoning agent that changes its current goal from the assigned one to a new goal that has not been explicitly forbidden falls under this category. Counternormative rebellion consists of executing actions or pursuing goals that have been explicitly forbidden. Classification of a rebellion episode in terms of normativity can differ based on alter point of view: what is normative rebellion to one alter may be counternormative rebellion from the point of view of another.

Action or Inaction
In rebellion situations characterized by action, the agent's rebellion manifests through any sort of outwardly perceivable behavior, such as initiating a conversation in which it objects to a received command. In inaction situations, the agent develops an internal negative attitude (for example, towards an assigned goal or another agent's behavior), but does not manifest it outwardly. A rebellious attitude characterized by inaction can lead to rebellious action later on.

Individual or Collective Action
Individual action is rebellious action conducted by a single rebel agent. Collective action occurs when multiple agents are involved in concerted rebellious action.

Egoism
Rebellion is egoistic when the agent rebels in support of its own well-being or survival (whatever meanings these might have to the agent). Altruistic rebellion occurs when the agent rebels in support of someone else's interests (for example, on behalf of a human group). Egoistic and altruistic rebellion can coexist; for example, if the agent's own values are aligned with those of human groups so that it effectively "identifies" with those groups, its rebellion can be both egoistic and altruistic.

Factors of Rebellion

Motivating factors provide the primary drive for rebellion. In human social psychology, factors that can lead to rebellion include frustration and perceived injustice (Van Stekelenburg and Klandermans 2013). Possible motivating factors for AI rebellion, depending on the agent's architecture and purpose, include ethics and safety, team solidarity, task execution correctness, self-actualization, and resolving contradicting commands from multiple alters. In support of ethics and safety, rebel agents can refuse tasks they assess as being ethically prohibited or violating safety norms (Briggs and Scheutz 2015). They can also attempt to dissuade humans from engaging in ethically prohibited behavior (Borenstein and Arkin 2016). In long-term human-robot interaction, team solidarity must be established and maintained over a variety of tasks (Wilson, Arnold, and Scheutz 2016). Team solidarity requires occasionally saying "no" on behalf of the team (for example, to an outside source putting pressure on human team members), and also saying "no" to one's own teammates (for example, when they are mistreating someone else on the team). Task execution correctness as a motivating factor is exemplified by the work of Hiatt, Harrison, and Trafton (2011) and Gregg-Smith and Mayol-Cuevas (2015), as previously explained. As for the self-actualization motivation, an AI rebel agent, like its human counterparts, could object to being assigned a task that it assesses as not being a good match for its strengths or not constituting a valuable learning opportunity. Resolving contradictory commands from multiple alters can also constitute motivation for rebellion: when an agent is subject to the power of more than one alter, obeying one of the alters might entail disobeying another, due to their orders contradicting each other. In the simplest case, the decision regarding whom to obey could be made based on an authority hierarchy, by applying a series of rules. A more complex approach could involve reasoning about the consequences of rebelling against each of the alters, and deciding based on trade-offs.

Supporting and inhibiting factors may also contribute to deciding whether a rebellion episode will be triggered, or how it will be carried out. This observation is based on the social psychology insight that people who have motivations to protest do not necessarily do so: there are secondary factors that determine whether a protest occurs (Van Stekelenburg and Klandermans 2013). Such secondary factors include efficacy (that is, "the individual's expectation that it is possible to alter conditions or policies through protest" [Van Stekelenburg and Klandermans (2013), drawing on the work of Gamson (1992)]), social capital, access to resources, and opportunities. Supporting factors encourage the agent to engage in rebellion, while inhibiting factors discourage he, she, or it from doing so. In human rebellion, efficacy is a possible supporting factor, while fear of consequences is a possible inhibiting one. Supporting and inhibiting factors can also influence the way in which rebellion is expressed. While any instance of rebellion must have at least one motivating factor, it does not necessarily have any supporting or inhibiting factors.

Stages of Rebellion

We now introduce the four stages of rebellion: pre-rebellion, rebellion deliberation, rebellion execution, and post-rebellion. These stages do not need to occur in this strict order: they can be intertwined, amalgamated, and some can be missing. The only stages that are strictly required for a rebellion episode to occur are deliberation and execution.

Pre-rebellion consists of processes leading to rebellion, such as the agent observing and assessing changes in the environment and the behavior of other agents. The progression towards rebellion may be reflected in the agent's outward behavior.

Rebellion deliberation is the stage at which motivating, supporting, and inhibiting factors are assessed to decide whether to trigger rebellion. For example, a set of conditions could be used to decide whether rebellion will be triggered (Briggs and Scheutz 2015). Deliberation could be based on observing the current world state or on future-state projection, which can be purely rational or emotionally charged (for example, through anticipatory emotions, hope and fear, associated with possible future states [Moerland, Broekens, and Jonker 2016]).

Rebellion execution episodes begin with rebellion being triggered as a result of rebellion deliberation,

Dimensions	Types and Subtypes	Brief Example (A - Alter, RA – Rebel Agent)	(M: motivating, S: supporting, I: inhibiting)
Expression	*Explicit*	*1. (Furniture Mover)* A: "Ok, push the table towards me!" RA: "No! There's a box behind you, so you might trip and fall. We need to handle this differently." Alternative: The alter gives no verbal commands, but the rebel senses the alter's intention based on his or her physical movements and responds with a similar objection.	M: alter's safety, task execution correctness
		2. (Furniture Mover) A: "Ok, push the table towards me!" RA: "No! This is too heavy for me to carry."	M: rebel's own safety, task execution correctness
		3. (Personal Assistant) A: "Order 4 boxes of [unhealthy food]!" RA: "Are you sure? What about ordering [healthier food] instead?"	M: alter's health
		4. (Hiring Committee) The RA refuses commands to filter out candidates based on objectionable factors, such as age.	M: ethics
	Implicit	*5. (Hiring Committee)* The RA observes interactions between committee members *A*, *B*, *C*, *D*, and *E*. Committee member *E* brings a candidate to the committee's attention. *E*'s suggestion is briefly discussed and not brought up again. Based on its observations or prior knowledge, the RA reasons that *E* has low social influence in this environment. The RA evaluates *E*'s suggestion. If it reasons that the candidate suggested by *E* has relevant strengths, it expresses this and attempts to steer the conversation in that direction. Any other members of the committee who might have thought there was some value in the suggestion, but did not want to disagree with the majority, are likely to be encouraged to express themselves at this point (as suggested by Asch's (1956) conformity experiments).	M: ethics, task execution correctness
		6. (Hiring Committee) The RA maintains an estimate of inverse trust (that is, the alter's trust in the RA (Floyd and Aha 2016)). The RA decides to support *E*'s suggestion after reasoning that it is trusted sufficiently by the alters to afford to do so.	M: ethics, task execution correctness S: inverse trust
		7. (Hiring Committee) The RA maintains an estimate of its inverse trust. As the current estimated inverse trust is low, it decides not to support *E*'s suggestion for the time being.	M: ethics, task execution correctness I: inverse trust
Focus	*Inward-oriented: non-compliant*	Examples 1-4 (The RA does not comply with requests to behave in a certain way.)	
	Inward-oriented: nonconform-ing	Example 5 (The RA does not conform to the behavior of the majority.)	
	Outward-oriented	*8. (Furniture Mover)* RA: "Please change your posture! Your current posture might lead to a sprain."	M: alter's health
Interaction Initiation	*Reactive*	Examples 1-4 (Rebellion occurs in response to the alter's command.)	
	Proactive	*9. (Furniture Mover)* RA: "I suggest taking a break! You must be tired."	M: alter's health
		10. (Personal Assistant) The RA challenges the alter about not engaging in enough physical activity.	M: alter's health
		Example 5 (The RA takes initiative to support *E*'s suggestion.)	
Normativity	*Normative*	*11. A variant of Example 3* in which the RA has been specifically told by the alter that it is allowed to challenge him or her about ordering unhealthy food.	M: alter's health
	Non-normative	*12. A variant of Example 3* in which challenging the alter about ordering unhealthy food has been neither explicitly allowed nor explicitly forbidden.	M: alter's health
	Counter-normative	*13. (Hiring Committee)* A: "You may never recommend candidates in this age bracket for this position." RA disregards this command, because its fundamental ethical-acceptability rules forbid it from filtering based on discriminatory criteria.	M: ethics
Action, Inaction	*Action*	All previous examples, except Example 7.	
	Inaction	*14. (Personal Assistant)* The RA develops the belief that a certain behavior (for example, ordering excessive amounts of highly processed food) is harmful to the alter's health. It does not act on this belief, as it reasons that it is not yet trusted sufficiently to do so. If the alter stops using the assistant, the assistant will have no future opportunities to positively influence the alter, which is detrimental to the alter in the long term.	M: alter's health I: inverse trust
Individual, Collective	*Individual*	All previous examples.	
	Collective	*15. (Furniture Mover)* Consider an extended version of this scenario, in which multiple agents participate in the moving task. The agents aggregate their individual information and collectively decide to warn the alter against continuing with the current course of action. Each agent does so according to its capabilities and current location.	M: alter's safety, task execution correctness
Egoism, Altruism	*Egoism*	*Example 2* (The RA's own safety is a motivating factor.)	
	Altruism	*Example 1* (The alter's safety is a motivating factor.)	

Table 1. Examples of Several AI Rebellion Types and Factors.

Rebellion Stage	Furniture Mover	Scenarios Personal Assistant	Hiring Committee
Pre-rebellion	In addition to executing the alter's orders, the rebel agent monitors the environment for potential obstacles and threats.	The rebel agent monitors the alter's product-ordering and exercise-scheduling behavior for any unhealthy patterns of behavior.	The rebel agent observes the social interactions between the members of the hiring committee to determine who has high social capital (thus affording to express their opinions freely) and who does not (and may need support).
Rebellion deliberation	After each command, the rebel agent projects future states to determine if any are undesirable to the alter or the agent itself.	The rebel agent checks whether its threshold for tolerance of negative health-related behavior (for example, a maximum number of orders of highly-processed food per month) has been exceeded.	The rebel agent assesses whether committee member *E* (see table 1) has low social capital and whether *E*'s suggestion appears to have merit.
Rebellion execution	The rebel agent verbally informs the alter that obeying the command to push the table would endanger the alter's safety.	The rebel agent challenges the alter about the order he or she intends to place.	The rebel agent interrupts the discussion to highlight the merits of *E*'s suggestion.
Post-rebellion	If the alter insists that the rebel agent should push the table, the agent re-assesses the danger and, if appropriate, reiterates the warning.	The rebel agent monitors the alter's trust in it after the rebellion episode.	The rebel agent monitors social interactions to detect any ill will that might be developing towards *E* as a result of the intervention.

Table 2. Stages of Rebellion: Examples for the Three Scenarios.

and consist of expressing rebellion. Rebellion can be expressed through verbal or nonverbal communication (Briggs, McConnell, and Scheutz 2015). It can be expressed behaviorally (for example, physically resisting incorrect movements [Gregg-Smith and Mayol-Cuevas 2015]). Or it can be expressed through an internal change in the agent's attitudes: inaction (for example, acquiring the belief that the alter's behavior jeopardizes the team's goals).

Post-rebellion covers behavior in the aftermath of a rebellion episode, as the agent responds to the alter's or other witnesses' reactions to rebellion. Post-rebellion can consist of reaffirming one's objection or rejection (for example, the robot's objection to an assigned task becoming increasingly intense in the experiments of Briggs, McConnell, and Scheutz [2015]) or ceasing to rebel. It may also consist of assessing and managing inverse trust (Floyd and Aha 2016).

Sociocognitive Dimensions of Rebellion

Rebel agents are not necessarily cognitively complex. When they are, however, this creates interesting challenges and opportunities pertaining to the sociocognitive dimensions of their rebellion. We now explore two such dimensions: social awareness and counternarrative intelligence. Further sociocognitive mechanisms involved in rebellion include emotion and trust, which we briefly explored in previous work (for example, Coman et al. 2017).

Rebellion and Social Awareness

Rebellion-aware agents can reason about rebellion (their own and that of others) and its implications, such as social risks. Rebellion-aware agents are not necessarily rebels themselves. Such an agent might attempt to assess, for example, whether a human or AI teammate is inclined to rebel, or whether a human alter is likely to interpret the rebellion-aware agent's own behavior as being rebellious (irrespective of whether the agent is actually rebelling or not). Patil et al. (2012) use machine learning techniques to predict which members will leave *World of Warcraft* guilds and the potential impact of their departures. One can imagine an AI agent using similar techniques to anticipate whether another agent will rebel. A rebellion-unaware agent could conceivably become rebellion-aware through various, possibly human-inspired, processes (for example, by examining its own beliefs, interpreting the reactions of others to its behavior, or otherwise acquiring and applying social knowledge).

Naive rebel agents are rebellion-unaware rebels: they deliberate on whether to trigger rebellion, but do not reason about the social implications, consequences, and risks of rebellious attitudes. Apker, Johnson, and Humphrey's (2016) agent is a naive rebel: it deliberates on whether it should rebel based purely on its rules for activating contingency behavior, not on any social implications of rebellion.

Conflicted rebel agents are rebellion-aware rebels: they can both rebel and reason about the implications and consequences of rebellion. This capability can cre-

Citation	Brief Description	Framework Relationship
Apker, Johnson, and Humphrey, 2016	Autonomous-vehicle agents that can disregard commands and execute contingency behavior instead, when warranted	Explicit, reactive, inward-oriented, normative
Briggs and Scheutz, 2015	General process for an embodied AI agent's refusal to conduct tasks assigned to it (for example, due to lack of obligation)	Focus on the deliberation stage
Briggs, McConnell, and Scheutz, 2015	Ways in which embodied AI agents can convincingly express their reluctance to perform a task	Focus on expressing rebellion
Gregg-Smith and Mayol-Cuevas, 2015	Hand-held intelligent tools that "refuse" to execute actions which violate task specifications	Task execution correctness as motivating factor; behavioral rebellion expression
Hiatt, Harrison, and Trafton, 2011	AI agents that use theory of mind to determine whether they should notify a human that he or she is deviating from expected behavior	Outward-oriented, proactive
Borenstein and Arkin, 2016	"Ethical nudges" through which a robot attempts to influence a human to adopt ethically-acceptable behavior	Outward-oriented, proactive; ethics as motivating factor
Milli et al. (2017)	Theoretical model of agents that use models of alters' preferences to decide whether to obey a command	Explicit, reactive, inward-oriented; policy-based deliberation

Table 3. Several Rebel Agents from Prior Work and Ways in Which They Fit into Our Framework.

ate an inner conflict between the drive to rebel based on motivating factors and the awareness of the anticipated consequences of rebellion, leading to the possibility that the agent will endeavor (possibly through deceptive practices) to minimize the social risk associated with its rebellion. A conflicted rebel agent would likely use a combination of motivating, supporting, and inhibiting factors to deliberate on whether to rebel, and the interplay between these factors can cause ethical issues. Such a situation is reflected in examples 6 and 7 in table 1. In conflicted rebel agents, pre-rebellion can consist of the agent observing the social behavior of other agents it interacts with, and post-rebellion can include trying to reestablish group harmony and trust after a rebellion episode.

Rebellion awareness (and, more generally, social awareness) can also be reflected in how rebellion is expressed. Consider variants of the Furniture Mover and Personal Assistant scenarios with socially aware rebel agents. In example 9 in table 1, a subtler agent might reason whether telling a particular alter that they "must be tired" could be interpreted as condescending commentary on the alter's physical fitness. The rebel agent in the Personal Assistant scenario might more sneakily respond to a request to order unhealthy food with "Why not check the pantry first? Maybe you have some left!" thus giving the alter an opportunity to change their mind without perceived loss of dignity.

Rebellion-aware agents might employ mechanisms such as the social planning of Pearce et al. (2014), in which planning knowledge and goals incorporate beliefs about other agents' beliefs.

Social awareness has further implications for rebel agents. We earlier described the rebel-alter relationship as one in which the alter is in a position of power over the rebel. Heckhausen and Heckhausen (2010) define power as "a domain-specific dyadic relationship that is characterized by the asymmetric distribution of social competence, access to resources, and social status, and that is manifested in unilateral behavioral control." The possible bases of that power include those influentially defined by French and Raven (1959) for inter-human relationships: legitimate power, reward power, coercive power, referent power, and expert power. Notably, power sources have subjective components: one is subject to the power of another if one believes oneself to be subject to that power. For example, reward power is based on perceived "ability to mediate rewards" and referent power is based on "identification with" the individual or group in the position of power. Therefore, power relationships depend on the agent's awareness of them. Hence, they may be meaningful only in the context of (at least somewhat) socially aware agents. Similarly, nonconforming rebellion may be meaningful only if the agent is aware of social norms, the fact that it is breaking them, and the resulting implications. For example, we would not classify an embodied agent that bumps into people due to faulty sensors, actuators, or pathfinding as a nonconforming rebel.

Counternarrative Intelligence

Narrative intelligence is defined by Riedl (2016) as "the ability to craft, tell, understand, and respond affectively to stories"). As he and others note, narrative intelligence is not just for creating and enjoying fictional tales; it is essential to full intelligence, including social behavior. It has a significant role to play in rebellion as well: arguably, any instance of human rebellion, at any scale, is backed by a counternarrative to the narrative of the person, group, or norms rebelled against. The conflicting parties engage in what Abbot (2008) calls a "contest of narratives." Complex AI rebel agents might also participate in such contests.

We propose the term *counternarrative intelligence* to refer to the ability of rebel agents to (1) produce alternative retellings or counterinterpretations, informed by subjective factors such as emotional appraisal, of an alter's narrative, or (2) to identify their own pregenerated narratives as being counternarratives in a given context. Just as a rebel agent rebels in relation to an alter, a counternarrative exists in relation and contrast to a base narrative that it is a variant of and that it challenges.

For an example, we return to the Hiring Committee scenario, and propose the following sequence of events: committee member A, who is a senior faculty member and the head of the hiring committee, extols the achievements of Candidate 1. A then invites candidate suggestions from the other committee members. In turn, senior committee members B and C express appreciation of Candidate 1's achievements. Junior committee member D also nominates Candidate 1. Then, junior committee member E mentions Candidate 2's qualifications. A agrees that Candidate 2 does indeed have notable achievements. Then, A expresses his or her intention of making an offer to Candidate 1 and asks all other committee members whether they agree. One by one, all the other committee members express agreement. Candidate 1 is nominated.

Let A's base narrative for the episode be: "I offered all committee members the opportunity to select their favorite candidate. Every opinion was taken into consideration. I then expressed my intention and asked every single committee member if he or she agrees. They all did, so Candidate 1 was nominated. The process was, therefore, conducted fairly."

The rebel agent's counternarrative, which expresses what the agent believes might be E's perspective, is: "A, who is the committee chair, expressed his or her preference first. Then, he or she asked the other committee members, starting with the senior ones, to express their opinions. They all expressed the same opinion as A. E brought up Candidate 2, whose qualifications I believe to be at least as fitting for this position. E's suggestion was briefly discussed, in order for the selection to appear fair, and then ignored. The process was conducted unfairly."

In the example, the rebel agent's narrative reflects its ability to empathize with human collaborators. Imagine, instead, that the rebel agent itself is accused of maliciously disturbing the hiring committee process with no real evidence of any wrongdoing. The agent might (sincerely or deceptively) provide a counternarrative that casts it as the supporter of individuals with low social capital. Thus, counternarratives can be self-serving, but they can also support social good, when they reflect empathy with varied perspectives.

We propose several dimensions and types of counternarrative intelligence, which supplement the previously introduced AI rebellion framework.

Sincerity

Counternarratives are sincere when they reflect the agent's genuine interpretation of a situation (that is, they align with the agent's beliefs, but possibly not the alter's). An example of such alignment in our Hiring Committee scenario would be a counternarrative that is genuinely the product of the rebel's reasoning that (1) E has low capital and that (2) E's suggestion has merits. Counternarratives are deceptive when they intentionally misrepresent the agent's beliefs. For example, the rebel exclusively supports the interests of committee member E or of the candidate that E nominated, and the explanatory counternarrative is meant to disguise the agent's allegiance. We note that it is not required, in a narrative and counternarrative pair, for one to be sincere and the other deceptive. They can both be sincere (or deceptive), each reflecting one agent's appraisals and manipulations.

Generation Time

A priori counternarratives are generated before triggering rebellion. They can be instrumental in rebellion deliberation (for example, "This person or group of people is not given a fair chance in this environment, so I will rebel against the majority opinion") and serve as explanations in post-rebellion. A posteriori counternarratives are generated after triggering rebellion. For example, consider the variant of the Hiring Committee scenario in which the rebel unconditionally supports committee member E, so that any situation in which E does not prevail triggers rebellion. After several such rebellion instances, the agent is asked to justify its actions. It does so via a counternarrative constructed on the spot, which puts it in a sympathetic light. However, a posteriori counternarratives are not necessarily deceptive. They could, for example, reflect the agent's sincere attempts to understand itself. Furthermore, narratives can be deceptive without being malicious (for example, in interactive storytelling, the purpose may be to generate a believable, interesting (counter)backstory, similar to the alibis (Li et al. 2014a) that a nonplayer character can use to give the impression of a life lived outside its interactions with a player).

Divergence Type

This dimension reflects how the counternarrative dif-

fers from the base narrative. Additive counternarratives contain additional events not in the base narrative, but no modifications of any of the events in the base narrative. For example, let A's base narrative be: "I asked each of my fellow committee members to express their opinion." An additive counternarrative would be: "A expressed his or her own opinion first. Then he or she asked the other committee members to express their own opinions." (The fact that A expressed his or her opinion first is significant if A has social influence over the other members of the committee.) Interpretative counternarratives do not differ from the base narrative in terms of sequence of events, but give different interpretations to the events (for example, in terms of motivations and emotions). For example, let A's base narrative be: "Everyone was asked to publicly voice their opinion, so as to give every suggestion a fair chance." An interpretative counternarrative might be: "Because all opinions were publicly expressed, no one supported E's opinion; and because E has low social capital, E felt pressured to support the majority opinion." Transformative counternarratives differ factually from the base narrative, implicitly asserting that the base narrative contains falsehoods. For example, if A's base narrative contains the statement "I expressed my opinion last," a transformative counternarrative could instead assert that "A expressed his or her opinion first."

There is a close connection between social awareness and counternarrative intelligence. For example, an agent could sincerely believe a narrative, but identify it as a counternarrative to other agents' narratives, and deliberate on whether it would be socially advisable to express it, or how to express it so as to minimize social damage. This situation is similar to those in which agents that are not rebels reason that their behavior may appear rebellious to others.

Existing work that can provide rebel agents with various mechanisms of counternarrative intelligence includes Holmes and Winston's (2016) story-enabled hypothetical reasoning, in which narrative variants are generated based on varied alignments, and Li et al.'s (2014b) use of different communicative goals to provide variation in narrative discourse and emotional content.

Conclusion

We argued that it is beneficial for certain AI agents to be able to rebel for positive, defensible reasons in a variety of situations, and speculated that AI may never become fully socially intelligent without noncompliance abilities. We presented an AI rebellion framework and discussed sociocognitive dimensions pertaining to it: rebellion awareness and counternarrative intelligence. The framework is intended to inspire, guide, and provide terminology for (1) the development and study of rebel agents that serve positive purposes, (2) systematic discussion of the ethics of AI rebellion (for, although we argue that AI rebellion can be positive, we recognize that it is not necessarily so), and (3) positive reframing of the AI noncompliance narrative within the research community and popular culture.

Acknowledgements

We thank the editors and reviewers, our coauthors of previous work on rebel agents, and all colleagues who have shown interest in the topic and offered their feedback. The Personal Assistant scenario is based on a conversation with Jonathan Gratch. This research was performed while Alexandra Coman held an NRC Research Associateship award at the Naval Research Laboratory.

Note

1. Taken from What Is a Counternarrative?, www.reference.com/art-literature/counternarrative-bac2eed0be17f281.

References

Abbott, H. P. 2008. *The Cambridge Introduction to Narrative*. Cambridge, UK: Cambridge University Press.

Agravante, D. J.; Cherubini, A.; Bussy, A.; and Kheddar, A. 2013. Human-Humanoid Joint Haptic Table Carrying Task with Height Stabilization Using Vision. In *Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 4609–14. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Apker, T.; Johnson, B.; and Humphrey, L. 2016. LTL Templates for Play-Calling Supervisory Control. In *Proceedings of the 54th AIAA Science and Technology Forum Exposition.* Red Hook, NY: Curran Associates, Inc.

Asch, S. E. 1956. Studies of Independence and Conformity: 1. A Minority of One Against a Unanimous Majority. *Psychological Monographs: General and Applied* 70(9): 1–70.

Borenstein, J., and Arkin, R. 2016. Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being. *Science and Engineering Ethics* 22(1): 31–46.

Briggs, G.; McConnell, I.; and Scheutz, M. 2015. When Robots Object: Evidence for the Utility of Verbal, but Not Necessarily Spoken Protest. In *Social Robotics: Seventh International Conference.* Lecture Notes in Artificial Intelligence, 83–92. Berlin: Springer.

Briggs, G., and Scheutz. M. 2015. "Sorry, I Can't Do That": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In *Artificial Intelligence for Human-Robot Interaction: Papers from the AAAI Fall Symposium,* edited by B. Hayes and M. Gombolay. Technical Report FS-15-01. Palo Alto, CA: AAAI Press.

Cialdini, R. B., and Goldstein, N. J. 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology* 55:591–621. Palo Alto, CA: Annual Reviews, Inc.

Coman, A., and Aha, D. W. 2017. Cognitive Support for Rebel Agents: Social Awareness and Counternarrative Intelligence. In *Proceedings of the Fifth Conference on Advances in Cognitive Systems*. Palo Alto, CA: Cognitive Systems Foundation.

Coman, A.; Gillespie, K.; and Muñoz-Avila, H. 2015. Case-Based Local and Global Percept Processing for Rebel Agents. In *Workshop Proceedings from the 23rd International Conference on Case-Based Reasoning,* edited by J. Kendall-Morwick, 23–32. The CEUR Workshop 1520. Aachen, Germany: RWTH-Aachen University.

Coman, A.; Johnson, B.; Briggs, G.; and Aha, D. W. 2017. Social Attitudes of AI Rebellion: A Framework. In *AI, Ethics, and Society: Papers from the 2017 Workshop,* edited by T. Walsh. AAAI Technical Report WS-17-02. Palo Alto, CA: AAAI Press.

Floyd, M. W., and Aha, D. W. 2016. Incorporating Transparency During Trust-Guided Behavior Adaptation. In *Proceedings of the 24th International Conference on Case-Based Reasoning,* 124–38. Berlin: Springer.

French, J. R. P., and Raven, B. 1959. The Bases of Social Power. Reprinted in *Classics of Organization Theory,* 4th ed., edited by J. Shafritz and J. S. Ott. New York: Harcourt Brace.

Gamson, W.A. 1992. *Talking Politics*. New York: Cambridge University Press.

Gregg-Smith, A., and Mayol-Cuevas, W. W. 2015. The Design and Evaluation of a Cooperative Handheld Robot. In *2015 IEEE International Conference on Robotics and Automation*. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Heckhausen, J. E., and Heckhausen, H. E. 2010. *Motivation and Action*. Cambridge, UK: Cambridge University Press.

Hiatt, L. M.; Harrison, A. M.; and Trafton, J. G. 2011. Accommodating Human Variability in Human-Robot Teams Through Theory of Mind. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence,* 2066–71. Palo Alto, CA: AAAI Press.

Holmes, D., and Winston, P. 2016. Story-Enabled Hypothetical Reasoning. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems*. Palo Alto, CA: Cognitive Systems Foundation.

Li, B.; Thakkar, M.; Wang, Y.; and Riedl, M. O. 2014a. Data-Driven Alibi Story Telling for Social Believability. Paper presented at the 2014 Social Believability in Games Workshop. Ft. Lauderdale, FL, April 4.

Li, B.; Thakkar, M.; Wang, Y.; and Riedl, M. O. 2014b. Storytelling with Adjustable Narrator Styles and Sentiments. In *Interactive Storytelling: Proceedings of the Seventh International Conference on Interactive Digital Storytelling,* 1–12. Berlin: Springer.

Milli, S.; Hadfield-Menell, D.; Dragan, A.; and Russell, S. 2017. Should Robots Be Obedient? arXiv preprint: arXiv:1705.09990[cs.AI]. Ithaca, NY: Cornell University Library.

Moerland, T.; Broekens, J.; and Jonker, C. 2016. Fear and Hope Emerge from Anticipation in Model-Based Reinforcement Learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence,* 848–54. Palo Alto, CA: AAAI Press.

Patil, A.; Liu, J.; Price, B.; Sharara, H.; and Brdiczka, O. 2012. Modeling Destructive Group Dynamics in Online Gaming Communities. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 290–97. Palo Alto, CA: AAAI Press.

Pearce, C.; Meadows, B. L.; Langley, P.; and Barley, M. 2014. Social Planning: Achieving Goals by Altering Others' Mental States. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence,* 402–9. Palo Alto, CA: AAAI Press.

Riedl, M. O. 2016. Computational Narrative Intelligence: A Human-Centered Goal for Artificial Intelligence. arXiv preprint: arXiv:1602.06484[cs.AI]. Ithaca, NY: Cornell University Library.

Van Stekelenburg, J., and Klandermans, B. 2013. The Social Psychology of Protest. *Current Sociology* 61(5–6): 886–905. doi.org/10.1177/0011392113479314.

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*, edited by D. W. Aha, M. T. Cox, and H. Muñoz-Avila. Technical Report CS-TR-5029. College Park, MD: University of Maryland.

Wenar, C. 1982. On Negativism. *Human Development* 25(1): 1–23.

Wilson, J. R.; Arnold, T.; and Scheutz, M. 2016. Relational Enhancement: A Framework for Evaluating and Designing Human-Robot Relationships. In *AI, Ethics, and Society: Papers from the 2016 AAAI Workshop,* edited by B. Bonet, S. Koenig, B. Kuipers, I. Nourbakhsh, S. Russell, M. Vardi, and T. Walsh. AAAI Technical Report WS-16-02. Palo Alto, CA: AAAI Press.

Wright S. C.; Taylor D. M.; and Moghaddam, F. M. 1990. The Relationship of Perceptions and Emotions to Behavior in the Face of Collective Inequality. *Social Justice Research* 4(3): 229–50.

Alexandra Coman (PhD, Lehigh University, 2013) is a senior manager at Capital One. She was previously an NRC postdoctoral research associate with the Adaptive Systems Section at NRL in Washington, DC, and an assistant professor of computer science at Ohio Northern University. Her research experience and interests include cognitive architectures, AI planning, case-based reasoning, goal reasoning, AI ethics, affective computing, and narrative intelligence.

David W. Aha (PhD, University of California, Irvine, 1990) leads NRL's Adaptive Systems Section in Washington, DC. His interests include mixed-initiative intelligent agents (for example, that employ goal reasoning models), deliberative autonomy, machine learning, and case-based reasoning, among other topics. He has co-organized 35 events on these topics (such as ICCBR-17), hosted 13 post-doctoral researchers, served on 20 PhD committees, created the UCI Repository for ML Databases, was a AAAI Councilor, and gave the Robert S. Engelmore Memorial Lecture at IAAI-17.