AI Rebel Agents

Alexandra Coman and David W. Aha

2018. *AI Magazine*, vol. 39, no. 3, pp. 16-26.

Presented by Leen-Kiat Soh

Abstract

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 (1) ***refuse*** assigned goals and plans, or (2) ***oppose*** the behavior or attitudes of other agents

- Can serve purposes such as ethics, safety, task execution correctness, and providing or supporting diverse points of view

Several ***examples***: potential benefits

A ***framework*** to help categorize and design rebel agents

- Social and ethical implications
- Potential benefits and risks
- Social awareness and counternarrative intelligence

Outline

Introduction: Human Noncompliance

Prior Work and Scenarios

An AI Rebellion Framework

Sociocognitive Dimensions of Rebellion

Conclusion

Introduction: Human Noncompliance

Introduction

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

- What if we gave agents commands that are in conflict with our own long-term goals, or with accurate knowledge that they have, or that have unethical implications not known to us?
- What if agents received contradictory commands from several humans?
- …

Introduction: Definitions

Rebel agents: 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

Rebellion: An umbrella term covering reluctance, protest, refusal, rejection of tasks, and similar attitudes or behaviors

Introduction: Non-Compliance in Humans

Why do we say “no”?

How do we decide whether, when, and how to say “no”?

What are the further social implications of saying “no”?

Introduction: Characteristics of Human Rebellion

Multiple types

Multiple possible motivations

Several stages (i.e., preliminary, deliberation, actual manifestation, aftermath)

Sociocognitive mechanisms at play

Introduction: Overview of Framework

Inspired by social psychology and designed to accommodate the variations

General: it does not assume any particular agent architecture

Counternarrative Intelligence*:* a mechanism that enables rebels to produce, express, and reason about counternarratives that support and justify rebellion

Introduction: Core of a Common Language 1

- Developing and implementing AI agents embodying various facets of rebellion
  - Potential research directions: (1) the development of *AI cognitive prostheses* that empower humans with *low social capital* to adopt positively motivated noncompliant behavior, and (2) *goal alignme*nt in mixed human and AI teams through cycles of noncompliance, negotiation, or agreement cycles
- 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
  - An example of an ethics question: whether an AI agent should always signal to humans that it is considering rebellion, even if it does not end up rebelling

Introduction: Core of a Common Language 2

- Identifying new possible directions of transdisciplinary research
  - for example, delving deeper into the psychological functions of noncompliance, and exploring their transferability to AI
- Promoting richer models of AI in popular culture
  - to offer a counterpoint to cliché representations of AI rebellion

Prior Work and Scenarios

Prior Work 1

- Cooperative handheld intelligent tools with task-specific knowledge that refuse to execute actions which violate task specifications
  - E.g., Disable painting function if user points the tool at a pixel that is not supposed to be painted
- Embodied AI agents to refuse based on reasons: knowledge, capacity, goal priority and timing, social role and obligation, and normative permissibility
- Embodied AI agents to express reluctance
  - E.g., robot protests repeatedly with increasingly intense emotions
- Autonomous-vehicle agents execute “health-preserving” actions to over-ride driver commands
  - E.g., after detecting faults such as insufficient fuel

Prior Work 2

- AI agents that use theory of mind to simulate what human teammates may be thinking
  - Notify them whether they are deviating from expected behavior
- Robots influence humans to adopt ethically acceptable behavior
  - Ethical nudges to get a human to stop neglecting a child
  - Subtly vs. directly influence: ethical issues
- Robot disobedience to augment irrational human behavior
- Goal reasoning agents can use rebellion to improve reasoning

Hypothetical Scenarios

- Furniture Mover
  - Cause for Rebellion: Ensuring safety
- Personal Agent
  - Cause for Rebellion: Preventing user from ordering too much unhealthy food
- Hiring Committee
  - Cause of Rebellion: Ensuring opinions are heard, interpreting information about the candidates

An AI Rebellion Framework

Dimensions and Types

Factors

Stages

Dimensions and Types of Rebellion 1

- **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)
- **Expression**
  - Explicit rebellion: the alter is clearly defined and the rebel agent's behavior is clearly identifiable as rebellious
  - Implicit rebellion: the alter is not clearly defined or the rebel agent's behavior suggests rebellion, but is not clearly expressed as such.
- **Focus**
  - Inward-oriented rebellion: rebel agent's own behavior (e.g., the agent refuses to adjust its behavior as requested by an alter)
  - Outward-oriented rebellion: the alter's behavior (e.g., the agent might confront a human alter whom it identifies as mistreating another human)

Dimensions and Types of Rebellion 2

- **Interaction Initiation**
  - Reactive: when an interaction within which rebellious behavior occurs is initiated by the alter (e.g., the alter making a request that the rebel agent rejects
  - Proactive: the rebel agent initiates the rebellious behavior, which may or may not occur within an explicit interaction (e.g., agents 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.”

Dimensions and Types of Rebellion 3

- **Normativity**
  - Normative: taking action within the confines of what has been explicitly allowed (e.g., questioning without disobeying, if questioning has been allowed)
  - Nonnormative: behavior that has been neither explicitly allowed nor explicitly forbidden, but diverges from the current command given to the agent (e.g., a goal reasoning agent that changes its current goal from the assigned one to a new goal that has not been explicitly forbidden)
  - Counternormative: 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***

Dimensions and Types of Rebellion 4

- **Action or Inaction**
  - Action: agent's rebellion manifests through any sort of outwardly perceivable behavior, such as initiating a conversation in which it objects to a received command
  - Inaction: agent develops an internal negative attitude (e.g., towards an assigned goal or another agent's behavior), but does not manifest it outwardly

Dimensions and Types of Rebellion 5

- **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**
  - Egoistic: agent rebels in support of its own well-being or survival (whatever meanings these might have to the agent)
  - Altruistic: agent rebels in support of someone else's interests (e.g., on behalf of a human group)
  - ***Egoistic and altruistic rebellion can coexist***: e.g., 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

- Human social psychology
  - Frustration, perceived injustice
- AI
  - Ethics and safety, team solidarity, task execution correctness, self-actualization, resolving contradicting commands from multiple alters
  - Tradeoffs
- Supporting and inhibiting factors may contribute to deciding whether a rebellion episode will be triggered, or how it will be carried out
  - Social psychology: people who have motivations to protest do not necessarily do so
  - Efficacy, fear of consequences, social capital, access to resources, opportunities

Stages of Rebellion 1

- **Pre-rebellion**
  - Processes that lead to rebellion (e.g., the agent observing and assessing changes in the environment and the behavior of other agents)
- **Rebellion deliberation**
  - When motivating, supporting, and inhibiting factors are assessed to decide whether to trigger rebellion.
  - Deliberation could be based on observing the *current* world state or on *future-state* projection, which can be purely *rational* or *emotionally charged*

Stages of Rebellion 2

- **Rebellion execution**
  - Begin with rebellion being triggered as a result of rebellion deliberation, and consist of expressing rebellion
  - Verbal or nonverbal communication
  - Can be expressed behaviorally
  - Can be expressed through an internal change in the agent's attitudes: inaction
- **Post-rebellion**
  - Behavior in the aftermath of a rebellion episode, as the agent responds to the alter's or other witnesses' reactions to rebellion
  - Reaffirming one's objection or rejection (e.g., the robot's objection to an assigned task becoming increasingly intense) or ceasing to rebel
  - Or, assessing and managing inverse trust

Stages of Rebellion: Example Scenarios

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.

Sociocognitive Dimensions of Rebellion

Social awareness

Counternarrative intelligence

Social Awareness

- **Rebellion-aware** agents can reason about rebellion (their own and that of others) and its implications, such as social risks
  - **Conflicted rebel agents**: they can both rebel and reason about the implications and consequences of rebellion
- **Rebellion-unaware**
  - **Naïve rebel agents**: they deliberate on whether to trigger rebellion, but do not reason about the social implications, consequences, and risks of rebellious attitudes
  - Could become aware through various processes (e.g., human inspired)

Counternarrative Intelligence

- **Narrative intelligence** is defined as “the ability to craft, tell, understand, and respond affectively to stories”
- **Counternarrative intelligence** refers to the ability of rebel agents to
  - produce alternative retellings or counterinterpretations, informed by subjective factors such as emotional appraisal, of an alter’s narrative, or
  - identify their own pregenerated narratives as being counternarratives in a given context
  - A counternarrative exists in relation and contrast to a base narrative that it is a variant of and that it challenges

Counternarrative Intelligence: Dimensions 1

- **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)
  - Counternarratives are **deceptive** when they intentionally misrepresent the agent's beliefs (e.g., 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)

Counternarrative Intelligence: Dimensions 2

- **Generation Time**
  - **A priori** counternarratives are generated before triggering 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.

Counternarrative Intelligence: Dimensions 3

- **Divergence Type:** This dimension reflects how the counternarrative differs 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
  - **Interpretative counternarratives** do *not* differ from the base narrative in terms of sequence of events, but give different interpretations to the events (e.g., in terms of motivations and emotions)
  - **Transformative counternarratives** differ factually from the base narrative, implicitly asserting that the base narrative contains falsehoods

Conclusion

Conclusion

- Argued: It is beneficial for certain AI agents to be able to rebel for positive, defensible reasons in a variety of situations
- Speculated: AI may never become fully socially intelligent without noncompliance abilities
- Proposed: An AI rebellion framework with sociocognitive dimensions: rebellion awareness and counternarrative intelligence
- The framework is intended to inspire, guide, and provide terminology for
  - the development and study of rebel agents that serve positive purposes
  - systematic discussion of the ethics of AI rebellion
  - positive reframing of the AI noncompliance narrative within the research community and popular culture