The AI Rebellion: Changing the Narrative

David W. Aha[1] and Alexandra Coman[2]

[1]Navy Center for Applied Research in AI; Naval Research Laboratory (Code 5514); Washington, DC

[2]NRC Postdoctoral Fellow; Naval Research Laboratory (Code 5514); Washington, DC

{david.aha, alexandra.coman.ctr.ro}@nrl.navy.mil

Abstract

Sci-fi narratives permeating the collective consciousness endow AI Rebellion with ample negative connotations. However, for AI agents, as for humans, attitudes of protest, objection, and rejection have many potential benefits in support of ethics, safety, self-actualization, solidarity, and social justice, and are necessary in a wide variety of contexts. We launch a conversation on *constructive* AI rebellion and describe a framework meant to support discussion, implementation, and deployment of AI Rebel Agents as protagonists of positive narratives.

Embracing Rebellion

Consider seven episodes in Cara's lifetime: (1) Cara as a visibly angry small child saying "no" in response to a request made by her parents; (2) as an actor in a high-school film project arguing with the director about the way in which her lines should be delivered; (3) in college, as a participant in a variant of Solomon Asch's line experiment (Asch, 1956) not conforming to the erroneous opinion expressed by the majority; (4) as an early-career employee expressing a preference not to take on a new position she has been offered within the same company, as she does not consider it a learning opportunity; (5) as an established engineer refusing to continue working on a project unless her supervisor delays the release date so that a flaw which could endanger the lives of final users can be remedied; (6) as a concerned teammate refusing a task that she believes would put too much strain on her long-time collaborators, whom she knows well; (7) as the leader of a social protest campaign in support of her ingroup (a social group she psychologically identifies as being a part of (van Stekelenburg and Klandermans, 2010)), the rights of which she believes are not upheld by the outgroup perceived as dominant in terms of social influence.

In all these scenarios, Cara acts autonomously and assertively, expressing an attitude divergent from those of one or more others. We can argue that her rebellion has solid, defensible purposes, as opposed to being arbitrary and unreasonable. In these seven scenarios, rebellion is in support of: (1) personal identity development (although child Cara does not know this, she is in the throes of the "negativism" (Wenar, 1982) of the "terrible twos", an early-childhood period often characterized by protest-oriented behavior instrumental to personal identity development); (2) character believability; (3) expressing what one believes to be the truth; (4) self-actualization; (5) ethics and safety; (6) team solidarity; and (7) social justice. We assume that, when possible and appropriate, explanation and negotiation attempts accompany protest. Such rebellion is primarily *in support of* something rather than *against* something: it is a "positive no" (Ury, 2007). Thus, in each narrative, rebel Cara can easily be cast as a protagonist, rather than an antagonist.

But what if Cara's name stands for "Collaborative Autonomous Rebel Agent", and she belongs to an AI agent class called "Rebel Agents"? Then the prevalent popular narrative about her rebellion defaults to a chilling version of the seventh scenario, the "robot uprising", promptly endowing Cara with a propensity for violence and human-level/superintelligence (Bostrom, 2014), which both inform her rebellion.

Is rebellion then always less necessary, beneficial, or welcome when the rebel is synthetic? We believe that attitudes of rebellion by various types of AI agents have clear potential benefits, some of them similar to the benefits of human protest while others are specific to AI agents of that type. For example, in human-AI collaboration contexts, AI and human team members should all be dedicated to maintaining safety, ethical behavior, and team solidarity; actions that undermine any of these key concerns should warrant resistance from any team member.

Rebel Agents are AI agents that can object to or even reject goals or associated courses of action assigned to them by other (human or artificial) agents, or challenge the general attitudes or behaviors of those agents. We call an

agent against which one rebels an *Interactor*. The Interactor role can be performed by a wide variety of agents with diverse degrees of authority over the Rebel Agent (e.g., commander, operator, or teammate). The Rebel Agent's regular stance toward the Interactor is generally not intended to be adversarial; rather, it must collaborate and even share goals fully or partially with the Interactor. A Rebel Agent has potential for rebellion that, based on external and internal conditions, may or may not manifest, but such an agent is generally not in a rebelling state by default. One may argue that "rebellion" is too strong and restrictive a term for the attitudes reflected in some of the scenarios above. We use "rebellion" as an umbrella term covering reluctance, protest, refusal, rejection of tasks, and similar attitudes.

The ability to survive and thrive, while not strictly following commands at all times, is an essential part of AI autonomy, and is exemplified by agents that react to unexpected events, deviate from initial plans, exploit opportunities, formulate their own new goals, and are driven to explore and learn by intrinsic motivation (Vattam et al., 2013; Van Der Krogt and De Weerdt, 2005; Singh et al., 2010). AI explanation and negotiation skills, which are necessary to express a well-founded "no" in a convincing, prosocial manner, have also been studied (Molineaux and Aha, 2015; Jonker et al., 2012; Gratch, Nazari, and Johnson, 2016). Hence, although not explicitly advertising themselves as such, agents incorporating various aspects of rebellion already exist and are proving beneficial. However, the issue of realistic AI Rebellion is, to our knowledge, being explicitly addressed only in very few cases, which we list in the next section.

Our research objectives pertaining to Rebel Agents include establishing a framework for AI agent rebellion that is informed by social and personality psychology, investigating potential benefits and challenges pertaining to Rebel Agents in selected applications, and implementing Rebel Agents driven by a variety of motivation models.

To the AI community, we propose a conversation about AI rebellion as a standalone process separate from (but combinable with) related ones, such as goal reasoning (Vattam et al., 2013). Why do we believe this conversation to be opportune? Because rebellion: (1) is a necessary and, we believe, unavoidable step in AI development; (2) is a rich source of interdisciplinary opportunity in that it shares characteristics with human protest attitudes as studied in psychology and sociology, but cannot and needs not replicate them in every way; (3) has deep implications for human-AI interaction and long-term collaboration; and (4) has its own set of challenges that need to be addressed. AI Rebellion is worthy of more varied, nuanced, and grounded narrative representations, including positive ones.

Existing AI Rebels

Coman, Gillespie, and Muñoz-Avila (2015) proposed Rebel Agents in a limiting context of goal reasoning. We expand and generalize their definition.

To our knowledge, a general formal framework for AI rebellion does not exist, although several authors have addressed (using varied terminology) what we refer to as "rebellion". We can classify their work, briefly described below, according to our framework, which we will introduce in the next section.

Gregg-Smith and Mayol-Cuevas (2015) developed hand-held smart tools that can refuse to execute actions which violate task specifications. Briggs and Scheutz (2015) proposed a general process in which an embodied AI agent refuses to conduct an assigned task due to reasons including: lack of obligation; goal priority and timing; and permissibility issues (e.g., safety requirements, ethical norms). Briggs, McConnell, and Scheutz (2015) demonstrated ways in which embodied AI agents can convincingly express their reluctance to perform a task. Hiatt, Harrison, and Trafton (2011) proposed agents that use theory of mind to determine whether they should notify a human that he/she is deviating from expected behavior. Similarly, in some ways, but on the level of ethics, rather than task execution correctness, Borenstein and Arkin (2016) explore the idea of "ethical nudges" through which a robot attempts to influence a human to adopt ethically-acceptable behavior.

In addition to the above, other types of AI agents have the potential of becoming rebels. Notable examples include motivated agents (Coddington, 2006), goal reasoning agents, and artificial moral agents (Wiegel, 2006).

A Framework for AI Rebellion

We propose an initial version of an AI Rebellion framework that includes types, stages, and factors of rebellion. As the latter constitute a significant and complex topic, we dedicate the entire next section to them.

Types of Rebellion

We propose rebellion types based on three dimensions: **expression** (explicit and implicit), **focus** (inward-oriented, with two subtypes: non-compliance and non-conformity, terms adapted from social influence theory (Cialdini and Goldstein, 2004); and outward-oriented), and **interaction initiation** (reactive and proactive).

Explicit rebellion is exemplified by Scenario 1, in which Cara's protest is expressed through outwardly visible anger. **Implicit rebellion** is exemplified by Scenario 3: the agent never refuses or criticizes anything

directly, but protest is implicit in its expression of an opinion that differs from that of the majority.

Inward-oriented rebellion focuses on the Rebel Agent's own behavior (e.g., as in Scenario 1, the agent refuses to adjust its behavior as requested by an Interactor or dictated by implicit norms). **Outward-oriented** rebellion instead focuses on the Interactor's behavior, to which the Rebel Agent objects (e.g., Scenario 7, and the work of Hiatt, Harrison, and Trafton (2011) and Borenstein and Arkin (2016)).

Rebellion is **reactive** when the interaction resulting in rebellion is initiated by the Interactor (e.g., the Interactor makes a request that the Rebel Agent rejects, as in scenario 1). In **proactive** rebellion (e.g., Scenarios 5 and 7; Hiatt, Harrison, and Trafton (2011)), the Rebel Agent initiates the interaction, which consists of objecting to behaviors, attitudes, or general contexts identified as problematic, rather than to specific requests.

Non-compliance is inward-oriented, reactive rebellion: the agent rejects requests to adjust its own behavior. Scenario 1 exemplifies this type, as does the work of Briggs and Scheutz (2015). **Non-conformity** is inward-oriented, proactive rebellion: refusing to adjust one's behavior in order to "fit in" (Scenario 3).

Stages of Rebellion

Pre-rebellion: This stage includes processes leading to rebellion, including observation and assessment of changes in the environment relevant to the agent's motivation. For example, in Scenario 6, pre-rebellion includes Cara observing her teammates' behavior over their long-term collaboration. The progression towards rebellion may be reflected in the Rebel Agent's outward behavior. The ways in which a possible "no" could be manifested can also be decided during this stage (e.g., how to frame the "no" so as not to jeopardize a long-term collaboration with the Interactor).

Rebellion deliberation: This refers to any episode within pre-rebellion or rebellion execution in which motivating and supporting factors of rebellion (see next section) are assessed to decide whether to trigger rebellion (e.g., Cara asking herself questions such as "Are my teammates under too much strain to handle an additional task?" in Scenario 6) or stop a rebellion.

Rebellion execution: Episodes begin with rebellion being **triggered** and consist of **expressing rebellion**. The main questions associated with this stage concern what triggers rebellion and how rebellion should be expressed. Is there a rebellion threshold for motivating factors (e.g., `if (discontent > 5) rebel();`)? Are there any occurrences that, if observed, are sufficient to immediately trigger rebellion, with no other preconditions? Is a set of conditions (as in the process proposed by Briggs and Scheutz (2015)) used to decide whether rebellion will be triggered? Is triggering based on observing the current world state or is it based on projection (either purely rational, such as reasoning about future states of the environment, or emotionally charged, such as through anticipatory emotions, like hope and fear, associated with possible future states (Moerland, Broekens, and Jonker (2016))? Is rebellion expressed through verbal or non-verbal communication (Briggs, McConnell, and Scheutz (2015)) or behaviorally (e.g., such as when a handheld intelligent tool physically resists a movement (Gregg-Smith and Mayol-Cuevas, 2015))?

Post-rebellion: This is the agent's behavior in the aftermath of a rebellion episode, as it responds to the Interactors' reactions to rebellion. Post-rebellion can consist of re-affirming one's objection or rejection (e.g., the robot's objection to an assigned task becoming increasingly intense in the experiments of Briggs, McConnell, and Scheutz (2015)) or deciding not to, and assessing and managing trust and relationships after rebellion. As in the case of pre-rebellion, some of these processes and concerns may be expressed in the agent's outward behavior.

These stages can be roughly mapped to the three steps of "a positive no" recommended by Ury (2007) to humans who need to reject or object: (1) preparing the "no", (2) delivering the "no", and (3) following through.

These stages can be interpreted in various ways (some can be intertwined or missing). They can be addressed in multiple areas of research, such as human-robot interaction (HRI) for expressing rebellion and post-rebellion, and can be used to categorize future research directions.

Emotion and Rebellion

In social psychology, Van Stekelenburg and Klandermans (2010) list emotion (notably anger, as expressed by Cara in Scenario 1) as a key factor of human protest. Should an AI agent's rebellion be emotional? What are the potential roles and implications (including problematic ones) of emotions in AI rebellion?

In AI, emotion is currently studied in the following contexts: (1) simulating displays of emotion; (2) acquiring and replicating models of human emotion ((1) and (2) fall under the subfield of affective computing (Picard, 2003)); and (3) as an integral component of cognitive processes such as learning (e.g., in some approaches to intrinsically-motivated reinforcement learning (Sequeira, Melo, and Paiva, 2011)).

In the context of our framework, roles for emotion include: displays of emotion can be used as outward manifestations in pre-rebellion, rebellion execution, and post-rebellion; modeled emotions can be used as motivating and/or supporting factors during rebellion

deliberation, while models of the Interactor's emotional states can be used to determine whether a rebellion episode would be opportune; and models of human teammates' emotional states can be used to decide whether to rebel on their behalf, against an outside Interactor. Emotional contagion (Saunier and Jones, 2014) can be used to spread rebellion to other agents (with possibly problematic implications), while anticipatory emotion (e.g., hope and fear (Moerland, Broekens, and Jonker, 2016)) can be experienced in pre-rebellion and used as a rebellion trigger.

Factors of Rebellion

We distinguish between two types of factors of rebellion. **Motivating factors** provide the primary drive for rebellion (e.g., striving for social justice in Scenario 7). **Supporting factors** instead contribute to assessing whether a rebellion episode will be triggered, and/or how it will be carried out. Efficacy, the individual's expectation that his/her rebellion can have the desired effect, has been shown to often fulfill such a role in human protest (van Stekelenburg and Klandermans, 2010)). Certain factors (e.g., emotion) can be in either category, depending on the context.

A discrepancy between the agent's motivation and the assigned task, or various observed conditions or behaviors, usually triggers rebellion. Even more generally, some form of divergent access to information of the Rebel Agents and the Interactors is usually at the root of rebellion episodes. This information could be objective, but only partially available to a proper subset of the agents in the environment at the time of the rebellion episode, or subjective (e.g., a Rebel Agent's own motivation, its autobiographical memory, knowledge about its teammates' past behavior, strengths, weaknesses, and needs).

Unlike humans, AI agents are not all based on the same general cognitive architecture. Thus, the motivating factors of rebellion will not be general, but depend on the agent's architecture, interaction context, and purpose. The following list provides examples of motivating factors that are based on the scenarios previously described and pertain to positive narratives of AI rebellion.

Ethics and safety: Rebel Agents can refuse tasks they assess as being ethically prohibited and/or violating safety norms, as in (Briggs and Scheutz, 2015). They can also attempt to dissuade humans from engaging in ethically-prohibited behavior (Borenstein and Arkin (2016)).

Team solidarity and trust: In long-term HRI, team solidarity must be established and maintained over a variety of tasks (Wilson, Arnold, and Scheutz, 2016). This requires occasionally saying "no" on behalf of the team (Scenario 6), and also saying "no", constructively, to one's teammates, when necessary. Rebellion puts a strain on trust and can affect it both negatively and positively; trust and distrust can be motivating and/or supporting factors of rebellion. Trust can increase after instances of rebellion, depending on what caused the rebellion, and how post-rebellion was conducted by the agents involved. For example, Cara's trust-worthiness as an expert in her field can increase after her rebellion in Scenario 5.

Believability and intentionality: Believability (Bates, 1994) is a key requirement for AI characters in interactive narratives such as computer games and training simulations. Just like a human actor might argue with the director about their character's arc, an AI actor could rebel against its drama manager (the AI director in various interactive storytelling systems (Sharma et al., 2010)). We believe that an agent which can assert itself convincingly encourages an intentional stance: the attitude that the agent is rational, and has beliefs, desires, and goals (Dennett, 1987). In HRI research, taking an intentional stance with regard to AI collaborators has been found to increase humans' cognitive performance in the collaborative tasks (Walliser et al., 2015; Wykowska et al., 2014)).

Self-actualization: Like its human counterparts, an AI Rebel Agent (that is, possibly, a "perpetual learner" (Roberts et al., 2016)) could object to an assigned task that it assesses as not playing up to its strengths or not constituting a valuable learning opportunity.

Social justice: The prospect of an AI agent protesting for social justice for its ingroup, which it considers to be oppressed by a dominant outgroup, can be problematic. But what if the ingroup and outgroup are not, simplistically, "AI" and "human", but the AI agent "identifies" or "sympathizes" with a human group, possibly a minority group the rights of which it can support? In this case, AI rebellion can act in support of constructive human protest and empowerment.

Further Issues and Questions

The seven scenarios we proposed all assume a certain anthropomorphism of a protagonist AI agent, which (who?) is involved in human-like social relationships. We can also investigate models of AI rebellion that do ***not*** emulate human rebellion, but are made possible and useful by the AI agent's architecture, abilities, and/or purpose.

The first scenario reflects protest in support of identity development. What would that even mean in the context of AI agents? Are there AI learning models that can accommodate and benefit from such rebellion?

How should AI and human acts of rebellion interact? Can the former support the latter? What methods could be used by an AI agent to model and detect rebellion in others, and leverage this in its decision making, and what are the ethical implications of doing so? We look forward to investigating these and related issues in our future work.

References

Asch, S.E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1-70.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, *37*(7), 122-125.

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.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford, UK: Oxford University Press.

Briggs, G., McConnell, I., and Scheutz, M. (2015). When robots object: Evidence for the utility of verbal, but not necessarily spoken protest. *Proceedings of the Eighth International Conference on Social Robotics* (pp. 83-92). Paris: Springer.

Briggs, G., and Scheutz. M. (2015). "Sorry, I can't do that": Developing mechanisms to appropriately reject directives in human-robot interactions. In B. Hayes et al. (Eds.) *AI for Human-Robot Interaction: Papers from the AAAI Fall Symposium* (Technical Report FS-15-01). Arlington, VA: AAAI Press.

Cialdini, R.B., and Goldstein, N.J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591-621.

Coddington, A. (2006). Motivations for MADbot: A motivated and goal directed robot. *Proceedings of the Twenty-Fifth Workshop of the UK Planning and Scheduling Special Interest Group* (pp. 39-46). Nottingham, UK: ISSN 1368-5708.

Coman, A., Gillespie, K., and Muñoz-Avila, H. (2015). Case-based local and global percept processing for rebel agents. In M. Floyd and D.W. Aha (Eds.) *Case-Based Agents: Papers from the ICCBR Workshop*. Frankfurt, Germany: CEUR 1520.

Dennett, D.C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Gratch, J., Nazari, Z., and Johnson, E. (2016). The misrepresentation game: How to win at negotiation while seeming like a nice guy. *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems* (pp. 728-737). Singapore: ACM Press.

Gregg-Smith, A., and Mayol-Cuevas, W.W. (2015). The design and evaluation of a cooperative handheld robot. *Proceedings of the International Conference on Robotics and Automation* (pp. 1968-1975). Seattle, WA: IEEE Press.

Hiatt, L.M., Harrison, A.M., and Trafton, J.G. (2011). Accommodating human variability in human-robot teams through theory of mind. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (pp. 2066-2071). Barcelona, Spain: AAAI Press.

Jonker, C.M., Hindriks, K.V., Wiggers, P., and Broekens, J. (2012). Negotiating agents. *AI Magazine*, *33*(3), 79-91.

Moerland, T., Broekens, J., and Jonker, C. (2016). Fear and hope emerge from anticipation in model-based reinforcement learning. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 848-854). New York: AAAI Press.

Molineaux, M., and Aha, D.W. (2015). Continuous explanation generation in a multi-agent domain. *Proceedings of the Third Conference on Advances in Cognitive Systems* (pp. 1-18). Atlanta, GA: Cognitive Systems Foundation.

Picard, R.W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies*, *59*(1), 55-64.

Roberts, M., Hiatt, L.M., Coman, A., Choi, D., Johnson, B., and Aha, D.W. (2016). ActorSim, A toolkit for studying cross-disciplinary challenges in autonomy. In L. Humphrey, U. Topcu, S. Singh, C. Miller, and M. Vardi (Eds.) *Cross-Disciplinary Challenges for Autonomous Systems: Papers from the AAAI Fall Symposium* (Tech. Rep. FS-16-04). Arlington, VA: AAAI Press.

Saunier, J., and Jones, H. (2014). Mixed agent/social dynamics for emotion computation. *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 645-652). Paris: ACM Press.

Sequeira, P., Melo, F.S., and Paiva, A. (2011). Emotion-based intrinsic motivation for reinforcement learning agents. *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction* (pp. 326-336). Memphis, TN: Springer.

Sharma, M., Ontañón, S., Mehta, M., and Ram, A. (2010). Drama management and player modeling for interactive fiction games. *Computational Intelligence*, *26*(2), 183-211.

Singh, S., Lewis, R.L., Barto, A.G., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 70-82.

Ury, W. (2007). *The power of a positive no: How to say no and still get to yes*. London, UK: Bantam.

Van Der Krogt, R., and De Weerdt, M. (2005). Plan repair as an extension of planning. *Proceedings of Fifteenth International Conference on Automated Planning and Scheduling* (pp. 161-170). Monterey, CA: AAAI Press.

Van Stekelenburg, J., and Klandermans, B. (2010). The social psychology of protest. *Sociopedia.isa*.

Vattam, S., Klenk, M., Molineaux, M., & Aha, D.W. (2013). Breadth of approaches to goal reasoning: A research survey. In D.W. Aha, M.T. Cox, & H. Muñoz-Avila (Eds.) *Goal Reasoning: Papers from the ACS Workshop* (Technical Report CS-TR-5029). College Park, MD: University of Maryland, Department of Computer Science.

Walliser, J., Tulk, S., Hertz, N., Issler, E., and Wiese, E. (2015). Effects of perspective taking on implicit attitudes and performance in economic games. *Proceedings of the Eighth International Conference on Social Robotics* (pp. 684-693). Paris: Springer

Wenar, C. (1982). On negativism. *Human Development*, *25*(1), 1-23.

Wiegel, V. (2006). Building blocks for artificial moral agents. In C. Allen, W. Wallach, and M. Brady (Eds.) *Ethical Agents: Papers from the Alife Workshop.* Bloomington, IN.

Wilson, J.R., Arnold, T, and Scheutz, M. (2016). Relational enhancement: A framework for evaluating and designing human-robot relationships. In T. Walsh (Ed.) *AI, Ethics, and Society: Papers from the AAAI Workshop* (Technical Report WS-16-02). Phoenix, AZ: AAAI Press.

Wykowska, A., Wiese, E., Prosser, A., and Müller, H.J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLoS One*, *9*(4), e94339.